Chapter 3
Communication in the presence of noise

3.1 Discrete recording of a continuous signal — concretization

3.1.1 Energy and mean power of a signal

Recall that the sampling theorem (the Kotel'nikov formula)

$$f(t) = \sum_{-\infty}^{\infty} f\left(\frac{k}{2W}\right) \frac{\sin 2\pi W(t - \frac{k}{2W})}{2\pi W(t - \frac{k}{2W})} \tag{3.1}$$

recovers the signal, which is a function $f \in L_2(\mathbb{R})$ with compactly supported spectrum of frequencies ν not exceeding W Hertz from the set of sample values $f(t_k)$ at the points $t_k = k\Delta$, where $\Delta = \frac{1}{2W}$ is the sampling time interval (Nyquist interval), which depends on W.

The wider the frequency band the more complex the function f can be and the more frequently one needs to take samples in order to adequately encode it discretely and recover it, but then the more information it (that is, the signal) can carry.

The function $\operatorname{sinc} t = \frac{\sin t}{t}$ is basic in the expansion (3.1); this function, as we already know, has constant spectrum equal to 1, on the unit interval of frequencies and is the instrumental function of an ideal low-frequency filter with unit pass-band. Thus the sampling function sinc is realized as the response of such a filter to a unit impulse realized at time $t = 0$.

The corresponding function $e_k(t) = \operatorname{sinc} 2\pi W(t - \frac{k}{2W}) = \frac{\sin 2\pi W(t - \frac{k}{2W})}{2\pi W(t - \frac{k}{2W})}$ has spectrum $\check{e}_k(\nu) = \frac{1}{2\pi W} \exp(-i\frac{\pi}{W}k\nu)$ and frequency band $0 \le \nu \le W$ ($|\nu| \le W$).

One can conclude from the orthogonality of the functions $\check{e}_k$ on the interval $[-W, W]$ (or on any interval of length $2W$) and Parseval's equality for the Fourier transform that the functions e_k themselves are orthogonal

V. Zorich, *Mathematical Analysis of Problems in the Natural Sciences*,
DOI 10.1007/978-3-642-14813-2_6, © Springer-Verlag Berlin Heidelberg 2011

in the space $L_2(\mathbb{R})$ and $\|e_k\|^2 = \frac{1}{2W}$. Hence we can infer from the equality $f = \sum_{-\infty}^{\infty} x_k e_k$ that $\|f\|^2 = \frac{1}{2W}\sum_{-\infty}^{\infty} x_k^2$.

In practice the signal f has a certain finite duration T, that is, $f(t) \equiv 0$ outside the interval $0 \le t \le T$. This condition is incompatible with the condition that the spectrum of f be compactly supported. However, one can assume that the values $f(t)$ of the function are small outside the interval $[0,T]$ and the sample values of the function outside this interval are set equal to zero.

Then the equality $f = \sum_{-\infty}^{\infty} x_k e_k$ is replaced by $f(t) = \sum_{k=1}^{2WT} x_k e_k(t)$, where $t \in [0,T]$, $x_k = f(k\Delta)$ and $\Delta = \frac{1}{2W}$. This signal f is written by the vector $x = (x_1,...,x_n) \in \mathbb{R}^n$ of its sample values, where $n = 2WT$

Under the same conditions Parseval's equality

$$\int_{-\infty}^{\infty} f^2(t)dt = \sum_{-\infty}^{\infty} x_k^2 \|e_k\|^2 = \frac{1}{2W}\sum_{-\infty}^{\infty} x_k^2$$

is replaced by the equality

$$\int_0^T f^2(t)dt = \sum_1^n x_k^2 \|e_k\|^2 = \frac{1}{2W}\sum_1^n x_k^2 = \frac{1}{2W}\|x\|^2.$$

Here, to within a specific uniform factor, the integral gives the energy (work) of the signal f (for example, when f is realized as the drop in voltage on a unit resistance). Hence the mean power P of the signal f over the time interval $[0,T]$ is

$$P = \frac{1}{T}\int_0^T f^2(t)dt = \frac{1}{2WT}\|x\|^2 = \frac{1}{n}\|x\|^2.$$

Thus, $\|x\|^2 = nP = 2WTP$ and P can be interpreted as the mean power required at one coordinate of the vector x, that is, one sample value of the signal f.

Thus, signals of duration T with compactly supported spectrum in a frequency band W whose mean power is at most P in the vector representation $x = (x_1,...,x_n)$ turn out to be located in a ball $B(0,r) = B(r) \subset \mathbb{R}^n$ of radius $r = \sqrt{2WTP} = \sqrt{nP}$ with centre at the origin of the Euclidean space $\mathbb{R}^n$ of dimension $n = 2WT$.

3.1.2 Quantization by levels

The measurement of the sample value of a signal f is performed from a certain threshold (limiting) precision ε. If the amplitude of any signal to be transmitted is not greater than A (that is, $|f|(t) \le A$ for $t \in [0,T]$), then, by endowing the interval $[-A,A]$ with a uniform network of points (levels)

with mesh size ε, for $f(t)$ we can take the point of this network nearest to $f(t)$. The values of $f(t)$ turn out to be quantized by levels, the number of which is $\alpha = \frac{2A}{\varepsilon}$ (we take α to be an integer greater than 1). The word $x = (x_1,...,x_n)$ corresponding to the signal f, which consists of n letters x_k will be written in an alphabet having α different characters. In all there are α^n such different words x. If $n = 2WT$ and W and T are large numbers, then α^n is enormous.

3.1.3 *Ideal multilevel communication channel*

Under these conditions after time T one can distinguish $M = \alpha^{2WT}$ (and not more) different signals $f \sim x = (x_1,...,x_n)$, that is, one can determine one definite signal-word-message out of the M possible ones.

The binary notation $x^0 = (x_1^0,\dots,x_m^0)$, which distinguishes M objects, requires $m = \log_2 M$ symbols 0, 1 (we take m to be an integer). The information about the next coordinate of the binary vector (if the coordinates are on an equal footing and their possible values 0, 1 are equally likely) is taken as the elementary unit of information and is called the *bit*. If we could without any error receive and transmit vectors (words) encoding our M messages, then in time T we could distinguish M objects (signals, messages). The speed of transmitting information (on the choice of one of the M possible objects) along such an ideal communication channel (and with such an encoding) measured in bits per second would be equal to $\frac{1}{T}\log_2 M = 2W\log_2\alpha$.

3.1.4 *Noise (white noise)*

We now work with vectors $x = (x_1,...,x_n) \in \mathbb{R}^n$. Here $n = 2WT \gg 1$ and we know that $\|x\|^2 = 2WTP = nP$, where P is the mean power at one coordinate of the vector x , that is, the mean power of the signal f corresponding to x.

Suppose (and this indeed usually happens) that there is noise in the communication channel. It gives rise to a noise vector $\xi = (\xi_1,...,\xi_n) \in \mathbb{R}^n$ and at the receiving end of the communication channel, instead of the vector x, the displaced vector $x + \xi$ is received. Thus around each point $x \in \mathbb{R}^n$ there occurs a region of uncertainty $U(x)$ at points of which noise can displace x.

Noise can be of different kinds; accordingly it can have various characteristics. We shall assume that our noise is random, is independent of x and is white (thermal) noise, that is, the vector $\xi \in \mathbb{R}^n$ is random and its coordinates are independent random quantities identically distributed in accordance with the normal Gaussian law (with zero mathematical expectation and variance σ^2). Let N be the mean (at the sample value) power of the noise. Then $\|\xi\|^2 = nN = 2WTN$ (N comes from the word "noise" and P

from the word "power") and $\sqrt{N} = \sigma$ is the standard deviation of the random value of each coordinate of the vector ξ. As before, we assume that $2WT = n \gg 1$.

3.2 Transmission capacity of a communication channel with noise

3.2.1 Rough estimate of the transmission capacity of a communication channel with noise

The combined mean power of the signal and the noise is at most $P + N$, therefore the coordinates of the vector $x + \xi$ in modulus should on average not exceed $\sqrt{P+N}$ and it should lie inside a ball of radius $\sqrt{n(P+N)}$.

Since the expected displacement of each of the coordinates of the vector (of information) x under the influence of the noise (white noise) is of order $\sigma = \sqrt{N}$, the number of well distinguishable values of each coordinate at the receiving end is proportional to $\sqrt{\frac{P+N}{N}}$. The coefficient of proportionality k depends on how one interprets the phrase "well distinguishable". If one is required to improve the resolution, then k needs to be made smaller.

In time T there are $n = 2WT$ independent values (of the samples) of the coordinates, therefore the total number K of distinguishable signals will be $\left(k\sqrt{\frac{P+N}{N}}\right)^{2WT}$. Hence, the number $\log_2 K$ of bits that can be transmitted in time T will be $WT \log_2 k^2 \frac{P+N}{N}$. This means that the speed of transmission will be $W \log_2 k^2 \frac{P+N}{N}$ bits per second.

3.2.2 Geometry of signal and noise

We now recall the following. On the vector $x \in \mathbb{R}^n$ an obstacle $\xi \in \mathbb{R}^n$ is imposed in the form of white noise. This means that we are given a vector $x \in \mathbb{R}^n$ and a random vector $\xi \in \mathbb{R}^n$ that is independent of x and has uniform distribution along the directions of $\mathbb{R}^n$. The dimension n of the space $\mathbb{R}^n$ is enormous. Then it follows from the principle of concentration, which we discussed in Chapter 2, that with negligibly small probability of error the vector ξ will be almost orthogonal to the vector x (that is, the scalar product and the correlation of the vectors x and ξ should be considered to be zero).

We add to this that, in view of the concentration of the main part of the volume of a multidimensional ball in a small neighbourhood of its boundary sphere, we can suppose that if a random point lies in such a ball, then it is

most likely to be situated almost on its boundary. Thus in our situation, when $n = 2WT \gg 1$, we are justified in supposing that $\|x\|^2 = nP$, $\|\xi\|^2 = nN$, $\|x+\xi\|^2 = n(P+N)$.

The regions of uncertainty $U(x)$ arising at the receiving end around each point $x \in \mathbb{R}^n$ as a result of the influence of noise can, in our case, be considered to be balls $B(x,r)$ of radius $r = \sqrt{n\sigma} = \sqrt{nN}$.

Under these conditions how many distinguishable signals are there in the ball $B(0,\sqrt{nP})$? Clearly, not more than the ratio of the volume of the ball $B(0,\sqrt{n(P+N)})$ to the volume of a ball of radius $\sqrt{nN}$. Thus, we have the following upper estimate for the number M of distinguishable signals:

$$M \leq \left(\sqrt{\frac{P+N}{N}}\right)^{2WT} = \left(\frac{P+N}{N}\right)^{WT}, \tag{3.2}$$

which means that we have the following estimate for the speed C of transmission of information:

$$C = \frac{\log_2 M}{T} \leq W \log_2 \frac{P+N}{N} = W \log_2 \left(1 + \frac{P}{N}\right). \tag{3.3}$$

Here it is worth pausing and making some observations. If one tries to pack as many balls of radius $\sqrt{nN}$ as possible in a ball of radius $\sqrt{n(P+N)}$ under the condition (as a presupposition) that the inserted balls are, as it were, rigid and non-intersecting, but can abut one another, then for $n = 2WT \gg 1$ the number of such balls will be catastrophically small by comparison with the ratio of volumes indicated above. Shannon's theorem (see below), whose proof we are now ready for, states that, nevertheless, for sufficiently large times T we can get the speed of transmission to be as near as we like to the upper-estimate quantity C indicated above and also having an arbitrarily small probability of an error when transmitting the message.

Just the possibility of making some errors, although as rare as one pleases, eliminates the condition that the inserted balls do not intersect. If the dimension of the space is large, then, as we noted in Chapter 2, the centres of the balls can get close to one another, the balls will intersect, but the relative volume of their intersection can be very small even when the centres of balls of the same radius are at a distance equal to the length of the radius. As the centres approach one another, the number of inserted balls increases, but then also does the probability of an error when decoding the received signal.

The calculation of the interaction of the above circumstances forms the geometric basis of Shannon's theorem.

3.2.3 Shannon's theorem

Theorem. *Let P be the mean power of the transmitter and suppose that we have white noise with power N in the frequency band W. Then by applying a sufficiently complex system of coding it is possible to transmit binary digits with speed*

$$C = W \log_2 \frac{P+N}{N}$$

with arbitrarily small frequency of errors. No method of coding can be transmitted with greater average speed and with arbitrarily small frequency of errors.

We take M to be the number on the left-hand side of the estimate (3.2). This is a large number and we assume that it is an integer (by ignoring its fractional part). In the ball $B(0, \sqrt{nP}) \subset \mathbb{R}^n$, where the vectors x (words, signals) to be transmitted are, we choose M points at random. Here by "at random" we mean that the points are chosen independently at random and the probability that a point hits some region is proportional to the volume of that region, that is, it is equal to the ratio of the volume of that region to the volume of the whole ball $B(0, \sqrt{nP})$. (If the random choice of M balls is repeated many times, then, as a rule, the points will be distributed in the above fashion.) We have $n = 2WT$ and $M = \left(\frac{P+N}{N}\right)^{WT}$, therefore one point will be arriving in a volume of size $\frac{1}{M}|B|$, where $|B|$ is the volume of the whole ball $B(0\sqrt{nP})$. Hence the probability that one of our M points will hit this same region is equal to $\frac{1}{M} = \left(\frac{N}{P+N}\right)^{WT}$. As $T \to +\infty$ this probability of course tends to zero independently of the ratio of the positive quantities P and N.

If our M points are chosen randomly, then, assuming that $n = 2WT \gg 1$ and the volume of the ball $B(0, \sqrt{nP})$ is concentrated near its boundary sphere, one can with negligible relative error (which is smaller the larger T is) assume that all the chosen points will be in an arbitrarily small neighbourhood of the boundary sphere.

We recall further that the noise vector ξ, as we showed earlier, when $n \gg 1$ is orthogonal to the signal vector x with probability arbitrarily close to 1. Thus, for $n = 2WT \gg 1$ (that is, as $T \to +\infty$) we have $\|x\|^2 = nP$, $\|\xi\|^2 = nN$, $\|x + \xi\|^2 = n(P+N)$.[1]

We now proceed with the concluding arguments and obtain a concrete estimate. Suppose that we have made a typical random selection of M points in the ball $B(, \sqrt{nP})$. Suppose that they correspond to M different messages that we intend to send along a communication channel. Such a choice of points in the ball together with their corresponding messages means a cer-

[1] If P and N are interpreted as the variances D_x, D_ξ of the signal and noise, then here we have the classical probability-theoretic relation for the variance of the sum of independent random quantities: $D^2_{x+\xi} = {D_x}^2 + {D_\xi}^2$.

tain encoding of the messages intended for transmission. We coordinate in advance the chosen code with the receiving device. If there were no noise, then the receiver, having received a signal x without corruptions, would uniquely decipher it into the message corresponding to it in accordance with the agreed code.

In the presence of noise in the channel, instead of x one will get $x + \xi$ at the receiving device. The receiver looks for the point in the ball $B(0, \sqrt{nP})$ among the points of the fixed code that is nearest to $x + \xi$ and takes it to be the transmitted signal. Here there is the possibility of error, that is, it is possible not to read the message that was sent. However, this is possible only if there is one of the M points of the code apart from x in the $(\|\xi\| = \sqrt{nN})$-neighbourhood of $x + \xi$.

We find an upper estimate of the probability of such an event. For this we estimate the volume of the intersection of a $(\sqrt{nN})$-neighbourhood of a point $x + \xi$ with the ball $B(0, \sqrt{nP})$. Since $\|x + \xi\|^2 = n(P + N)$, this is a simple geometrical problem. Consider a two-dimensional plane passing through the origin 0 and the points $a = x$ and $b = x + \xi$. The triangle $0ab$ is right-angled with right angle at the vertex a and with side lengths of the legs $|0a| = \sqrt{nP}$ and $|ab| = \sqrt{nN}$ and hypotenuse $|0b| = \sqrt{n(P + N)}$. By calculating its area by two methods we easily find the length h of the perpendicular drawn from the vertex a to the hypotenuse: $h = \sqrt{n\frac{PN}{P+N}}$. If we now take a ball of radius h and centre at the base of this perpendicular, then clearly it covers the entire region (of interest to us) of intersection of the ball $B(0, \sqrt{nP})$ and the $(\sqrt{nN})$- neighbourhood of the point $b = x + \xi$. Hence, the probability that, along with x, which lies on the boundary of this region, there will also be one of the M code points is less than the ratio of the volume of a sphere of radius h to the volume of a sphere of radius $\sqrt{nP}$. Thus, this probability is less than $\left(\frac{N}{P+N}\right)^n = \left(\frac{N}{P+N}\right)^{WT}$, and this tends to zero as $T \to +\infty$.

Thus, for sufficiently large values of T, with arbitrarily small probability of error in such a communication channel one can distinguish one of $M = \left(\frac{P+N}{N}\right)^{WT}$ different objects; more precisely, one can identify one of the M possible different messages in time T. In terms of binary units this is $\log_2 M$ bits of information in time T. Hence we can indeed achieve the speed of transmission arbirtrarily close to the upper bound estimation indicated in inequality (3.3).

This completes the proof of the theorem.

3.3 Discussion of Shannon's theorem, examples and supplementary remarks

3.3.1 Shannon's commentary

The best brief commentary shedding light on certain aspects of the theorem which at first reading are seldom noticed is the following commentary by Shannon himself [2]:

"We shall call a system that transmits with speed C and without errors an ideal system. Such a system cannot be realized by any finite encoding process but can be approximated as closely as one pleases. The following happens when the approximation approaches the ideal: 1. The speed [2] of transmission of binary numbers approximates to $C = W\log_2(1 + P/N)$. 2. The frequency of errors approximates to zero. 3.The signal to be transmitted approximates to white noise in its statistical properties. Roughly speaking this is true because the coding points are randomly distributed inside a ball of radius $\sqrt{2WTP}$. 4. The threshold effect becomes sharp. If the noise exceeds the value for which the system was constructed then the frequency of errors increases very rapidly. 5. The required delays in the transmitter and receiver increase unboundedly. Of course in a broad-band system a delay of one millisecond can already be considered to be infinite."

Here perhaps an explanation is required only for the first sentence in item 5, which at the same time also explains the real meaning of the quantity C featuring in the theorem as the speed of transmission. To write down in bits each of M different objects requires $\log_2 M$ bits. An individual message of the M possible ones is sent or arrives only after what is transmitted (respectively, received) takes the whole of its binary code of length $\log_2 M$. For this time $2T$ is required too, which also implies a delay of the same message, while at the same time, as $T \to +\infty$, the mean speed of transmission of the bits (bits per second) in fact approximates to the upper limit indicated in the theorem.

Later on we shall give some examples which possibly will explain certain aspects of the range of questions touched upon here.

3.3.2 Weak signal in a large amount of noise

It is clear from the construction of an optimal code (and this is explicitly pointed out by Shannon in item 3 of the above quotation) that in its statistical properties such a code is similar to white noise. This means that establishing

[2] *Author's remark.* Like the Meshcherskiĭ–Tsiolkovskiĭ formula for the speed of a rocket!

(or discovering) contact with a very intelligent extraterrestrial civilization sending us signals indistinguishable from noise is fairly difficult.

But let us consider the following situation. Arriving towards us is a weak periodic signal in the background of a large amount of noise that is, however, random. For example, in the channel there is a large amount of white noise. Is it possible to separate the useful signal f from the noise? Suppose that we know or could know the period T of the signal f. Let us listen to and record the signal with the noise $n \gg 1$ times. We then reproduce all these signals synchronously, that is, we put them together. Then the random noise will itself be dampened, while the signal will be strengthened. This means that sometimes one can combat the noise by actually making use of it.

3.3.3 Redundancy of language

In a communication channel we very often cope with noise in a similar way, which the science calls complicating the code or its redundancy.

In item 4 of his commentary Shannon noted the sharp threshold effect of the optimal code. We shall return to this a little later, but meanwhile we give some explanation with an example.

If in a telephone conversation you are dictating something to the person at the other end and there is some word that he could not make out or does not know, then you start to repeat or communicate letter by letter, while the letters are communicated by pronouncing entire words such as Anna, Maria, Booby, Aristotle, and so on.

You fight against the noise by encoding A, M, B, etc., with a code that is certainly redundant. An optimally economical code is, of course, splendid, but also dangerous, as is every maximum of potential possibilities — it is unstable. Any spoken language, as we can easily observe, is redundant (roughly by 50%) but, on the other hand, good for everyday intercourse.

3.3.4 Precise measurements in a crude piece of apparatus

How does one measure the thickness of a sheet of paper on an apparatus where you have measured your height only to within 0,5cm? Recall the example given above of a weak signal in a communication channel with a large amount of noise. If there is the possibility of taking some packet of this paper, and by adding them one can find, for example, that one thousand sheets of paper have a thickness of 20cm with absolute error within the limits of 0,5cm, then, assuming that the sheets are all roughly the same, we find

that the thickness of one sheet is 0,2mm with possible error in the limits of 0,005mm.

The idea of the above examples can be extended to create accurate constructs (devices, apparatuses) from inaccurate elements.

3.3.5 Shannon–Fano code

Hidden in the numerous details of the proofs the probabilistic structure of an optimal code in Shannon's theorem is clearly distinguishable in all its detail in the following naked idea of an optimal code, called the Shannon–Fano code. For economy of space and words we consider a simple demonstrative example, the possible generalization of which is obvious.

We have an alphabet of four letters from which words are formed. Along a communication channel the bits 0 and 1 can be transmitted with the same speed and accuracy. One can encode the letters of our alphabet as follows: (0,0), (0,1), (1,0), (1,1). After this one can transmit text in these letters. Meanwhile we forget about the noise and concern ourselves with the economy of the code, which influences the speed of the transmission of information.

We suppose that a statistical analysis of the language establishes that the four letters of the alphabet are encountered with different frequencies; for example, suppose that their probabilities are $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$, respectively. Then it is most reasonable to proceed as follows. First we divide the letters into two equiprobable groups (here it is the first letter and all the remaining letters), which we distinguish by the symbols 0 and 1, respectively. We then repeat the same procedure with each of the groups and their subgroups as long as the subgroups do not reduce to a single element. And this is the idea of the Shannon–Fano code. In our case the code looks like this: (0), (1,0), (1,1,0), (1,1,1). We compare the above two codes on a sufficiently long text of T letters. In the first case we need to send $2T$ bits. In the second case it is $(\frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3)T = \frac{7}{4}T$ bits. Moreover, even without punctuation signs in the optimal code one can recover the sequence of letters (10011000111110 — decipher that). But one only has to make an error in the transmitter or receiver of just one bit and the text becomes unreadable.

3.3.6 Statistical characteristics of an optimal code

The above example demonstrating the idea behind the Shannon–Fano code shows that an optimal code tends to distribute the information uniformly in terms of the symbols transmitted. This can be achieved for the transmission of long messages provided that their statistical processing is dealt with beforehand. The dangers of an optimal code are now also clear.

We note further that Shannon's theorem relates to noise assumed to be white noise. Noise can be of various kinds, both random and deterministic. Moreover, even random noise can have various statistical characteristics. Obviously in a specific situation one needs to act in a specific manner. The general theory gives indications on the reasonable order of such actions but does not solve the whole problem in one go.

3.3.7 *Encoding and decoding — ε-entropy and δ-capacity*

Earlier we already mentioned the quantization of a continuous signal into a discrete number of levels. The standard procedure enabling one to go over to a discrete finite description of any compact subset of a metric space consists in the construction of a finite ε-net, that is, a finite collection of points such that any point of the compact set can to within an ε-shift be approximated (replaced) by one of the points of this net. The quantity ε characterizes the allowable accuracy of the approximation, or the allowable error. If the devices being used are not capable of distinguishing objects in scales less than ε, then, without special need, it makes no sense to bother with all the points of the compact set; it suffices to replace the compact set by some ε-net of it. The ε-net itself can be considered to be a discrete code describing the compact set to within an accuracy of ε.

Of course it is desirable to have an ε-net with the greatest possible economy, that is, containing the fewest possible number of points. When ε tends to zero the number N_ε of points in such an ε-net of greatest economy increases without bound in general. Its rate of increase is related to the specific character of the compact set and the metric space.

Kolmogorov called the quantity $\log_2 N_\varepsilon$ the *ε-entropy* of the compact set. If, for example, one takes the unit cube I^n or any bounded region of Euclidean space $\mathbb{R}^n$, then, as is easily verified, the limit of the ratio of $\log N_\varepsilon$ to $\log \frac{1}{\varepsilon}$ as ε tends to 0 gives the dimension n of the space.

Incidentally one can exploit the above circumstance to redefine dimension and thus have the possibility of talking about dimensions that are not necessarily integers.

Another more interesting example of the use of ε-entropy which is worth recalling relates to Hilbert's thirteenth problem [3]. Roughly speaking, the question can be restated in the following eye-catching form: do functions of several variables exist? More precisely: can every function of several variables be assembled from functions of a smaller number of variables, that is, can it be represented as a superposition (composite) of finitely many such functions?

A.N. Kolmogorov and V.I. Arnold proved that every continuous function of several variables can be represented as a superposition of continuous functions of one and two variables; here, as Kolmogorov noted, for the

function of two variables it suffices to have only the function $x + y$ (see [5], [6a], [6b]).

But even before that A.G. Vitushkin had proved that not every smooth function of several variables is the superposition of functions of a smaller number of variables and enjoying the same amount of smoothness; see [4]. To state this precisely, after Vitushkin we consider a number $v = \frac{n}{p}$ — the ratio of the number of variables to the order of its highest-order continuous derivatives. It serves as the index of complexity of the function in the sense of Vitushkin. As always, we denote by $C_n^{(p)}$ the class of p-smooth functions of n variables defined on the unit n-dimensional cube $I^n \subset \mathbb{R}^n$. Let $k < n$. The question is: when can any function of class $C_n^{(p)}$ be represented as a superposition of functions of class $C_k^{(q)}$? Vitushkin showed that this is possible only if $(\frac{n}{p} = v) \leq (\tilde{v} = \frac{k}{q})$.

Vitushkin's proof used, in particular, Oleĭnik's estimates of the Betti numbers of the algebraic manifolds obtained in connection with investigations by her and Petrovskiĭ of Hilbert's 16th problem on the number and positioning of ovals of a real algebraic curve.

Kolmogorov gave another direct and, seemingly the most natural, explanation (proof) of Vitushkin's result precisely in connection with information and entropy [7a], [7b].

The spaces $C_n^{(p)}$ and $C_k^{(q)}$ are infinite-dimensional but, as Kolmogorov showed, if in these spaces one takes the compact sets consisting of all functions whose derivatives are bounded by some fixed constant, then as $\varepsilon \to 0$ their ε-entropy will increase as $(\frac{1}{\varepsilon})^{\frac{n}{p}}$ and $(\frac{1}{\varepsilon})^{\frac{k}{q}}$, respectively. If all the functions in $C_n^{(p)}$ can be represented as a superposition of finitely many functions of class $C_k^{(q)}$, then $\left(\frac{1}{\varepsilon}\right)^{n/p} = O\left(\left(\frac{1}{\varepsilon}\right)^{k/q}\right)$. Thus, the inequality $(\frac{n}{p} = v) \leq (\tilde{v} = \frac{k}{q})$ must hold.

As regards Hilbert's problem itself it is worth noting nevertheless that within the framework of algebraic functions (which possibly Hilbert himself also had in mind when speaking about his thirteenth problem on the representation of solutions of a seventh-degree algebraic equation) that the problem is still open. In this connection see the sources [3], [7a] and [7b].

We do not intend here to delve too deeply into these questions and we have only mentioned an example of another non-trivial use of the notion of discrete code and ε-entropy.

Again we return briefly to discrete encoding and add a few words about decoding. An economical ε-net can serve as an economical discrete code of an object (compact metric space) describing it to within an accuracy of ε. Suppose that a specimen of such a code is at both ends of a communication channel. If there is no error in the transmission of the message, then at the receiving end a signal is obtained about that point of the ε-net that was selected at the transmitting end. But if for some or other reason errors

in the communication channel are possible, and the transmitted point can be interpreted within the limits of its δ-neighbourhood, then clearly errors are possible during the decoding process; and we already spoke about that earlier. If we are going to exclude the possibility of errors completely, then we have to refrain from using an economical code in the form of an ε-net. By contrast we now have to seek a maximal collection n_δ points of the compact set separated from each other by a distance of at least 2δ. Only such a code (it clearly will be a 2δ-net) can, under the above conditions, guarantee error-free transmission.

While the quantity $\log_2 N_\varepsilon$ is called the *ε-entropy*, as we already know, the quantity $\log_2 n_\delta$ is called the *δ-capacity*.

Calculation of the ε-enropy and δ-capacity of various classes of functions can be found in [8], [9]. Some further information relating to signal processing can be found in [9]–[14].

3.4 Mathematical model of a channel with noise

3.4.1 Simplest model and formulating the problem

As usual, for economy of everything we consider to begin with the simplest model, which however already contains almost everything of most value for our needs for the moment and can easily be generalized if one wishes.

In a communication channel the transmitter sends to the receiver the symbols 0 and 1. The noise results in the possibility that from time to time the receiver deciphers the sent symbol 0 as 1, and 1 as 0. Let p be the probability of a correct passage of the transmitted symbol.

Sent along the channel are messages (text, words) consisting of successive letters (symbols 0, 1 of our two-letter alphabet). We suppose that the channel acts on each letter of the word independently, that is, it is a *channel without memory*. What is the transmission capacity of such a channel ?

So that is our problem. Intuitively it is clear that it is a reasonable problem. At the same time it is clear that for its answer it needs to be made clear what exactly one has in mind.[3]

Earlier, before the proof of Shannon's theorem, (after Shannon) we had to sort out the meaning and precise content of certain terms and concepts which our intuition allowed us to use. We now implement this task (again after Shannon). Of course our earlier attempts will considerably lighten this task. Properly speaking it will largely be an abstract formulation of it.

[3] It is recounted that one of the visitors of the celebrated Princeton Institute of Advanced Study was housed in Gödel's study, which was temporarily vacated. On leaving, the visitor left a thank-you note on the table expressing his regret that he had not made a closer acquaintance with Gödel. After a while he received a polite letter from Gödel, who had read the note, which asked him to clarify what exactly he had in mind.

As regards a more general abstract model of a communication channel with noise, it is clear that instead of an alphabet of two letters one can have any finite (but not one-letter) alphabet and have the probability that the ith sent letter is converted to the jth at the receiving end. The matrix (p_{ij}) of conversion probabilities models a communication channel with noise.

If the letters are corrupted with different probabilities, then the speed of transmission of information can depend (and does depend) on the cleverness of the code used for writing the messages to be transmitted. Clearly it is best to use most often those symbols that are least subject to corruption. Furthermore, as we already know by experience of the Shannon–Fano code, it is helpful to take into account the statistical peculiarities of the text of the very message subject to transmission.

Apparently, the transmission capacity of a channel of given matrix (p_{ij}) must simply be the upper bound of the possible speed of transmission with respect to everything that the channel (device) itself does not depend on, for example, the upper bound of all possible encodings of the texts to be transmitted. Clearly different users can use one and the same device with different degrees of efficiency. The capabilities of the device itself must be evaluated under the assumption that it is used with maximum efficiency.

After maximal speed under an optimal code has been achieved there can clearly emerge new problems. For example, we saw what dangers are hidden behind codes of maximal economy. But let us lay all this to one side. Just now we need to gradually investigate the question of the speed of transmission of information and what in general we mean by the terms *information* and *quantity of information*.

3.4.2 Information and entropy (preliminary considerations)

As we have already remarked, the appearance of the telegraph and wireless communication stimulated the development of the concept of information and its quantitative description.

It would seem that the measure of information can reasonably be considered to be the measure of the change of uncertainty associated with the information received.

In the simplest situation when there are two possibilities on an equal footing, for example, when the random quantity has exactly two equiprobable values 0 and 1 (off, on), the information about its concrete value (state) liquidates the uncertainty. Recall that the measure of such information deals with what is called the *bit* (short for binary digit).

To identify one of the M objects by putting questions to which the answers are only "yes" (1) or "no" (0) requires, as is well known, $\log_2 M$ binary symbols (the repeated-bisection algorithm). Such a system (random quantity) is capable of storing $m = \log_2 M$ bits of information (correspond-

ing to the measure of its uncertainty). More precisely, if all M possible values (states) of the random quantity under consideration are equally likely, then the identification (selection, information on realization) of one of them under the indicated correspondence is equivalent to giving $\log_2 M$ bits, that is a message of $\log_2 M$ bits of information.

We now state this more formally. Let X be an arbitrary discrete random quantity that can take M different values x_i with probabilities p_i, respectively. How does one take into account the probabilities? What measure of uncertainty (and information) is it reasonable to associate with such a random quantity?

We write down the result just obtained by us in the following form:

$$m = \log M = M \cdot \frac{1}{M} \log M = \sum \frac{1}{M} \log M = -\sum \frac{1}{M} \log \frac{1}{M},$$

where we treat $\frac{1}{M}$ as the probability of the appearance (realization, selection) of a specific one of these M objects. (Here and in what follows $\log = \log_2$.)

Then, surely, in general we should arrive at the quantity $-\sum_{i=1}^{M} p_i \log p_i$. We now substantiate this assertion. The quantity $H(X) = -\sum p_i \log p_i$ is called the *entropy* of the discrete random quantity X. (By continuity we suppose that $0 \log 0 = 0$.)

Let us experiment with this. If the probability p_i of an event x_i is small, then the information that this very rare event has occurred can be taken to be the very large number $-\log p_i$. On the other hand, if the event is rare, then over a long period of observations it appears with its information in its teeth extremely rarely (a fraction p_i of the whole time of observations). Therefore the information averaged over a large time interval of observations which this event yields (the value x_i of the random quantity X) is equal to $-p_i \log p_i$.

Thus, if $-\log p_i$ is the measure of uncertainty and information associated with the event x_i whose probability is p_i, then $-p_i \log p_i$ is the average statistical quantity of information that the appearance of such an event yields, and then $H(X) = -\sum_{i=1}^{M} p_i \log p_i$ (mathematical expectation of $-\log p_i$) is the average quantity of information that a single event (value) of the random quantity X carries.

Be aware of the fact that we are not interested in what exactly the real event x_i consists in, although for other purposes it may be that this is the most important thing.

We now settle on a precise notation for the statistical character of entropy: for any positive numbers ε, δ there is a number $n_{\varepsilon\delta}$ such that for $n \geq n_{\varepsilon\delta}$ we have the inequality

$$P\{|-\frac{1}{n}\sum_{i=1}^{n} \log p_{x_i} - H(X)| < \delta\} > 1 - \varepsilon, \tag{3.4}$$

where, as usual, P is the probability of the event indicated in the curly brackets, but now the x_i, $i = 1,...,n$, are n independent values of the random quantity X and the p_{x_i} are the probabilities of these values.

How is the entropy related to the encoding ?

Consider the message-word-vector $\bar{x} = (x_1,...,x_n)$ formed by n successive independent values of the random quantity X. The probability $p_{\bar{x}}$ of the appearance of the word $\bar{x}$ is equal to $p_{\bar{x}} = p_{x_1} \cdot ... \cdot p_{x_n}$. In view of formula 3.4 for $n \geq n_{\varepsilon\delta}$, with probability $1 - \varepsilon$ we have

$$2^{-n(H(X)+\delta)} \leq p_{\bar{x}} \leq 2^{-n(H(X)-\delta)}. \tag{3.5}$$

The word $\bar{x}$ is called *δ-typical* if these estimates hold for it. Clearly there exist at most $2^{n(H(X)+\delta)}$ such δ-typical words, while if $n \geq n_{\varepsilon\delta}$, then there are at least $(1-\varepsilon)2^{n(H(X)-\delta)}$ of them and the entire set of non-δ-typical words has probability at most ε.

In principle, now we can already use binary sequences of length $n(H(X) + \delta)$ to encode all δ-typical words. Even if all the remaining words are encoded with one symbol, the probability of an error in transmitting words $\bar{x}$ of length n invoking such a code will be less than ε.

On the other hand (and we have already mentioned this effect of the instability of economical codes), any code using in the same situation binary sequences of relatively slightly smaller length $n(H(X) - \delta)$ (for example, $2\delta n$ out of the $n(H(X) + \delta)$ sent symbols were lost in the noise) will have an asymptotically non-vanishing probability of an error, which tends to one as $n \to +\infty$.

Thus the relation between entropy and encoding of information consists, for instance, in the fact that as $n \to +\infty$ an efficient encoding requires $N \sim 2^{nH(X)}$ words and the entropy $H(X)$ can be interpreted as a measure of the quantity of information in bits in the symbol being transmitted, that is, in one value of the random quantity X.

Hence it follows, in particular, that the entropy of the source of the information should not exceed the capacity of the communication channel if we wish adequately and without delays to transmit the information at hand along this communication channel.

3.4.3 Conditional entropy and information

We turn step-by-step to the transmission of information along a communication channel. The transmitter sends the message $\bar{x} = (x_1,...,x_n)$ and the receiver receives $\bar{y} = (y_1,...,y_n)$. How does one recover what was sent from what was received ? If there are no corruptions, that is, $y_i = x_i$ always, then there is no problem. We therefore assume that the channel is characterized

by some matrix (p_{ij}) of probabilities that the transmitted signal x_i will be converted to the received signal y_j.

We put the question another way. What information about the message $\bar{x}$ is contained in the message $\bar{y}$? Or in other words: how is the uncertainty of $\bar{x}$ changed (decreased) when we know $\bar{y}$?

We turn to conditional probabilities and introduce the concept of conditional entropy $H(X|Y)$ of a random quantity X at the input of a communication channel with respect to the random quantity Y at the output. Then, after Shannon, we consider the quantity

$$I(X;Y) = H(X) - H(X|Y), \tag{3.6}$$

and regard it as the effective *quantity of information* that, on average, is transmitted by one sent signal (value of the random quantity X) in this communication channel.

Hence, the *capacity of the communication channel* is defined as

$$C = \sup_{\{p_x\}} I(X;Y), \tag{3.7}$$

where the supremum is taken over all possible codes, that is, over all possible probability distributions $\{p_x\}$ of the input random quantity X, which has a fixed finite set of values (alphabet).

Thus, we define the conditional entropy $H(X|Y)$ of one random quantity X with respect to another random quantity Y.

Let $\{p_x\}$, $\{p_y\}$ and $\{p_{x,y}\}$, respectively, be the probability distributions of random quantities X, Y and the joint random quantity $Z = (X,Y)$. If the probability of the appearance of the value x_i at the input of the random quantity X is equal to p_{x_i}, and the probability $p(y_j|x_i)$ of conversion of x_i to y_j is given and is equal to p_{ij}, then the probability p_{x_i,y_j} of the combined event $z_{ij} = (x_i,y_j)$ is equal to $p(y_j|x_i)p_{x_i}$, and the total probability p_{y_j} of the appearance at the output of the value y_j of the random quantity Y is equal to $\sum_i p(y_j|x_i)p_{x_i}$.

To ease the notation and without losing any clarity we shall no longer write the extra lower indices. For example, we shall write the standard formula for conditional probability as follows: $p_{x,y} = p(y|x)p_x$ or $p_{x,y} = p(x|y)p_y$, since $p_{x,y} = p_{y,x}$.

First we find the conditional entropy $H(X|Y = y)$ of a random quantity X under the condition that the random quantity Y has the value y. In other words, we now find what the entropy (uncertainty) X becomes under the condition that the random quantity Y has taken the value y.

$$H(X|Y = y) = -\sum_x p(x|y)\log p(x|y) = -\sum_x \frac{p_{x,y}}{p_y}\log\frac{p_{x,y}}{p_y}.$$

We can now find what we are interested in, namely, the conditional entropy $H(X|Y)$ of the random quantity X with respect to the random quantity Y:

$$H(X|Y) = -\sum_y p_y H(X|Y=y) = -\sum_y p_y \sum_x \frac{p_{x,y}}{p_y} \log \frac{p_{x,y}}{p_y} =$$

$$= -\sum_{x,y} p_{x,y} \log p_{x,y} + \sum_y p_y \log p_y = H(X,Y) - H(Y).$$

Here $H(X,Y)$ is the combined entropy of the pair $Z = (X,Y)$ of random quantities X and Y; the probability distribution of the pair is $\{p_{x,y}\}$.

We have found that $H(X|Y) = H(X,Y) - H(Y)$. But since $p_{x,y} = p_{y,x}$ and $p_{x,y} = p(y|x)p_x = p(x|y)p_y = p_{y,x}$, we also have the relations $H(X,Y) = H(Y,X)$ and $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$. Thus,

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X) = H(Y,X). \tag{3.8}$$

Taking into account formula (3.6) (Shannon's definition) for the quantity of information we find that

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y). \tag{3.9}$$

Since $H(X,Y) = H(Y,X)$, it follows that

$$I(X;Y) = I(X;Y). \tag{3.10}$$

3.4.4 *Interpretation of loss of information in a channel with noise*

We pause briefly to summarize informally the purport of the concepts introduced above and the interactions that have been uncovered.

The entropy $H(X) = -\sum_x p_x \log p_x$ of a discrete random quantity X is a certain statistical average of its character. If $-\log p_x$ is treated as the measure of uncertainty of the rareness of the event (of the value x of the random quantity X), expressed in bits, and measuring the quantity of information contained in the message about the occurrence of the event x is proportional to this, then $H(X)$ will be the mathematical expectation of this quantity $-\log p_x$.

The entropy is some average measure of the uncertainty of a random quantity X taking on one of its values. Put another way, it is the average measure of information that arrives at one value of the random quantity. Here it is assumed that we obtain a linear series of independent values of the random quantity X and we average the quantity of information obtained from the number of received values of the random quantity. We also tac-

itly assume that the values are transmitted and dealt with uniformly — one value per unit of time. Therefore in the case when we are dealing with the transmission of information along a communication channel one prefers to treat the entropy of the source as the average quantity of information that the source creates in a unit of time.

If the channel is capable of transmitting this flow of information without corruptions, then everything is fine. If, on the other hand, errors can arise, then we have new problems on our hands. In the example of Shannon's theorem it is clear that in a concrete situation it is necessary to deal with concrete physical parameters (frequency band, signal level, noise level, statistical characteristics of the noise, and so on). Properly taking into account and handling these parameters is an important problem in its own right.

We considered some abstract model of the communication channel and arrived at the useful concept of conditional entropy $H(X|Y)$. Its meaning is to enable one to estimate the average level of uncertainty of the random process X remaining if one has the possibility of observing the state of the random quantity Y. If X and Y are independent, then clearly an observation on Y says nothing about X and $H(X|Y) = H(X)$. On the other hand, if $X = Y$ (for example, when there is error-free transmission along the communication channel), then $H(X|Y) = 0$.

Thus, in the problem of transmission of information along a communication channel, the quantity $H(X|Y)$ can be treated as the average loss of information per transmitted value (per symbol or in unit time) in this communication channel. This means that it is natural to take $I(X;Y) = H(X) - H(X|Y)$ as the average measure of information that passes along the communication channel when the values of the random quantity X encoding the original messages are sent along it.The informative part of the messages is of no interest. We measure the information in bits and we measure the speed of its creation or reception in bits per symbol or in bits per unit time.

The values that a random quantity X can take can be considered to be the alphabet in which the messages subject to transmission are written (encoded). The messages are assumed to be long enough so that statistical characterizations can be used in general for the problem. This can be arranged alphabetically in different ways, as we have seen from the Shannon–Fano code. The optimal code for transmission is chosen with the characteristics of the communication channel to be used being taken into account.

The transmission capacity (3.7) of the communication channel is the maximal average speed of transmission along this communication channel that can be attained or that can be arbitrarily closely approximated in transmitting long texts of messages using sensible encodings beforehand in the channel alphabet.

3.4.5 Calculating the transmission capacity of an abstract communication channel

We have defined the capacity of an abstract communication channel by formula (3.7). We have just discussed the content of these ideas in broad outline. Now, in conclusion we nevertheless carry out a concrete calculation. We shall find the capacity of our abstract communication channel in the simplest example from which we began this abstract calculation. We recall the conditions.

In the communication channel the transmitter sends to the receiver the symbols 0 and 1. As a result of noise, the receiver occasionally deciphers the transmitted signal 0 as 1, and 1 as 0. Let p be the probability that the symbol is passed through correctly.

Along the channel messages (text, words) are sent consisting of sequences of letters, which are the symbols 0,1 of our very simple two-letter alphabet. We assume that the channel acts independently on each letter of the word, that is, it is a channel without memory.

What is the transmission capacity of such a communication channel?

In the present case the matrix of the probabilities of conversion is simplified to the limit not merely because we have a two-letter alphabet, but also because both transmitted symbols 0 and 1 have the same probability of being passed through without corruption. Thus the random quantity X at the input of the channel can take two values. Suppose that the encoding of the message to be sent is such that the probabilities of the appearance of the values 0, 1 are p_0, p_1, respectively.

At the output of the channel the random quantity Y can also take these two values 0 or 1, but possibly with different probabilities q_0, q_1. Let us find them.

The value 0 is obtained at the output with probability p when 0 is at the input, and with probability $1-p$ when 1 is at the input. In turn, at the input we have 0 with probability p_0 and 1 with probability p_1. Therefore the probability of getting 0 at the output is $pp_0 + (1-p)p_1$. Correspondingly, 1 appears at the output with probability $pp_1 + (1-p)p_0$.

The probability distribution of the combined random variable $Z = (X,Y)$ is also easy to write down: $(0,0) \sim pp_0$, $(0,1) \sim (1-p)p_1$, $(1,0) \sim (1-p)p_0$, $(1,1) \sim pp_1$.

We can now calculate the entropies $H(X)$, $H(Y)$, $H(X,Y)$ and via the second of formulae (3.9) find the speed of transmission of information. In our case we find that

$$I(X;Y) = H(Y) - h(p),$$

where $h(p) = -p\log p - (1-p)\log(1-p) = H(X,Y) - H(X) = H(Y|X)$.

The maximal value of the quantity $I(X;Y)$ is attained when $H(Y) = 1$, that is, when the distribution at the output is uniform: $q_0 = q_1 = \frac{1}{2}$. But $q_0 =$

$pp_0 + (1-p)p_1$, and $q_1 = pp_1 + (1-p)p_0$, therefore the condition $q_0 = q_1 = \frac{1}{2}$ holds precisely when the input distribution is uniform: $p_0 = p_1 = \frac{1}{2}$.

(Here we have used the following fact, which is easily verified: using the convexity of the logarithm function, if a discrete random quantity X has M different values, then $0 \leq H(X) \leq \log M$, where the left-hand relation holds with equality for the degenerate distributions when one value is taken with probability 1 and the others with probability 0, while the right-hand relation holds with equality for a uniform distribution.)

Thus, we have found that the channel transmission capacity of our simplest model communication channel with noise is equal to $C = 1 - h(p)$. Here $h(p) = H(Y|X)$ characterizes the loss of information on the symbol being transmitted. (See also [16].)

The calculation of the speed is, of course, always carried out to within a constant coefficient corresponding to the choice of the unit of time. For example (see [15]), suppose that the channel is physically capable of transmitting 100 bits 0, 1 in unit time, where each bit to be transmitted can be replaced by the opposite bit with probability 0,01. In this case, $h(p) = h(1-p) = h(0{,}01) \approx 0{,}0808$ and $C = 100(1 - 0{,}0808) = 91{,}92 \approx 92$ bits per unit of time. Take note the result is not equal to 99.

Armed with one's accumulated experience one can now try to prove the following intuitively clear Theorem of Shannon.

Theorem. *Suppose that there are a source of information X whose entropy per unit of time is equal to $H(X)$ and a communication channel of capacity C. If $H(X) > C$, then it is impossible to have an encoding delivering messages without delay or corruption. If, on the other hand, $H(X) < C$, then it is always possible to encode sufficiently long messages so that they are transmitted without delay; furthermore the probability of errors could be made arbitrariily close to zero.*