MATHEMATICAL ANALYSIS FOR DATA ENGINEERS

YAROSLAV BAZAIKIN

1. Preliminaries

2. Optimization Problem

2.1. **General concept.** Let $K \subset \mathbb{R}^n$ and consider a function $f : K \to \mathbb{R}$. Point $p \in K$ is called a local minimum (local maximum) point of function f if there exists such a neighborhood U of point p that $f(q) \geq f(p)$ ($f(q) \leq f(p)$) for every $q \in U \cap K$. If these inequalities hold for every $q \in K$ then we say that p is (global) minimum (maximum) of f in K. We will speak further about minimum (maximum) points supposing local minimum (maximum) ones.

The fundamental result lying in the base of all theory of optimization is the following.

Theorem 1 (Weierstrass)**.** *Let function $f : K \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}^n$ is closed and bounded set. Then there exits minimum (maximum) point of f in K.*

The following two examples show that both conditions on K in Theorem 1 are important.

Example 1. *Let $K = [-1, 0) \cap (0, 1]$ (it is not closed) and $f(x) = 1/x$ for $x \in K$. We see that*

$$\lim_{x \to -0} f(x) = -\infty, \lim_{x \to +0} f(x) = +\infty$$

and f has neither minimum nor maximum.

Example 2. *Let $K = \mathbb{R}$ (it is not bounded), $f(x) = x$. Its evident that f has neither minimum nor maximum.*

Weierstrass theorem however gives no constructive method of finding minimum (maximum) points. In the remaining part of section we will assume that all functions have continuous partial derivatives up to order two which is a subject of classical optimization theory and in this situation we can build more constructive theory.

Date: October – December 2022.

2.2. **Multivariable optimization without constrains.** Let consider particular case of optimization problem where there are no any constraints. This means that we have function

$$f : \mathbb{R}^n \to \mathbb{R}$$

defined on whole coordinate space $\mathbb{R}^n$. Such a function can have no minimum or maximum points at all (see previous examples), but we have following necessary condition of their existence.

Theorem 2 (Necessary condition of minimum (maximum))**.** *Let $p \in \mathbb{R}^n$ is a maximum or minimum point of function f. Then*

$$\frac{\partial f}{\partial x_1}(p) = \frac{\partial f}{\partial x_2}(p) = \ldots = \frac{\partial f}{\partial x_n}(p) = 0. \tag{1}$$

The point p satisfying condition (1) is called *stationary*. Theorem 2 shows that we need to search minimum and maximum points among stationary ones. It is sometime more convenient to rewrite equation (1) in other form. Remind that *first differential* of function f in point p is the following linear form of variables $h_1, h_2, \ldots, h_n$:

$$df(p) = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(p) h_i. \tag{2}$$

Notice that variables h_i are often replaced with coordinate differentials: $h_i = dx_i$, $i = 1, \ldots, n$. This is rather informal but very convenient in computational practice. Then formula (2) becomes

$$df(p) = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(p) dx_i.$$

Then necessary condition (1) of point to be minimum (maximum) point can be rewrited as

$$df(p) = 0.$$

When stationary points are indeed maximum or minimum points? To formulate sufficient conditions we need to define *second differentials* of function f at a point p: a quadratic form of variables $h_1, \ldots, h_n$

$$d^2 f(p) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x^i \partial x^j}(p) h_i h_j.$$

Let $Q(h_1, \ldots, h_n)$ be a quadratic form defined by symmetric coefficients $q_{ij} = q_{ji} \in \mathbb{R}$:

$$Q(h_1, \ldots, h_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} h_i h_j.$$

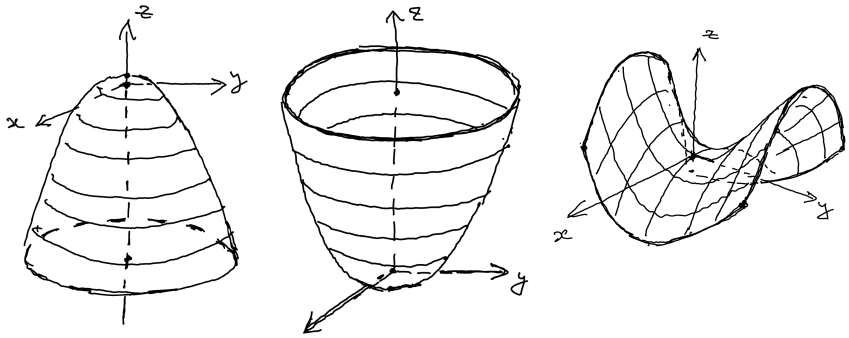

FIGURE 1. Paraboloids: maximum, minimum and saddle point.

We say that Q is positive-definite (negative-definite) if $Q(h_1, \ldots, h_n) > 0 (< 0)$ for all $h_1, \ldots, h_n \in \mathbb{R}$ such that $h_1^2 + h_2^2 + \ldots + h_n^2 \neq 0$.

The following well-known theorem gives an effective instrument for checking if quadratic form is positive- (negative-) definite.

Theorem 3 (Sylvester's criterion)**.** *Quadratic form Q defined by symmetric matrix $(q_{ij})_{i,j=1}^n$ is positive-definite (negative-definite) if and only if all upper left k-by-k corners of Q have positive determinants (have determinants of sign $(-1)^k$), for $k = 1, \ldots, n$.*

Second differential of function f in point p is a quadratic form which is defined by symmetric matrix:

$$d^2 f(p) \sim Hess(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(p) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(p) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(p) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(p) & \frac{\partial^2 f}{\partial x_2^2}(p) & \ldots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(p) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(p) & \ldots & \frac{\partial^2 f}{\partial x_n^2}(p) \end{pmatrix}$$

matrix $Hess(f)$ is called *Hesse matrix* of function f in point f.

Theorem 4 (Sufficient condition of minimum (maximum))**.** *Let p is stationary point of f. If quadratic form $d^2 f(p)$ is positive-definite (negative-definite) then p is a minimum (maximum) point.*

By Theorem 4 to check if $d^2 f(p)$ is positive- (negative-) definite we can use Sylvester's criterion for Hesse matrix of f in point p.

Example 3 (Three paraboloids)**.**

$$z = f(x, y) = -x^2 - y^2.$$

Here we have only one stationary point $O(0,0,0)$ *and second differential in* O *is:*

$$d^2f(O) = -2h_1^2 - 2h_2^2 < 0$$

for all $(h_1, h_2) \neq (0,0)$. *So* $d^2f(O)$ *is negative-definite and* O *is a maximum point (Fig. 1, left).*

$$z = f(x,y) = x^2 + y^2.$$

Here we also have only one stationary point $O(0,0,0)$ *and second differential in* O *is:*

$$d^2f(O) = 2h_1^2 + 2h_2^2 > 0$$

for all $(h_1, h_2) \neq (0,0)$. *So* $d^2f(O)$ *is positive-definite and* O *is a minimum point (Fig. 1, middle).*

$$z = f(x,y) = -x^2 + y^2.$$

Here we also have only one stationary point $O(0,0,0)$ *and second differential in* O *is:*

$$d^2f(O) = -2h_1^2 + 2h_2^2 > 0$$

for all $(h_1, h_2) \neq (0,0)$. *We see that second differential has no definite sign:* $d^2f(O) > 0$ *if* $h_1 = 0, h_2 \neq 0$ *and* $d^2f(O) < 0$ *if* $h_1 \neq 0, h_2 = 0$. *This quadratic form is called indefinite and this point is called "saddle point"; this is neither maximal nor minimum point (Fig. 1, right).*

Situation we yet consider is very restrictive: indeed, almost always in real problems we have constraints. But we already can consider very useful example which is very important in many applications including machine learning.

Example 4 (Least squares and linear regression model)**.** *Let in some experiment the observation consists of vector* $\bar{x} = (x_1, \ldots, x_p)$ *of* p *parameters (regressors) and a scalar respond* y. *For instance, we measure weight* y *of person depending of two parameters: height* x_1 *and age* x_2. *Usually data consists of a collection of observations* y_i *which depends of collection of regressors* $\bar{x}_i$, $i = 1, \ldots, n$, *where* n *is a relatively big number comparing to number of parameters. In our example we can have more than thousand people tested for weight.*

In the linear regression model we assume that response variable is the linear function of regressors:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots, n.$$

Coefficients β_i *are unknown parameters of a model,* ε_i *are errors of observations.*

We can not expect that all errors can vanish; always there are errors during measurement, or there are some extra factors which are not included in model. The problem is: how to estimate unknown parameters in the most reasonable way using observations we have?

In the least squares method we postulate that unknown parameters minimize the following sum of squares of errors:

$$S(\beta_1, \beta_2, \ldots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left| y_i - \sum_{k=1}^{p} \beta_k x_{ik} \right|^2$$

Applying necessary condition theorem we have:

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \left(y_i - \sum_{k=1}^{p} \beta_k x_{ik} \right) = 0, j = 1, \ldots, p.$$

Collect all observations of parameter with number k to one vector: $X_k = (x_{1k}, x_{2k}, \ldots, x_{nk})$ and let $Y = (y_1, y_2, \ldots, y_n)$ be vector of responses over all observations. Let denote:

$$g_{kl} = \langle X_k, X_l \rangle = \sum_{i=1}^{n} x_{ik} x_{il},$$

$$a_k = \langle X_k, Y \rangle = \sum_{i=1}^{n} x_{ik} y_i.$$

Then we can rewrite last equations as follows:

$$g_{j1}\beta_1 + g_{j2}\beta_2 + \ldots g_{jp}\beta_p = a_j, j = 1, \ldots, p. \tag{3}$$

Coefficients g_{jk} consists of all pairwise scalar products of vectors X_1, ..., X_p and generates Gram matrix G of system of vectors X_k. From linear algebra we know that this matrix is positive-definite (and invertible) if there is no any nontrivial linear combination of vectors $X_1, \ldots, X_p$. In this case

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \ldots \\ \hat{\beta}_p \end{pmatrix} = G^{-1} \begin{pmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{pmatrix} \tag{4}$$

In practice we do not need to compute G, usually more reasonable methods (Gauss method for example) are used for solution equation (3). The symbol $\hat{\beta}_k$ is used to emphasize that we find only estimate of unknown coefficients β_k.

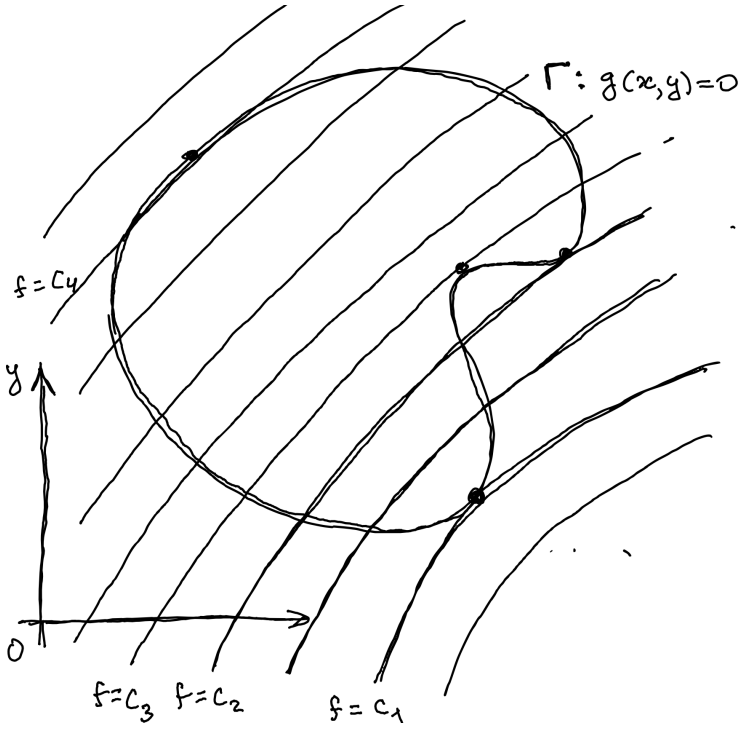

FIGURE 2. Paraboloids: maximum, minimum and saddle point.

Finally we need to check that (4) gives minimum of S. Applying Theorem 4 we obtain:

$$\frac{\partial^2 S}{\partial\beta_k\partial\beta_l} = 2\sum_{i=1}^{n} x_{ik}x_{il} = 2g_{kl}, k, l = 1, \ldots, p.$$

So quadratic form d^2S is defined by Gram matrix $2G$ which is positive-definite and (4) describes minimum point indeed.

2.3. **Multivariable optimization with equality constraints by method of Lagrange multipliers.** Let we need to minimize (maximize) $f(x, y)$ under the constraint $g(x, y) = 0$ (let us begin with two-dimensional case).

Geometrical interpretation: $g(x, y) = 0$ is a curve Γ on the plane with coordinates (x, y). Minimum principle: if $p \in \Gamma$ is a minimum point of f then level set $f^{-1}(f(p))$ is tangent to Γ in point p. Look at Fig. 2, where illustration of this principle is done. Here $c_1 < c_2 < c_3 < c_4$ and we see intersections of curve Γ with level-sets $f(x, y) = const$ of function f. Function increases from right bottom corner to left upper corner; we see four local extremum point where level-sets of f are

tangent to curve Γ. There are: one global maximum point, lying on the level-set $f(x,y) = c_4$; one global minimum point lying on level-set $f(x,y) = c_1$; one local minimum point lying on the level-set $f(x,y) = c_2$ and one local maximum point lying on the level-set $f(x,y) = c_3$).

Let formulate this in analytic way. Normal to Γ in a point p is

$$\nabla_p g = \left(\frac{\partial g}{\partial x}(p), \frac{\partial g}{\partial y}(p)\right),$$

normal to level set $f = f(p)$ is

$$\nabla_p f = \left(\frac{\partial f}{\partial x}(p), \frac{\partial f}{\partial y}(p)\right).$$

Tangent condition means $\nabla_p g$ and $\nabla_p f$ are proportional. Hence there exists $\lambda \in \mathbb{R}$ such that

$$0 = \nabla_p f + \lambda \nabla_p g = \nabla_p (f + \lambda g). \tag{5}$$

We come to some new function (which is called *Lagrange function*) $L(x,y,\lambda)$:

$$L(x,y,\lambda) = f(x,y) + \lambda g(x,y),$$

which depends of *three* variables x, y and λ. New variable λ is called *Lagrange multiplier*. Condition (5) means that

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} = 0.$$

One can express constraint $g(x,y) = 0$ as

$$\frac{\partial L}{\partial \lambda} = 0.$$

So we obtain the following necessary condition for minimum problem with one equation constraint.

Theorem 5. *Let p is minimum (maximum) point of f under constraint $g = 0$. Then there exists some $\lambda \in \mathbb{R}$ such that p, λ is a stationary point for Lagrange function*

$$L(p,\lambda) = f(p) + \lambda g(p).$$

More specifically, the following equations hold:

$$\frac{\partial L}{\partial x}(p) = \frac{\partial L}{\partial y}(p) = g(p) = 0.$$

Now formulate theorem for more general case of several equality constraints.

Theorem 6 (Necessary conditions for n variables and k constraints)**.** *Let $f, g_1, \ldots, g_k : \mathbb{R}^n$ be continuously differentiable functions. Let point $p = (x_1, \ldots, x_n)$ is a minimum (maximum) point of function f with constraints:*

$$g_1(p) = g_2(p) = \ldots = g_k(p) = 0.$$

Then there exist real numbers $\lambda_1, \lambda_2, \ldots, \lambda_k$ (Lagrange multipliers) such that point $(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k)$ is a stationary for Lagrange function

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k) = f(x_1, \ldots, x_n) + \\ \lambda_1 g_1(x_1, \ldots, x_n) + \ldots + \lambda_k g_k(x_1, \ldots, x_n).$$

More specifically, the following equations hold:

$$\frac{\partial L}{\partial x_1}(p) = \ldots = \frac{\partial L}{\partial x_n}(p) = g_1(p) = \ldots = g_k(p) = 0.$$

Example 5. *Let*

$$f(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}, \\ g(x, y, z) = x^2 + y^2 + z^2 - 1, \\ a > b > c > 0.$$

Consider a problem of finding minimum and maximum points of function f under the constrain $g = 0$. Geometrically, minimum point is a tangent point of minimal ellipsoid intersected with unit sphere and with axes proportional to triple a, b, c; maximum point is a tangent point of maximal ellipsoid intersected with unit sphere and with axes proportional to triple a, b, c.

Applying Theorem 6 we construct Lagrange function

$$L = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} + \lambda(x^2 + y^2 + z^2 - 1)$$

and obtain system of algebraic equations:

$$\frac{x}{a^2} + \lambda x = 0, \\ \frac{y}{b^2} + \lambda y = 0, \\ \frac{z}{c^2} + \lambda z = 0, \\ x^2 + y^2 + z^2 = 1.$$

Solving this system we find following stationary points of form

$$(x, y, z, \lambda) : \left(\pm 1, 0, 0, -\frac{1}{a^2}\right), \left(0, \pm 1, 0, -\frac{1}{b^2}\right), \left(0, 0, \pm 1, -\frac{1}{c^2}\right)$$

Investigating these points we can understand that the first is minimum, the last is maximum points and the middle neither minimum nor maximum.

Now formulate sufficient conditions of minimum (maximum) problem with equation constraint.

Theorem 7. *A sufficient condition for $f(p)$ to have minimum (maximum) in stationary point p under the constraints $g_1(p) = \ldots = g_k(p) = 0$ is that the quadratic form*

$$d^2L(p,\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 L}{\partial x^i \partial x^j}(p,\lambda_1,\ldots,\lambda_k)h_i h_j$$

is positive-definite (negative-definite) for all non-trivial $(h_1, h_2, \ldots, h_n)$ for which the following constraints hold:

$$dg_l(p) = \sum_{i=1}^{n}\frac{\partial g_l}{\partial x^i}(p)h_i = 0, l = 1,\ldots,k.$$

Returning to Example 5 we have constraint $xdx+ydy+zdz = 0$ and

$$d^2L = 2\left(\frac{(dx)^2}{a^2}+\frac{(dy)^2}{b^2}+\frac{(dz)^2}{c^2}+\lambda((dx)^2+(dy)^2+(dz)^2)\right).$$

Then

$$d^2L(\pm1,0,0) = 2\left(\frac{1}{b^2}-\frac{1}{a^2}\right)(dy)^2 + 2\left(\frac{1}{c^2}-\frac{1}{a^2}\right)(dz)^2 > 0$$

and

$$d^2L(0,0,\pm1) = 2\left(\frac{1}{a^2}-\frac{1}{c^2}\right)(dx)^2 + 2\left(\frac{1}{b^2}-\frac{1}{c^2}\right)(dy)^2 < 0$$

(remind that $a > b > c > 0$). Therefore $(\pm1,0,0)$ are minimum and $(0,0,\pm1)$ are maximum points. For remaining two points we can see that

$$d^2L(0,\pm1,0) = 2\left(\frac{1}{a^2}-\frac{1}{b^2}\right)(dx)^2 + 2\left(\frac{1}{c^2}-\frac{1}{b^2}\right)(dz)^2$$

is indefinite and points are neither maximum nor minimum ones.

2.4. **Multivariable optimization with inequality constraints.** In this subsection consider the problem of minimizing (maximizing) function $f(p)$, $p = (x_1,\ldots,x_n) \in \mathbb{R}^n$ under the inequality constraints $g_1(p) \le 0, g_2(p) \le 0, \ldots, g_k(p) \le 0$.

There is very simple trick that reduces this problem to equality constraint. Let introduce new set of variables $q = (y_1,\ldots,y_k) \in \mathbb{R}^k$. Then

previous problem is equivalent to a problem of minimizing (maximizing) of $f(p)$ with equality constraints

$$g_l(x_1, \dots, x_n) + y_l^2 = 0, l = 1, \dots, k.$$

For this problem we can apply previous necessary and sufficient conditions with Lagrange multipliers $\lambda_1, \dots, \lambda_k$ and Lagrange function

$$L(p, q, \lambda) = f(x_1, \dots, x_n) + \lambda_1(g_1(x_1, \dots, x_n) + y_1^2) + \\ \dots + \lambda_k(g_k(x_1, \dots, x_n) + y_k^2).$$

We can use of cause combination equality and inequality constraints by the same trick.

Example 6. *Let find maximum point of function*

$$f(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}, a > b > c$$

under the constraint

$$x^2 + y^2 + z^2 \leq 1.$$

Introduce new variable w and replace old inequality constraint with new equality:

$$x^2 + y^2 + z^2 + w^2 - 1 = 0.$$

Now consider Lagrange multiplier λ and Lagrange function

$$L(x, y, z, w) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} + \lambda(x^2 + y^2 + z^2 + w^2 - 1).$$

Necessary conditions are:

$$\begin{aligned} \frac{x}{a^2} + \lambda x &= 0, \\ \frac{y}{b^2} + \lambda y &= 0, \\ \frac{z}{c^2} + \lambda z &= 0, \\ \lambda w &= 0, \\ x^2 + y^2 + z^2 &= 1 - w^2. \end{aligned}$$

Solving this system of algebraic equation we obtain following points of form (x, y, z, w, λ): $(0, 0, 0, \pm 1, 0)$, $(\pm 1, 0, 0, 0, -\frac{1}{a^2})$, $(0, \pm 1, 0, 0, -\frac{1}{b^2})$ $(0, 0, 0, \pm 1, -\frac{1}{c^2})$. If we compare with the example 5 we can see than replacing equality constraint with inequality one adds new stationary point.

Now let check sufficient conditions.

$$d^2L(\pm 1, 0, 0, 0) = 2\left(\frac{1}{b^2} - \frac{1}{a^2}\right)(dy)^2 + 2\left(\frac{1}{c^2} - \frac{1}{a^2}\right)(dz)^2 > 0$$

and

$$d^2L(0,0,\pm1,0)=2\left(\frac{1}{a^2}-\frac{1}{c^2}\right)(dx)^2+2\left(\frac{1}{b^2}-\frac{1}{c^2}\right)(dy)^2<0$$

give minimum and maximum points.

$$d^2L(0,\pm1,0,0)=2\left(\frac{1}{a^2}-\frac{1}{b^2}\right)(dx)^2+2\left(\frac{1}{c^2}-\frac{1}{b^2}\right)(dz)^2$$

is indefinite as in previous example. Finally

$$d^2L(0,\pm1,0,0)=2\left(\frac{(dx)^2}{a^2}+\frac{(dy)^2}{b^2}+\frac{(dz)^2}{c^2}\right)>0$$

corresponds to minimum point. So we have only two maximum points $x=0, y=0, z=\pm1$.

2.5. **Linear Programming.** One particular case of multivariable optimization is a linear programming problem which can be stated in the following form:

$$\begin{aligned}
f(x_1,x_2,\ldots,x_n)=c_1x_1+c_2x_2+\ldots+c_nx_n &\longrightarrow \min\\
\text{under the constraints}&\\
a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1,\\
a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2,\\
&\vdots\\
a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n&=b_m;\\
x_1&\geq 0,\\
x_2&\geq 0,\\
&\vdots\\
x_n&\geq 0,
\end{aligned}\tag{6}$$

where c_j, b_j and a_{ij}, $i=1,\ldots,m$, $j=1,\ldots,n$ are known constants and x_j are unknown variables.

It is convenient to reformulate this problem in matrix form:

$$\begin{aligned}
f(x)=c^Tx &\longrightarrow \min\\
\text{under the constraints}&\\
Ax&=b,\\
x&\geq 0,
\end{aligned}$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \; c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix},$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

We say that x is *feasible solution* if $Ax = b$ and $x \geq 0$, that is x satisfies to constraints. A feasible solution that optimizes the function f is said *optimal* solution.

Remarks. The above classical statement of linear programming problem may seem limited but it is not, as the following remarks show.

1. Replacing f by $-f$ we can consider maximization problem instead of minimization one.

2. In some real problems there are no restrictions $x_i \geq 0$ for some of variables x_i. But for any variable x_i unrestricted in sign we can express it as $x_i = x_i' - x_i''$ replacing x_i by new variables $x_i' \geq 0$ and $x_i'' \geq 0$.

3. One can formulate problem involving inequalities of type

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \leq b_i. \tag{7}$$

together with equalities. Adding new nonnegative variable $x_{n+1} \geq 0$ we can convert (7) to equality

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + x_{n+1} \leq b_i. \tag{8}$$

The same idea can be used in case inequality with opposite sigh with respect to (7).

Consider geometric interpretation of linear programming problem which is the most obvious in case $k = n - m = 2$.

Example 7. *Let $k = n - m = 2$. In this case we always can choose independent variables, say x_1, x_2 such that all other variables $(x_3, \ldots, x_n)$ can be be found as linear combinations of x_1, x_2 from linear constraint $Ax = b$:*

$$\begin{aligned} x_3 &= \alpha_{31}x_1 + \alpha_{32}x_2 + \beta_3 \geq 0, \\ x_4 &= \alpha_{41}x_1 + \alpha_{42}x_2 + \beta_4 \geq 0, \\ &\vdots \\ x_n &= \alpha_{n1}x_1 + \alpha_{n2}x_2 + \beta_n \geq 0, \end{aligned} \tag{9}$$

for some constants α_{ij}, β_i. Now we can represent values of independent variables x_1, x_2 by a point on the plane with coordinates (x_1, x_2).

Because of constraint $x \geq 0$, any feasible solution has point (x_1, x_2) lying in the upper right quadrant Q_1 of the plane. Inequalities in (9) implies that we have some finite number of m half-planes and intersection of all this half-planes with Q_1 gives the set F of feasible solutions (may be empty).

Now let come back to function f. Using (9) we can express f as function of independent variables x_1, x_2:

$$f(x_1, x_2) = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \longrightarrow \min .$$

The level-sets of function f are lines, orthogonal to vector $n = (\gamma_1, \gamma_2)$. Going along to n we increase function f and going in opposite direction $(-n)$ we decrease function f. Now it is clear that if minimum exists then set of optimal solutions (x_1, x_2) has to contain some "corner" point of F (such a point is said extreme point of F). If it is not unique then F contains segment connecting to corner points.

returning from geometry to algebra let us note that corner point lies at least on two lines defined by equations in (9). It follows that unique optimal solution satisfies equations $x_i = x_j = 0$ for some indexes i, j. Not unique solution is defined by to corner points, that is by two solutions of type $x_i = x_j = 0$ for some i, j.

So the principal scheme of solution of linear programming problem is following: iterate over all pairs of indices i, j, substitute $x_i = x_j = 0$ and solve obtaining system of equations, choosing point with minimal f.

The next theorem shows that general case works in the same way.

Theorem 8. *Let $k = n - m$. If an optimal solution of linear programming problem (?) exists, then minimum of function f is attained on those $x = (x_1, \ldots, x_n)$ for which at least k variables are equal to zero and all other are positive.*

In real problems one has a huge numbers n of variables and m of constraints. In this case iteration over all selections of $k = n - m$ variables form the set of n variables which equals to a extremely big number

$$\left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{(n-m)!m!}.$$

There are exist very efficient algorithms (simplex-method, genetic algorithms, etc.) of finding optimal solution of linear programming problem but these questions are outside the scope of our course.

2.6. **Dynamic programming.** Dynamic programming is a method of solution of some optimization problems which can be divided into more simple sequential steps. We consider here only several examples of using dynamical programming.

Let variable t changes between values 0 and $T > 0$ (usually t represents a time variable). Consider two monotonic continuous functions $a \leq x(t) \leq b$, $c \leq y(t) \leq d$. Consider the following optimization problem:

$$F(T, x(t), y(t)) = \int_0^T f(x(t), y(t))dt \longrightarrow \min$$

under the constraints

$$x(0) = a, x(T) = b, a \leq x(t) \leq b, \text{ for all } 0 \leq t \leq T,$$
$$y(0) = c, y(T) = d, c \leq y(t) \leq d, \text{ for all } 0 \leq t \leq T.$$

Remark that variables which need to be varied for optimization of F are two functions $x(t)$, $y(t)$ and one variable T.

Example 8 (The fastest trajectory problem)**.** *Consider plane with coordinates x and y. Let $A = (a, c)$ and $B = (b, d)$ be two fixed points on the plane. Let $\Omega = \{(x, y)|a \leq x \leq b, c \leq y \leq d\}$ be rectangular domain in the plane.*

Introduce function $v(x, y) > 0$ defined in Ω which represents the signal propagation speed in domain Ω. Define function f:

$$f(x, y) = \frac{1}{v(x, y)}.$$

Then

$$F = \int_0^S \frac{ds}{v(x(s), y(s))}$$

is a time of propagation of signal from point $A = (x(0), y(0))$ to point $B = (x(S), y(S))$ which has to be minimized in the above problem statement (note that here $t = s$ and $T = S$ have physical dimension of a distance, not a time; and x, y are arbitrary coordinates on the plane, not necessarily flat!).

Example 9 (Two well logs correlation problem)**.** *In the notations of previous example let interpret x and y as depth coordinate in two oil wells. Consider result of measurement of some geophysical field along the well (say electrical resistance, radioactivity, etc.). We can represent two such measurement by two functions $\alpha(x)$ and $\beta(y)$. Now the problem arise to correlate this two wells: to understand what segments in x and y correspond to the same oil formation.*

Mathematically we correlate variables x and y by some parameter t, $0 \leq t \leq T$ such that point $x(t)$ corresponds to a point $y(t)$. We want to find such a correlation that $\alpha(x(t))$ is the most closed to $\beta(y(t))$. So let

$$f(x,y) = \frac{1}{T}|\alpha(x) - \beta(y)|^2.$$

We come to a problem of minimizing of a function

$$F = \frac{1}{T}\int_0^T |\alpha(x(t)) - \beta(y(t))|^2 dt$$

over all feasible $T, x(t), y(t)$.

In most applications we need not find exact solution, we only want to find good approximation. So we can first replace our problem by their discrete analogue and then try to solve it.

Let $0 \leq i_k \leq N$, $0 \leq j_k \leq M$, $0 \leq k \leq T$. Consider some function $f(x,y)$ and state the following optimization problem

$$\begin{aligned} F = \sum_{k=0}^{T} f(i_k, j_k) &\longrightarrow \min, \\ &\text{under the constraints} \\ i_0 = N, i_T = 0, 0 \leq i_k \leq N &\text{ for all } 0 \leq k \leq T, \\ j_0 = M, j_T = 0, 0 \leq j_k \leq M &\text{ for all } 0 \leq k \leq T, \\ i_k \geq i_{k+1} \geq i_k + 1, &\text{ for all } 0 \leq k \leq T-1, \\ j_k \geq j_{k+1} \geq j_k + 1, &\text{ for all } 0 \leq k \leq T-1. \end{aligned} \tag{10}$$

The variables which are needed to be optimized are sequences i_k, j_k and number T. The next theorem describes algorithm of dynamical programming for this problem (here function $argminf$ denotes a some point in which function f has minimal value).

Theorem 9. *The following formulas allows to find optimal solution to problem (10).*

$$\begin{aligned} S_{0,0} &= 0, \\ S_{1,0} &= f(1,0), \\ S_{0,1} &= f(0,1), \\ S_{1,1} &= \min\{S_{1,0}; S_{0,1}\} + f(1,1), \\ S_{i,j} &= \min\{S_{i-1,j}; S_{i,j-1}; S_{i-1,j-1}\} + f(i,j), \text{ for } 1 \leq i \leq N, 1 \leq j \leq M, \\ i_0 &= N, j_0 = M, \\ (i_{t+1}, j_{t+1}) &= argmin\{S_{i_t-1,j_t}; S_{i_t,j_t-1}; S_{i_t-1,j_t-1}\}, 0 \leq t \leq T, \\ &\text{where } T \text{ is defined by condition } i_T = 0, j_T = 0. \end{aligned}$$

3. Lebesgue integration

Lebesgue integration is the most convenient method of integration of functions which allows to consider the most wide spaces of functions used in applications. We consider also Riemann integration because of its very intuitive meaning. Both methods of integration are based on concepts of Jordan and Lebesgue measures, for which we will begin.

3.1. **Jordan and Lebesgue measures in $\mathbb{R}$.** Jordan measure (which intuitively is the most clear) is just a accurate formulation of concept of length (or area in dimension two, or volume in dimension three) which dates back to antiquity.

We will define Jordan measure λ going from simplest subsets of $\mathbb{R}$ to more complicated. First of all let J is a segment of one of form: $[a, b]$, $[a, b)$, $(a, b]$ or (a, b), $a \leq b$. For such simplest subset we put

$$\lambda(J) = b - a.$$

Now if J is a union of pairwise non-intersecting segments $J_1, \ldots, J_n$ then

$$\lambda(J) = \sum_{i=1}^{n} \lambda(J_i).$$

We will call such sets *elementary*.

Now let $A \subset \mathbb{R}$ is arbitrary subset. Set A is said *Jordan measurable* if for every $\varepsilon > 0$ there exists such elementary sets $A_\varepsilon, B_\varepsilon$ which satisfies

$$A_\varepsilon \subset A \subset B_\varepsilon,$$
$$\lambda(B_\varepsilon \backslash A_\varepsilon) < \varepsilon.$$

It can be proved that for Jordan measurable set $\lambda(A_\varepsilon)$ and $\lambda(B_\varepsilon)$ have limits if $\varepsilon \to 0$ and

$$\lim_{\varepsilon \to 0} \lambda(A_\varepsilon) = \lim_{\varepsilon \to 0} \lambda(B_\varepsilon).$$

This common limit is said to be equal to $\lambda(A)$, the *Jordan measure* of A.

Example 10. *Let $A = \{\frac{1}{n} | n \in \mathbb{N}\}$ is infinite countable set. Prove that A is Jordan measurable and $\lambda(A) = 0$.*

Lemma 1 (Properties of Jordan measure)**.** *(a) If A and B are Jordan measurable and have empty intersection, then $A \cup B$ is Jordan measurable and*

$$\lambda(A \cap B) = \lambda(A) + \lambda(B).$$

(b) If A_1, $A_2, \dots, A_n$ are Jordan measurable, $A_i \cap A_j = \emptyset$ for every $i \neq j$ then $A = \cup_{i=1}^{n} A_i$ is Jordan measurable and

$$\lambda(A) = \sum_{i=1}^{n} \lambda(A_i).$$

(c) If A_1, $A_2, \dots, A_n, \dots$ are Jordan measurable (infinite countable family of sets), $A_n \cap A_m = \emptyset$ for every $m \neq n$ and $A = \cup_{n=1}^{\infty} A_n$ is Jordan measurable then

$$\lambda(A) = \sum_{n=1}^{\infty} \lambda(A_n).$$

Notice the difference between (b) and (c): the (b) guarantees that union of finite family of non-intersecting Jordan measurable sets is measurable, but in (c) we need to state in addition the Jordan measurability of infinite union. This in fact the main inconvenience of Jordane measure: the infinitely countable union of measurable sets may be not measurable. The next two example shows that this indeed may happens.

Example 11. *If $\subset \mathbb{R}$ is Jordan measurable then A is bounded (prove it). In particular, the set $\mathbb{Q}$ of all rational numbers is not Jordan measurable.*

Example 12. *Let $A = \mathbb{Q} \cap [0, 1]$ is the set of all rational numbers belonging to segment $[0, 1]$. It obviously is bounded and infinite countable. But A is not Jordan measurable. Indeed, assuming A is Jordan measurable and taking $\varepsilon = 0.5$ we can find two elementary sets B_1, B_2 such that $B_1 \subset A \subset B_2$ and $\lambda(B_2 \backslash B_1) < 0.5$. Set $B_2 \backslash B_1$ is again elementary and $\lambda([0, 1]) = 1$ implies there exists $0 < x < y < 1$, $[x, y] \cap B_2 = \infty$. Then there exists $r \in \mathbb{Q}$, $x < r < y$. Therefore $r \notin B_2$ and $r \notin \mathbb{Q}$ — a contradiction.*

The Lebesgue measure μ improves this defect of Jordan measure. For segments and for elementary sets J we put

$$\mu(J) = \lambda(J),$$

that is Lebesgue measure is just a length of elementary set.

Now define the *outer measure* μ^* of any $A \subset \mathbb{R}$:

$$\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \mu(C_n) | \text{ for all countable (possibly infinite!)} \right.$$

$$\left. \text{family of elementary sets } C_n, n = 1, \dots, \text{ such that } A \subset \cup_{n=1}^{\infty} C_n \right\}$$

The most important thing here is that we consider infinite countable covers of A by elementary sets, but not only finite ones.

Now we can say that set $A \subset \mathbb{R}$ is Lebesgue measurable if for every $\varepsilon > 0$ there exists elementary set A_ε which satisfies condition:

$$\mu^*(A \triangle A_\varepsilon) < \varepsilon \tag{11}$$

(remind that $A \triangle A_\varepsilon = A \backslash A_\varepsilon \cup A_\varepsilon \backslash A$ is a symmetric difference of two sets). It can be proved that under the condition (11) there exists limit $\lambda^*(A_\varepsilon)$ at $\varepsilon \to 0$ and we set Lebesgue measure μ of A to be equal

$$\mu(A) = \lim_{\varepsilon \to 0} \mu^*(A_\varepsilon).$$

It is important in this definition that we allow set A_ε to be located in relation to A in an arbitrary way and consider then symmetric difference $\triangle$ of two sets A and A_ε.

Example 13. *The Lebesgue measure of set $\mathbb{Q}$ of rational numbers is equal to zero.*

The best way "to fill" difference between Jordan and Lebesgue measures is to understand what is set which has Lebesgue measure zero.

Lemma 2. *(a) A set $A \subset \mathbb{R}$ has Lebesgue measure zero (or simply is a null set) if and only if for every $\varepsilon > 0$ there exists* **infinite countable** *family of intervals $J_n = (a_n, b_n)$, $n \in \mathbb{N}$ such that*

$$A \subset \cup_{n=1}^{\infty} J_n,$$

$$\sum_{n=1}^{\infty} |b_n - a_n| < \varepsilon.$$

(b) A set $A \subset \mathbb{R}$ has Jordan measure zero if and only if for every $\varepsilon > 0$ there exists **finite** *family of intervals $J_n = (a_n, b_n)$, $n \in \mathbb{N}$ such that*

$$A \subset \cup_{n=1}^{\infty} J_n,$$

$$\sum_{n=1}^{\infty} |b_n - a_n| < \varepsilon.$$

In particular, any countable set is null set (prove it!).

Lemma 3. *The following properties of Lebesgue measure take place.*

(a) If A_n, $n \in \mathbb{N}$ are Lebesgue measurable and pairwise non-intersecting then $\cup_{n=1}^{\infty} A_n$ is Lebesgue measurable and

$$\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n).$$

(b) Compliment, countable union and countable intersection of Lebesgue measurable sets are again Lebesgue measurable.

(c) Any subset of null set is again a null set (and in particular again Lebesgue measurable).

3.2. **Jordan and Lebesgue measures in $\mathbb{R}^n$.** All constructions in general case of $\mathbb{R}^n$ are the same as in case of $\mathbb{R}$ with the only difference that we need to re-define the elementary sets.

Now the simplest set is "parallelepiped":

$$P = I_1 \times I_2 \times \ldots \times I_n = \{(x_1, x_2, \ldots, x_n) | x_i \in I_i, i = 1, \ldots, n\},$$

where I_i is a segment of one of forms: $[a_i, b_i], [a_i, b_i), (a_i, b_i], (a_i, b_i)$, $a_i \leq b_i$. Then we call elementary set a union $J = P_1 \cup P_2 \cup \ldots \cup P_m$ of parallelepipeds $P_1, P_2, \ldots, P_m$ such that $P_i \cap P_j = \emptyset$.

We define Jordan measure of elementary sets in the following way:

$$\lambda(P) = |b_1 - a_1| \cdot |b_2 - a_2| \cdot \ldots \cdot |b_n - a_n| \text{ for parallelepiped } P,$$

$$\lambda(J) = \sum_{k=1}^{m} \lambda(P_k) \text{ for elementary set } J.$$

Remaining definitions and properties of Jordan and Lebesgue are the same as in previous section.

I conclusion we give a statement showing the connection between Lebesgue measures in $\mathbb{R}^n$ for different n.

Lemma 4. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ are Lbesgue measurable sets, then Cartesian product $A \times B \subset \mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$ is measurable and*

$$\mu(A \times B) = \mu(A)\mu(B).$$

3.3. **Riemann integral.** Let $f : [a, b] \to \mathbb{R}$ be a function defined on a segment $[a, b]$, $-\infty < a < b < \infty$. Define a tagged partition $P(x, t)$ of $[a, b]$ to be two sequences of numbers $\{x_i\}_{i=0}^n$ and $\{t_j\}_{j=1}^n$ of the form

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b,$$

$$x_{i-1} \leq t_i \leq x_i, i = 1, \ldots, n.$$

Segments $[x_{i-1}, x_i]$ are called sub-intervals of partition. Define the mesh of partition $P(x, t)$:

$$\Delta(P) = \max_{i=1,\ldots,n} |x_i - x_{i-1}|. \tag{12}$$

Riemann sum of function f with respect to partition $P(x, t)$ is

$$\sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}).$$

This sum has following geometric meaning. If f is positive on $[a, b]$ then Riemann sum is a sum of rectangles generated by two sides: subinterval $[x_{i-1}, x_i]$ and vertical segment of length $f(t_i)$ passing through coordinate t_i. This sum approximates area of part of plane between graph of function f and segment $[a, b]$. The more detailed partition the more precise approximation should be. If we consider function f having both positive and negative values the Riemann sum is an alternating sum of areas with signs depending of up or below coordinate axe the graph of f lies.

We say that Riemann integral of function f is equal to S if for every $\varepsilon > 0$ there exist $\delta > 0$ such that for any tagged partition $P(x, t)$ of segment $[a, b]$ whose mesh is less than δ we have

$$\left| \sum_{i=1}^{n} f(t_i)(x_i - x_{i-1}) - S \right| < \varepsilon.$$

If Riemann integral for f exists then we say that f is Riemann integrable and write

$$\int_a^b f(x)dx = S. \tag{13}$$

This definition of Riemann integral differs from traditional one: it is much more complicated technically to prove usually properties of Riemann integral using this definition. We choose it because of more transparency of geometrical idea.

Theorem 10. *Function $f : [a, b] \to \mathbb{R}$ is Riemann integrable if and only if it is bounded and it is continuous everywhere except (Lebesgue) null set.*

Riemann integrability of function $f : [a, b] \to \mathbb{C}$ is equivalent to integrability of every function $Re(f), Im(f) : [a, b] \to \mathbb{R}$. In this case

$$\int_a^b f(x)dx = \int_a^b Re(f(x))dx + i \int_a^b Im(f(x))dx.$$

One can define an Riemann integral for functions $f : A \subset \mathbb{R}^n \to \mathbb{R}$ defined on Jordan measurable set A. To do this it is necessary only to define tagged partition of A, the remaining part of definition is the same. Usual way to do this is to consider A as finite union of Jordan measurable sets such that every two set can intersect each other only by common boundary. The accurate definition become complicated but less meaningful. The similar problem is to extend definition to Jordan non-measurable sets A or to whole $\mathbb{R}$ (this leads to concept of improper

integral). We prefer stop here to speak about Riemann integral and continue with more important Lebesgue integral.

3.4. **Lebesgue integral.** The main difference between Lebesgue integral and Riemann integral is that in case of Lebesgue integral we need to use a finite partition of range of function f. Then Lebesgue sum (analogue of Riemann sum) becomes more complicated: every term in the sum may consist of union of rectangles with the same height. The area of this union can be computed using notion of Lebesgue measure. Now let give this definition in more details.

Let $A \subset \mathbb{R}$ is Lebesgue measurable. Function $f : A \to \mathbb{R}$ is said *measurable* if for any $y \in \mathbb{R}$ the pre-image $f^{-1}(y, \infty) = \{x \in A | f(x) > y$ is Lebesgue measurable set.

Tagged partition $P(y, t)$ of range of measurable function $f : A \to \mathbb{R}$ is given by two finite consequences y_i, $i = 0, \ldots, n$ and t_j, $j = 1, \ldots, n$ such that

$$-\infty < y_0 < y_1 < \ldots < y_n < +\infty,$$
$$y_{j-1} < t_j < y_j, j = 1, \ldots, n.$$

As before, the mesh of $P(y, t)$ is

$$\Delta(P(y, t)) = \max_{i=1,\ldots,n} |y_j - y_{j-1}|.$$

Lebesgue sum of f with respect to tagged partition $P(y, t)$ is

$$\sum_{k=1}^{n} \mu(\{x \in A | f(x) > t_i\})|y_i - y_{i-1}|$$

(remark that this definition is correct because f is measurable).

Now we say that Lebesgue integral of function f is equal to S if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for any tagged partition $P(y, t)$ of range of f whose mesh is less than δ we have

$$\left| \sum_{k=1}^{n} \mu(\{x \in A | f(x) > t_k\})(y_k - y_{k-1}) - S \right| < \varepsilon.$$

If S exists then f is said Lebesgue integrable and we call S Lebesgue integral, denoting it by

$$\int_A f(x) d\mu(x).$$

or even simply

$$\int_A f(x) dx.$$

The set $\{x \in A | f(x) > t_k\}$ in this definition may have complicated structure but it has well defined Lebesgue measure (by the way it

is possible that this set is not Jordan measurable). This causes the Lebesgue integral to be defined on a much wider class of functions than Riemann integral. Note that the main reason for the superiority of the Lebesgue integral over the Riemann integral is the assumption of infinite countable unions when determining the measure of a set.

Because we have well defined concept of Lebesgue measure in $\mathbb{R}^n$, we can directly generalize definition of measurable function and Lebesgue integral to the case of functions $f : A \subset \mathbb{R}^n \to \mathbb{R}$ or $f : A \subset \mathbb{R}^n \to \mathbb{C}$.

Example 14. *Let $X \subset \mathbb{R}$ is measurable and $\mu(X) = 0$. Let $f(x) = g(x)$ for $x \notin X$ and f is Lebesgue integrable then g is also Lebesgue intagrable and*

$$\int_{-\infty}^{+\infty} f(x)d\mu(x) = \int_{-\infty}^{+\infty} g(x)d\mu(x).$$

4. Introduction to functional analysis

4.1. **Vector space.** Set X is a vector space if there are two operations on its elements: vector addition "$+$" and multiplication "$\lambda\cdot$" by scalar $\lambda \in \mathbb{C}$. These operations should satisfy the following three groups axioms:

(a1) $(x + y) + z = x + (y + z)$ (associativity);

(a2) $x + y = y + x$ (commutativity);

(a3) there exists an element $0 \in X$ such that $x + 0 = 0 + x = x$ (existence of zero vector);

(a4) for every $x \in X$ there exists $-x \in V$ such that $x + (-x) = (-x) + x = 0$ (existence of inverse vector);

(m1) $1 \cdot x = x$;

(m2) $\lambda \cdot (\mu \cdot x) = (\lambda\mu) \cdot x$;

(am1) $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$;

(am2) $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot y$

for all $x, y, z \in X; \lambda, \mu \in \mathbb{C}$.

We say that set sequence $e_1, e_2, \ldots, e_n$ is a basis of X if for any vector $x \in X$ there exists unique sequence of numbers $c_1, c_2, \ldots, c_n \in \mathbb{C}$ such that

$$x = c_1e_1 + c_2e_2 + \ldots + c_ne_n.$$

Coefficients $c_1, \ldots c_n$ are called coordinates of x in the basis $e_1, \ldots, e_n$ and integer number n is a dimension of vector space X. It can be proven that if basis exists then n does not depend of choice of basis. In this case we speak about *finite-dimensional* vector space.

The correspondence $x \mapsto (c_1, c_2, \ldots, c_n)$ identifies finite-dimensional space X with the coordinate space $\mathbb{C}^n$. Therefore space $\mathbb{C}^n$ is the universal example of finite-dimensional vector space.

In many application the infinite-dimensional vector spaces are useful. Let consider several examples of vector spaces.

Example 15 (Vector space of all sequences $\mathbb{C}^\infty$)**.** *May be the simplest case of infinite-dimensional vector space is the space $X = \mathbb{C}^\infty$ consisting of all sequences $z = \{z_i\}_{i=1}^\infty$. Every term z_i is a coordinate of z and operations $+$ and $\lambda\cdot$ are defined coordinate-wise:*

$$(z + w)_i = z_i + w_i,\ (\lambda \cdot z)_i = \lambda z_i,\ i \in \mathbb{N}.$$

Example 16 (Vector space of all functions $M(a, b)$)**.** *Let $-\infty \le a \le b \le +\infty$ (a, b can be infinite numbers). Let $X = M(a, b)$ consist of all functions $f : [a, b] \to \mathbb{C}$. Operations $+$ and $\lambda\cdot$ are defined point-wise:*

$$(f + g)(t) = f(t) + g(t),$$
$$(\lambda \cdot f)(t) = \lambda f(t),$$
$$a \le t \le b,\ f, g \in M(a, b).$$

We can think that $f(t)$ is a t-coordinate of vector $f \in X$, that is we have "infinite continuum number" of coordinates in the vector space $M(a, b)$.

4.2. **Normed space.** Let X be a vector space. X is said to be a *normed space* if for every $x \in X$ there is associated a nonnegative real number $\|x\|$, called the *norm*, in such a way that

(a) $\|x + y\| \le \|x\| + \|y\|$ for all $x, y \in X$,
(b) $\|\alpha x\| = |\alpha|\|x\|$, if $x \in X$ and α is a scalar,
(c) $\|x\| > 0$ if $x \ne 0$.

If (a) and (b) hold only then we speak about *semi-norm*.

We can define distance between vectors in normed space: $d(x, y) = \|x - y\|$, for $x, y \in X$. Normed space X with a metric d satisfies axioms of metric space:

(i) $0 \le d(x, y) \le \infty$ for all $x, y \in X$,
(ii) $d(x, y) = 0$ if and only if $x = y$,
(iii) $d(x, y) = d(y, x)$ for all $x, y \in X$,
(iv) $d(x, z) \le d(x, y) + d(y, z)$ for all $x, y, z \in X$.

Having a metric on X we can say about convergence of sequences of elements of X, fundamental sequences, open and closed subsets etc.

Example 17 (Normed space $\mathbb{C}_p^n$)**.** *Consider complex coordinate space*

$$\mathbb{C}^n = \{\xi = (z_1, z_2, \ldots, z_n) | z_i \in \mathbb{C}, i = 1, \ldots, n\}.$$

Consider classical euclidean norm

$$\|\xi\|_2 = \sqrt{\sum_{i=1}^{n} |z_i|^2}, \xi \in \mathbb{C}^n$$

This norm can be generalizing: define a family of norms depending of parameter p, $1 \leq p < \infty$*:*

$$\|\xi\|_p = \sqrt[p]{\sum_{i=1}^{n} |z_i|^p}, \xi \in \mathbb{C}^n$$

We can extend last formula to a case $p = \infty$*:*

$$\|\xi\|_\infty = \lim_{p\to\infty} \sqrt[p]{\sum_{i=1}^{n} |z_i|^p} = \max_{i=1,\dots,n} |z_i|, \xi \in \mathbb{C}^n$$

Vector space $\mathbb{C}^n$ *together with norm* $\|\cdot\|_p$ *gives to us finite-dimensional normed space* $\mathbb{C}_p^n$, $1 \leq p \leq \infty$.

Proposition 1. *The convergence in space* $\mathbb{C}_p^n$ *does not depend of* p*. This means that any sequence* $\{z_i\}_{i1}^{\infty}$ *is (or is not) convergent simultaneously for all* $1 \leq p \leq \infty$.

We can generalized Example 17 to infinite-dimensional case.

Example 18 (Space l^∞)**.** *Define vector space* $l^\infty \subset \mathbb{C}^\infty$ *as space of all infinite bounded sequences* $\xi = \{z_i\}_{i=1}^{\infty}$*. This means that for every* ξ *there exits constant* C *(which depends of* ξ*) such that* $|z_n| \leq C$ *for all* $n \in \mathbb{N}$*. The operations summation and multiplication by a scalar are defined coordinate-wise (see Example 15). We regard* l^∞ *as normed space by equipping it with the norm*

$$\|\xi\|_\infty = \sup\{|z_n|, n \in \mathbb{N}\} < \infty. \tag{14}$$

Example 19 (Spaces l_c^∞ and l_0^∞)**.** *In the space* l^∞ *consider subspaces* l_c^∞ *consisting of convergent sequences* ξ *and* l_0^∞ *consisting of sequences converging to a zero. Certainly we consider* l_c^∞ *and* l_0^∞ *with the same norm (14).*

Example 20. *Fix real number* p *such that* $1 \leq p < \infty$ *and consider vector space* l^p *consisting of all infinite consequences* $\xi = \{z_n\}_{i=1}^{\infty}$ *which satisfies to condition:*

$$\sum_{i=1}^{\infty} |z_i|^p < \infty.$$

As in Example 15, the operations summation and multiplication by scalar are defined coordinate-wise. Equip l^p with the norm

$$\|\xi\|_p = \sqrt[p]{\sum_{i=1}^{\infty} |z_i|^p}.$$

Normed spaces l^p are natural generalizations of spaces $\mathbb{C}_p^n$ from Example 17.

Example 21 (Uniform norm)**.** *Consider to real numbers $-\infty \leq a \leq b \leq +\infty$ and space $C[a,b] \subset M(a,b)$ of all continuous bounded functions $f : [a,b] \to \mathbb{C}$. Space of all such functions with point-wise operations defined in Example 16. Define uniform (or sup-norm) norm on the space $C[a,b]$:*

$$\|f\|_\infty = \sup\{|f(t)||t \in C[a,b]\} < \infty$$

for $f \in C[a,b]$. Remark that case $a = -\infty$ and $b = +\infty$ is also possible and we denote this space $C(\mathbb{R})$.

Example 22. *Consider two subspace of $C(\mathbb{R})$ with the same uniform norm $\|\cdot\|_\infty$: subspace $C_0(\mathbb{R})$ of functions $f(t)$ converging if $|t| \to \infty$ and subspace $C_{00}(\mathbb{R})$ of functions $f(t)$ with compact support (this means that there exists $T > 0$ such that $f(t) = 0$ for all $|t| > T$).*

Example 23 (Space $L^p(\mathbb{R})$)**.** *Let $1 \leq p < \infty$. Consider vector space $L_0^p(\mathbb{R})$ of measurable functions $f : \mathbb{R} \to \mathbb{C}$ such that*

$$\int_{-\infty}^{+\infty} |f(t)|^p d\mu(t) < \infty.$$

Define

$$\|f\|_p = \sqrt[p]{\int_{-\infty}^{+\infty} |f(t)|^p d\mu(t)}.$$

Note that $\|\cdot\|_p$ is not a norm on the space $L_0^p(\mathbb{R})$! Indeed, we can consider function

$$f(t) = \begin{cases} 0, & (t \in \mathbb{R}\backslash\mathbb{Q}) \\ 1, & (t \in \mathbb{Q}). \end{cases} \tag{15}$$

Set $\mathbb{Q}$ has Lebesgue measure zero and therefore $\|f\|_p = 0$ but f is not a zero element of $L_0^p(\mathbb{R})$. Thus the last axiom of norm is not satisfied and $\|\cdot\|_p$ is a semi-norm on $L_0^p(\mathbb{R})$.

To make $\|\cdot\|_p$ a norm we say that f and g are equivalent and write $f \overset{\text{a.e.}}{=} g$ if and only if $f = g$ almost everywhere. Remind that this means that there exists zero Lebesgue measure set $Z \subset \mathbb{R}$ (null-set) such that

for any $t \in \mathbb{R}$, $t \notin Z$ we have $f(t) = g(t)$. Let $L^p(\mathbb{R})$ be the space of all equivalence classes of element from $L_0^p(\mathbb{R})$.

This definition means that an element of $L^p(\mathbb{R})$ can be represented by some function $f : \mathbb{R} \to \mathbb{C}$ but we can freely replace f by any function g, $f \overset{\text{a.e.}}{=} g$ without replacing the element of $L^p(\mathbb{R})$. In particular, it is make no sense to speak about value $f(t_0)$ in some particular point t_0. However, one can check that operations $+$, $\lambda\cdot$ are defined correctly and norm $\|f\|_p$ does not depend of choice of representing function f. Now $\|\cdot\|_p$ is indeed a norm: if $\|f\|_p = 0$ then $f \overset{\text{a.e.}}{=} 0$ and 0 is representing the same element as f in the space $L^p(\mathbb{R})$.

Example 24 (Space $L^\infty(\mathbb{R})$)**.** *We have construction which is analogues to previous example. Let $L_0^\infty(\mathbb{R})$ consists of all functions $f : \mathbb{R} \to \mathbb{C}$ which are measurable and bounded almost everywhere. For $f \in L_0^\infty(\mathbb{R})$ we have semi-norm $\|f\|_\infty$ (one can consider the same function as in 15 to see that $\|\cdot\|_\infty$ is not a norm) and again consider equivalence classes $f \overset{\text{a.e.}}{=} g$ as in previous example. The resulting normed space is denoted by $L^\infty(\mathbb{R})$.*

Example 25 (Comparison of spaces $L^1(\mathbb{R})$, $L^2(\mathbb{R})$ and $L^\infty(\mathbb{R})$)**.** *Consider the following functions $f_i : \mathbb{R} \to \mathbb{R}$, $i = 0, 1, 2, 3$ and $g_j : \mathbb{R} \to \mathbb{R}$, $j = 1, 2$:*

$$
\begin{aligned}
f_0(t) &= \frac{1}{1+t^2},\\
f_1(t) &= \frac{1}{\sqrt{|t|}(1+t^2)},\\
f_2(t) &= \frac{\sin t}{t} + \frac{1}{\sqrt[3]{|t|}(1+t^4)},\\
f_3(t) &= \frac{1}{\sqrt{1+|t|}},\\
g_1(t) &= \frac{1}{\sqrt[3]{|t|}(1+t^4)},\\
g_2(t) &= \frac{\sin t}{t}.
\end{aligned}
$$

One can easily check that the diagram on the Fig. takes place, so spaces L^1, L^2 and L^∞ are all different.

4.3. **Banach spaces.** We say set normed space is *Banach space* if it is complete as metric space, that is every fundamental sequence has a limit.

Example 26. *Any finite-dimensional normed space is Banach. In particular all spaces $\mathbb{C}_p^n$ are Banach.*

Proposition 2. *(a) Every absolutely convergent series in a Banach space converges;*

(b) If every absolutely convergent series converges in a normed space, then it is Banach.

Let X, Y are two normed spaces. Map $T : X \to Y$ is a linear operator if the following two properties are satisfied:

(l1) $T(u + v) = T(u) + T(v)$, $u, v \in X$;

(l2) $T(\lambda u) = \lambda T(u)$ $\lambda \in \mathbb{C}, u \in X$.

We say that linear operator $T : X \to Y$ is bounded if

$$\|T\| = \sup\{\|Tx\| | x \in X, \|x\| \leq 1\} < \infty. \tag{16}$$

Let $B(X, Y)$ be the vector space of all bounded linear operators $T : X \to Y$ with operations:

$$(f + g)(u) = f(u) + g(u), u \in X, f, g, \in B(X, Y),$$
$$(\lambda \cdot f)(u) = \lambda f(u), u \in X, \lambda \in \mathbb{C}, f \in B(X, Y).$$

Proposition 3. *Formula (16) defines a norm on the space $B(X, Y)$, which is called operator norm.*

Theorem 11. *Let X, Y are normed spaces and Y is Banach. Then vector space $B(X, Y)$ equipped with operator norm (16) is Banach.*

The one important case is $Y = \mathbb{C}$. The space $B(X, \mathbb{C})$ is denoted X' and is called dual space to X. By the previous theorem X' is Banach space.

Proposition 4. *Normed spaces l^p, l^∞, L^p are Banach spaces, $1 \leq p \leq \infty$.*

For $1 < p < \infty$ let $1 < q < \infty$ be the unique *dual* number such that

$$\frac{1}{p} + \frac{1}{q} = 1$$

Proposition 5. *For $1 < p < \infty$ the dual space to L^p and l^p are L^q and l^q:*

$$(L^p)' = L^q,$$
$$(l^p)' = l^q.$$

Proposition 6. *Space l^∞ is dual to l^1 and vice versa:*

$$\left(l^1\right)' = l^\infty,$$
$$(l^\infty)' = l^1.$$

4.4. **Hilbert spaces.** We know from previous section that l^2, L^2 are Banach spaces. But these spaces posses much more interesting structure: the structure of Hilbert space.

Scalar product on the vector space X is a map $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ with the properties:

(a) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$, $x, y, z \in X$, $\alpha, \beta \in \mathbb{C}$;
(b) $\langle x, \alpha y + \beta z \rangle = \overline{\alpha} \langle x, y \rangle + \overline{\beta} \langle x, z \rangle$, $x, y, z \in X$, $\alpha, \beta \in \mathbb{C}$;
(c) $\langle y, x \rangle = \overline{\langle x, y \rangle}$, $x, y \in X$;
(d) $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$, $x \in X$.

Scalar product defines the norm on X:

$$\|x\| = \sqrt{\langle x, x \rangle}, x \in X.$$

We say that space X with scalar product is Hilbert space if it is Banach space with respect to norm $\|x\|$.

Example 27. *The norm in l^2 is defined from the scalar product*

$$\langle z, w \rangle = \sum_{i=1}^{\infty} z_i \overline{w_i}.$$

Example 28. *The norm in L^2 is defined from the scalar product*

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) \overline{g(t)} dt.$$

So the spaces l^2 and L^2 are Hilbert.

Theorem 12 (Riesz theorem)**.** *Let X is Hilbert space. Then every vector $x \in X$ defines bounded linear operator $f_x : X \to \mathbb{C}$ by the rule $f_x(y) = \langle y, \rangle x$. Vice versa, every bounded linear operator $f : X \to \mathbb{C}$ is equal f_x for some $x \in X$. Moreover, this bijection $x \mapsto f_x$ preserves norms on X and X'.*

This theorem illustrates the relation $(L^2)' = L^2$.

Theorem 13. *Let X is Hilbert space. which is in addition is separable (this means that there exists dense countable subset in X). Then there exists basis $x_\alpha \in X$, $\alpha \in A$ for some countable set A with the property;*

$$\langle x_\alpha, x_\beta \rangle = \begin{cases} 1 & (\alpha = \beta) \\ 0 & (\alpha \neq \beta) \end{cases}$$

Moreover, for any $x \in x$ we have

$$x = \sum_{\alpha \in A} c_\alpha x_\alpha$$

and

$$c_\alpha = \langle x, x_\alpha \rangle.$$

Basis $x\alpha$ in the theorem is called Hilbert basis (or orthogonal system) in X.

Examples of Hilbert basis: harmonics e^{ikt}, Legendre polynomials, wavelets, etc.

5. Fourier analysis

5.1. **Heat equation and Fourier series.** Fourier series arise very naturally while solving the heat equation:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}. \tag{17}$$

Here $u = u(t, x)$ is a temperature in the heat conducting rod of length L, t is time coordinate and x is the coordinate in the rod, $x \in [0, L]$ and a be some constant related to rod material.

The usual statement of problem is the following: to find solution $u(t, x)$ of (17) for which 1) the initial temperature is given and is determined by some function $u(0, x) = f(x), 0 < x < L$; 2) boundary conditions during heating are controlled by two functions; we consider particular case when boundary temperature is kept zero: $u(t, 0) = u(t, L) = 0$.

One of standard method of solution of this problem is separation of variables approach. We suppose

$$u(t, x) = A(x)B(t),$$

This gives equation:

$$AB' = a^2 A''B.$$

It is natural to expect that A is proportional to A'' and B' is proportional to B. Thus we have $A(x) = \sin(\alpha x + \beta)$ for some $\alpha, \beta \in \mathbb{R}$. Condition $A(0) = A(L) = 0$ implies

$$A(x) = \sin\left(\frac{k\pi x}{L}\right)$$

for some $k \in \mathbb{Z}$. Then

$$B' = -a^2 \frac{k^2\pi^2}{L^2} B$$

and

$$B(t) = e^{-\frac{k^2\pi^2a^2}{L^2}t}$$

At last we obtain following sequence of solutions of heat equation with trivial boundary condition:

$$U_k(t, x) = \sin\left(\frac{k\pi x}{L}\right) e^{-\frac{k^2\pi^2a^2}{L^2}t},$$

$k \in \mathbb{Z}$ (to get positive solution we need to add condition $k \geq 1$).

Therefore the general solution of (17) vanishing at the ends of the rod can be expressed as series

$$u(t,x) = \sum_{k=1}^{\infty} c_k U_k(t,x)$$

with some unknown coefficients. Now initial condition $f(x) = u(0,x)$ gives

$$f(x) = \sum_{k=1}^{\infty} c_k \sin\left(\frac{k\pi x}{L}\right)$$

The right side of the last equation is exactly the Fourier series of function $f(x)$. For solving (17) we need to be able to find coefficients c_k, which are called Fourier coefficients of function f. Further we consider this series in more mathematical details.

5.2. **Function space $L^2(\mathbb{R}/2\pi)$.** The points of space $L^2(\mathbb{R}/2\pi)$ are represented by measurable 2π-periodic functions $f : \mathbb{R} \to \mathbb{C}$:

$$f(t + 2\pi) = f(t), \forall t \in \mathbb{R},$$

for which integral

$$\frac{1}{2\pi}\int_0^{2\pi} |f(t)|^2 dt$$

is finite. In the space $L^2(\mathbb{R}/2\pi)$ we identify those functions f and g for which set

$$Z_0 = \{t \in \mathbb{R} | f(t) \neq g(t)\}$$

has zero Lebesgue measure. Remark also that $f(t) \in L^2(\mathbb{R}/2\pi)$ can be not defined in some zero Lebesgue measure set of values t. Because of this it is make no sense to speak about value $f(t)$ for for fixed t; you can freely replace $f(t)$ by any other value without changing f as element of $L^2(\mathbb{R}/2\pi)$; only the integrals $\int_a^b f(t)dt$ are correctly and uniquely defined for all $a < b$. You can find more rigorous definitions in previous section.

Now define scalar product $\langle , \rangle$ in $L^2(\mathbb{R}/2\pi)$:

$$\langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}dt$$

(remind that $\overline{z}$ denotes complex conjugation of complex number $z = x + iy$: $\overline{z} = x - iy$). Scalar product defines norm on $L^2(\mathbb{R}/2\pi)$:

$$\|f\|_2 = \sqrt{\langle f, f\rangle} = \left(\frac{1}{2\pi}\int_0^{2\pi} |f(t)|^2 dt\right)^{\frac{1}{2}}$$

Theorem 14. *Space $L^2(\mathbb{R}/2\pi)$ equipped by norm $\|\cdot\|_2$ is a (complex) Hilbert space.*

5.3. **Pure harmonics.** Consider following family of 2π-periodic functions $\mathbf{e}_k(t)$ for $k \in \mathbb{Z}$:

$$\mathbf{e}_k(t) = e^{ikt} = \cos(kt) + i\sin(kt).$$

Lemma 5. *Functions $\mathbf{e}_k(t)$ form orthogonal system in $L^2(\mathbb{R}/2\pi)$, that is:*

$$\langle \mathbf{e}_k, \mathbf{e}_l \rangle = \frac{1}{2\pi}\int_0^{2\pi} e^{i(k-l)t}dt = \begin{cases} 1, & \text{if } k = l, \\ 0, & \text{if } k \neq .l \end{cases}$$

Remark that factor $\frac{1}{2\pi}$ in definition of scalar product is chosen to satisfy condition $\langle \mathbf{e}_k, \mathbf{e}_k \rangle = 1$.

Functions $\mathbf{e}_k$, $k \in \mathbb{Z}$ are called *pure harmonics*.

5.4. **Fourier coefficients.** If function $f \in L^2(\mathbb{R}/2\pi)$ satisfies

$$f(x) = \sum_{k=-\infty}^{k=\infty} c_k \mathbf{e}_k$$

(convergence in the norm $\|\cdot\|_2$; we understand now this equality just formally, not thinking if convergity take place or not) then the right side is called Fourier series of f and coefficients c_k are Fourier coefficients of f.

To find Fourier coefficients we consider scalar product of left and right sides with harmonic $\mathbf{e}_k$ and use orthogonality:

$$\langle f, \mathbf{e}_k \rangle = c_k.$$

Therefore we obtain the following formula which we will consider as definition of *Fourier coefficients of function $f \in L^2(\mathbb{R}/2\pi)$*:

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}dt.$$

Lemma 6 (Riemann–Lebesgue lemma)**.**

$$\lim_{k\to\pm\infty} c_k = 0.$$

This lemma shows that "high frequency harmonics" becomes more and more insignificant and most of information about function is containing in finite part of harmonics.

To formulate the next very important result, let denote k-th Fourier coefficient of function f as $\hat{f}(k)$:

$$\hat{f}(k) = c_k = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}dt$$

(the real meaning of this notation we will understand later, when will study Fourier transform).

Lemma 7 (Parseval's formula)**.** *The following equality is valid for every $f, g \in L^2(\mathbb{R}/2\pi)$:*

$$\langle f, g\rangle = \sum_{k=-\infty}^{k=+\infty} \hat{f}(k)\overline{\hat{g}(k)}.$$

In particular,

$$\|f\|_2^2 = \sum_{k=-\infty}^{k=+\infty} |\hat{f}(t)|^2 = \sum_{k=-\infty}^{k=+\infty} |c_k|^2.$$

Parseval's lemma shows that we can interpret Fourier coefficients as elements of Hilbert space $l^2(\mathbb{R})$ and norm in $L^2(\mathbb{R}/2\pi)$ corresponds to standard Hilbert norm in $l^2(\mathbb{R})$. In some sense all information about f is contained in Fourier coefficients (we will see that it is not really true if we we wish to discretize f!).

5.5. **Example: The Unknown Strength Problem.** Consider some distant rod sending signal which is characterized by amplitude and frequency. The examples of such signal could be: light or radio wave from sun; reflected sunlight from the moon; radio waves in radar, etc.

We assume that rod has coordinate $x \in [-L, L]$, $2L$ is a length of the rod. Signal sent by point x is

$$f(x)e^{i\omega t}.$$

Here $f(x)$ is strength of signal (it depends of point on the rod); ω is frequency, t is a time.

Now consider a situation when rod is very distant, that is we receive signal in some point P which is on the distance D from the center of the rod. Making *far-field assumption* we can approximate the distance from P to point x by

$$D - x\cos\theta,$$

where $\theta \in [0, \pi]$ is an angle between the rod axe and vector passing from P to center of rod.

The signal receiving by P at time t was sent from point x at time $t - \frac{1}{c}(D - x\cos\theta)$, where c is a speed of propagation of the signal. This follows that signal received in point P from point x at time t is equal to

$$f(x)e^{i\omega(t-\frac{1}{c}(D-x\cos\theta))} = e^{i\omega(t-\frac{D}{c})}f(x)e^{i\frac{\omega\cos\theta}{c}x}$$

Point P receives signals from all points $x \in [-L, L]$ and resulting signal received by P at time t is

$$e^{i\omega(t-\frac{D}{c})} \int_{-L}^{L} f(x) e^{i\frac{\omega \cos\theta}{c}x} dx$$

Therefore the quantity which we can measure in P is

$$\int_{-L}^{L} f(x) e^{i\frac{\omega \cos\theta}{c}x} dx$$

Now if we could able to choose θ such that

$$\frac{\omega \cos\theta}{c} = -\frac{k\pi}{L}$$

then the result of our measurement is Fourier coefficient $\hat{f}(k)$ of function f.

Remark that the last equation has solution only if

$$|k| \leq \frac{L\omega}{\pi c}.$$

So we can measure only finite number of Fourier coefficients of f.

In signal processing the wavelength is defined as

$$\lambda = \frac{2\pi c}{\omega}.$$

Therefore one can measure $2N + 1$ Fourier coefficients, where N is maximal integer number such that $N \leq \frac{2L}{\lambda}$.

There is the way to increase N: to use higher frequency ω. Larger frequency can decrease wavelength λ and increase quotient $2L/\lambda$.

Technically, resolution of radar is proportional to wavelength λ, so to get more detailed information about $f(x)$ we need higher frequency ω. This problem of shortest possible wavelength was extremely important during WW2: "the side with the shortest wavelength would win the war" (Körner notes). Usual radar has wavelength counting by meters. The invention of cavity magnetron by British scientists during the WW2 made possible microwave radar having a wavelength 10cm, which could mounted on planes.

5.6. **Convergence of Fourier series.** Now let discuss in what sense we have

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \mathbf{e}_k(t) = \sum_{k=-\infty}^{\infty} c_k e^{ikt}.$$

Consider partial sums:

$$s_N(t) = \sum_{k=-N}^{N} c_k e^{ikt}$$

Remark that if we consider finite-dimensional subspace $L_N^2 \subset L^2(\mathbb{R}/2\pi)$ generated by vectors $\mathbf{e}_{-N}, \dots, \mathbf{e}_0 = 1, \dots, \mathbf{e}_N$, then s_N is exactly the orthogonal projection of f to L_N^2.

Then we have at $N \to \infty$ (using orthogonality of harmonics and Pythagoras' theorem)

$$\|f - s_N\|^2 = \|f\|^2 - \|s_N\|^2 = \|f\|^2 - \sum_{k=-N}^{k=N} |c_k|^2 \to 0$$

(the last limit follows from Parseval's formula). We obtain:

Theorem 15. *The Fourier series of function f converges to f in the sense of metric $\|\cdot\|_2$ on the space $L^2(\mathbb{R}/2\pi)$.*

The following theorem is much more hard to prove.

Theorem 16 (Carleson's theorem)**.** *The partial sums $s_N(t)$ of function $f \in L^2$ converges to $f(t)$ for almost t.*

The following two theorems allow to us to estimate the convergence rate of Fourier series using an information about smoothness of f.

Theorem 17. *Let r-derivative $f^{(r)}$ of function f exists, is continuous and the following integral is finite:*

$$V(f) = \int_0^{2\pi} |f'(t)|^2 dt < \infty,$$

then

$$|c_k| \leq \frac{V(f)}{2\pi |k|^{r+1}}, \forall k \neq 0.$$

Theorem 18. *If Fourier coefficients satisfy*

$$c_k = O\left(\frac{1}{2\pi |k|^{r+1+\varepsilon}}\right)$$

for some $\varepsilon > 0$ then function $f(t) = \sum_{k=-\infty}^{\infty} c_k e^{ikt}$ is at least r times continuously differentiable.

It is very important observation: differentiability of function f is related to asymptotic behaviour of its Fourier coefficients.

5.7. **Case of arbitrary segment** $[0, L]$**.** If we consider arbitrary segment $[0, L]$ then one can obtain analogous formulas for Fourier series. We state the most important of them in the next theorem.

Theorem 19. *Let $f : \mathbb{R} \to \mathbb{C}$ is a periodic function with period $L > 0$. Suppose $\int_0^L |f(t)|^2 dt < \infty$. Then Fourier coefficients and Fourier series is given by*

$$c_k = \frac{1}{L}\int_0^L f(t)e^{-\frac{2k\pi i}{L}x}dt, f(t) = \sum_{k=-\infty}^{k=\infty} c_k e^{\frac{2k\pi i}{L}t}dt,$$

and Prseval's formula is

$$\sum_{k=-\infty}^{k=\infty} |c_k|^2 = \frac{1}{L}\int_0^L |f(t)|^2 dt.$$

6. Fourier transform

6.1. **Fourier transform in** $L^1 = L^1(\mathbb{R})$**.** In this section we will work with functions defined on whole real line:

$$f : \mathbb{R} \to \mathbb{C}.$$

We will think of these functions as time signals.

Remind that space $L^1 = L^1(\mathbb{R})$ consists of the measurable functions f for which the integral

$$\|f\|_1 = \int_{-\infty}^{\infty} |f(t)|dt$$

is finite (certainly, we need to identify all functions which differ on the zero measure set only). Details are in section ?.

The Fourier transform of function $f \in L^1$ is defined by integral

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-i\xi t}dt, \xi \in \mathbb{R}. \tag{18}$$

We can interpret $\hat{f}$ in the following way: for fixed ξ value $\hat{f}(\xi)$ is the (complex) amplitude with which the "pure harmonics" $\mathbf{e}_\xi$ is represented in f. In case ξ being integer we actually obtain amplitude of pure harmonic, as in previous lectures. But we also obtain some amplitudes for harmonics with fractional and even irrational frequencies.

Theorem 20 (A version of Riemann–Lebesgue lemma for Fourier transform)**.** *The Fourier transform $\hat{f}$ of a function $f \in L^1$ is continuous and*

$$\lim_{\xi \to \pm\infty} \hat{f}(\xi) = 0.$$

6.2. **Translation, dilation and convolution.** For any time signal $f : \mathbb{R} \to \mathbb{C}$ define its translation by h to the right $T_h f$:

$$T_h f(t) = f(t-h), t \in \mathbb{R}.$$

Theorem 21.

$$\widehat{T_h f}(\xi) = e^{-i\xi h}\hat{f}(\xi).$$

Proof. Indeed,

$$\widehat{T_h f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} T_h f(t)e^{-i\xi t}dt =$$
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t-h)e^{-i\xi(t-h)}e^{-i\xi h}dt =$$
$$e^{-i\xi h}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t-h)e^{-i\xi(t-h)}d(t-h) = e^{-i\xi h}\widehat{f}(\xi).$$

Theorem 22.

$$\widehat{(e^{i\omega t}f)}(\xi) = \widehat{f}(t-\omega) = T_\omega \widehat{f}(\xi).$$

Proof.

$$\widehat{(e^{i\omega t}f)}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{i\omega t}f(t)e^{-i\xi t}dt =$$
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{-i(\xi-\omega)t}dt = \widehat{f}(\xi-\omega).$$

Now define dilation along time axe by factor $a \neq 0$:

$$D_a f(t) = f\left(\frac{t}{a}\right), t \in \mathbb{R}.$$

Theorem 23.

$$\widehat{D_a f}(\xi) = |a| D_{\frac{1}{a}}\widehat{f}(\xi).$$

Proof. Indeed, we have:

$$\widehat{D_a f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} D_a f(t)e^{-i\xi t}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f\left(\frac{t}{a}\right)e^{-i\xi t}dt.$$

Consider the change of variable $t = au$:

$$\widehat{D_a f}(\xi) = \frac{|a|}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(u)e^{-ia\xi u}du = |a|\widehat{f}(a\xi) = |a|D_{\frac{1}{a}}\widehat{f}(\xi).$$

Let $f, g \in L^1$. Define convolution $f * g$:

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(t-x)g(x)dx, t \in \mathbb{R}.$$

To understand what convolution means take g_ε be the function, supported in some interval $(-\varepsilon, \varepsilon)$, $\varepsilon > 0$ with the property

$$\int_{-\varepsilon}^{\varepsilon} g_\varepsilon(t)dt = 1.$$

Then convolution $f = f * g_\varepsilon$ is "smoothed" version of f, where every value $f(t)$ is replaced by the weighted mean of function f in the "window" $(t-\varepsilon, t+\varepsilon)$ with weight g.

Theorem 24.

$$\widehat{(f * g)}(\xi) = \sqrt{2\pi}\widehat{f}(\xi)\widehat{g}(\xi).$$

Proof.

$$\widehat{(f * g)}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f * g(t) e^{-i\xi t} dt =$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f(t-x)g(x)dx \right) e^{-i\xi t} dt =$$
$$\int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t-x) e^{-i\xi(t-x)} d(t-x) \right) g(x) e^{-i\xi x} dx =$$
$$\sqrt{2\pi}\widehat{f}(\xi) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} g(x) e^{-i\xi x} dx = \sqrt{2\pi}\widehat{f}(\xi)\widehat{g}(\xi).$$

Example 29 (Fourier transform of density of normal distribution)**.** *Consider density function of normal distribution with zero mean and unit standard deviation:*

$$N_{0,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, t \in \mathbb{R}.$$

Let $h(\xi) = \widehat{N_{0,1}}(\xi), \xi \in \mathbb{R}$. *Then*

$$\frac{dh}{d\xi} = \frac{1}{\sqrt{2\pi}} \frac{d}{d\xi} \int_{-\infty}^{+\infty} N_{0,1}(t) e^{-i\xi t} dt =$$
$$\frac{-i}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} t e^{-i\xi t} dt = \frac{i}{2\pi} \int_{-\infty}^{+\infty} e^{-i\xi t} d(e^{-\frac{t^2}{2}}) =$$
$$\frac{i}{2\pi} e^{-\frac{t^2}{2}} e^{-i\xi t} \Big|_{-\infty}^{+\infty} - \frac{i}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} d(e^{-i\xi t}) =$$
$$-\frac{\xi}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} e^{-i\xi t} dt = -\frac{\xi}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} N_{0,1}(t) e^{-i\xi t} dt = -\xi h(\xi).$$

Thus function $h(\xi)$ *satisfies to differential equation*

$$\frac{dh}{d\xi} = -\xi h \tag{19}$$

Equation (19) has general solution

$$h(\xi) = Ce^{-\frac{\xi^2}{2}}$$

for some constant C. *We have*

$$C = h(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} N_{0,1}(t) dt = \frac{1}{\sqrt{2\pi}}$$

by property of probability density function and therefore $C = 1$ *and we obtain:*

$$\widehat{N_{0,1}}(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}},$$

that is Fourier transform of normal distribution density is identical copy of itself:

$$\widehat{N_{0,1}} = N_{0,1}.$$

Example 30 (Fourier transform of indicator function)**.** *For* $a > 0$ *consider indicator function*

$$1_{[-a,a]}(t) = \begin{cases} 1, & -a \le t \le a \\ 0, & \textit{otherwise.} \end{cases}$$

Then for $\xi \neq 0$ *we have*

$$\widehat{1_{[-a,a]}}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-i\xi t} dt = \frac{i}{\xi\sqrt{2\pi}} \int_{-a}^{a} de^{-i\xi t} =$$
$$\frac{i}{\sqrt{2\pi}} \frac{e^{-i\xi a} - e^{i\xi a}}{\xi} = \sqrt{\frac{2}{\pi}} \frac{\sin(a\xi)}{\xi}.$$

Formally we can not use last formula for $xi = 0$. But Theorem 20 state that Fourier transform is always continuous therefore

$$\widehat{1_{[-a,a]}}(0) = \lim_{\xi\to 0}\widehat{1_{[-a,a]}}(\xi) = \sqrt{\frac{2}{\pi}}\lim_{\xi\to 0}\frac{\sin(a\xi)}{\xi} = a\sqrt{\frac{2}{\pi}}.$$

Define sinc function:

$$sinc(t) = \begin{cases} sin(t)/t, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Then we can conclude that

$$\widehat{1_{[-a,a]}}(\xi) = a\sqrt{\frac{2}{\pi}}sinc(a\xi).$$

Example 31 (Fourier transform of exponential tail)**.** *For $a > 0$ let*

$$f(t) = e^{-a|t|}, t \in \mathbb{R}.$$

Then

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-a|t|}e^{-i\xi t}dt =$$

$$\frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} e^{-(a+i\xi)t}dt + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{(a-i\xi)t}dt =$$

$$\frac{1}{\sqrt{2\pi}}\left(-\frac{1}{a+i\xi}e^{-(a+i\xi)t}\Big|_0^{+\infty} + \frac{1}{a-i\xi}e^{(a-i\xi)t}\Big|_{-\infty}^{0}\right) =$$

$$\frac{1}{\sqrt{2\pi}}\left(\frac{1}{a+i\xi} + \frac{1}{a-i\xi}\right) = \frac{1}{\sqrt{2\pi}}\frac{2a}{a^2+\xi^2}.$$

Therefore

$$\widehat{(e^{-a|t|})}(\xi) = \frac{1}{\sqrt{2\pi}}\frac{2a}{a^2+\xi^2}.$$

6.3. **Fourier transform in $L^2 = L^2(\mathbb{R})$.** Remind that norm in $L^2 = L^2(\mathbb{R})$ comes from scalar product:

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(t)\overline{g(t)}dt,$$

and L^2 consists of those signals for which L^2-norm

$$\|f\|_2 = \left(\int_{-\infty}^{\infty}|f(t)|^2dt\right)^{\frac{1}{2}}$$

is finite (we again identify "equivalent functions", details in Section ?).

Theorem 25 (Schwarz's inequality)**.** *For any $f, g \in L^2$*

$$\langle f, g\rangle \leq \|f\|_2\|g\|_2.$$

Trying formally define Fourier transform for $f \in L^2$ we find it difficult: function $e^{-i\xi t}$ is not an element of L^2. Corresponding integral can be infinite. Using the fact that $L^1 \cap L^2$ is dense in L^2 we can expand formula (18) from $L^1 \cap L^2$ to L^2 and obtain one-to-one map

$$\mathcal{F} : L^2 \to L^2$$

which is called Fourier transform on L^2. Here we formulate main properties of this transform.

Theorem 26 (Parceval-Plancherel's)**.**

$$\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle,$$

in particular,

$$\|f\|^2 = \|\hat{g}\|^2.$$

Both formulas can be written in integral form:

$$\int_{-\infty}^{\infty} f(t)\overline{g}(t)dt = \int_{-\infty}^{\infty} \hat{f}(t)\overline{\hat{g}}(t)dt,$$
$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} \|\hat{f}(t)\|^2 dt.$$

The following *inversion formula* shows that signal f can be restored from all pure oscillations of all possible frequencies $\xi \in \mathbb{R}$.

Theorem 27 (Inversion formula)**.** *If f and $\hat{f}$ are both in L^1 then*

$$f(t) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi t} d\xi$$

almost everywhere (and especially this equality holds in those points t in which f is continuous).

We can conclude from the last theorem that

$$f(t) = \widehat{\hat{f}}(-t). \tag{20}$$

Example 32. *In Example 31 we showed that*

$$\widehat{(e^{-a|t|})}(\xi) = \frac{1}{\sqrt{2\pi}} \frac{2a}{a^2 + \xi^2}.$$

Combining this with (20) we have

$$\left(\widehat{\frac{2a}{a^2 + t^2}}\right)(\xi) = \sqrt{2\pi} e^{-a|\xi|}, \xi \in \mathbb{R}, a > 0.$$

6.4. **Fourier transform, differentiating and Schwarz's space.** Let f be C^1-function and $f, f' \in L^1$. Remark that this implies

$$\lim_{t\to\pm\infty} f(t) = 0.$$

Then

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f'(t)e^{-i\xi t}dt = \frac{1}{\sqrt{2\pi}} f(t)e^{-i\xi t}\Big|_{-\infty}^{+\infty} + \frac{i\xi}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{-i\xi t}dt = \frac{i\xi}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} f(t)e^{-i\xi t}dt$$

Therefore

$$\widehat{f}'(\xi) = i\xi\widehat{f}(\xi), \xi \in \mathbb{R}.$$

We can continue differentiating assuming $f \in C^r$ and obtain:

Theorem 28. *If f is C^r-function and $f^{(k)} \in L^1$ for $0 \le k \le r$ then*

$$\widehat{f^{(r)}}(\xi) = (i\xi)^r\widehat{f}(\xi), \xi \in \mathbb{R}, r \ge 0.$$

In particular (using Theorem 20) we have

$$\lim_{t\to\pm\infty} |\xi|^r\widehat{f}(\xi) = 0.$$

What one can say about differentiating of Fourier transform itself? We have

$$(\widehat{f})'(\xi) = \frac{1}{\sqrt{2\pi}}\frac{d}{d\xi}\int_{-\infty}^{+\infty} f(t)e^{-i\xi t}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(-it)f(t)e^{-i\xi t}dt = \frac{-i}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} tf(t)e^{-i\xi t}dt = -i\widehat{(tf)}(\xi)$$

So we proved

$$\widehat{(tf)}(\xi) = i(\widehat{f})'(\xi), \xi \in \mathbb{R}.$$

Repeating this by induction we obtain:

Theorem 29. *Let $f \in L^1$ and*

$$\int_{-\infty}^{+\infty} |t|^r|f(t)|dt < \infty.$$

Then

$$\widehat{(t^r f)}(\xi) = i^r(\widehat{f})^{(r)}(\xi), \xi \in \mathbb{R}.$$

Theorem 27 seems to be very hard for use. To formula for restoring signal f by oscillations $\hat{f}$ we need already to know some good properties $\hat{f}$ which is not yet has been restored!

We can avoid this difficulty if consider subspace of much more "good" signals. The idea here is following: the smoother the signal $f(t)$ for $|t| \to \infty$, the the faster decay of $\hat{f(\xi)}$ for $|\xi| \to \infty$. Inverse statement is also true: the faster decay of signal, the smoother oscillations (compare these principles to Theorems 17 and 18). So we come to idea to consider space of signals with very fast decay of the time signal $f(t)$ for $|t| \to \infty$ — Schwarz's space.

By definition Schwarz's space $\mathcal{S}$ consists of those functions f which have derivatives of all orders and for $|t| \to \infty$ all derivatives decay faster to zero than any power $\frac{1}{|t|^n}$, $n \in \mathbb{N}$:

$$\sup_{t \in \mathbb{R}} |t|^n |f^{(r)}(t)| < \infty \text{ for any } n, m \in \mathbb{N}.$$

We have the following result.

Theorem 30. *If $f \in \mathcal{S}$ then $\hat{f} \in \mathcal{S}$ and corresponding map*

$$\mathcal{F} : \mathcal{S} \to \mathcal{S}$$

is a bijection.

In other words Fourier transform is a one-to-one transformation of Schwarz's space.

Theorem 31. *Let $f, g \in L^1$ are C^1-functions. Then*

$$(f * g)' = f' * g = f * g'.$$

Proof. We have

$$(f * g)'(t) = \frac{d}{dt} \int_{-\infty}^{+\infty} f(t-x)g(x)dx =$$

$$\int_{-\infty}^{+\infty} f'(t-x)g(x)dx = (f' * g)(t) =$$

$$-f(t-x)g(x)\Big|_{x=-\infty}^{x=+\infty} + \int_{-\infty}^{+\infty} f(t-x)g'(x)dx = (f * g')(t).$$

6.5. **Application Fourier transform for solution of some differential equations.**

Example 33. *Let's solve a problem*

$$-u'' + u = f(t), \quad \lim_{t \to \pm\infty} u(t) = 0.$$

Applying Fourier transform to both sides of equation we have

$$\xi^2 \widehat{u} + \widehat{u} = \widehat{f}.$$

Equation becomes purely algebraic and we obtain

$$\widehat{u}(\xi) = \frac{\widehat{f}(\xi)}{1+\xi^2} = \widehat{f}(\xi)\frac{1}{1+\xi^2} = \frac{\sqrt{2\pi}}{2}\widehat{f}(\xi)\widehat{e^{-|t|}}(\xi) = \frac{1}{2}\widehat{(e^{-|t|} * f)}(\xi).$$

Then

$$u(t) = \frac{1}{2}(e^{-|t|} * f)(t) = \frac{1}{2}\int_{-\infty}^{+\infty} e^{-|t-x|}f(x)dx.$$

Example 34 (Laplace equation)**.** *Consider following problem for Laplace equation:*

$$\Delta u = u_{xx} + u_{yy} = 0,\ x \in \mathbb{R}, y > 0,\ u(x,0) = f(x),\ \lim_{y\to+\infty} u(x,y) = 0.$$

Consider Fourier transform in variable x:

$$\widehat{u}(\xi, y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} u(x,y)e^{-i\xi x}dx. \tag{21}$$

Applying Fourier transform to equation we obtain

$$-\xi^2\widehat{u} + \widehat{u}_{yy} = 0,\ \widehat{u}(\xi,0) = \widehat{f}(\xi),\ \lim_{y\to+\infty}\widehat{u}(\xi,y) = 0.$$

This differential equation has the following general solution:

$$\widehat{u}(\xi, y) = A(\xi)e^{-|\xi|y} + B(\xi)e^{|\xi|y}$$

for some functions $A(\xi), B(\xi)$. Asymptotic condition for $u(x,y)$ implies $B = 0$ and boundary condition gives

$$\widehat{u}(\xi, y) = \widehat{f}(\xi)e^{-|\xi|y} = \frac{1}{\sqrt{2\pi}}\widehat{f}(\xi)\widehat{\left(\frac{2y}{y^2+\xi^2}\right)}$$

(we used Example 32 here). Then

$$u(x,y) = \frac{1}{2\pi}\left(\frac{2y}{y^2+\xi^2}\right) * f = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{yf(z)dz}{y^2+(x-z)^2}.$$

Example 35 (Transport equation)**.** *Let's solve the following problem for transport equation:*

$$u_t + cu_x = 0,\ x \in \mathbb{R}, t > 0,\ u(x,0) = f(x).$$

Consider the same Fourier transform as in 21 (replacing y by t). Then

$$\widehat{u}_t + ic\xi\widehat{u} = 0,\ \widehat{u}(\xi,0) = \widehat{f}(\xi).$$

General solution of this equation gives

$$\widehat{u}(\xi, t) = A(\xi)e^{-ic\xi t}$$

for some function $A(\xi)$. Boundary condition gives

$$\widehat{u}(\xi, t) = \widehat{f}(\xi)e^{-ic\xi t}.$$

We see that right side of the last equation looks like translation formula (Theorem 21). Then

$$\widehat{u}(\xi,t) = \widehat{T_{ct}f}(\xi).$$

Therefore

$$u(x,t) = T_{ct}f(x) = f(x-ct).$$

Example 36 (Wave equation)**.** *Consider the following problem for wave equation:*

$$u_{tt} = c^2 u_{xx},\ x \in \mathbb{R}, t > 0,\ u(x,0) = f(x),\ u_t(x,0) = g(x).$$

Considering Fourier transform 21 (replacing y by t) we obtain:

$$\widehat{u}_{tt} + c^2\xi^2\widehat{u} = 0,\ \widehat{u}(\xi,0) = \widehat{f}(\xi),\ \widehat{u}_t(\xi,0) = \widehat{g}(\xi).$$

Then

$$\widehat{u}(\xi,t) = A(\xi)\cos(c\xi t) + B(\xi)\sin(c\xi t)$$

for some functions $A(\xi), B(\xi)$. Applying initial conditions we have

$$\widehat{u}(\xi,t) = \widehat{f}(\xi)\cos(c\xi t) + \frac{\widehat{g}(\xi)}{\xi}\sin(c\xi t) =$$
$$\frac{1}{2}\widehat{f}(\xi)\left(e^{ic\xi t} + e^{-ic\xi t}\right) + \frac{\widehat{g}(\xi)}{2i\xi}\left(e^{ic\xi t} - e^{-ic\xi t}\right) =$$

To compute the first part of the last formula we use Theorem 21:

$$\frac{1}{2}\widehat{f}(\xi)\left(e^{ic\xi t} + e^{-ic\xi t}\right) = \frac{1}{2}\widehat{T_{-ct}f}(\xi) + \frac{1}{2}\widehat{T_{ct}f}(\xi).$$

To compute second part we rewrite formula from Theorem 28:

$$\frac{\widehat{h}(\xi)}{i\xi} = \int_{-\infty}^{\xi}\widehat{h}(\eta)d\eta$$

for any function $h \in L^2$. Then

$$\frac{\widehat{g}(\xi)}{2i\xi}\left(e^{ic\xi t} - e^{-ic\xi t}\right) = \frac{\widehat{T_{-ct}g}(\xi)}{2i\xi} - \frac{\widehat{T_{ct}g}(\xi)}{2i\xi} =$$
$$\frac{1}{2}\int_{-\infty}^{\xi}\left(\widehat{T_{-ct}g}(\eta) - \widehat{T_{ct}g}(\eta)\right)d\eta$$

Thus

$$u(x,t) = \frac{1}{2}\left(f(x+ct) + f(x-ct)\right) + \frac{1}{2}\int_{-\infty}^{\xi}\left(g(\eta+ct) - g(\eta-ct)\right)d\eta =$$
$$\frac{1}{2}\left(f(x+ct) + f(x-ct)\right) + \frac{1}{2}\int_{-\infty}^{\xi+ct} g(\eta)d\eta - \frac{1}{2}\int_{-\infty}^{\xi-ct} g(\eta)d\eta =$$
$$\frac{1}{2}\left(f(x+ct) + f(x-ct)\right) + \frac{1}{2}\int_{\xi-ct}^{\xi+ct} g(\eta)d\eta.$$

6.6. **The Heisenberg uncertainty principle.**

Theorem 32 (The Heisenberg uncertainty principle)**.** *Let $f \in L^2$. Then*

$$\|tf\|_2 \cdot \|\xi\hat{f}\|_2 \geq \|f\|_2^2.$$

I other words, a time signal f and its Fourier transform $\hat{f}$ cannot be simultaneously localized in a small domains of t- and ξ-axes.

This principle manifests itself in many situations. We give several examples:

1. Trying localize f we can horizontally compress graph f, that is replace f with $D_a f$ for small enough $a > 0$. Theorem 23 shows that graph of $\hat{f}$ then is stretched in horizontal directions with a factor $\frac{1}{a}$ and additionally is flattened by vertical scaling.

2. If we cutoff signal $f(t)$ for $|t| > A > 0$ then its Fourier transform $\hat{f}$ is non-zero for whole $\mathbb{R}$, and moreover, $\hat{f}$ is not absolutely integrable for $|\xi| \to \infty$.

3. A bandlimited signal is a signal $f(t)$ with compactly supported Fourier transform, that is $f(\xi) = 0$ for $|\xi| > \Omega$ for some $\Omega > 0$. Then signal f can not have compact support.

7. Wavelet transform

7.1. **Mother wavelet.** A *mother wavelet* (or simply *wavelet*) is a function $\psi : \mathbb{R} \to \mathbb{C}$ satisfying the conditions:

$$\psi \in L^2, \|\psi\|_2 = 1,$$
$$C_\psi = \int_{\mathbb{R},\xi\neq 0} \frac{\|\hat{\psi}(\xi)\|^2}{|\xi|} d\xi < \infty. \tag{22}$$

Speaking non-formally mother wavelet gives the template for "key patterns" of our signal. All such key patterns are dilated and translated

copies of mother wavelet. Of course in case Fourier transform we have sinusoid as mother wavelet.

When the second condition of (22) holds? We have the following useful lemma.

Lemma 8. *If $\psi \in L^2, t\psi \in L^1$, then the second condition of (22) is equivalent to any of the next two conditions:*

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0,$$

$$\widehat{\psi}(0) = 0.$$

We see that the mean of mother wavelet on $\mathbb{R}$ is equal to zero.

7.2. **Wavelet transform.** Fix mother wavelet ψ. Let $f \in L^2$ be a time signal. The function

$$\mathcal{W}f(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)}dt$$

is called *the wavelet transformation* of f, $a \neq 0$. Remark that often wavelet transform is considered only for $a > 0$.

Note that for one-dimensional time signal f the wavelet transform is a function Wf of two variables a and b (in contrast to Fourier transform which depends of one variable ξ).

Define

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)$$

Then $\psi_{a,b}$ is a "key pattern" we told about. Remark that

$$\|\psi_{a,b}\|_2 = \int_{-\infty}^{+\infty} |\psi_{a,b}(t)|^2 dt = \frac{1}{|a|}\int_{-\infty}^{+\infty} \left|\psi\left(\frac{t-b}{a}\right)\right|^2 dt =$$
$$\frac{1}{|a|}\int_{-\infty}^{+\infty} \left|\psi\left(\frac{u}{a}\right)\right|^2 du = \int_{-\infty}^{+\infty} |\psi(v)|^2 dv = 1.$$

and

$$Wf(a,b) = \langle f, \psi_{a,b}\rangle.$$

Let try to compute Fourier transform of wavelet $\psi_{a,b}$. By translation and dilation rules we have:

$$\widehat{\psi_{a,b}}(\xi) = \frac{1}{\sqrt{|a|}}\widehat{D_a T_b \psi}(\xi) = \sqrt{|a|}D_{\frac{1}{a}}\widehat{T_b\psi}(\xi) =$$
$$\sqrt{|a|}e^{-i\xi b}D_{\frac{1}{a}}\widehat{\psi}(\xi) = \sqrt{|a|}e^{-i\xi b}\psi(a\xi).$$

Then (using Parxebal's formula)

$$\mathcal{W}f(a,b) = \langle f, \psi_{a,b}\rangle = \langle \widehat{f}, \widehat{\psi_{a,b}}\rangle =$$
$$\sqrt{|a|}\int_{-\infty}^{+\infty} \widehat{f}(\xi)e^{i\xi b}\overline{\widehat{\psi}(a\xi)}d\xi = \widehat{F_a}(b),$$

where

$$F_a(\xi) = \sqrt{|a|}\sqrt{2\pi}\widehat{f}(\xi)\overline{\widehat{\psi}(a\xi)}$$

So we prove

Theorem 33. *For fixed $a \neq 0$ the function*

$$\mathcal{W}f(a,\cdot) : b \mapsto \mathcal{W}f(a,b)$$

is the Fourier transform of the function F_a defined in (7.2). In particular, $\mathcal{W}f$ is continuous on vertical lines $a = const$ and

$$\lim_{b\to\pm\infty} \mathcal{W}f(a,b) = 0.$$

Example 37 (Haar wavelet)**.** *The Haar wavelet is the following mother wavelet:*

$$\psi_{Haar}(t) = \begin{cases} 1 & (0 \le t \le \frac{1}{2}) \\ -1 & (\frac{1}{2} \le t \le 1) \\ 0 & (otherwise) \end{cases}$$

Denoting $\psi = \psi_{Haar}$ and assuming $a > 0$ for simplicity we have

$$\sqrt{|a|}\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) = \begin{cases} 1 & (b \le t \le b + \frac{a}{2}) \\ -1 & (b + \frac{a}{2} \le t \le b + a) \\ 0 & (otherwise) \end{cases}$$

Then

$$\mathcal{W}f(a,b) = \frac{1}{\sqrt{|a|}}\left(\int_b^{b+\frac{a}{2}} f(t)dt - \int_{b+\frac{a}{2}}^{b+a} f(t)dt\right) =$$
$$\sqrt{|a|}\left(\frac{2}{a}\int_b^{b+\frac{a}{2}} f(t)dt - \frac{2}{a}\int_{b+\frac{a}{2}}^{b+a} f(t)dt\right)$$

We see from the last expression that $\mathcal{W}f(a,b)$ (up to factor $\sqrt{|a|}$) is the difference of two integral means of function f: first integral mean is taken on the left half of interval $(b, b+a)$; the second is taken on the right half.

Example 38 (Mexican hat)**.** *Consider mother wavelet*

$$\psi(t) = \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}(1-t^2)e^{-\frac{t^2}{2}}$$

(factor is taken in such a way that $\|\psi\|_2 = 1$).

It easy to check that $\psi(t) = -cN''_{0,1}(t)$ *for appropriate constant* $c > 0$. *Then*

$$\widehat{\psi}(\xi) = -c(i\xi)^2\widehat{N_{0,1}}(\xi) = c\xi^2 N_{0,1}(\xi).$$

Then $\widehat{\psi}(0) = 0$ *and by Lemma 8 the* ψ *is indeed a wavelet. This wavelet is called Mexican hat because of geometric shape of graph.*

7.3. **Plancherel theorem.** In order to formulate analogue of Plancherel formula for wavelets we need to have scalar product of functions depending of parameters a, b. To do it we need a measure on the plane $H = \{(a, b)|a \neq 0\}$. This measure should be invariant with respect to all translations and dilations. To find this measure let remember that every element $(a, b) \in H$ acts on t as:

$$(a, b) : t \mapsto at + b.$$

So we can identify H with the group of all affine transformations of $\mathbb{R}$. Let understand what is a composition of two affine transformations $(a, b) \in H$ and $(c, d) \in H$? We have

$$t \mapsto at + b \mapsto c(at + b) + d = (ac)t + (bc + d).$$

We see that composition of (a, b) and (c, d) is pair $(ac, bc+d) \in H$. We denote this composition as $(a, b) \cdot (c, d) = (ac, bc + d)$.

We assume that measure on H has the form $h(a, b)dadb$ for some integrable function h on H. Then if $U \subset H$ is measurable then we can find the measure of U by formula:

$$|U| = \int_U h(a, b)dadb.$$

We say that $h(a, b)dadb$ is the Haar measure if for every $U \subset H$ and for every $(c, d) \in H$ the measure $|U|$ is invariant with respect to all compositions by element (c, d):

$$|U| = |U \cdot (c, d)|$$

Theorem 34. *Every Haar measure is proportional to the measure*

$$d\mu_0 = \frac{dadb}{a^2}.$$

We will use measure $d\mu_0$ as natural measure on H.

Proof. First of all let check that $d\mu_0$ is the Haar measure. Indeed, for avery $U \subset H$ and $(c, d) \in H$ we have:

$$|U \cdot (c, d)| = \int_{U\cdot(c,d)} \frac{dadb}{a^2}.$$

Consider change of variables: $(a_1, b_1) = (a, b) \cdot (c, d) = (ac, bc + d)$. Then $da_1 = cda$, $db_1 = cdb$ and $da_1 db_1 = c^2 dadb$ and we obtain:

$$|U \cdot (c, d)| = \int_U \frac{da_1 db_1}{c^2 a^2} = \int_U \frac{da_1 db_1}{a_1^2} = |U|.$$

Now let show that any other Haar measure is proportional to $d\mu_0$. Take $U \subset H$ and consider some Haar measure $d\mu_1$. Then there exists some function $h(a, b)$ on H such that $d\mu_1 = h(a, b)d\mu_0$ and we have

$$\int_U h(a, b)d\mu_0 = \int_U d\mu_1 = \int_{U\cdot(c,d)} d\mu_1 = \int_{U\cdot(c,d)} h(ac, bc + d)d\mu_0.$$

Thus $h(a, b) = h(ac, bc+d)$ for all $(c, d) \in H$. Then for any $(a_1, b_1) \in H$ we take $c = a_1/a$, $d = b_1 - b * c$ and have

$$h(a_1, b_1) = h(ac, bc + d) = h(a, b).$$

Therefore h is a constant and $d\mu_1$ is proportional to $d\mu_0$.

Now we can define scalar product of two functions $u = u(a, b)$ and $v = v(a, b)$ by the following formula:

$$\langle u, v\rangle_H = \int_H u(a, b)\overline{u(a, b)}\frac{dadb}{a^2}.$$

This gives a structure of Hilbert space on the space $L^2(H, d\mu_0)$.

Theorem 35. *Let ψ is arbitrary mother wavelet and W is wavelet transform. Then*

$$\langle Wf, Wg\rangle_H = C_\psi \langle f, g\rangle,$$

where

$$C_\psi = \int_{\mathbb{R},\xi\neq 0} \frac{\|\hat{\psi}(\xi)\|^2}{|\xi|}d\xi$$

Theorem 36 (Inversion formula)**.**

$$f(x) = \frac{1}{C_\psi}\int_H \mathcal{W}f(a, b)\psi_{a,b}(x)\frac{dadb}{a^2}.$$

Theorem 37 (Decay of the wavelet transform)**.** *Fix wavelet ψ, such that $t\psi \in L^1$. Let f be a time signal and $f \in L^2$ is bounded globally and is Hölder continuous in some point b (that is there is $\alpha \in [0, 1]$ such that in a neighbourhood of b an estimate of the form*

$$|f(t) - f(b)| \leq C|t - b|^\alpha$$

holds). Then

$$|\mathcal{W}f(a, b)| \leq C'|a|^{\alpha+\frac{1}{2}}.$$

Theorem 38. *Fix wavelet ψ with a compact support (this menas that set $\{t|\psi(t) \neq 0\}$ is bounded). Let $f \in L^2$ be a time signal whose wavelet transform satisfies an estimate of the form*

$$|\mathcal{W}f(a,b)| \leq C|\alpha|^{\alpha+\frac{1}{2}}$$

for some $\alpha \in [0,1]$. Then f is globally Hölder continuous with exponent α.

7.4. **Sampling of wavelet transform.** One of the key features of wavelet transform is a good adaptation to logarithmic scale. Let choose zoom step $\sigma > 1$ (usually $\sigma = 2$). Consider grid

$$a_r = \sigma^r, r \in \mathbb{Z}.$$

Choose now base step $\beta > 0$ on axis b. The step on the b-grid depends on a(!):

$$b_{r,k} = k\sigma^r\beta, k \in \mathbb{Z}.$$

So the goal of *fast wavelet transform* (FWT) is to compute coefficients

$$c_{r,k} = \mathcal{W}f(a_r, b_{r,k})$$

and conversely, to reconstruct signal f by its coefficients $c_{r,k}$.

Most used "key pattern":

$$\psi_{r,k}(t) = \frac{1}{2^{\frac{r}{2}}}\psi\left(\frac{t-2^rk}{2^r}\right). \tag{23}$$

7.5. **Example: the Haar wavelet.** The *Haar wavelet* is the following mother wavelet:

$$\psi(t) = \begin{cases} 1 & (0 \leq t \leq \frac{1}{2}) \\ -1 & (\frac{1}{2} \leq t \leq 1) \\ 0 & (\text{otherwise}) \end{cases}$$

It is evident that:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0, \ \int_{-\infty}^{+\infty} |\psi(t)|^2 dt = 1.$$

Formula (23) shows that wavelet $\psi_{r,k}$ is non-zero on the segment

$$I_{r,k} = [2^rk, 2^r(k+1)).$$

The length of $I_{r,k}$ is equal to 2^r; wavelet $\psi_{r,k}$ is positive and equal to $2^{-r/2}$ on the left half of $I_{r,k}$; and is negative and equal to $-2^{-r/2}$ on the right half of $I_{r,k}$. The larger r, the longer interval, the wider "key pattern".

Theorem 39. *Functions $\psi_{r,k}$, $k, r \in \mathbb{Z}$ constitutes an orthogonal basis in the spece $L^2(\mathbb{R})$.*