On the distribution of the stationary point of significance level for empirical distribution function

A.A. Kislitsyn
Department of Kinetic Equations
Keldysh Institute of Applied Mathematics of RAS
Moscow, Russian Federation
yuno@kiam.ru

Yu.N. Orlov
Department of Kinetic Equations
Keldysh Institute of Applied Mathematics of RAS
Department of Applied Probability and Informatics
Peoples Friendship University of Russia (RUDN University)
Moscow, Russian Federation
orlmath@keldysh.ru

D.A. Moltchanov
Applied Probability and Informatics Department
Peoples Friendship University of Russia (RUDN University)
Moscow, Russian Federation
Laboratory of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland
molchanov_da@rudn.university

A.K. Samuylov
Applied Probability and Informatics Department
Peoples Friendship University of Russia (RUDN University)
Moscow, Russian Federation
Laboratory of Electronics and Communications Engineering
Tampere University of Technology
Tampere, Finland
samuylov_ak@rudn.university

A.V. Chukarin
Applied Probability and Informatics Department
Peoples Friendship University of Russia (RUDN University)
Moscow, Russian Federation
chukarin_av@rudn.university

Yu.V. Gaidamaka
Applied Probability and Informatics Department
Peoples Friendship University of Russia (RUDN University)
Institute of Informatics Problems
of Federal Research Center "Computer Science and Control"
Moscow, Russian Federation
gaydamaka_yuv@rudn.university

***Abstract*—We consider empirical distribution functions of non-stationary time-series, depending on set length. The local self-consistent significance level is introduced. The class of time-series, for which the distribution function of significance level is stationary, is considered. For example, the signal-to-interference ratio for random walking subscribers in D2D model of wireless connection belongs to this class of random processes. We introduce also the so-called Chernoff equivalence of the self-consistent significance level and derive the formula of averaging levels for various sets.**

***Index Terms*—D2D communications, non-stationary distribution function, significance level, stationary point, Chernoff equivalence**

I. Introduction

One of the distinguishing features of 5th generation networks is the extremely high density of communicating devices — up to 1 million per square km [1]. The examples for such devices are sensors transmitting data on machine-to-machine (M2M) communications technology, users equipment for device-to-device (D2D) communications, etc., while the transceivers can be either fixed or mobile. The network heterogeneity as well as the mutual location of transceiver devices should be taken into account when analyzing the quality of data transmission in a radio channel. Features of mathematical modeling of M2M traffic servicing in the 5th generation networks were investigated in [2]–[5], while in [2] the combination of traffic from M2M-devices with traditional human-to-human (H2H) traffic was analyzed.

To estimate the Quality of Service for wireless communications we register the movement of devices that affect the interference significaly. We consider the signal-to-interference ratio (SIR) as a key channel quality parameter and trace its dynamic as a stochastic process. In this paper we construct a statistical indicator of the disorder in a non-stationary stochastic process. Most statistical criteria either operates directly with the elements of time-series or assumes the stationarity of the corresponding distribution functions (see, for example, [6], [7]). The functional under consideration is presented as a certain quantile of the sample distribution function of distances between sample distribution functions (SDF) of the process in the supremum norm, constructed on disjoint samples of the time-series. Such approach is effectively applied to the large length data samples (around several million data).

The goal of constructing such "disorder indicator" for large data series is to make possible to predict changes in the behavior of the analyzed system, based on assumption that first marker of change can be revealed through an analysis

of the evolution of the SDF [8]. In practice, this can be applied for the time-series of functional values of subscribers positions, such as signal-to-interference ratio (SIR). These data represents array of vectors, where each vector corresponds to a given sender-receiver wireless connection.

The statistical methods, traditionally used to predict disorder effect, based on the analysis of the power spectrum of signal [9], [10], give a very large error, because these methods can be correctly applied only to stationary stochastic processes, whereas the SIR series strongly are nonstationary (will be shown below). In this case, "disorder" must be considered as a change in the level of non-stationarity of the time-series with the assumption that each state (such as random walk in a malls, movement by public transport, etc.) corresponds with its own unique level of non-stationarity of the SIR data.

SIR is defined as a ratio of the useful signal power at a given spatial point to sum of signal powers from other sources. In traditional approach (see e.g. [11], [12]) the signal power is proportional to r^{-2}, where r is a non-zero distance between given transmitter and receiver. So the SIR value is defined as $r^{-2}/\sum_k r_k^{-2}$, where k_r is a distance between given receiver and other transmitter. One can see, that SIR indicator is nonlinear with respect to subscribers positions distribution function, so the theoretical results may be obtained in some relatively simple cases, e.g. for a given spatial subscribers distribution without any moving effects. But in practice we need to describe non-stationary random walk of receivers and transmitters, so that the appropriate method for modeling of statistical characteristics of corresponding ensemble of trajectories should be developed.

If the SIR value becomes less then definite minimum level, corresponding to service under consideration, the wireless connection for associated pair ceases to meet the Quality of Service (QoS) requirements. The possibility of interruption of the wireless connection and duration of the valid intervals of interruption of connection in the provision of a service is determined by the Quality of Experience (QoE) requirements, which are defined by international standards. The effect of interruption can be treated as disorder. More precisely, the variation of interruption probability is disorder of statistical properties of subscribers system.

The kinetic approach to non-stationary SIR modeling for ensemble of subscribers has been developed in the works [13]–[15]. It was supposed, that he distribution function density (DFD) of the differences of subscribers positions is satisfied to the Fokker-Planck equation. This equation is used to generate subscribers trajectories and corresponding SIR values. The non-stationary effects were controlled by special parameters of kinetic equation, so-called drift and diffusion. In practice these parameters are unknown and we must estimate the closeness of empirical SDF for various SIR time-series samples.

The problem under consideration in the present paper is to analyze the dependence of non-stationary level on the sample length. It is rather essential problem, because the disorder should be recognized very quickly, i.e. for the samples of small lengths, but to SDF estimation it is necessary to use sufficiently large lengths. So we need to formulate the optimization method to determine the sample length or to compare the statistical properties of non-stationary samples of different lengths.

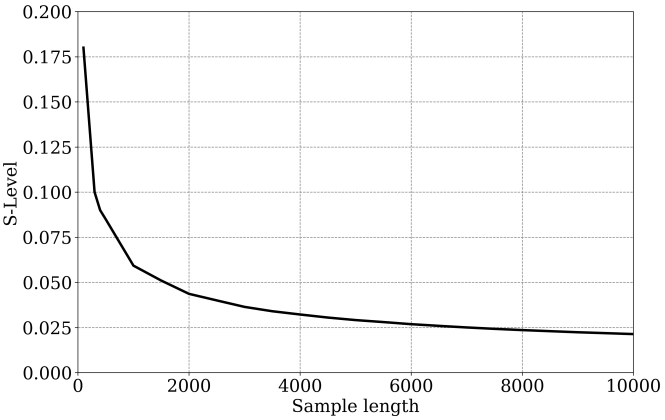

Fig. 1. Self-consistent significance level for stationary distributions

II. Self-consistent significance level

It is well known [16], [17], that the problem of belonging to two SDF for the same general set can be solved with the use of non-parametric statistics of Kolmogorov-Smirnov test

$$S_N = \sup_x |F_{1,N}(x) - F_{2,N}(x)|, \tag{1}$$

for which the following asymptotic representation is valid:

$$\lim_{N\to\infty} P\left\{0 < \sqrt{\frac{N}{2}} S_N < z\right\} = K(z), \tag{2}$$

where $K(z)$ is a tabulated Kolmogorov function (see e.g. [17]) and N is a sample length. Here $F_N(x)$ is a sample distribution function of a stochastic variable ξ, which represents a value x from the sequence of events in a sample window of length N. In formula (2) the significance level Q is approximated by $1 - K(z)$. Let ε be a distance between samples in the supremum norm, defined by (1). Now consider a self-consistent significance level for the sample of length N:

$$Q(\varepsilon) \equiv 1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) = \varepsilon. \tag{3}$$

In the stationary case self-consistent significance level $\varepsilon = \varepsilon_0(N)$ will be a solution of the equation (3) and doesn't depend on a form of F. This solution is unique because of the monotonicity of the function $K(z)$. Tabulated values of $\varepsilon_0(N)$ can be found in [8] (see fig. 1).

For non-stationary SDF, the distribution of distances between samples with defined length is different from the statistics (2). We can construct the empirical distribution function $G_N(\rho)$ for the distances $\rho(N)$ between two independent samples with length N:

$$\rho(N) = \|F_{1,N}(x) - F_{2,N}(x)\|_C. \tag{4}$$

Let us call the numerical solution of the following equation

$$G_N(\rho) = 1 - \rho \tag{5}$$

as self-consistent stationary level (SCSL) $\rho^*(N)$. It will be the probability that distance between samples of length N is more than ρ^*. If it happens, that $\rho^*(N) > \varepsilon^*(N)$, then such series are non-stationary. The value $\rho^*(N)$ is the correct significance level for statistical hypothesis about properties of these time-series samples.

If we assume that the value of the SCSL is a characteristic of this stochastic process, then a change in it can be interpreted as a "disorder". In this case, the practical question arises about the length of the window where such index is considered.

III. Chernoff equivalence for significance level

Suppose that we have self-consistent significance level for the time-series of length L_{tot} as a solution of equation (5) for two nearest independent samples from the main data set. It should be taken into account that $L_{tot} \gg 2N$ and it is possible to fit enough N segments with length N on L_{tot}, so that we can actually collect statistics for building the distribution $G_N(\rho)$. For this length L_{tot} in some cases of $\rho^*(N)$ the distance between two nearest samples from the main data set is greater than $\rho^*(N)$. If now in a certain sliding window of length L the proportion of events where distances between two nearest samples from the main data set of length N turned out to be larger than the value $\rho^*(N)$, then in this window we can register a disorder situation. To achieve such conclusion, it is required that the SCSL of the nonstationary series must be a stochastic variable with a stationary distribution.

Studying the local value of the SCSL – that is, the value obtained over the whole interval L, which is substantially smaller than the original data set L_{tot}, the question can arise about the fluctuations of this local SCSL relative to the base self-consistent stationary level of the whole data set. It is important to understand that the SCSL of the whole data set $\rho^*_{tot}(N)$ is not the average value of the sequence of SCSL $\rho^*_n(N, L)$ (each element of the sequence $\rho^*_n(N, L)$ is constructed for a certain group of two nearest samples of length N from the main data set in windows of length L). Let us consider the function

$$\Psi_N(\rho) = 1 - G_N(\rho). \tag{6}$$

Suppose that SDF $G_N(\rho)$ approximate the differentiable function $G(\rho)$ of the corresponding general data set. Then the function $\Psi_N(\rho)$ has such properties: $\Psi_N(0) = 1$, $\Psi'_N(0) = -a_N \leq 0$. In this case, there exist a limit

$$\lim_{n\to\infty} \left(\Psi_N\left(\frac{\rho}{n}\right)\right)^n = e^{-\rho a_N} \equiv \Phi_N(\rho). \tag{7}$$

Following the definitions given in [18]–[20] for operator functions, the limit function $\Phi_N(\rho)$ will be called Chernoff equivalent function to $\Psi_N(\rho)$ (that is, the significance level of the distribution $G_N(\rho)$). We will write this equivalence as $\Phi_N(\rho) \overset{Ch}{\propto} \Psi_N(\rho)$. Obviously, $|\Phi_N(\rho) - \Psi_N(\rho)| = o(\rho)$.

According to [19], if there is a set (finite or infinite) of functions $\Psi^k_N(\rho)$ in the form of (6), each of them is equivalent in the sense of (7) to the function $\Phi^k_N(\rho)$ with a coefficient $(-a^k_N)$ in the exponent. If we specify a set of corresponding non-negative coefficients p_k so, that $\sum_k p_k = 1$, then the mean function $\bar{\Psi}_N(\rho) = \sum_k p_k \Psi^k_N(\rho)$ is equivalent in the sense of (7) to the function $\Phi_N(\rho) = e^{-\rho \bar{a}_N}$, where

$$\bar{a}_N = \sum_k p_k a^k_N. \tag{8}$$

Consider a sequence of disjoint intervals of length L. For each interval k, we construct an empirical distribution $G^k_N(\rho, L)$ of distances between two nearest samples of length N from the main data set. Let it be n of such intervals, so that $nL = L_{tot}$ is the total length of the sequence. Then the distribution of distances, constructed over the entire data, is the average distribution obtained by averaging over individual samples:

$$\begin{gathered} G_N(\rho, L_{tot}) = \frac{1}{n}\sum_{k=1}^{n} G^k_N(\rho, L), \\ \bar{\Psi}_N(\rho) = \sum_k p_k \Psi^k_N(\rho) = 1 - G_N(\rho, L_{tot}), \quad p_k = \frac{1}{n}. \end{gathered} \tag{9}$$

Using the results of [19], we can prove the following theorem about stationary points of distributions with continuous densities.

Theorem. Let the distributions of random variables have continuous densities and let some non-negative measure be given on the set of these stochastic variables. Then the stationary point of the function will be Chernoff equivalent to the average level of significance of the given distributions, with accuracy up to second order of an infinitesimal, coincides with invers value of the mean value of the reciprocals of the stationary points of the functions Chernoff equivalent to levels of the significance of these distributions:

$$\frac{1}{\tilde{\bar{\rho}}(N)} = \sum_{k=1}^{n} \frac{p_k}{\tilde{\rho}_k(N)} \equiv \frac{1}{\tilde{\bar{\rho}}(N)}. \tag{10}$$

The formula (10) is important, because it allows to reduce considerably the number of calculations while studying the behavior of SCSL on unions of data sets in different tasks of Big Data analysis. Let us briefly describe the results of a numerical analysis for simulating SIR data.

IV. Numerical simulation results

We generate a set of $L_{tot} = 10^6$ data for $n = 10$ subscriber positions in a unity plane with reflecting boundary conditions according to kinetic method, developed in works [13]–[15]. The time step is equal to unit, and distribution function of coordinate differences at initial moment is taken to be symmetrical with respect to zero with small dispersion, so that the average time of boundary destination is equal to 1000 steps. The one half of the data corresponds to certain values of drift $u_1(x, t)$ and diffusion $\lambda_1(t) \geq 0$ in Fokker-Plank equation

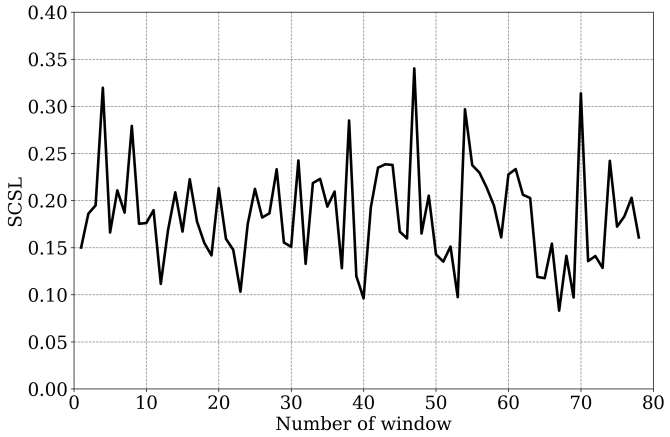

Fig. 2. SCSL time-series fragment of average SIR value for $\rho_k^*(5000, 30000)$

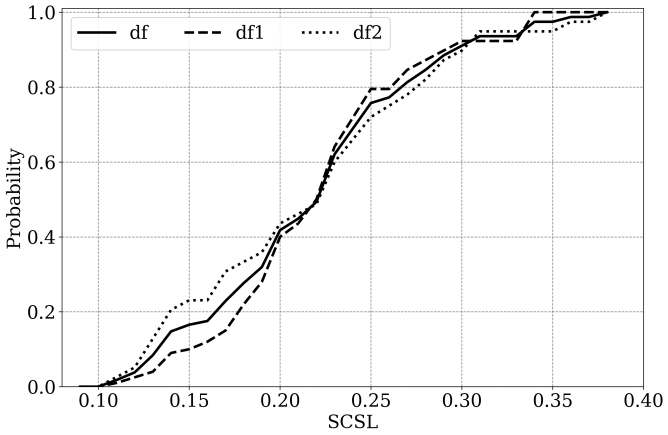

Fig. 3. Empirical distributions SCSL for time-series $\rho_k^*(5000, 30000)$

for the sample distribution function density of coordinates differences x, y, which varies independently:

$$\begin{aligned} &\frac{\partial f(x,t)}{\partial t} + div(u(x,t)f(x,t)) - \frac{\lambda(t)}{2}\frac{\partial^2 f(x,t)}{\partial x^2} = 0, \\ &\frac{\partial f(y,t)}{\partial t} + div(u(y,t)f(y,t)) - \frac{\lambda(t)}{2}\frac{\partial^2 f(y,t)}{\partial y^2} = 0. \end{aligned} \tag{11}$$

The second half of data is generated under drift parameter $u_2(x,t)$ and diffusion coefficient λ_2. It is noticeable that the self-consistent stationary level of coordinate differences data is 4 times higher than the stationary level of significance $\varepsilon_0(N)$ for the first part and 2 times higher for the second. For each pair of generated non-stationary trajectories we calculate SIR value as a function of time step according to formula

$$s_{12}(t) = S\left(\mathbf{r}_1(t), \mathbf{r}_2(t)\right) = \frac{1/\left|r_{12}(t)\right|^2}{\sum_{j=3}^{n} 1/\left|r_{1j}(t)\right|^2}, \tag{12}$$

where $r_{ij}^2(t)$ is a quadrat of distance between two points of separate trajectories at a given time moment t in a plane, subscriber "1" is receiver and subscribers "2" and others are transmitters. One should note, that $s_{12}(t) \neq s_{21}(t)$, so the SIR function is not symmetrical with respect to subscriber numbers. We shall consider average over ensemble SIR value, defined as

$$\bar{s}(t) = \frac{1}{n(n-1)} \sum_{i,j} s_{ij}(t). \tag{13}$$

We take a minimal window length, in which the disorder indicator is calculated, equals to $L = 30000$. The sample length for calculation SDF is assumed to be $N = 5000$. The typical example of SCSL time series for SIR values is presented in fig. 2.

It is sufficient to notice, that for each part of initial data of coordinate differences the SCSL time-series is stationary in accordance with the method of trajectories generation. The empirical distributions of SCSL for the first quarter of data and for the second quarter are closely enough (see fig. 3).

The C-norm differences between df1 (first quarter of data), df2 (second quarter) and df (first half of data) are satisfied

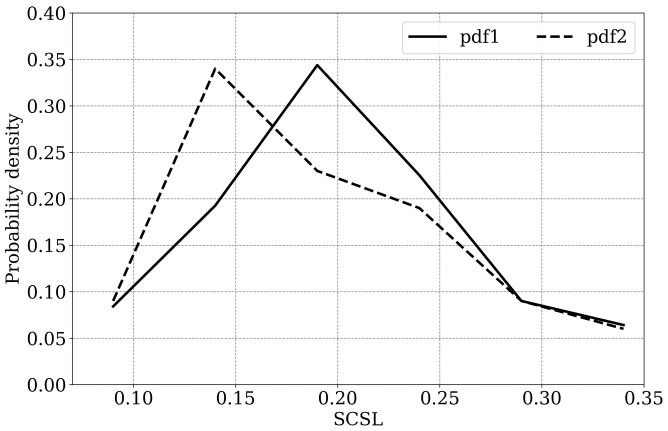

Fig. 4. Stationary SIR probability densities

to Kolmogorov criterion (2-3). For the second half of data we have analogous picture. But the difference between DF for the first and the second parts is sufficiently large. The stationary densities, corresponding to the first and the second parts of data set, are presented in fig. 4.

The stationary point of significance level of SIR distribution $\Psi_{5000}(\rho)$, corresponding to first part of data, is equal to $\rho_{tot}^*(5000) \approx 0,21$. The stationary point for Chernoff equivalent function $\Phi_{5000}(\rho)$ is equal to $\tilde{\rho}_{tot}(5000) \approx 0,20$, and its estimation, obtaining by the averaging formulas (8),(10), is approximately equal to $\tilde{\tilde{\rho}}_{tot}(5000) \approx 0,19$. For the second part of data we have analogously $\rho_{tot}^*(5000) \approx 0,15$, $\tilde{\rho}_{tot}(5000) \approx 0,14$ and also $\tilde{\tilde{\rho}}_{tot}(5000) \approx 0,14$. So we numerically demonstrated, that Chenoff equivalence leads to approximately the same results, as the exact formula (6), applying to various subsets.

V. Conclusions

In this paper we proved the theorem, concerning statistical application of the special quantum theoretical notation, which is known as Chernoff equivalence. It allows us to significantly reduce the volume of calculation for analysis of very complex

statistical object: variation between sample distribution function of the distances between sample distribution functions of non-linear functional of stochastic trajectories.

In particularly, it appears, that two sets of data with various non-stationary levels can be statistically separated in sufficiently narrow sliding window. The disorder indicator in the form of Chernoff equivalent self-consistency level $\tilde{\tilde{\rho}}$ has a stationary distribution function for each subset with the same non-stationary level, as a main sample of data. In other words, the type of non-stationary behavior of the data does not change during the period of time when the physical system is in the same (may be, non-stationary) condition.

This disorder indicator can be used for the practical solution of the problem of optimal non-stationary stochastic control.

ACKNOWLEDGEMENT

The publication has been prepared with the support of the "RUDN University Program 5–100" and funded by RFBR according to the research projects No. 17-07-00845, 18-37-00380. This work has been developed within the framework of the COST Action CA15104, Inclusive Radio Communication Networks for 5G and beyond (IRACON).

REFERENCES

[1] *ITU-R Recommendation M.2083 (09/2015): IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[2] I. Gudkova, K. Samouylov, I. Buturlin, V. Borodakiy, M. Gerasimenko, O. Galinina, and S. Andreev, "Analyzing impacts of coexistence between M2M and H2H Communication on 3GPP LTE System," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8458, pp. 162–174, 2014.

[3] K. Samouylov, V. Naumov, E. Sopin, I. Gudkova, and S. Shorgin, "Sojourn time analysis for processor sharing loss system with unreliable server," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9845, pp. 284–297, 2016.

[4] V. Borodakiy, K. Samouylov, I. Gudkova, and E. Markova, "Analyzing Mean Bit Rate of Multicast Video Conference in LTE Network with Adaptive Radio Admission Control Scheme," *Journal of Mathematical Sciences*, vol. 218, no. 3, pp. 257–268, 2016.

[5] V. Naumov and K. Samouylov, "Analysis of multi-resource loss system with state-dependent arrival and service rates," *Probability in the Engineering and Informational Sciences*, vol. 31, no. 4, pp. 413–419, 2017.

[6] V. S. Koroluk, N. I. Portenko, A. V. Skorohod, and A. F. Turbin, *A handbook on probability theory and mathematical statistics (in Russian)*. Moscow: Nauka, 1985.

[7] A. I. Kobzar, *Applied Mathematical Statistics (in Russian)*. Moscow: Fizmatlit, 2006.

[8] Y. N. Orlov, *Kinetic methods for non-stationary time series analysis (in Russian)*. Moscow: MIPT, 2014.

[9] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications*. Cambridge, Mass.: Technology Press and John Wiley & Sons and Chapman & Hall, 1949.

[10] R. N. Bracewell, *The Fourier Transform and Its Applications (3rd ed.)*. Boston: McGraw-Hill, 2000.

[11] V. Petrov, D. Moltchanov, P. Kustarev, J. M. Jornet, and Y. Koucheryavy, "On the use of integral geometry for interference modeling and analysis in wireless networks," *IEEE Communications Letters*, vol. 20, pp. 2530–2533, 2016.

[12] A. Samuylov, A. Ometov, D. Moltchanov, S. Andreev, V. Begishev, R. Kovalchukov, Y. Gaidamaka, K. Samouylov, and Y. Koucheryavy, "Analytical performance estimation of network-assisted D2D communications in urban scenarios with rectangular cells," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 2, 2017.

[13] Y. Gaidamaka, Y. Orlov, D. Moltchanov, and A. Samuylov, "Modeling the signal-interference ratio in a mobile network with moving devices," *Informatika i ee Primeneniya*, vol. 11, no. 2, pp. 50–58, 2017.

[14] Y. Orlov, S. Fedorov, A. Samoulov, Y. Gaidamaka, and D. Molchanov, "Simulation of Devices Mobility to Estimate Wireless Channel Quality Metrics in 5G Network," in *AIP Conference Proceedings*, vol. 1863, 090005, 2017.

[15] Y. Orlov, S. Fedorov, A. Samoulov, and et al., "SIR Distribution in D2D Environment with Non-stationary Mobility of Users," in *31st European Conference on Modelling and Simulation, ECMS*, 2017, pp. 720–725.

[16] A. N. Kolmogoroff, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell' Istituto Italiano degly Attuari*, vol. 4, no. 1, pp. 83–91, 1933.

[17] B. V. Gnedenko, *Course of the theory of probability*. Moscow: Fizmatlit, 1961.

[18] O. G. Smolyanov, H. Weizsacker, and O. Wittin, "Chernoff's theorem and discrete time approximation of Brownian motion on manifolds," *Potential Anal.*, vol. 26, pp. 1–29, 2007.

[19] Y. N. Orlov, V. J. Sakbaev, and O. G. Smolyanov, "Feynman formulas as a method of averaging random Hamiltonians," *Trudy Matematicheskogo Instituta imeni V. A. Steklova*, vol. 285, pp. 232–243, 2014.

[20] Y. A. Butko and O. G. Smolyanov, "Feynman's formulas in stochastic and quantum dynamics," *Modern problems of mathematics and mechanics*, vol. 6, no. 1, pp. 61–75, 2011.