Physics – Uspekhi **64** (6) 617 – 634 (2021)

METHODOLOGICAL NOTES PACS numbers: 03.67.Dd, 42.50.Ex

Extraction of quantum randomness

I M Arbekov, S N Molotkov

DOI: https://doi.org/10.3367/UFNe.2020.11.038890

Contents

Abstract. The nature of randomness and constructive and provable methods to obtain (extract) it from observations of physical systems are discussed. True randomness, which exists only in a microcosm in the quantum-mechanical description of physical systems, is a fundamental property of quantum systems, which manifests itself in the outcomes of measurements upon quantum systems. The classical description of physical systems does not include any randomness and, in fact, it is introduced 'manually' by means of uncertainty — unknown initial conditions. Methods to really 'feel' quantum randomness are discussed using the example of a quantum device, a random number generator. Issues related to the 'proof' of randomness — testing of numerical sequences — are reviewed, and logical constructions that underlie such testing are analyzed. A mathematical apparatus is used to this end, which does not require special academic training, so standard knowledge from university courses on quantum mechanics and probability theory is sufficient. The authors aim to track a unified logical path from the origin of randomness in the quantum domain to its extraction, physical implementation, and testing.

Keywords: quantum random number generators, randomness extraction

I M Arbekov [(1)], **S N Molotkov** [(1, 2, 3, 4, a)]
[(1)] Academy of Cryptography of the Russian Federation, ul. Yartsevskaya 30, 121552 Moscow, Russian Federation
[(2)] Institute of Solid State Physics, Russian Academy of Sciences, ul. Akademika Osip'yana 2, 142432 Chernogolovka, Moscow region, Russian Federation
[(3)] Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie gory 1, str. 52, 119991 Moscow, Russian Federation
[(4)] Lomonosov Moscow State University, Quantum Technology Center, Leninskie gory 1, str. 35, 119991 Moscow, Russian Federation
E-mail: [(a)] molotkov@issp.ac.ru

Received 15 May 2020, revised 28 October 2020
Uspekhi Fizicheskikh Nauk **191** (6) 651 – 669 (2021)
Translated by M Zh Shmatikov

The most incomprehensible thing about the universe is that it is comprehensible.
Albert Einstein

1. Introduction. On the nature of randomness and the basics of constructing quantum generators of random numbers

Random numbers are widely used in various fields of science and technology, for example, in the simulation of physical processes by the Monte Carlo method. Every person comes across random numbers in everyday life: computer access passwords and PIN codes of smart cards and other electronic devices.

Random numbers are especially widely used in cryptography. A random number generator (RNG) is an integral part of cryptographic information protection systems; its quality largely determines the cryptographic robustness of such systems.

Encryption of large information arrays requires frequent changes of secret keys generated using random number generators. If the key changes frequently, the amount of information that is encrypted on individual keys will be small, thus making the secret communication reliable.

Secret keys are changed in *classical symmetric encryption systems* at the transmitting and receiving sides using devices

of input from digital media of a limited volume. This, in turn, limits the rate of changing keys, since it requires regular replacement of key information carriers. Frequent key changes during the entire life of the cryptosystem becomes virtually impossible. Therefore, in *classical systems*, secret keys are used as a rule in the form of master keys to obtain *session* keys derived from them, which, in general, does not provide high cryptographic robustness. Modern systems of *quantum cryptography* can provide a frequent change of secret keys, but, at the same time, the generation of each secret key requires a large number of random numbers.

How many should there be?

The quantum cryptography system is a distributed system of matching and hashing into a secret final key of random bit sequences on the transmitting and receiving sides, which are formed by transferring quantum quasi-single-photon (ideally single-photon) states via an optical channel [1].

We estimate for illustration the number of random numbers — random bits — required to obtain a secret key for a fiber-optic line length of, for example, 100 km. The 100-km distance to the receiver can be passed in a standard optical fiber with a specific loss of ≈ 0.2 dB km^{-1} by, on average, one in 10^2 photons.

Since there are no exactly single-photon sources, highly attenuated coherent laser radiation is used instead. A coherent state is a superposition of Fock states with the number of photons $k = 0, 1, \ldots$, with the corresponding weights; only the average number of photons μ is specified in the state. The attenuation of the coherent state to the average number of photons per pulse at the level $\mu \approx 0.1$ leads to the situation where approximately only one out of 10 radiation pulses contains a single-photon Fock state, while the remaining nine pulses contain a vacuum state of the field.

Quasi-single-photon states are detected using avalanche photodetectors, whose efficiency is much less than unity. Typical values of the quantum efficiency of single-photon avalanche photodetectors are $\eta \approx 0.1$, a factor which further reduces the rate of secret key generation by about 10 times.

To ensure the cryptographic secrecy of the general final 256-bit-long key, random bit sequences on the transmitting and receiving sides approximately 10^4 bits in length should be processed and compressed (hashed).

The generation of a single 256-bit secret key requires as a result a random sequence from the RNG output with a length no less than

$$10^2(\text{losses in the line}) \times 10^1(\mu) \times 10^1(\eta) \times 10^4(\text{hashing}) = 10^8 \text{ bit}.$$

This example shows that, to implement quantum cryptography systems, random number generators with a high generation rate and a provable 'randomness' of the output sequence are required.

Randomness understood at the intuitive level as a process in which each next step is unpredictable seems to be quite comprehensible, but a closer examination shows that the concept of *randomness* turns out to be far from trivial. How can 'good' or, more accurately, *true* randomness be created, and how can it be determined whether that randomness is good and, as a maximum, is it possible to obtain *true* randomness?

It is fundamentally important that, in developing random number generators, it is not enough that the sequences generated by them be tested for randomness using some criterion. This is just a necessary condition. Of fundamental importance is the source of primary randomness that is used to obtain an equidistributed sequence of 0 *and* 1 *and which would really be a source of randomness for reasons independent of the recommended tests* (*for example, a source of randomness as a measurement process in a quantum system* (*see Eqns* (3)–(8) *below*). *Many pseudo-random random number generators, typical of which are closed-loop shift registers, have been successfully tested but are not truly random.*

The necessary information about the concepts and methods of testing finite bit sequences with regard to randomness are presented in Sections 11 and 12, which are quite independent and can be recommended to the reader who is not familiar with the subject of this study, independently of other sections.

Summarizing the above, we come to the conclusion that, strictly speaking, the very concept of *randomness* requires an additional mathematical definition. We discuss first the situation at a qualitative level.

Random numbers arise as a result of the RNG operation. Random number generators[1] can be divided into two classes: *mathematical* and *physical*. It should be noted from the very beginning that, to generate keys in symmetric encryption systems that claim *high* cryptographic robustness, *physical* RNGs alone are used.

Mathematical generators are transformations, usually recursive ones:

$$x_i = \mathcal{F}(x_{i-1}) = \mathcal{F}\big(\mathcal{F}\big(\mathcal{F}(\ldots\mathcal{F}(x_0))\big)\big)\,, \tag{1}$$

where $\mathcal{F}$ is some function and x_0 is the initial value (seed), which is selected 'manually'.

Is the sequence of numbers $\{x_i\}$ random? Apparently not, because, if the seed and the transformation $\mathcal{F}$ itself are known, the entire sequence is known. It is for this reason that such generators are referred to as pseudo-random, since they completely depend on the initial conditions: the entire 'randomness' is concentrated in the unknown 'seed' value x_0 and partly in the transformation $\mathcal{F}$, possibly hidden from an illegitimate party. Thus, a mathematical transformation cannot provide *true* randomness.

Physical generators are based on measuring the state of a physical system, and they can also be divided into two types: *classical* and *quantum*.

The first type of generator is *classic*.

The output bit sequence in classical generators is a function of actually observed physical quantities generated by a nondeterministic physical system. From the standpoint of the axiomatic construction of the theory of probability, a rigorous proof of randomness (independence and equiprobability) of the output sequence requires setting the initial probabilistic space in which the actually observed physical quantities would be random variables (functions of space elements) and, through appropriate transformations of a random number generator, would generate either a provably independent and equiprobable bit sequence or a sequence

[1] In the Russian-language literature on cryptography and various documents, the phrase 'random number transmitter' is often used instead of 'generator'.

close to it within the limits of established probability-theoretic requirements. [2]

If we now abstract from the axiomatic theory of probabilistic assumptions and, in turn, hypothesize that the system evolves according to the laws of classical physics, i.e., evolution is described by differential equations, the *randomness* of the measurement result will only be associated with the unknown initial conditions. As above, it can be said that the sequence of measurement results is in this case pseudorandom, since it is determined under the known law of evolution by the uncertainty of the initial conditions alone.

It is often argued in favor of classical physical systems that their evolution is complex, and the trajectories corresponding to close initial conditions diverge exponentially rapidly in the phase space. Nevertheless, they are still trajectories, and if the initial conditions are known, the trajectories are exactly predictable.

A good example that illustrates the 'determinism' of classical systems is the 'Galton board' [2], used to demonstrate the law of the normal distribution of probabilities as a result of the application of the central limit theorem. [3]

The Galton board is a system with hard metal balls that fall from the center of the upper part of the board through a large number of thin pins located below in a checkerboard pattern (Fig. 1). The system is a purely classical one. Falling down, the ball undergoes elastic reflections (deflections) in one direction or the other from the pins it encounters on its way and eventually falls into one of the boxes located horizontally below. The resulting horizontal displacement is interpreted as the sum of individual random deviations (as the sum of a large number of random variables), whose probability distribution according to the central limit theorem should be normal. This is confirmed by the visual similarity of the final picture of the distribution of balls by boxes (histograms) with the density of the normal distribution.

Can such a system that behaves according to the laws of classical physics lead to the generation of randomness? Apparently it can't.

The trajectory of each ball and its final position, namely the box in which it arrives, can be reliably predicted if the angle and speed at which the ball enters the first row are known. Small but known deviations in the initial angle of incidence and initial velocity of each ball lead to divergence of trajectories and, ultimately, to a distribution over the boxes, similar to the normal one.

We emphasize once again that if all the initial angles and velocities are known, the entire distribution over the boxes is unambiguously predictable. The apparent 'randomness' is only associated with the unreliability of the initial conditions.

It is of importance — and this is a property of any classical system — that, if it is prepared at the initial moment in the same initial conditions and undergoes the same evolution, it evolves to the same final result. Classical physical generators are in this sense pseudo-random. The evolution of any complex classical system is completely predictable at the *fundamental level* (can be calculated and predicted) provided initial conditions are known.

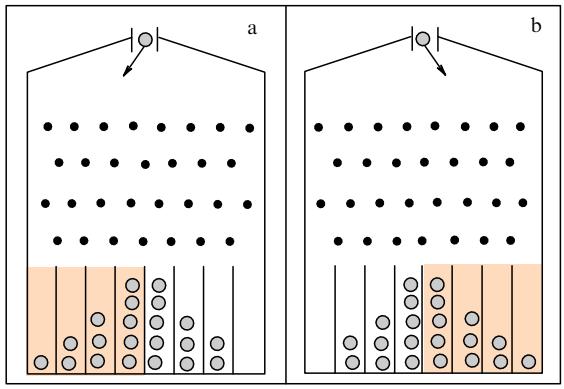

Figure 1. Galton board is an example of a classical physical system that illustrates the emergence of 'randomness' under unknown initial conditions. A speculative example of the extraction of 'randomness': (a) 'random' bit 0 is assigned if, after dropping balls, the number of balls in the left half of the cells is larger than in the right half; (b) 'random' bit 1 is assigned if, after dropping balls, the number of balls in the right half of the cells is larger than in the left half. If such a generator of 'random' 0s and 1s operates as a black box (only the distribution of balls among the boxes is available at the output, but not the initial conditions), the attacker can control the primary randomness at his/her discretion (set the initial position and speed of the balls) and the sequence of 0s and 1s will be compromised; as a result, the key will be fully known rather than being secret.

The 'output' of the Galton board in this example is not *truly* random. However, if the origin of 'randomness' is not *known a priori*, i.e., it is not known in advance that the distribution of balls over boxes is governed by known initial conditions, such a source can be taken as truly random. This example also shows why it is of importance to know and control the source of primary randomness.

Special attention to the origin of primary randomness extracted from physical generators is given in symmetric cryptography systems.

In conditions where no cryptographic methods of decreasing the strength have been found that use the algorithmic weaknesses of the encryption transformation, the strength of the cryptosystem is determined solely by the *secrecy* of the key, i.e., by how close the selection of a key from the key set is to a random and equiprobable choice.

A situation may be imagined in which an equiprobable bit key is obtained, for example, using a reduction in sample values (final deviations) of the Galton board: if the sample value is greater than the average value, the bit is assumed to be 1; if less, the bit is 0. However, if the generator operates as a black box, i.e., outputs only the distribution of balls without control of the initial data, and the attacker is able to control these data, a key obtained in this way can apparently be compromised (see also the explanations of Fig. 1).

Thus, the output of any random number generator that uses measurements in the classical system as primary randomness cannot be considered truly random.

The very concept of probability is absent in classical physics in the sense that mathematical probability theory is a separate mathematical discipline, in no way 'linked' to classical physics. More precisely, the theory of probability is one external to classical physics: a theory that is introduced

[2] In our opinion, such a fully natural approach to justifying the randomness of the output sequence of physical generators of random numbers, apparently due to the complexity of the problem, has not been presented in research publications, including reviews, on this topic (see, for example, [2]). The authors of publications usually limit themselves to issues of practical implementation, using the unpredictability of a particular physical process (for example, the phenomenon of jitter [3] or metastability [4]) as the basis for constructing an RNG.

[3] The Galton board was initially used for illustration in questions concerning the inheritance of genetic traits [5].

into classical physics 'manually' in an explicit or implicit way through the uncertainty of the initial conditions. The laws of classical physics are true for 'macro-objects' and lose their validity as the size of the system decreases.

We now explain this assertion using the example of the Galton board. The positions of the fixed pins definitely undergo zero fluctuations, which, if strictly taken into account, should lead to a deviation from the trajectory calculated in the 'classical' way. Will such a contribution to the 'classical' trajectory be of significance? The velocity inaccuracy in comparison with classical estimates after collision with the pin due to its position Δx can be estimated from the uncertainty relation

$$\Delta v \approx \frac{\hbar}{\Delta x} \approx 10^{-27} \text{ cm s}^{-1} \qquad (2)$$

with a pin mass of 1 g. Such a contribution of quantum effects, of course, cannot be noticed. As the size of the system decreases, neglecting this contribution results in increasingly larger errors.

We now turn to the analysis of physical generators of the second type — quantum ones (see, for example, review [6]). Extraction of randomness in such a generator is based on the measurement of a quantum system.

Unlike measurements in classical physics, measurements in a quantum system, each time prepared in a certain and the same state, yield a random result, which is a fundamental law of nature in the microcosm. Therefore, only quantum random number generators can be truly random.

The evolution of a quantum system is, generally speaking, also described by differential equations and depends on the initial conditions. However, even if initial conditions are known, it is fundamentally impossible to predict the outcome of measurements in a quantum system. The result of observations or measurements is fundamentally unpredictable under the same initial conditions and the same evolution of a quantum system. This is the main difference between quantum systems and classical ones. *There are no fundamental prohibitions in classical physics on measuring the state of a classical system without perturbing it.*

The following example is usually given as a speculative situation that illustrates the fundamental unpredictability of the result of measurement in a quantum system (Fig. 2).

The source emits each time the same single-photon packet, which hits a 50/50 symmetric beam splitter, behind which are located two detectors, D_0 and D_1. The response of *only one* detector can be recorded, and with the same initial conditions — preparation of a single-photon packet and its evolution — it is fundamentally impossible to predict which of the detectors will be activated. True randomness only takes place in the quantum domain, in which probability is *built into* the apparatus of quantum mechanics, in contrast to probability in classical physics, into which it is introduced from the outside.

Details of this assertion should be clarified.

The result of measurements in a quantum system in state $|\psi\rangle$ is reduced to projecting the state of the system onto one of the states of the measurement basis $\{|\phi_i\rangle\}_{i=1}^{N}$, where $\{|\phi_i\rangle\}_{i=1}^{N}$ are orthonormal states and N is the number of measurement outcomes (we only consider von Neumann's orthogonal measurements). The squared modulus of scalar product

$$P_\psi(i) = \left|\langle\phi_i|\psi\rangle\right|^2 \qquad (3)$$

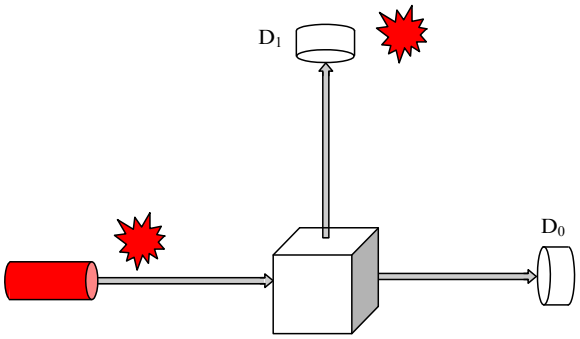

Figure 2. Simple example of a quantum system that illustrates the fundamental unpredictability of the results of measurements in a quantum system.

is interpreted as a probability. This actually reflects the Born interpretation of the squared modulus of the state vector — the wave function. The sum of probabilities for all measurement outcomes is equal to one:

$$\sum_{i=1}^{N} P_\psi(i) = \sum_{i=1}^{N} \left|\langle\phi_i|\psi\rangle\right|^2 = \sum_{i=1}^{N} \langle\psi|\phi_i\rangle\langle\phi_i|\psi\rangle$$
$$= \langle\psi|\left(\sum_{i=1}^{N} |\phi_i\rangle\langle\phi_i|\right)|\psi\rangle = \langle\psi|(I)|\psi\rangle = 1\,, \qquad (4)$$

where I is the unit operator. Probability is in this sense built into the apparatus of quantum mechanics.

Thus, the measurement of quantum systems yields a probability built into the measurement process itself, and the measurement result cannot be predicted in principle, in contrast to that in classical systems.

An important step in developing a quantum random number generator is to find a suitable quantum system and a way to make a measurement in it to extract quantum randomness in the 'purest' form. Examples of such quantum processes are α-decay and the photoelectric effect. It should be noted that quantum effects were used to create random number generators in cryptography earlier, for example, based on sources of radioactive radiation. The resulting generators were technically flawed and slow. It was difficult to reconcile conflicting requirements: to preserve the quantum nature of the process and provide a high generation rate. This is already possible, however, at the modern technological level. We now consider the current situation in more detail.

The photoelectric effect was discovered by A G Stoletov at Moscow University even before the very concept of quanta had appeared; therefore, it was not explained at that time. A consistent explanation of the photoelectric effect on the basis of quantum theory was given by Albert Einstein [7].[4] The root cause of the randomness (Poisson statistics) of photocounts in the detection of laser radiation being fundamentally quantum in nature is due to the absorption of photons by atoms (see details in [8, 9]).

In creating quantum generators of random numbers in a real situation, it is necessary to attenuate the laser radiation to a quasi-single-photon level to make photocounts not too frequent. This requirement is associated, on the one hand, with the need to obtain for measurements a truly quantum (if possible, single-photon) process, and, on the other hand, with

[4] Although the word 'quantum' itself was not used explicitly in [7], the discreteness of the radiation energy was employed.

the finite recovery time of the photodetector after registration. The recovery of the photodetector before the next registration event ensures the statistical independence of successive photocounts.

We consider first the ideal situation. We choose photodetection of an attenuated coherent state as a suitable quantum process. A coherent state is a superposition of states with different Fock numbers of photons:

$$|\alpha\rangle = \exp\left(-\frac{|\alpha|^2}{2}\right)\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle\,, \tag{5}$$

where $|\alpha|^2 = \mu$ is the average number of photons in a coherent state.

The detection process is formally reduced to projection onto a subspace with the Fock number of photons $k \geqslant 1$. Real avalanche photodetectors do not distinguish the number of photons and only register either the presence or absence of a photocount. Such a measurement is described according to the projection postulate of measurements by the decomposition of the unit:

$$I = |0\rangle\langle 0| + \sum_{k=1}^{\infty}|k\rangle\langle k|\,. \tag{6}$$

Each elementary event is associated with its own projection operator, while measurement (6) has two outcomes: the absence and occurrence of a photocount.

The absence of a photocount is a projector $\mathcal{P}_0 = |0\rangle\langle 0|$ onto a subspace with a zero Fock number of photons (the vacuum component of the state), while the occurrence of a photocount is a projector $\mathcal{P}_{k\geqslant 1} = \sum_{k=1}^{\infty}|k\rangle\langle k|$ onto a subspace with the Fock number of photons $k \geqslant 1$.

The probability of photocount $(*)$ is expressed as (see, for example, [8])

$$P(*) = \mathrm{Tr}\left\{|\alpha\rangle\langle\alpha|\mathcal{P}_{k\geqslant 1}\right\} = 1 - \exp(-\mu)\,; \tag{7}$$

consequently, the probability of the absence of a photocount $(\sqcup)$

$$P(\sqcup) = \mathrm{Tr}\left\{|\alpha\rangle\langle\alpha|\mathcal{P}_0\right\} = \exp(-\mu)\,, \quad P(\sqcup) + P(*) = 1\,. \tag{8}$$

Strictly speaking, state (5) is monochromatic (formally infinitely extended). If measurements are carried out within a finite time window T, a substitution $\mu \to \mu_T$ should be made in Eqns (7) and (8) (see details in [8]). Such a substitution is intuitively understandable, since μ_T is the fraction of the average number of photons which accumulate in the time window[5] T. The inequality $\mu_T \ll 1$ should definitely be fulfilled.

Events $*$ and $\sqcup$ — the occurrence and absence of a photocount — will be taken as the basis for extracting randomness.

It is shown below that only the statistical independence of photocounts in time windows T is required in the procedure for extracting randomness, while the probabilities $P(*)$ and $P(\sqcup)$ per se can be arbitrary. If the requirement of independence is fulfilled, the photodetection procedure leads to a Bernoulli test scheme based on quantum phenomena.

It is fundamentally important that the probability of outcomes be of a quantum nature — the outcomes are fundamentally unpredictable, statistically independent, and truly random.

Any classical scheme obtained in any way from a classical physical system (for example, by reducing the final deflection of a ball for the Galton board) is not truly random — the outcome is exactly predictable if the initial conditions and evolution of the classical system are known.

The main experimental problem in implementing a high-speed quantum random number generator based on the detection of photocounts is the need to fulfil contradictory requirements: on the one hand, the quantum nature of the signal with a small average number of photons per pulse should be ensured, and, on the other hand, a high rate of generation of random numbers should be obtained. A real avalanche photodetector has a quantum efficiency $\eta < 1$; in this case, probabilities (7) and (8) retain their form, but with the replacement $\mu_T \to \eta\mu_T$.

The absorption of an individual photon in the solid-state structure of an avalanche photodetector leads to the production of an electron-hole pair, which is 'amplified', i.e., gives rise to an avalanche of charge carriers, whose current pulse is detected. After an avalanche is triggered, it dissolves, which takes a certain amount of time. The photodetector is not ready for a new detection event until the avalanche dissolves; otherwise, it will lead to a correlation of photocounts and distortion of Poisson statistics, i.e., the photocounts, especially those detected in close time windows, will no longer be independent.

The recovery of the photodetector before the next detection event is the first of two conditions under which the statistical independence of successive photocounts is ensured. The second condition is the stability of the intensity of the laser optical field. The distribution $(P(*), P(\sqcup))$ will be in this case stationary, and successive photodetection events (registration of photons) will be strictly independent [10].

The avalanche dissolution time is an intrinsic characteristic of the photo detector. This time sets a limitation on the photodetection rate and, consequently, on the rate with which a random sequence is generated. Typical times range from several ten to several hundred nanoseconds, which sets a limitation on the generation frequency, even in the optimistic case, varying from 10 to 100 MHz.

The first problem consists in a controllable and provable method of obtaining a primary 'quantum' randomness — a Poisson random process. The closest approach to solving this problem at the physical level is to use an SiPM (Silicon Photo Multiplier) array of photodetectors for detection (see Sections 9 and 10).

The second problem is to efficiently extract a uniformly distributed random sequence of 0s and 1s from a Poisson random process.

The second problem is divided into two stages. The first is the implementation of a physical device that provides a 'quantum' randomness in the form of a Bernoulli sequence of events $*$ and $\sqcup$ — the occurrence or absence of a photocount. The second stage is the extraction of a random and equiprobable sequence of 0s and 1s from the Bernoulli sequence.

[5] The quantities μ and μ_T are dimensionless. The physical meaning of μ is the number of photons (for small $\mu \ll 1$) which can be detected if a coherent state of length $L \approx c/(\delta\omega)$ is completely available for measurement (c is the speed of light and $\delta\omega$ is the width of the spectrum of the state). For the monochromatic state $\delta\omega \to 0$, the length is $L \to \infty$. If only the length ΔL of the entire length is available (consequently, the time interval $T = \Delta L/c$), the fraction $\mu_T = \mu\Delta L/L$ will be detected in this window.

2. Relationship between the amount of information and the amount of randomness

It is intuitively clear that a certain number of truly random and equiprobable bits can be extracted from any finite random sample. In measuring a physical process, it is of importance to know the upper bound of this true randomness to establish the efficiency of each particular method.

We define a discrete random variable A with the distribution $P_A(a)$, $a \in \{a_1, \ldots, a_m\}$. Measurements in the physical system are represented in the form of an n-multiple sample,

$$L_n = (a_{i_1}, \ldots, a_{i_n}),\ a_{ij} \in \{a_1, \ldots, a_m\},\ j = \overline{1,n}\,,$$

from the distribution of the random variable A.

Let $v_1, \ldots, v_m$ be the frequencies of occurrence of the outcomes $\{a_1, \ldots, a_m\}$ in the sequence L_n. For large n, in accordance with the law of large numbers, the frequencies become close (nonstrictly) to the probabilities, i.e., $v_1 \approx nP_A(a_1), \ldots, v_m \approx nP_A(a_m)$. Then, the probability

$$P(L_n) = \prod_{k=1}^{m} P_A^{v_k}(a_k) \approx \prod_{k=1}^{m} \big(P_A(a_k)\big)^{nP(a_k)} = 2^{-nH(A)}\,,$$

where $H(A)$ is the Shannon entropy of the distribution of the random variable A,

$$H(A) = -\sum_{k=1}^{m} P_A(a_k) \log P_A(a_k)\,;$$

all logarithms here and below are taken to base 2.

This implies at the qualitative level that almost all possible sequences $(a_{i_1}, \ldots, a_{i_n})$ that can be obtained by measuring a physical process are equiprobable, and their number is $2^{n'}$, $n' = nH(A)$ (we consider it to be an integer). These are the sequences, which we refer to as typical, in which the number of places occupied by the outcomes $\{a_1, \ldots, a_m\}$ is virtually equal to $\{nP_A(a_1), \ldots, nP_A(a_m)\}$, $\sum_{k=1}^{m} nP_A(a_k) = n$.

We now arrange in order (number) the typical sequences as $(a_{i_1}^{(j)}, \ldots, a_{i_n}^{(j)})$, $j = \overline{0, 2^{n'}-1}$. We then assign to each typical sequence $(a_{i_1}^{(j)}, \ldots, a_{i_n}^{(j)})$ a binary sequence $(\varepsilon_1^{(j)} \ldots \varepsilon_{n'}^{(j)})$ — a binary expansion of the number j, populating in this way the entire set of bit sequences:

$$\left\{ \begin{array}{ccc} a_{i_1}^{(0)}, \ldots, a_{i_N}^{(0)} & \to & 0 \ldots 0 \\ \ldots & \to & \ldots \\ a_{i_1}^{(j)}, \ldots, a_{i_n}^{(j)} & \to & \varepsilon_1^{(j)} \ldots \varepsilon_{n'}^{(j)} \\ \ldots & \to & \ldots \\ a_{i_1}^{(2^{n'}-1)}, \ldots, a_{i_n}^{(2^{n'}-1)} & \to & 1 \ldots 1 \end{array} \right\}. \qquad (9)$$

Then, when measuring the physical process after obtainment of $(a_{i_1}^{(j)}, \ldots, a_{i_n}^{(j)})$ and selection of the corresponding $(\varepsilon_1^{(j)} \ldots \varepsilon_{n'}^{(j)})$, we obtain an *equiprobable* selection of bit sequences of length n', i.e., extract a truly *random* sequence.

Suppose that, in implementing some actual algorithm for extracting randomness, we extract n' random equiprobable bits from a random sample of size n. It is reasonable to adopt the relative value $\lambda = n'/n$ as the 'amount of randomness' measured in bits per measurement of the physical process in extracting a binary equiprobable sequence.

The algorithm for extracting randomness described above is optimal; the 'amount of randomness' is the maximum possible for this algorithm,

$$\lambda_{\max} = \frac{n'}{n} = H(A)\,,$$

but it cannot be effectively implemented in practice, since it requires enormous memory to build and store Table (9). The quantity $\lambda_{\max} = H(A)$ is the upper bound which can be used to assess the effectiveness of a particular actual algorithm.

The result on the equipartition and the corresponding cardinality of the set of typical sequences, formulated above at the qualitative level, is one of the fundamental results of information theory presented in Shannon's theorems for a discrete source of messages without memory [11–14]. The entropy $H(A)$ is also known as the amount of information contained in the probabilistic scheme A.

In the terminology of information theory, the outcomes $\{a_1, \ldots, a_m\}$ are letters (of the alphabet), typical sequences are random messages of a discrete source, and binary decomposition (3) is a set of coded messages. The value $H(A)$, Shannon's entropy, is the amount of information measured in bits per letter of the message that characterizes the minimum length of the binary representation of messages for their transmission over a communication channel without errors.

Thus, the Shannon entropy per letter for a discrete source of random messages is the measure of the maximum amount of true randomness.

The amount of randomness (in the asymptotic limit) for an alphabet $\{\sqcup, *\}$ with a probability distribution $\{P(\sqcup), P(*)\}$ is

$$h\big(P(*)\big) = h\big(P(\sqcup)\big) < 1\,, \qquad (10)$$

where it is taken into account that $P(\sqcup) + P(*) = 1$ and $h(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is the Shannon binary entropy function.

The length of the processed block n is in a real situation always finite. In addition to typical sequences, atypical sequences will emerge in such a situation with, naturally, a lesser, but not vanishing, probability. Such sequences also contain some truly random number, 0 and 1, which is preferable not to lose.

The asymptotic case considered above provides a general way to extract random 0s and 1s. The main idea of the extraction of randomness is based on the fact that typical sequences are equiprobable, since they contain (asymptotically) the same set of frequencies $v_1 = nP_A(a_1), \ldots, v_m = nP_A(a_m)$ of the emergence of symbols $a_1, \ldots, a_m$, and only differ by their permutation.

All sequences can be divided for finite n into several classes that differ in the set of frequencies $v_1, \ldots, v_m$: typical sequences form one of these classes. The sequences differ within each class by permutation of symbols and are equiprobable. The sequences can be numbered for each class, and an attempt can be made to extract truly random bits from equiprobable numbers.

The next question to be answered is: how can an efficient way to number sequences be found within each class? The number of sequences is enormous (exponential in n) even for sufficiently small block lengths n. Straightforward numbering is exponentially complex, and therefore practically impossible to implement. However, we show in Sections 9 and 10 that there is an efficient way to extract truly random bits, which has polynomial complexity in terms of the block length n, provided that the original symbols are independent. This method is implemented in actual devices [15–18], an example of which is given in Section 9.

3. Von Neumann method

Before presenting a polynomial algorithm for extracting random 0s and 1s, which guarantees true randomness and extracts it from all, not just typical, sequences, it is convenient to present a particular method for extracting random 0s and 1s from a Bernoulli sequence that was proposed by von Neumann[6] back in 1951.

We now describe the von Neumann method using the example of the alphabet $\{\sqcup, *\}$ with the probability distribution $\{P(\sqcup), P(*)\}$.

A sequence of events $*$, $\sqcup$ of length n is divided into consecutive pairs without engagement, which are viewed 'on the fly'. The encountered pair combination $(*, \sqcup)$ is replaced by zero, while the combination $(\sqcup, *)$ is replaced by one. The other two paired combinations, $(*, *)$ and $(\sqcup, \sqcup)$, are discarded. The resulting sequence of 0s and 1s is an *equiprobably* distributed random sequence, since the probability of zero $P(0) = P(*)P(\sqcup)$ is equal to the probability of unity $P(1) = P(\sqcup)P(*)$. Thus, the equalities $P(0) = P(1) = 1/2$ are valid in the remaining part, regardless of the values of $P(*)$ and $P(\sqcup)$.

It is of importance to emphasize that this method is applicable for any initial probabilities $P(*)$ and $P(\sqcup)$.

It is easy to see that in the von Neumann method, even in the most favorable case, when $P(*), P(\sqcup)$ are close to $1/2$, the amount of randomness extracted is no more than $1/4$ bit per character of the sequence of events $*, \sqcup$, while the maximum amount of randomness that, in principle, can be extracted from a sequence of events $*, \sqcup$ is equal to $\lambda_{\max} = h(P(*)) = h(\approx 1/2)$, i.e., very close to 1.

Some much more general and important results can be can be obtained from this obvious, simple, and elegant method.

The method may be conveniently presented in the form of a table:

$$\begin{aligned} (\sqcup, \sqcup) &\to \text{discarded}, \\ (\sqcup, *) &\to 0, \\ (*, \sqcup) &\to 1, \\ (*, *) &\to \text{discarded}. \end{aligned} \tag{11}$$

The following steps of the randomness extraction algorithm can be derived from representation (11) of the von Neumann method.

(1) Step 1: the length n of the processed block of the original sequence is selected; in this case, $n = 2$.

(2) Step 2: all blocks of length n are split into different classes so that all representatives (blocks) from the same class have the same number of $\sqcup$ and $*$, and, therefore, the same probability.

(3) Step 3: the classes that consist of one block are discarded — the $\sqcup\sqcup$ class and the $**$ class.

(4) Step 4: all equiprobable blocks within the class are numbered and presented as a correspondence table: $(\sqcup, *) \to 0$, $(*, \sqcup) \to 1$; here, 0 and 1 are block numbers.

(5) Step 5: the block of length $n = 2$ obtained in the experiment is compared with the table; the block number is determined; a binary representation of the block number, in this case 0 or 1, is output.

[6] Von Neumann was also the first to propose a software-based pseudo-random number generator [19].

The fundamental step of the von Neumann method, which we use in what follows, is the division of all possible blocks into classes of equiprobable blocks that contain the same number of $*$ and $\sqcup$ and only differ in the permutation of elements. Equation (11) shows three classes are obtained in the von Neumann method. Class 1 and class 3 contain one element each. These classes are discarded. Class 2 contains two equiprobable elements: $(\sqcup, *)$ and $(*, \sqcup)$. The number of elements in class 2 is equal to a power of two, namely 2^1, and the number of truly random bits extracted is $\log 2^1 = 1$.

It is shown in Section 2 that, to extract all randomness close to the asymptotic limit, large blocks should be numbered. Attempts to solve this problem in a straightforward way using a table of the form (9) turn out to be infeasible; namely, they are exponentially complex in terms of the length n of the processed block.

If, for example, $n = 64$, to record the table, a storage size of $n \times 2^n = 2^{37}$ GB is required. The search for the required number based on the composition of the observed block and, consequently, the binary decomposition of its number requires viewing the entire table and consists on average of $2^n \approx 10^{19}$ steps.

We discuss in Sections 5–7 an efficient polynomial way of numbering and extracting true randomness 'on the fly'. The method requires a storage size of $n^3 = 2^5$ kB.

4. Limiting the number of equiprobable bits. Accurate statements

We now formulate an accurate statement about the limiting number of equiprobable bits that can be extracted from a nonequiprobable sequence that consists of events $*$ and $\sqcup$ and has a length n with $n \to \infty$.

All sequences of length n can be divided into disjoint classes $\mathcal{R}_n(k)$, $k = \overline{0, n}$, within which the sequences contain k events $*$ and, consequently, $n-k$ events $\sqcup$. The number of sequences in the class is

$$\left|\mathcal{R}_n(k)\right| = C_n^k = \frac{n!}{k!(n-k)!}; \tag{12}$$

all sequences from a given class have the same probability:

$$P_n(k) = (1-p)^{n-k} p^k, \quad p = P(*), \quad 1 - p = P(\sqcup). \tag{13}$$

We denote

$$\ell_n(k) = \left[\log\left(C_n^k\right)\right], \tag{14}$$

where $[x]$ is the integer part of the number x.

We associate each sequence from the class $\mathcal{R}_n(k)$ with a binary sequence from the set $\{0,1\}^{\ell_n(k)}$. If $\log C_n^k$ is not an integer, we discard the 'unnecessary' sequences from the class $\mathcal{R}_n(k)$, i.e., delete them from the real sample.

This choice is equiprobable provided that a sequence of length n is chosen from the class $\mathcal{R}_n(k)$ (taking into account the deleted sequences); therefore, the selection of binary sequences of length $\ell_n(k)$ also becomes equiprobable.

The average number of random equiprobable bits is defined as

$$\mathcal{L}_n = \sum_{k=0}^{n} \ell_n(k) C_n^k (1-p)^{n-k} p^k = \sum_{k=0}^{n} \ell_n(k) C_n^k P_n(k). \tag{15}$$

Proposition 1. The limit

$$\lim_{n\to\infty}\frac{1}{n}\,\mathcal{L}_n = h(p)\,, \tag{16}$$

where $h(p)$ is the Shannon binary entropy function,

$$h(p) = -p\log(p) - (1-p)\log(1-p)\,.$$

Proof. We use Stirling's formula

$$\begin{aligned} &n! = \sqrt{2\pi n}\,n^n\exp(-n)\bigl(1+o(1)\bigr)\,,\\ &\log(n!) = n\log n\bigl(1+o(1)\bigr) \end{aligned} \tag{17}$$

to estimate binomial coefficients.

There is also an asymptotic representation for the sum of the probabilities of the Bernoulli distribution beyond the deviation of $\ln n\sqrt{n}$ from the np-mathematical expectation, $n\to\infty$ [20]:

$$\sum_{k\notin(np-\ln n\sqrt{n},\,np+\ln n\sqrt{n})} C_n^k P_n(k) = o(1)\,. \tag{18}$$

We also present

$$\ell_n(k) = \log C_n^k - \varepsilon_n(k)\,, \tag{19}$$

where $0\leqslant\varepsilon_n(k)\leqslant 1$. Also, since

$$\sum_{k=0}^{n} C_n^k = 2^n\,, \tag{20}$$

for any $k=\overline{0,n}$, the following inequality holds true:

$$\log C_n^k \leqslant n\,. \tag{21}$$

It is easy to follow then the following chain of relations:

$$\begin{aligned} \frac{1}{n}\mathcal{L}_n &= \frac{1}{n}\sum_{k=0}^{n}\ell_n(k)C_n^kP_n(k) = \frac{1}{n}\sum_{k=0}^{n}\bigl(\log C_n^k\bigr)C_n^kP_n(k)\\ &+ O\left(\frac{1}{n}\right) = \frac{1}{n}\sum_{k\in(np-\ln n\sqrt{n},\,np+\ln n\sqrt{n})}\bigl(\log C_n^k\bigr)C_n^kP_n(k)\\ &+ o(1) = \frac{1}{n}\bigl(n\log n - \bigl(np\log np + n(1-p)\log n(1-p)\bigr)\bigr)\\ &\times\sum_{k\in(np-\ln n\sqrt{n},\,np+\ln n\sqrt{n})} C_n^kP_n(k)\bigl(1+o(1)\bigr) = h(p)\bigl(1+o(1)\bigr)\,. \end{aligned} \tag{22}$$

Therefore, limit (22) is equal to

$$\lim_{n\to\infty}\frac{1}{n}\,\mathcal{L}_n = h(p)\,. \tag{23}$$

Proposition 1 proved is closely related to Shannon theorems [11–13] for a discrete source of memoryless messages, which are fundamental results of classical information theory.

It follows from Shannon theorems that, for a source of random memoryless messages, i.e., with an independent choice of letters (in our case, $*$ and $\sqcup$), virtually all messages have (asymptotically) the same probability $2^{-nh(p)}$ (1st theorem), and their number is equal to $2^{nh(p)}$ (2nd theorem). These are the so-called typical sequences that make up a set of sequences, the probability of which is very close to 1. The length of the binary encoding or binary numbering of all typical sequences apparently requires $\log 2^{nh(p)} = nh(p)$ binary digits, or $h(p)$ binary digits per letter of the message. This corresponds to the limiting mean value in (23) (see a qualitative illustration of the set of all sequences and the set of typical sequences in Fig. 3). It is apparent that we obtain for an equiprobable choice of typical sequences an equiprobable choice of binary vectors that correspond to the binary representation of the numbers of typical sequences.

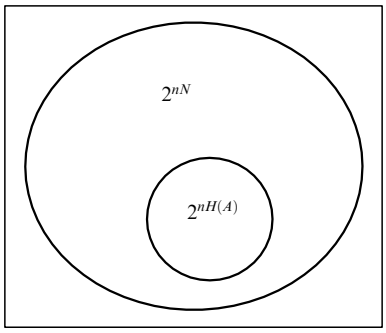

Figure 3. Illustration of the complete set and the set of typical sequences generated by a discrete memoryless source: each point in the set is a sequence. Both sets have an exponentially large size in the length of the sequences; nevertheless, the probability of getting into the set of typical sequences for a large series of usage of the source tends to one; consequently, the probability that the sequence falls into an exponentially large set of atypical sequences tends to zero.

Source entropy $h(p)<1$ is the limiting (maximum) number of truly random bits per symbol of the primary Bernoulli sequence of events $*$ and $\sqcup$.

5. Practical numbering

The assertions made in Section 4 are valid in the asymptotic limit $n\to\infty$. Asymptotic results are insufficient for the construction of quantum generators of random numbers for two reasons.

Reason 1. The Shannon coding theorem for a source is, in fact, an existence theorem, since it does not provide a constructive, algorithmically efficient — polynomial — way of numbering.

Reason 2, or rather the question: how do we proceed with sequences that are not typical? It should be recalled that typical sequences (asymptotically) are those in which the numbers of events $*$ and $\sqcup$ are very close to the mean values $nP(*)$ and $nP(\sqcup)$. The number of events $*$ and $\sqcup$ at a large but finite length n 'breathes', i.e. experiences noticeable fluctuations relative to their mathematical expectations. Such sequences, of course, also contain a certain number of truly random bits that should not be lost. Atypical sequences emerge with a much lower overall probability than typical ones, but, nevertheless, this probability is not zero.

For this reason, in order to extract all the randomness that is contained in all Bernoulli sequences (blocks) of finite length, it is desirable to obtain a provable method for extracting truly random 0s and 1s from all sequences — with any number of $*$ and $\sqcup$ — and not only from typical ones.

To solve this problem, we use the binary coding method discovered by V F Babkin [21] in 1971, a technique for enumerating Bernoulli sequences with polynomial resources in time and memory. The method originated in the theory of arithmetic coding (another name is lossless coding), and, in our opinion, it should long ago have attracted attention in the development of random number generators. This method is undoubtedly a gem of the coding theory.

We now proceed to a description of the method. We consider a block of length n, in which k events $*$ occurred at places $i_1, i_2, \ldots, i_k$, $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$.

Assigned to the block is a number,

$$\mathrm{Num}\,(i_1, i_2, \ldots, i_k) = C^1_{i_1-1} + C^2_{i_2-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\,, \tag{24}$$

where, as usual, we set $C^i_j = 0$ if $j < i$. This equality is the essence of the numbering method proposed by Babkin.

Given the exceptional importance of creating a high-speed RNG and to promote Babkin's scientific heritage, we present here two propositions that establish a one-to-one correspondence between the composition $(i_1, i_2, \ldots, i_k)$ of the processed block and its number $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$ calculated as the sum of binomial coefficients (24) [21].

Proposition 2. The following relations hold:

$$\begin{aligned} &\min_{i_1, i_2, \ldots, i_k} \mathrm{Num}\,(i_1, i_2, \ldots, i_k) = 0\,,\\ &\max_{i_1, i_2, \ldots, i_k} \mathrm{Num}\,(i_1, i_2, \ldots, i_k) = C^k_n - 1\,. \end{aligned} \tag{25}$$

Proof. Indeed, it is easy to verify that

$$\min_{i_1, i_2, \ldots, i_k} \mathrm{Num}\,(i_1,\, i_2, \ldots,\, i_k) = \mathrm{Num}(1, 2, \ldots, k) = 0\,. \tag{26}$$

Next,

$$\begin{aligned} &\max_{i_1, i_2, \ldots, i_k} \mathrm{Num}\,(i_1, i_2, \ldots, i_k)\\ &\quad= \max_{i_1, i_2, \ldots, i_k} \left(C^1_{i_1-1} + C^2_{i_2-1} + \ldots + C^k_{i_k-1}\right)\\ &\quad= \mathrm{Num}\,(n-k, n-k+1, \ldots, n)\\ &\quad= C^1_{n-k-1} + C^2_{n-k} + \ldots + C^k_{n-1} = C^k_n - 1. \end{aligned} \tag{27}$$

The last equality can be easily obtained by the consistent application of the well-known relation

$$C^k_n = C^{k-1}_{n-1} + C^k_{n-1}\,. \tag{28}$$

Proposition 3. The relationship between blocks with k events $*$ at places $i_1, i_2, \ldots, i_k$ and numbers $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$ is a one-to-one correspondence.

Proof. Consider two blocks with numbers

$$\begin{aligned} &\mathrm{Num}\,(i^{(1)}_1, \ldots, i^{(1)}_s, i_{s+1}, \ldots, i_k)\,,\\ &\mathrm{Num}\,(i^{(2)}_1, \ldots, i^{(2)}_s, i_{s+1}, \ldots, i_k)\,, \end{aligned} \tag{29}$$

and let $i^{(1)}_s > i^{(2)}_s$ be the first number (starting from the right) at which the blocks are 'separated' by the positions of event $*$.

To prove a one-to-one correspondence using *Proposition* 2, it is sufficient to show that the numbers of two different blocks in Eqn (29) do not coincide for any values of the positions of event $*$. This will be the case if the difference between the numbers in (29) is positive. We have

$$\begin{aligned} &\mathrm{Num}\,(i^{(1)}_1, \ldots, i^{(1)}_s, i_{s+1}, \ldots, i_k)\\ &- \mathrm{Num}\,(i^{(2)}_1, \ldots, i^{(2)}_s, i_{s+1}, \ldots, i_k)\\ &\geqslant \min_{i^{(1)}_1, \ldots, i^{(1)}_{s-1}} \mathrm{Num}\,(i^{(1)}_1, \ldots, i^{(1)}_{s-1}, i^{(1)}_s, i_{s+1}, \ldots, i_k)\\ &- \max_{i^{(2)}_1, \ldots, i^{(2)}_{s-1}} \mathrm{Num}\,(i^{(2)}_1, \ldots, i^{(2)}_{s-1}, i^{(2)}_s, i_{s+1}, \ldots, i_k)\,. \end{aligned} \tag{30}$$

We find

$$\begin{aligned} &\min_{i^{(1)}_1, \ldots, i^{(1)}_{s-1}} \mathrm{Num}\left(i^{(1)}_1, \ldots, i^{(1)}_{s-1}, i^{(1)}_s, i_{s+1}, \ldots, i_k\right)\\ &= \mathrm{Num}\left(1, \ldots, s-1, i^{(1)}_s, i_{s+1}, \ldots, i_k\right)\\ &= 0 + C^s_{i^{(1)}_s - 1} + \left(C^{s+1}_{i_{s+1}-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\right). \end{aligned} \tag{31}$$

We use formula (27) with $n = i^{(2)}_s$, $k = s$ and take into account that $i^{(2)}_{s-1} < i^{(2)}_s$. We then obtain

$$\begin{aligned} &\max_{i^{(2)}_1, \ldots, i^{(2)}_{s-1}} \mathrm{Num}\left(i^{(2)}_1, \ldots, i^{(2)}_{s-1}, i^{(2)}_s, i_{s+1}, \ldots, i_k\right)\\ &= \max_{i^{(2)}_1, \ldots, i^{(2)}_{s-1}} \left\{C^1_{i^{(2)}_1 - 1} + C^2_{i^{(2)}_2 - 1} + \ldots + C^{s-1}_{i^{(2)}_{s-1} - 1}\right\}\\ &+ \left(C^{s+1}_{i_{s+1}-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\right)\\ &= C^1_{i^{(2)}_s - s - 1} + C^2_{i^{(2)}_s - s} + \ldots + C^{s-1}_{i^{(2)}_s - 2} + C^s_{i^{(2)}_s - 1}\\ &+ \left(C^{s+1}_{i_{s+1}-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\right)\\ &= \left(C^s_{i^{(2)}_s} - 1\right) + \left(C^{s+1}_{i_{s+1}-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\right). \end{aligned} \tag{32}$$

Consequently, taking into account equalities (31) and (32) and the inequality $i^{(1)}_s > i^{(2)}_s$, we have

$$\begin{aligned} &\mathrm{Num}\left(i^{(1)}_1, \ldots, i^{(1)}_s, i_{s+1}, \ldots, i_k\right)\\ &- \mathrm{Num}\left(i^{(2)}_1, \ldots, i^{(2)}_s, i_{s+1}, \ldots, i_k\right)\\ &\geqslant \min_{i^{(1)}_1, \ldots, i^{(1)}_{s-1}} \mathrm{Num}\left(i^{(1)}_1, \ldots, i^{(1)}_{s-1}, i^{(1)}_s, i_{s+1}, \ldots, i_k\right)\\ &- \max_{i^{(2)}_1, \ldots, i^{(2)}_{s-1}} \mathrm{Num}\left(i^{(2)}_1, \ldots, i^{(2)}_{s-1}, i^{(2)}_s, i_{s+1}, \ldots, i_k\right)\\ &= C^s_{i^{(1)}_s - 1} - C^s_{i^{(2)}_s} + 1 \geqslant 1\,. \end{aligned} \tag{33}$$

This fact establishes a one-to-one correspondence between the blocks from set $\mathcal{R}_n(k)$ and their numbers [21].

We now give another simple heuristic proof of *Proposition* 2 based on recursion.

Let the number of the sequence with k events $*$ in positions $i_1, i_2, \ldots, i_k$ be $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$. Further, let the number of the sequence with $k-1$ events $*$ in positions $i_1, i_2, \ldots, i_{k-1}$ be $\mathrm{Num}\,(i_1, i_2, \ldots, i_{k-1})$.

The sequence numbers $(i_1, i_2, \ldots, i_k)$ and $(i_1, i_2, \ldots, i_{k-1})$ differ by a certain number of sequences. We count this number of sequences: it is equal to the number of ways to place sequences with k $*$, in which the kth photocount $*$ is located in position $i_k - 1$, the position that is previous compared to the sequence in which event $*$ is located in the position i_k. The

number of such sequences is equal to the number of ways to place photocounts in $i_k - 1$ boxes, i.e., $C^k_{i_k-1}$.

Thus, we obtain a recurrent formula 'descending' from the major numbers

$$\begin{aligned}\mathrm{Num}\,(i_1, i_2, \ldots, i_k) &= \mathrm{Num}\,(i_1, i_2, \ldots, i_{k-1}) + C^k_{i_k-1}\\ &= \mathrm{Num}\,(i_1, i_2, \ldots, i_{k-2}) + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1} = \ldots\\ &= C^1_{i_1-1} + C^2_{i_2-1} + \ldots + C^{k-1}_{i_{k-1}-1} + C^k_{i_k-1}\,. \end{aligned} \tag{34}$$

6. Complexity of numbering. Pascal's triangle

Blocks are numbered sequentially, 'on the fly', as events $*$ and $\sqcup$ arrive. The block size n is set, and the table of binomial coefficients (Table 1) with the size $(n-1)\times n$ is calculated once. The value of k is not fixed in advance; cases $k = 0$ and $k = n$ are excluded from consideration as unlikely. No more than n binary digits are required to store each binomial coefficient, which directly follows from the relations $\log C^k_n \leqslant \log 2^n = n$. Thus, the total memory for storing Table 1, which is 'Pascal's triangle', or more accurately, half 'triangle', requires no more than n^3 binary digits.

Table 1. 'Pascal's triangle'.

	1	2	3	4	5	...	$n-1$	n
i_1	0	1	C^1_2	C^1_3	C^1_4	...	C^1_{n-2}	C^1_{n-1}
i_2	0	0	1	C^2_3	C^2_4	...	C^2_{n-2}	C^2_{n-1}
i_3	0	0	0	1	C^3_4	...	C^3_{n-2}	C^3_{n-1}
...	...	...	...	...	...	...	...	...
i_{n-1}	0	0	0	0	0	...	0	1

Numbering of the sequence is reduced to motion along a certain trajectory on Pascal's triangle with successive summation of binomial coefficients (see the example in Fig. 4).

If event $*$ was encountered for the first time at the place m_1, the value of the binomial coefficient is taken at the intersection of the row with the number i_1 (first event $*$) and the column with the number m_1.

If event * is encountered for the second time at the place m_2 ($m_2 > m_1$), the value of the binomial coefficient is taken at the intersection of the row with the number i_2 and the column with the number m_2 and added to the previous value of the binomial coefficient.

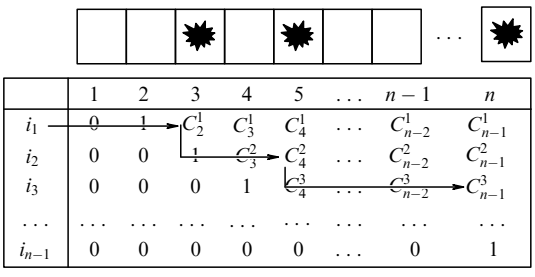

Figure 4. Example of 'Pascal's triangle' that illustrates the calculation of the number of the sequence of photocounts 'on the fly', as the photocounts emerge. An example of a sequence with three photocounts $*$. The numbering is reduced to motion along a trajectory on 'Pascal's triangle' and the sequential summation of the binomial coefficients in the table. In the example, the position number of the first photocount is $i_1 = 3$, the second $i_2 = 5$, and the third $i_3 = n$.

If the event occurs for the sth time at the place m_s ($m_s > m_{s-1}$), the value of the binomial coefficient at the intersection of the row number i_s and the column number m_s is taken and added to the previous sum of the binomial coefficients.

The process halts when the entire block of size n has been scanned. In accordance with Section 5, the number of block with events $*$ and $\sqcup$ is obtained in the form of a binary representation, but these bits are not yet random.

After the number of a specific sequence consisting of $\sqcup$ and $*$ is obtained, a block of truly random 0s and 1s is derived from its binary representation.

7. Extracting randomness

The cardinality of the set of blocks with k events $*$ and $n-k$ events $\sqcup$ is $|\mathcal{R}_n(k)| = C^k_n$. According to Eqn (24), the blocks $\mathcal{R}_n(k) = C^k_n$ are numbered from 0 to $C^k_n - 1$.

Let n be even, which is convenient for computer implementation. The method described below is operable for any n. We consider the representation of $|\mathcal{R}_n(k)|$ as a sum:

$$|\mathcal{R}_n(k)| = 2^{r_m} + \ldots + 2^{r_1} + 2^{r_0}\,, \quad r_m > r_{m-1} > \ldots r_1 > r_0\,. \tag{35}$$

Suppose now that a block has been realized that has the composition $(i_1, i_2, \ldots, i_k)$ of events $*$. The block number has a binary decomposition of the form

$$\begin{aligned}\mathrm{Num}\,(i_1, i_2, \ldots, i_k) &= \varepsilon_{r_m+1}2^{r_m+1} + \varepsilon_{r_m}2^{r_m}\\ &\quad + \varepsilon_{r_m-1}2^{r_m-1} + \ldots + \varepsilon_1 2^1 + \varepsilon_0 2^0, \quad \varepsilon_r \in \{0,1\}\,, \end{aligned} \tag{36}$$

and the corresponding binary representation $(\varepsilon_{r_m+1}, \varepsilon_{r_m}, \varepsilon_{r_m-1}, \ldots, \varepsilon_1, \varepsilon_0)$.

The block $\{\varepsilon\}$ of random 0s and 1s is extracted from the binary representation $(\varepsilon_{r_m+1}, \varepsilon_{r_m}, \varepsilon_{r_m-1}, \ldots, \varepsilon_1, \varepsilon_0)$; the extraction is carried out in different ways, depending on the range of numbers between 0 and $C^k_n - 1$ into which the number $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$ of the current block falls. Namely:

$$\begin{array}{ll} \text{Number} & \text{Block } \{\varepsilon\} \text{ of random 0 and 1}\\ 0 \leqslant \mathrm{Num}\,(i_1, i_2, \ldots, i_k) \leqslant 2^{r_0} - 1\,, & \varepsilon_{r_0-1}, \ldots, \varepsilon_0\,,\\ 2^{r_0} \leqslant \mathrm{Num}\,(i_1, i_2, \ldots, i_k) \leqslant 2^{r_0} + 2^{r_1} - 1\,, & \varepsilon_{r_1-1}, \ldots, \varepsilon_0\,,\\ 2^{r_0} + 2^{r_1} \leqslant \mathrm{Num}\,(i_1, i_2, \ldots, i_k) \leqslant 2^{r_0} + 2^{r_1} + 2^{r_2} - 1\,, & \varepsilon_{r_2-1}, \ldots, \varepsilon_0\,,\\ \ldots & \ldots\\ 2^{r_0} + \ldots + 2^{r_m} \leqslant \mathrm{Num}\,(i_1, i_2, \ldots, i_k) \leqslant 2^{r_0} + \ldots + 2^{r_m} - 1\,, & \varepsilon_{r_m-1}, \ldots, \varepsilon_0\,. \end{array} \tag{37}$$

We now number the rows (inequalities) as $0, \ldots, j, \ldots, m$. Then, the jth row, the subclass, contains 2^{r_j} different equiprobable numbers $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$, which correspond in a unique way to binary vectors from the space $\{0,1\}^{r_j}$. Then, for each current number $\mathrm{Num}\,(i_1, i_2, \ldots, i_k)$, the corresponding block $\{\varepsilon\}$ of random 0s and 1s is output (Table 2).

We consider an example that illustrates the general method for $n = 8$ and $k = 2$. In this case,

$$\begin{aligned}&|\mathcal{R}_n(k)| = \frac{8!}{2!6!} = 28 = 2^4 + 2^3 + 2^2,\\ &m = 2, \quad r_m = 4, \quad r_1 = 3, \quad r_0 = 2\,. \end{aligned} \tag{38}$$

Table 2. Example of extraction of truly random blocks.

Positions of $*$ and $\sqcup$ (i_1, i_2)	Number $N(i_1, i_2)$	Binary representation	Random block $\{\varepsilon\} = \varepsilon_{r_{j-1}}, \ldots, \varepsilon_0$
$**\sqcup\sqcup\sqcup\sqcup\sqcup\sqcup$	0	00000	00
	1	00001	01
$j = 0$	2	00010	10
	$3 = 2^{r_0} - 1$	00011	11
	4	00100	100
	5	00101	101
	6	00110	110
	7	00111	111
$j = 1$	8	01000	000
	9	01001	001
	10	01010	010
	$11 = 2^{r_1} + 2^{r_0} - 1$	01011	011
	12	01100	1100
	13	01101	1101
	14	01110	1110
	15	01111	1111
	16	10000	0000
	17	10001	0001
$j = 2$	18	10010	0010
	19	10011	0011
	20	10100	0100
	21	10101	0101
	22	10110	0110
	23	10111	0111
	24	11000	1000
	25	11001	1001
	26	11010	1010
$\sqcup\sqcup\sqcup\sqcup\sqcup\sqcup**$	$27 = 2^{r_2} + 2^{r_1} + 2^{r_0} - 1$	11011	1011

8. True randomness

We show below that, if Babkin's numbering method is implemented and blocks $\{\varepsilon\}$ of random 0s and 1s are extracted from it as its result, any output binary sequence of any length L will be equiprobable, i.e., truly random. It should be noted once again that this is only possible under the initial assumption that the sequence of events consisting of $*$ and $\sqcup$ (at the physical level, the photocounts) is the Bernoulli type, i.e., independent.

We consider a random independent sequence of events consisting of $*$ and $\sqcup$, which we split into a sequence of blocks of length n. We then obtain an independent sequence of pairs of numbers (k_i, j_i), $i = 1, 2, \ldots$, where k_i is the number of events $*$ in the ith block, and j_i is the number of the subclass in which the block number falls with a joint distribution $P(k, j)$, whose exact form, as is shown below, is insignificant. *It is only important that the sequence of pairs* (k_s, j_s), $s = 1, 2, \ldots$ *be statistically independent*.

We fix k, the number of events $*$ in a block of length n. Then, the block number $\mathrm{Num}(i_1, i_2, \ldots, i_k)$ falls randomly into one of the subclasses (see Eqn (37) and Table 2), which are numbered (see Section 7) with numbers $0, \ldots, j, \ldots, m$. We denote this subclass as $\mathcal{R}_n(k, j)$; its size (cardinality) by construction (see Table 2) is equal to 2^{r_j} (r_j, generally speaking, depends on k). The subsets of the numbers $\mathcal{R}_n(k, j)$ do not overlap and are a partition of the entire set $\mathcal{R}_n(k)$ of numbers $\mathrm{Num}(i_1, i_2, \ldots, i_k)$: $\mathcal{R}_n(k) = \bigcup_{j=0}^{m} \mathcal{R}_n(k, j)$.

So, the entire randomness is now set on the original Bernoulli sequence of events $*$ and $\sqcup$.

Let an event — a block of size n with a fixed pair (k, j) — occur. The question may be asked then: what is the probability that a specific number $\mathrm{Num}(i_1, i_2, \ldots, i_k)$ from the jth subclass of numbers $\mathcal{R}_n(k, j)$ will be chosen provided that it falls there?

The probability of interest to us is the conditional probability, which has the form

$$P\left(\mathrm{Num}(i_1, i_2, \ldots, i_k) | \mathrm{Num}(i_1, i_2, \ldots, i_k) \in \mathcal{R}_n(k, j)\right) = \frac{P\left(\mathrm{Num}(i_1, i_2, \ldots, i_k)\right)}{\sum_{\mathrm{Num}^*(i_1, i_2, \ldots, i_k) \in \mathcal{R}_n(k,j)} P\left(\mathrm{Num}^*(i_1, i_2, \ldots, i_k)\right)} = \frac{P^k(*)P^{n-k}(\sqcup)}{2^{r_j}P^k(*)P^{n-k}(\sqcup)} = 2^{-r_j}. \quad (39)$$

Since, by construction, each number $\mathrm{Num}(i_1, i_2, \ldots, i_k)$ from the jth subclass is associated with the corresponding binary block $\varepsilon_{r_j-1}, \ldots, \varepsilon_0$, it follows from (39) that, for fixed pair (k, j), an equiprobable scheme for choosing binary vectors $(\varepsilon_{r_j-1}, \ldots, \varepsilon_0)$ from the space $\{0, 1\}^{r_j}$ is realized. It is also easy to see that, if the pair (k, j) is fixed, the bits of any part of the $(\varepsilon_{r_j-1}, \ldots, \varepsilon_0)$ segment will also appear in a random and equiprobable way.

We now show that for any binary sequence $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L)$ extracted from the original sequence of events $*$ and $\sqcup$ its probability is

$$P(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L) = \frac{1}{2^L}.$$

By construction (see Section 7), for a fixed block size n, a specific binary sequence $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L)$ can be obtained from any sequence of events $*$ and $\sqcup$, the length of which is limited by the value of nM for some maximum M, where n is block size. Each such sequence generates a sequence of pairs $(k_1, j_1), (k_2, j_2), \ldots, (k_M, j_M)$, which are realizations of possible values of k_s, the number of events $*$ in blocks, and j_s, the numbers of subclasses, $s = \overline{1, M}$; the pairs (k_s, j_s) form individual binary segments of the $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L)$ sequence.

It is easy to see from this that a specific binary sequence $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L)$ can be obtained from any sequence of pairs $(k_1, j_1), (k_2, j_2), \ldots, (k_M, j_M)$.

Let the pair (k_1, j_1) generate a bit segment $\boldsymbol{\varepsilon}_1 = (\varepsilon_1, \ldots, \varepsilon_{m_1})$ of length m_1, the pair (k_2, j_2) generate a bit segment $\boldsymbol{\varepsilon}_2 = (\varepsilon_{m_1+1}, \ldots, \varepsilon_{m_1+m_2})$ of length m_2, etc., and the pair (k_{M^*}, j_{M^*}) generate a bit segment $\boldsymbol{\varepsilon}_{M^*} = (\varepsilon_{m_1+1+\ldots+m_{M^*-1}+1}, \ldots, \varepsilon_{m_1+\ldots+m_{M^*}})$ of length m_{M^*} so that

$$m_1 + m_2 + \ldots + m_{M^*} = L,\ M^* \leqslant M. \quad (40)$$

We denote by KJ the set of all sequences $(k_1, j_1), (k_2, j_2), \ldots, (k_M, j_M)$, $P(KJ) = 1$. We then have the representation

$$P(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L) = \sum_{\left((k_1,j_1),(k_2,j_2),\ldots,(k_M,j_M)\right) \in KJ} \times P\left(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_{M^*}, (k_1, j_1), (k_2, j_2), \ldots, (k_M, j_M)\right). \quad (41)$$

*Due to the independence of the original sequence, random pairs

$$(k_1, j_1), (k_2, j_2), \ldots, (k_{M^*}, j_{M^*}) \quad (42)$$

are independent, and hence independent are the 'triplets'

$$\left(\boldsymbol{\varepsilon}_1, (k_1, j_1)\right), \left(\boldsymbol{\varepsilon}_2, (k_2, j_2)\right), \ldots, \left(\boldsymbol{\varepsilon}_{M^*}, (k_{M^*}, j_{M^*})\right). \quad (43)$$

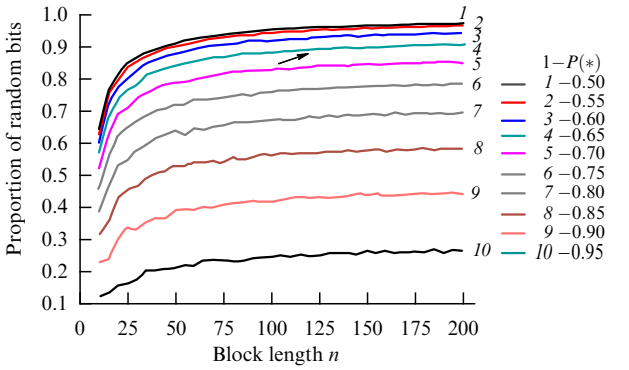

Figure 5. (Color online.) Efficiency of extraction of random bits from the Bernoulli sequence of $*$ and $\sqcup$ at various values of the probabilities $P(*)$ and $P(\sqcup)$ as a function of the length of the processed block n. The sequences of photocounts of $*$ and $\sqcup$ was generated using a mathematical generator of pseudorandom numbers. These dependences are required for a preliminary estimate of the block length n and an estimate of the onset of the asymptotic regime $n \to \infty$ for given experimental values of $P(*)$ and $P(\sqcup)$.

The conditional equiprobability of the choice of binary vectors obtained in (39) implies that

$$P(\boldsymbol{\varepsilon}_s|k_s, j_s) = 2^{-m_s},\ s = \overline{1, M^*}.$$

Hence, using the independence of 'triplets' (43), we obtain

$$\begin{aligned}
&P(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L) \\
&= \sum_{\left((k_1,j_1),(k_2,j_2),\ldots,(k_M,j_M)\right)\in KJ} \prod_{s=1}^{M^*} P\left(\boldsymbol{\varepsilon}_s, (k_s, j_s)\right) \prod_{s=M^*+1}^{M} P(k_s, j_s) \\
&= \sum_{\left((k_1,j_1),(k_2,j_2),\ldots,(k_M,j_M)\right)\in KJ} \prod_{s=1}^{M^*} P(\boldsymbol{\varepsilon}_s|k_s, j_s) P(k_s, j_s) \prod_{s=M^*+1}^{M} P(k_s, j_s) \\
&= \sum_{\left((k_1,j_1),(k_2,j_2),\ldots,(k_M,j_M)\right)\in KJ} \prod_{s=1}^{M^*} 2^{-m_s} P(k_s, j_s) \prod_{s=M^*+1}^{M} P(k_s, j_s) \\
&= 2^{-L} \sum_{\left((k_1,j_1),(k_2,j_2),\ldots,(k_M,j_M)\right)\in KJ} \prod_{s=1}^{M} P(k_s, j_s) = 2^{-L}. \qquad (44)
\end{aligned}$$

Thus, we have shown that any output binary sequence of any length extracted from sequential blocks is equiprobable, i.e., truly random. The equiprobability of a binary sequence is a consequence of the Bernoulli nature (independence) of the initial physical sequence of events $*$ and $\sqcup$, which is provided at the physical implementation level (see Sections 9–12).

To assess the efficiency of extraction of 0s and 1s, depending on the values of $P(*)$, $P(\sqcup)$ and the block length n, a computer simulation was carried out, the results of which are shown in Fig. 5.

9. Physical implementation of a quantum random number generator

To illustrate the ideas outlined in Sections 2–8, we provide in this section an example of a quantum random number generator.

In creating a high-speed quantum random number generator with a sequence of photocounts controlled by quantum laws, a compromise among mutually contradictory requirements has to be found.

For the source of primary randomness to have a quantum nature, the coherent state must be a quasi-single-photon one, i.e., the average number of photons should be $\mu_T \ll 1$. The small average number of photons leads to a low probability of detection in the time window T. The probability of detection is proportional to $\eta\mu_T$, where $\eta < 1$ is the quantum efficiency of the photodetector.

We consider the photodetection process, omitting technical details. The generation of random numbers in photodetection is actually a rather subtle physical experiment in the sense that, as in any physical experiment, verification of any theoretical hypothesis requires elimination of the factors that introduce undesirable distortions. As applied to our situation, the main problem is to implement quantum measurements — photodetection of quasi-single-photon states of radiation — in such a way that the probability really reduces to projecting (see Eqns (3)–(8)). The only devices acceptable for this purpose are avalanche photodetectors.

The event of detecting a photon (Fig. 6) looks like a current (or voltage) pulse — a 'click' from an avalanche of carriers at the photodetector output generated by the absorption of the photon.

Suppose that there is a single-photon state at the input of the photodetector. The photon is absorbed by a particular atom inside the semiconductor structure of the detector, which results in the emergence of an electron-hole pair. However, it is virtually impossible to detect a current pulse from a single electron-hole pair due to its small value. Therefore, the initial electron-hole pair is accelerated in the semiconductor structure and generates an avalanche of nonequilibrium carriers, or a current surge, which is detected.

This circumstance leads to the appearance of dead time — the time of avalanche dissipation. The detector is not ready to register the next photon until dissipation has completed. For this reason, the frequency (clock frequency) of polling the detector cannot be less than the dead time. In addition, in solid-state avalanche detectors, there are so-called afterpulsing effects, which are false alarms after the detection of a real photon. Nonequilibrium carriers can 'stick' to structural defects and then recombine and emit a false photon, which is detected, i.e., parasitic counts occur after detecting a real photon.

Thus, the rate with which a sequence of photocounts is generated is limited by both the small average number of photons and the dead time of the detector.

To eliminate parasitic counts, rather than using single avalanche photodetectors, we use a matrix — a silicon photomultiplier (SiPM) [22] that contains more than a thousand avalanche detectors. The average number of photons in the time window T, which is determined by the clock frequency, does not exceed one thousandth of a photon per pixel, i.e., per individual detector in the SiPM. Therefore, after a photon is detected by a particular individual photodetector, the probability that the next photon will hit the same detector is extremely small. The dead time of a single photodetector does not affect in this case the detection of photons by other detectors in the array, which makes it possible to increase the clock frequency. In fact, only one event of SiPM detection takes place in each time window. Owing to this, it is possible to achieve the Bernoulli property or independence of the photocount sequence in a controlled manner, which can be reliably verified experimentally.

Detection in one pixel may in principle affect detection in another pixel due to electrical interference in the SiPM circuit

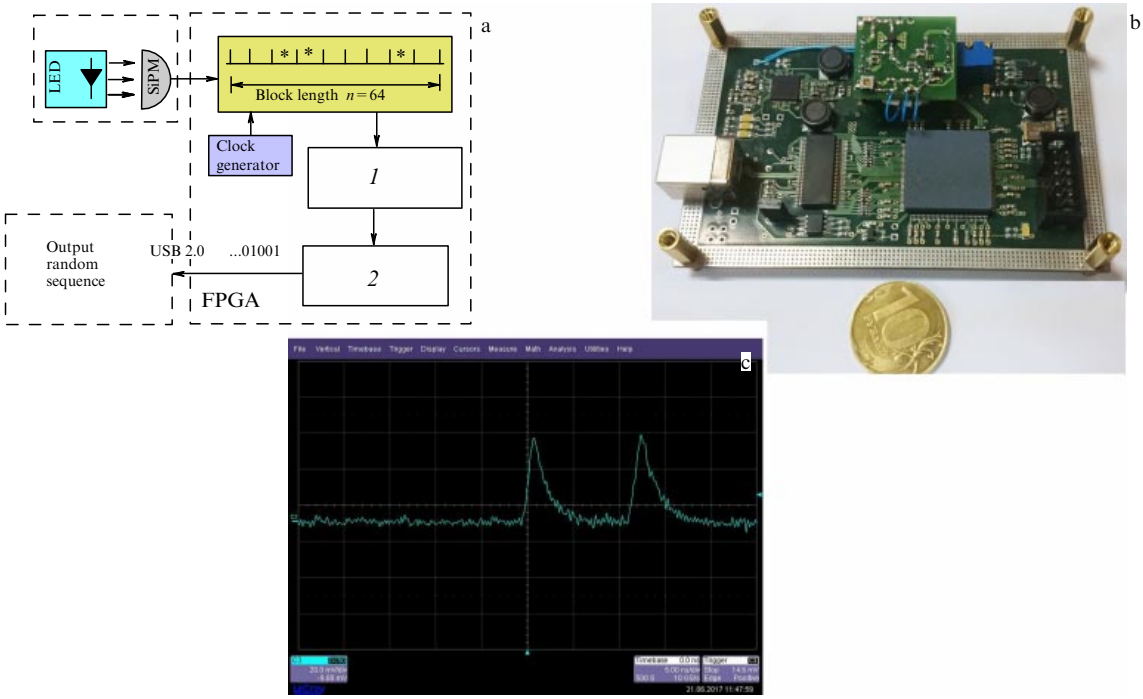

Figure 6. (a) Functional diagram of a quantum random number generator: *1* — conversion of counts into a sequence number using a table with a Pascal triangle, *2* — conversion of the block number into random bits; FPGA is a programmable integrated logic circuit. (b) External view of the device. (c) Example of current pulses that represent the photodetection process.

(so-called crosstalk), which can lead to a distortion of photocount statistics. It should be noted that the possible parasitic effect of cross-talk between neighboring pixels on the statistics of photocounts was studied earlier [23], and no distortion of statistics has been found.

A schematic functional diagram and the appearance of the quantum random number generator are displayed in Fig. 6.

The probability of photon detection in the generator under consideration is $P(*) = 0.3$ in the time interval that corresponds to the clock frequency $f = 200$ MHz of the electronics. In the asymptotic limit, when the length of the processed block is $n \to \infty$, the theoretical limit on the rate of generation of a random sequence is

$$h\big(P(*)\big) f = 0.88 \times 200 \approx 176 \text{ Mbit s}^{-1}.$$

A detector array was used as the SiPM, the technology of which was developed at MEPhI-Pulsar (Moscow, Russia). The matrix was manufactured at the Technological Center of the National Research University Moscow Institute of Electronic Technology (MIET) (Zelenograd, Russia). The SiPM array, whose sensitive area is approximately 1×1 mm, consisted of $N_{\text{pix}} = 1156$ pixels with an active area of 32×32 μm. The operating voltage (several volts above breakdown) was 40 V [22]. The detector temperature was stabilized at 25°C. A Sony laser light-emitting diode (SLD3143VL) with an operating wavelength of 405 nm was used as a radiation source. Post-processing, in which mathematical algorithms were implemented, was performed using an FPGA (Field Programmable Gate Array, Intel FPGA (Altera)) with a clock frequency of 200 MHz. USB 2.0 was used as the external interface for power supply and output in a continuous mode of the resulting binary random sequence. An additional advantage of the SiPM matrix is that it has a large pull-up resistance R_{q}, which exceeds 1 MΩ. This leads to rapid avalanche dissipation and a low probability of post-pulse effects. Another very important feature of such an SiPM is the rather short signal from a pixel, whose duration is about 1 ns.

10. Statistics of photocounts, estimated average number of photons per pixel

To be confident that the generator actually operates in the quantum mode, the average number of photons incident on a single SiPM pixel per clock cycle should be estimated. In the case of a purely quantum regime and Poisson statistics of the number of photons in a pulse, successive records (photocounts) form a Bernoulli sequence. Integer intervals in the number of clock cycles k between successive records are a random variable $\xi \in \{0, 1, \ldots\}$ with the geometric distribution

$$P(\xi = k) = \big(1 - P(*)\big)^k P(*). \tag{45}$$

The logarithm of the probability $\ln P(\xi = k)$ (k is the number of clock cycles) should be, in the case of Poisson statistics, a linear function of k. Figure 7 shows the experimental histogram (dependence $\ln(N(k))$, $N(k)$ is the number of counts in the kth 'box' of the histogram) obtained at the 'effective' sample length of 6×10^9 clock cycles.

The linearity of the plot in Fig. 7 shows the Poisson statistics. If the distance between the photocounts is large, the probability $P(\xi = k)$ is very small, which gives a noticeable error on the graph for values $k > 30$.

To extract randomness, the block length was chosen equal to $n = 64$ clock cycles, a value which is convenient for practical implementation based on the FPGA architecture.

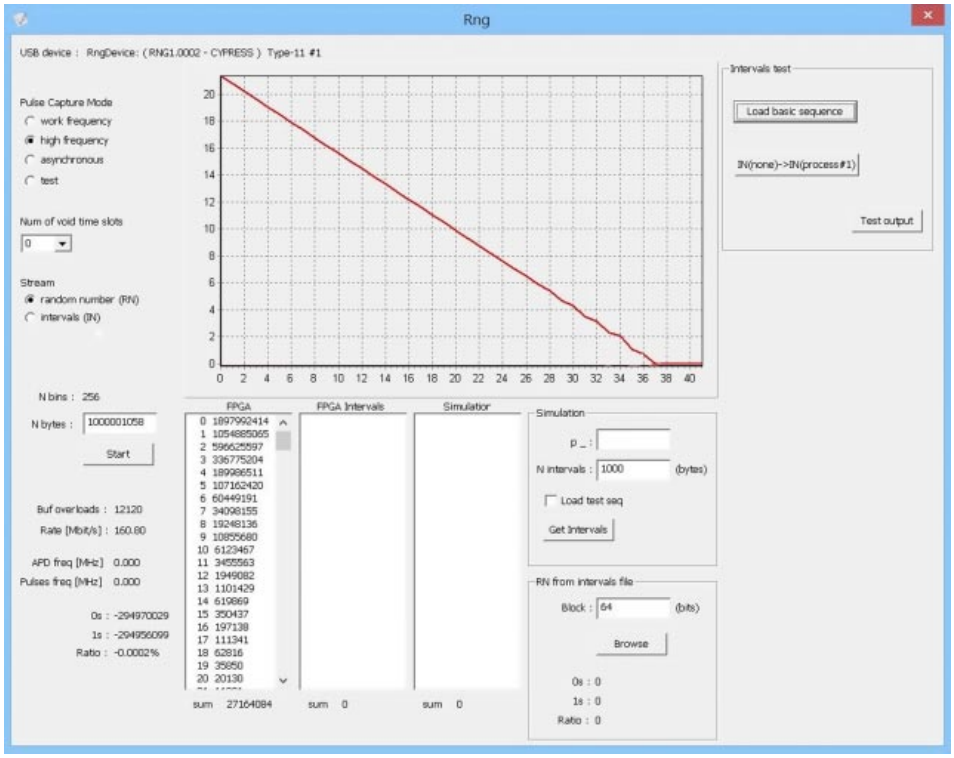

Figure 7. Example of 'real' experimental data used to check the Bernoulli character of photocount sequences. The FPGA column shows the number of photocounts in the histogram versus the number of 'blank' pulses between sequential photocounts. The histogram corresponds to the logarithm of the geometric distribution. The rate of random bit generation from a sequence of photocounts is 160.80 Mbit s^{-1}. The length of the processed block is $n = 64$.

We now estimate μ_T, the average number of photons per clock cycle, which follows from the experimental data.

The experimental histogram shown in Fig. 7 is a linear relationship. The total sample size is $N_{\text{tot}} = 6 \times 10^9$; the number of counts in the histogram zero box $N(0) =$ 1,897,992,414. Given these values and taking into account Eqn (45), the following estimate can be obtained:

$$P(*) = P(\xi = 0) \approx \frac{N(0)}{N_{\text{tot}}} = \frac{1,897,992,414}{6 \times 10^9} \approx 0.3\,. \quad (46)$$

This probability may be represented as

$$P(*) = \mu_T \eta N_{\text{pic}}\,,$$

where μ_T is the average number of photons per pixel in the SiPM per cycle, $\eta = 0.1$ is the quantum efficiency of a pixel, and $N_{\text{pic}} = 1156$ is the number of pixels in the matrix.

We obtain as a result $\mu_T = P(*)/(\eta N_{\text{pic}}) \approx 2.6 \times 10^{-3}$ photons in one clock cycle per pixel, i.e., there are several thousandths of 'photon fractions' per pixel per clock cycle. For a coherent state with a Poisson distribution with the parameter μ_T, the probability of the emergence of one photon is $P_1 = \exp(-\mu_T)\mu_T \approx \mu_T \approx 2.6 \times 10^{-3}$, and the probability of the emergence of two photons is $P_2 = \exp(-\mu_T)\mu_T^2/2 \approx 3.4 \times 10^{-6}$. Thus, it can be argued that a virtually single-photon regime is implemented in the quantum part of the generator.

We now estimate the efficiency of extracting random 0s and 1s in comparison with the theoretical asymptotic limit $h(P(*))f \approx 0.88 \times 200 \approx 176$ Mbit s^{-1}. For a block length of the processed sequence $n = 64$, a clock frequency of 200 MHz, and the rate of generation of random 0s and 1s at a photocount probability $P(*) \approx 0.3$ (see Fig. 5, curve *5*, which corresponds to $1 - P(*) \approx 0.7$), we obtain ≈ 160 Mbit s^{-1}, a value which is close to the theoretical asymptotic limit of 176 Mbit s^{-1}.

11. How to check for randomness. Statistical tests of random sequences

Absence of evidence is not evidence of absence.
Cryptographic slang [7]

The 'aphorism' cited as an epigraph to this section, which is often quoted in discussing randomness tests, fully reflects a fundamental principle — the absence of a 'ruler' that may be used to measure randomness. It is fundamentally impossible to prove that a given sequence of 0s and 1s is truly random, i.e., the fact that the events corresponding to the occurrence of 0 and 1 are strictly equiprobable and independent; it can only be proved that this sequence does not contradict the hypothesis of randomness by some statistical criterion. This implies that the assertion that the sequence is random depends in this sense on the choice of the randomness criterion.

We are dealing with studies with a certain finite sequence or sample that consists of 0s and 1s. The general concept of testing a sequence for randomness is reduced to the following.

We assume that the sequence of 0s and 1s under investigation originated from a source of true randomness, where the probabilities are $P(0) = P(1) = 1/2$, and the choice of 0 and 1 is independent.

[7] Another saying popular in relation to this topic is: "*The absence of proof of guilt is not proof of innocence*," implying that there is no presumption of randomness for the tested sequence.

The latter assumption is axiomatic and only implies that the probability of any binary sequence $(\varepsilon_1, \ldots, \varepsilon_n)$, $\varepsilon_i \in \{0, 1\}$ is set equal to

$$P(\varepsilon_1, \ldots, \varepsilon_n) = \prod_{i=1}^{n} P(\varepsilon_i) = 2^{-n}. \tag{47}$$

This is a fundamental issue. All other mathematical results, one way or another related to the development of a randomness testing system, follow from two equalities (47): the first implies independence, while the second, equiprobability.

What is the essence of the testing system?

If we only have available a binary sequence, generally speaking, of a very large size, and nothing else, we cannot say anything definite about this sequence. It is desirable to obtain some manageable set of values closely related to the binary sequence, which would represent it in a concentrated form and allow development of a reasonable criterion regarding whether this sequence is 'good' or not.

For example, the first value that is most natural is the number of ones in the S_1 sequence. The intuitive understanding of equiprobability is that the value of S_1 should be, for a 'good' sequence, close to $n/2$.

Binary sequences apparently have a different composition of zeros and ones; therefore, the calculated value of S_1 will deviate from $n/2$. In relation to this, we handle the value of S_1 as a statistic, i.e., as a variable defined in a set of observations — equiprobable binary sequences. It is clear that it is not possible to obtain exactly $n/2$; therefore, the next question that arises in developing a criterion is to decide which deviations in statistics can be considered natural for a 'good' binary sequence, and which can not.

This issue is solved using probabilistic methods. An asymptotic $n \to \infty$ probability distribution of the statistics S_1 is found under the conditions of independence and equiprobability (47): in this case, a normal distribution. The asymptotic distribution is 'good' for two reasons:

— its analytical form is known, and there are not so many limiting distributions in the probability theory at all;

— it 'works' for arbitrary but, of course, sufficiently large n.

The asymptotic normal distribution enables obtainment of the probability of deviation for the statistic S_1 in the form

$$P\left(\left|S_1 - \frac{n}{2}\right| > t\frac{\sqrt{n}}{2}\right) = 2\big(1 - \Phi(t)\big), \ t \geqslant 0, \tag{48}$$

where $\Phi(t)$ is the standard normal distribution function,

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t.$$

Let probability (48) be small. This implies that the deviation of the S_1 statistics from $n/2$ by more than $t\sqrt{n}/2$ is unlikely, i.e., such a deviation is unacceptable if the tested sequence is 'good'.

We set a small value of the probability α and find the value of t_α from the equation $2(1 - \Phi(t)) = \alpha$.

We are now ready to formulate a criterion of agreement with hypothesis $\mathcal{H}$ regarding the independence and equiprobability of a binary sequence:

— if $|S_1 - n/2| \leqslant t_\alpha\sqrt{n}/2$, the hypothesis $\mathcal{H}$ is accepted, and the sequence is 'good';

— if $|S_1 - n/2| > t_\alpha\sqrt{n}/2$, the hypothesis $\mathcal{H}$ is rejected, and the sequence is 'bad'.

The problem seems to have been resolved: we set a specific numerical value α, which is the level of significance of the criterion (usually $\alpha = 0.001 - 0.1$), and start using the criterion to test the sequences output from the random number generator.

However, a circumstance emerges that somewhat ruins our harmonious picture.

A sequence is considered 'bad', albeit with a small but not zero probability α (hypothesis $\mathcal{H}$ is rejected), while actually it is 'good'. Thus, performing the test, for example, N times, we go beyond the criterion in about αN cases, provided that the sequence is 'good'. How do we proceed with these sequences? Should they be excluded from consideration? And what about the entire combined sample of size nN? Should it be allowed for 'use' if, for example, the obtained relative number of events of going beyond the criterion boundary is actually close to α? What degree of closeness should be considered 'acceptable'? We are at risk of going in circles.

There is another subtle point. What should be done if in all N tests the statistics never go beyond the criterion, i.e., the agreement is 'too good'? For example, the deviation for a regular sequence 0101010101... will be zero in all tests....

The methodological recommendations published by the US National Institute of Standards and Technology (NIST) [24], which are currently generally accepted practice, suggest the following concept of testing. We consider it using as an example the statistics ζ of deviation from the average number of ones of the form

$$\zeta = \frac{|S_1 - n/2|}{\sqrt{n/4}},$$

which has the distribution function $F_\zeta(t) = 2\Phi(t) - 1$. The NIST recommendations are based on the mathematical observation that the distribution of the random variable $\xi = 1 - F_\zeta(\zeta)$ is uniform in the interval $[0, 1]$. This result remains valid for any random variable ζ with its own distribution function $F_\zeta(t)$.

Let N binary sequences of length n each be tested, as a result of which N deviations $\zeta^{(i)} = |S_1^{(i)} - n/2|/\sqrt{n/4}$, $i = \overline{1, N}$, and, consequently, N values $\xi^{(i)} = 1 - F_\zeta(\zeta^{(i)})$ are calculated, which are referred to as the *p-value*.

The deviations of $\zeta^{(i)}$ include, of course, the values of $\zeta^{(i')}$ that are large or even go beyond the t_α criterion. It is easy to see that the *p-value* $\xi^{(i')} = 1 - F_\zeta(\zeta^{(i')})$ for them 'tends' to zero. If the values of $\zeta^{(i'')}$ are small, i.e., the agreement is 'very good', the values $\xi^{(i'')} = 1 - F_\zeta(\zeta^{(i'')})$ 'tend' to unity.

No individual segments of the binary sequence, even the 'bad' ones, are discarded; *p-values* are not excluded.

A histogram is plotted of the frequencies of the events when the values of the *p-value* $\xi^{(i)} = 1 - F_\zeta(\zeta^{(i)})$, $i = \overline{1, N}$ fall in the intervals of the division of the unit segment into 10 equal parts $[0, 0.1)$, $[0.1, 0.2), \ldots, [0.9, 1)$:

$$\nu_1, \nu_2, \ldots, \nu_{10}, \ \sum_{k=1}^{10} \nu_k = N.$$

If the agreement is 'bad', the histogram is skewed to the left; if 'good', to the right.

Now, it is this histogram that is tested for uniformity according to the χ-square goodness test.

Statistics are calculated,

$$\chi = \sum_{k=1}^{10} \frac{(v_k - Np_k)^2}{Np_k}\,, \quad p_k = \frac{1}{10}\,,$$

which have an asymptotic χ-squared distribution with $m = 9$ degrees of freedom and the distribution function

$$F_\chi(z) = \frac{1}{2^{m/2}\Gamma(m/2)} \int_0^z \exp\left(-\frac{x}{2}\right) x^{m/2-1}\,\mathrm{d}x\,,$$

where $\Gamma(y)$ is the gamma function. The value $\alpha = 0.0001$ is set, and z_α is calculated from the equation $1 - F_\chi(z) = \alpha$. The probability that the statistics χ go beyond the boundary z_α is equal to α.

If

$$\chi = \sum_{k=1}^{10} \frac{(v_k - Np_k)^2}{Np_k} > z_\alpha\,, \tag{49}$$

then the cumulative binary sequence of size nN is considered to have failed the criterion based on the statistics of the number of units S_1.

A similar procedure is proposed in the NIST guidelines [24] for a number of statistics presented there, each of which is aimed at 'detecting' a certain type of deviation of the distribution of the binary sequence from hypothesis $\mathcal{H}$; the recommended values are $n = 10^6, N = 10^2$.

As a result of the study, all the goodness-of-fit criteria are listed; those criteria where statistics χ go beyond the z_α boundary, i.e., the criterion is not passed, are noted. It is left to the experimentalist to make a general conclusion regarding the suitability of the random number generator for usage.

Several recommended sets of tests (goodness of fit, statistics) for randomness have been developed [24–26] to date. The NIST test set [24], which is the minimum required, is the basis for examining sequences using other special test suites.

From the point of view of passing the randomness criteria by a binary sequence, a situation cannot be ruled out in which a sequence can be disguised as a random one — a limited set of statistics behaves in the same way as a set of statistics for a truly random sequence.

In particular, it is well known that the output binary sequences of all state-of-the-art means of cryptographic information protection pass any criteria for randomness, but they are not such in the true sense. With a key length of 256 bits, they reflect the randomness (equiprobability) of choosing from a set of 2^{256} binary vectors, but in no way from 2^n vectors, where n is the length of the output sequence.

True randomness can only be guaranteed by a quantum random number generator with proper tuning of its technical parameters.

12. Checking the results of various tests (statistics) for homogeneity. Experimental results

The level of significance α of a criterion (test, statistics) can be understood as the probability with which an ideal generator can generate sequences that will look nonrandom. We set the level α ourselves, usually $\alpha \in [0.001, 0.1]$ [24].

Each test tries to find its own 'evidence of nonrandomness', for example, unequal probability, the presence of correlations, hidden periods, Markov dependence, or an enhanced probability of the occurrence of some binary 'pattern'. It may be said that each test is focused on detecting its own, specific, deviation from a truly random sequence. This deviation is also called a competing hypothesis.

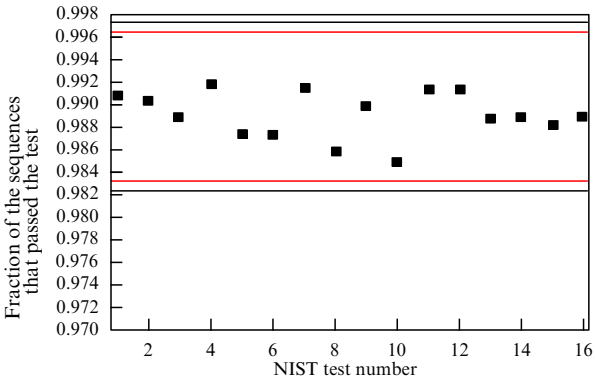

Figure 8. (Color online.) Example of a *p-value* uniformity test for 16 NIST tests — *p-value* histogram for each of the 16 tests. $M = 2000$ sequences were tested, each of length $L = 4 \times 10^6$ bits. The red horizontal lines show the 'three sigma' interval. The plot shows that the test for uniformity of values is passed with a margin.

If the sequence is truly random, and the significance level is the same for all tests, then it should pass different tests in a 'uniform way'. This implies from a more formal point of view that, for a large number of tests, the proportion of sequences that passed each test for a given α should be approximately the same, namely, approximately equal to $1 - \alpha$, and not depend on the test.

Therefore, an additional check for randomness can be a check for 'uniformity' (Fig. 8) of the number of sequences that passed the test for a set of criteria. For this reason, this secondary validation (often referred to in Russian-language publications as secondary labeling) is sometimes referred to as a 'test of tests'. This procedure does not have a strict rationale and is based, in fact, on qualitative considerations.

Let each test be applied to M different binary sequences. The relative proportion of sequences that have passed the test should fit in the 'three sigmas' interval for each test [27]:

$$(1 - \alpha) \pm 3\sqrt{\frac{\alpha(1-\alpha)}{M}}\,.$$

An example of the calculation of the relative proportion of sequences that have passed testing for a set of criteria is displayed in Table. 3. The total length is 8×10^9 bits (more precisely, the value was 8,000,000,464) (see Fig. 1), the number of sequence blocks is $M = 2000$, each with a length of $L = 4 \times 10^6$ bits, and $\alpha = 0.01$. The 'three sigma' interval is $[0.983, 0.997]$.

13. Conclusions

Algorithms for obtaining random binary sequences and methods for testing them were until recently primarily the domain of specialists in cryptography. The physical scientific community also used randomness for its own purposes, for example, to model processes in nuclear and statistical physics. Due to the objective isolation of cryptographers and physicists, each research community developed its own algorithms and formed its own understanding of the essence

Table 3. Proportion of the sequences that passed various tests.

No.	Name of the test	Proportion of sequences $M = 2000$ $L = 4{,}000{,}000$
1	Frequency Test	0.9910
2	Block Frequency	0.9905
3	Cumulative Sums	0.9890
4	Cumulative Sums Reverse	0.9920
5	Runs	0.9875
6	Longest Runs	0.9875
7	Rank	0.9915
8	Fast Fourier Transform (FFT)	0.9860
9	Non-overlapping Template	0.9910
10	Overlapping Template	0.9850
11	Universal	0.9915
12	Approximate Entropy	0.9915
13	Random Excursions	0.9899
14	Random Excursions Variant	0.9893
15	Serial	0.9890
16	Linear Complexity	0.9890

of randomness. The emergence of quantum informatics, in particular, quantum cryptography, has recently initiated the inevitable mutual penetration of these research communities and the methods they develop.

We made an attempt in this publication to show that obtaining quantum randomness and testing its properties do not essentially differ from any physical experiment in which a theoretical law is tested. Any experiment of this kind is reduced to isolating in a 'pure form' those factors that should be checked, while eliminating undesirable external effects that distort the results of the experiment. Any physical experiment is repeated a finite number of times under the same conditions, after which a conclusion is reached about the confirmation or refutation of the law under study. However, there are no logical grounds to believe that a repeated experiment will lead to the same results.

For clarification, it is reasonable to draw a parallel between the 'test of tests' for checking randomness and a physical experiment. Let a large series of experiments be carried out to test a physical law, which confirms its validity. However, any physical experiment contains some error. The agreement of the experiment with the theory is accepted if the results of the experiment fit into an error range. The choice of the permissible error is carried out, in fact, 'manually' and has no mathematical justification. Let each series of experiments be repeated many times. If the result of a series goes beyond the margin of error, should we consider the physical law not to be valid, or should this be attributed to the 'impurity' of a specific series? If the proportion of such unsuccessful series is small according to some criterion, these outliers should be discarded. The criterion for smallness is again selected 'manually'. Deviations from the ideal (theoretical model) should be uniform across the series in the rest of the successful series. A clear analogy with the 'test of tests' used in checking for randomness can be seen here.

This fully applies to experiments to obtain randomness but, possibly, in an even more distinct and concentrated form.

Let a 10^9-bit-long sequence of 0s and 1s be generated. There are in total 2^{10^9} such sequences. It should be kept in mind that the estimated number of atoms in the visible part of the Universe is $2^{256} \approx 10^{77}$. A speculative approach implies that, to check the equiprobability of all 2^{10^9} sequences, it is necessary to generate at least such a number of these sequences and find the frequency with which they occur. This is apparently not possible. By testing only one sequence, we are actually trying to infer the properties of an exponentially large 'set'.

The source (*laser*) *prepares a quantum* (*quasi-single-photon*) *state*[8] *which is subject to measurements. If the projection postulate* (*Eqn* (3)) *is valid and the experiment was carried out 'purely', the primary sequence of photocounts is the Bernoulli-type one. If the sequence is the Bernoulli-type one, a truly random sequence of* 0*s and* 1*s is extracted from it* (*and this is a strictly provable mathematical fact*). *According to this logic, checking the sequences for randomness actually implies checking the Bernoulli nature of the sequence of photocounts, which is a consequence of the projection postulate* (*Eqn* (3)) — *one of the fundamental postulates of quantum mechanics.*

Checking for randomness is in this sense in no way different from the interpretation of any physical experiment. Nevertheless (see the epigraph at the beginning of the notes), not even being able to write down such enormous sequences, we can make judgments about their properties.

Research in this area has been initiated by fairly practical goals. The above example of a quantum generator is on *the conceptual level a simple physical device*, whose principles of operation are based on the fundamental laws of quantum physics and can be understood without special knowledge. In this regard, the authors indulge in the pleasure of citing the following aphorism: ''Everything you need is simple; what is complicated is not needed'' (M T Kalashnikov).

Acknowledgments

We are grateful to our colleagues at the Academy of Cryptography of the Russian Federation for the numerous discussions and support. We also thank our colleagues at the Center for Quantum Technologies at Lomonosov State University (Moscow) K A Balygin, I B Bobrov, V I Zaitsev, V A Kiryukhin, A N Klimov, S P Kulik, and I V Sinil'shchikov for their helpful and active cooperation, and Elena Popova and Sergey Vinogradov for the kindly provided SiPM samples and discussions.

References

1. Bennett C H, Brassard G, in *Proc. of the IEEE Intern. Conf. on Computers, Systems, and Signal Processing, Bangalore, 10–12 December 1984* (Piscataway, NJ: IEEE, 1984) p. 175
2. Koç Ç K (Ed.) *Cryptographic Engineering* (New York: Springer, 2009)
3. Vasilenko V V *Informatsionnye Voiny* (3) 23 (2012)
4. Srinivasan S et al., in *2010 IEEE Symp. on VLSI Circuits, 16–18 June 2010, Honolulu, HI, USA* (Piscataway, NJ: IEEE, 1984) p. 203, https://doi.org/10.1109/VLSIC.2010.5560296
5. Galton F *Natural Inheritance* (London: Macmillan, 1894)
6. Herrero-Collantes M, Garcia-Escartin J C *Rev. Mod. Phys.* **89** 015004 (2017)
7. Einstein A *Ann. Physik* **17** 132 (1905)
8. Klyshko D N *Photons and Nonlinear Optics* (New York: Gordon and Breach, 1988); Translated from Russian: *Fotony i Nelineinaya Optika* (Moscow: Nauka, 1980)

[8] Ideally, to check the projection postulate, it would be desirable to use a strictly single-photon Fock state. However, despite numerous experiments, a strictly single-photon source is currently unavailable. The second-order correlation function must have for a strictly single-photon source a dip exactly to zero [8, 10]. The drop of the correlation function exactly to zero has not been demonstrated in experiments, which means that nonsingle-photon Fock components are present in the radiation. The use of a highly attenuated coherent state is for this reason preferable and more reliable.

9. Klyshko D N, Masalov A V *Phys. Usp.* **38** 1203 (1995); *Usp. Fiz. Nauk* **165** 1249 (1995)
10. Mandel L, Wolf E *Optical Coherence and Quantum Optics* (Cambridge: Cambridge Univ. Press, 1995); Translated into Russian: *Opticheskaya Kogerentnost' i Kvantovaya Optika* (Moscow: Fizmatlit, 2000)
11. Shannon C E *Bell Syst. Tech. J.* **27** 379 (1948)
12. Shannon C E *Bell Syst. Tech. J.* **27** 623 (1948)
13. Shannon C *Raboty po Teorii Informatsii i Kibernetike* (Scientific Works on Information Theory and Cybernetics) (Moscow: IL, 1963)
14. Cover T M, Thomas J A *Elements of Information Theory* (New York: Wiley, 1991)
15. Molotkov S N *JETP Lett.* **105** 395 (2017); *Pis'ma Zh. Eksp. Teor. Fiz.* **105** 374 (2017)
16. Balygin K A et al. *J. Exp. Theor. Phys.* **126** 728 (2018); *Zh. Eksp. Teor. Fiz.* **153** 879 (2018)
17. Balygin K A et al. *Laser Phys. Lett.* **14** 125207 (2017)
18. Balygin K A et al. *JETP Lett.* **106** 470 (2017); *Pis'ma Zh. Eksp. Teor. Fiz.* **106** 451 (2017)
19. Von Neumann J, in *Applied Mathematics Series* Vol. 12 (Washington, DC: U.S. National Bureau of Standards, 1951) p. 36; Reprinted in *Neumann's Collected Works* Vol. 5 (Oxford: Pergamon Press, 1963) p. 768
20. Feller W *An Introduction to Probability Theory and Its Applications* 2nd ed. (New York: Wiley, 1957); Translated into Russian: *Vvedenie v Teoriyu Veroyatnostei i Ee Prilozheniya* Vol. 1 (Moscow: Mir, 1964)
21. Babkin V F *Probl.y Peredachi Informatsii* **7** (4) 13 (1971)
22. Buzhan P et al. *Nucl. Instrum. Meth. Phys. Res. A* **567** 78 (2006)
23. Kalashnikov D A, Tan S-H, Krivitsky L A *Opt. Express* **20** 5044 (2012)
24. Computer Security esource Center, http://csrc.nist.gov/rng/SP800-22b.pdf
25. Knuth D E *The Art of Computer Programming* Vol. 2 (Cambridge: Addison Wesley, 1981)
26. Marsaglia G, http://stat.fsu.edu/pub/diehard
27. Cramer H *Mathematical Methods of Statistics* (Princeton, NJ: Princeton Univ. Press, 1946)