DISCRETE AND CONTINUOUS MODELS AND APPLIED COMPUTATIONAL SCIENCE

Volume 30 Number 2 (2022)

Founded in 1993
Founder: PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA

DOI: 10.22363/2658-4670-2022-30-2

Edition registered by the Federal Service for Supervision of Communications, Information Technology and Mass Media
Registration Certificate: ПИ № ФС 77-76317, 19.07.2019

ISSN 2658-7149 (online); 2658-4670 (print)
4 issues per year.
Language: English.

Publisher: Peoples' Friendship University of Russia (RUDN University).
Indexed by Ulrich's Periodicals Directory (http://www.ulrichsweb.com), Directory of Open Access Journals (DOAJ) (https://doaj.org/), Russian Index of Science Citation (https://elibrary.ru), EBSCOhost (https://www.ebsco.com), CyberLeninka (https://cyberleninka.ru).

Aim and Scope

Discrete and Continuous Models and Applied Computational Science arose in 2019 as a continuation of RUDN Journal of Mathematics, Information Sciences and Physics. RUDN Journal of Mathematics, Information Sciences and Physics arose in 2006 as a merger and continuation of the series "Physics", "Mathematics", "Applied Mathematics and Computer Science", "Applied Mathematics and Computer Mathematics".

Discussed issues affecting modern problems of physics, mathematics, queuing theory, the Teletraffic theory, computer science, software and databases development.

It's an international journal regarding both the editorial board and contributing authors as well as research and topics of publications. Its authors are leading researchers possessing PhD and PhDr degrees, and PhD and MA students from Russia and abroad. Articles are indexed in the Russian and foreign databases. Each paper is reviewed by at least two reviewers, the composition of which includes PhDs, are well known in their circles. Author's part of the magazine includes both young scientists, graduate students and talented students, who publish their works, and famous giants of world science.

The Journal is published in accordance with the policies of COPE (Committee on Publication Ethics). The editors are open to thematic issue initiatives with guest editors. Further information regarding notes for contributors, subscription, and back volumes is available at http://journals.rudn.ru/miph.

E-mail: miphj@rudn.ru, dcm@sci.pfu.edu.ru.

EDITORIAL BOARD

Computer Design: *A. V. Korolkova*, *D. S. Kulyabov*
English text editors: *Nikolay. E. Nikolaev*, *Ivan. S. Zaryadov*, *Konstantin P. Lovetskiy*

Address of editorial board:
Ordzhonikidze St., 3, Moscow, Russia, 115419
Tel. +7 (495) 955-07-16, e-mail: publishing@rudn.ru
Editorial office:
Tel. +7 (495) 952-02-50, miphj@rudn.ru, dcm@sci.pfu.edu.ru
site: http://journals.rudn.ru/miph

Paper size 70×100/16. Offset paper. Offset printing. Typeface "Computer Modern".
Conventional printed sheet 6,61. Printing run 500 copies. Open price. The order 419.
PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA
6 Miklukho-Maklaya St., 117198 Moscow, Russia
Printed at RUDN Publishing House:
3 Ordzhonikidze St., 115419 Moscow, Russia,
Ph. +7 (495) 952-04-41; e-mail: publishing@rudn.ru

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)
2022, 30 (2) 104
http://journals.rudn.ru/miph

Contents

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 105–114
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-105-114

Numerical solution of Cauchy problems with multiple poles of integer order

Aleksandr A. Belov[1, 2], **Nikolay N. Kalitkin**[3]

[1] *Lomonosov Moscow State University,*
1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation
[2] *Peoples' Friendship University of Russia (RUDN University),*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[3] *Keldysh Institute of Applied Mathematics RAS,*
4 A, Miusskaia Sq., Moscow, 125047, Russian Federation

(received: March 12, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. We consider Cauchy problem for ordinary differential equation with solution possessing a sequence of multiple poles. We propose the generalized reciprocal function method. It reduces calculation of a multiple pole to retrieval of a simple zero of accordingly chosen function. Advantages of this approach are illustrated by numerical examples. We propose two representative test problems which constitute interest for verification of other numerical methods for problems with poles.

Key words and phrases: Cauchy problem, singularities, continuation through a pole, multiple poles

1. Introduction

There are a number of important applied problems in which the solution has multiple singularities. In such problems, it is required to find a chain of sequentially located singularities. Similar problems are often found in the theory of special functions (elliptic functions, gamma function, etc.).

Numerical methods are widely used to compile tables of special functions [1] and for standard direct calculation programs [2]. Standard schemes (for example, Runge–Kutta schemes) allow one to calculate smooth sections of the solution with good accuracy. However, near the singularity, the error of such schemes increases catastrophically. Direct continuation of the solution beyond the pole, as a rule, is impossible. Therefore, the solution is continued beyond the pole with some artificial techniques. Continuation through a number of poles is an even greater problem and requires the development of special procedures.

The literature describes methods based on the Pade approximation [3]–[5] and on the approximation of the solution by chain fractions [6]. Abramov and

Yukhno proposed a special replacement for an unknown function that translates the solution into a non-singular one, see [7] and the bibliography there. However, these methods are applicable only for calculating the transcendental Painleves, for which there is a lot of a priori information. In addition, the coefficients of the Pade approximation are calculated from the coefficients of the Taylor series, and to find the latter, you need to solve the original problem with some difference scheme. The problems that arise along this path are described above.

In [8], we have constructed the reciprocal function method which for the first time allowed to perform highly accurate calculations through a sequence of first-order poles. However, for poles of order $k > 1$, accuracy sharply deteriorated. The reason was as follows: the reciprocal function had a zero of order $k > 1$. Calculation of such zero is an ill-conditioned problem conjuncted with considerable loss of accuracy.

In the present work, we propose the generalized reciprocal function method which overcomes the mentioned difficulty. It provides high accuracy in computation of a sequence of poles with multiplicity $k > 1$ if the differential equation is autonomous.

2. Generalized reciprocal function

2.1. Method

Let us write down the Cauchy problem for an ordinary differential equation of the first order

$$du/dt = f(u,t), \quad u(0) = u^0. \tag{1}$$

Its solution is assumed to have a sequence of poles at points t^*_m of integer orders k_m. The orders of the different poles may not be the same. At the same time, we assume that the solution does not have special points of other types.

Let us introduce some fine enough mesh t_n. Let us choose some one-step method of numerical integration. A large number of such methods is given in the monographs [9], [10]. One can detect approach to the nearest pole by rapid increase of the numerical solution u_n. However, this does not allow us to determine the position of the pole with sufficient accuracy, calculate the solution in its vicinity, and continue the solution beyond the pole.

To overcome this difficulty in the case of first-order poles, we proposed the reciprocal function method [8]. Let adjusting parameter $U > 0$ be introduced. If the condition $|u_n| > U$ is met, then the calculation proceeds from the function $u(t)$ to the reciprocal function $v(t) = [u(t)]^{-1}$. It satisfies the following equation:

$$dv/dt = -v^2 f(v^{-1}, t). \tag{2}$$

The initial condition at the transition point is assumed to be $v_n = (u_n)^{-1}$. Note that such a transition at any mesh node is possible only when using one-step schemes (for example, explicit Runge–Kutta methods).

The pole of the original function $u(t)$ of multiplicity k corresponds to the zero of the reciprocal function $v(t)$ of the same multiplicity. For $k = 1$, this is a simple zero, in which the solution of the equation (2) does not present

any problem. This is illustrated by examples of numerical calculations in [8]. In this case, the solution is calculated with good accuracy in the vicinity of the pole and continues beyond it. This makes it possible to perform through calculations of the sequence of poles of the first order with good accuracy.

However, for multiplicity $k > 1$, the zero $v(t)$ turns out to be a special point of the equation (2). At this point, the reciprocal function itself and all its derivatives up to the $(k-1)$th inclusive turn to zero. Numerical solution through this feature leads to a strong decrease in accuracy and even failure of the calculation. The solution cannot be confidently continued even beyond the first pole.

To overcome this difficulty, we propose to introduce a generalized reciprocal function $w(t)$. Suppose the multiplicity of the nearest pole k is known. Then for any k, we can put

$$w(t) = [v(t)]^{1/k}. \tag{3}$$

This expression has k complex branches. We choose the only real one from them. The generalized reciprocal function satisfies the following differential equation:

$$dw/dt = -k^{-1} w^{1+k} f(w^{-k}, t). \tag{4}$$

For it, this zero turns out to be simple, and its calculation does not cause fundamental difficulties. After passing this zero, one can return to calculation of the $u(t)$ function.

2.2. Multiplicity determination

Sometimes, from a theoretical study of the Cauchy problem, it is possible to determine a priori the multiplicities of the poles k_m. In general case, one has to find k_m a posteriori in the course of calculation. To do this, we propose the following procedure.

Near the pole, the following relation holds: $v(t) \approx A(t^* - t)^k$. Then in two adjacent nodes, $v_n \approx A(t^* - t_n)^k$, $v_{n+1} \approx A(t^* - t_{n+1})^k$, $f_n =$, f_{n+1}. This is an over-determined system in unknowns A, t^*, k. Excluding A and t_*, one obtains

$$k \approx \left[1 - \frac{\ln(f_n f_{n+1}^{-1})}{\ln(v_n v_{n+1}^{-1})}\right]^{-1}. \tag{5}$$

If the resulting k is close enough to some integer on several sequential mesh steps, this integer number can be taken as the pole multiplicity.

Note that in order to apply the formula (5), the following conditions are necessary (although not sufficient):

$$v_n v_{n+1} > 0, \quad f_n f_{n+1} > 0, \quad v_n f_n < 0, \quad |v_n| > |v_{n+1}|. \tag{6}$$

2.3. Test problem

Let us construct Cauchy problem with the following exact solution:

$$u(t) = \sin t \cos^{-k} t. \tag{7}$$

This function has poles at $t^*_m = 0.5\pi + \pi m$. Its derivative equals

$$du/dt = \cos^{1-k} t + k \sin^2 t \cos^{-k-1} t. \tag{8}$$

In the intervals between neighboring poles, the derivative preserves the sign. For odd k, the derivative is always positive, and for even k, its signs are opposite in neighboring intervals separated by a pole. Therefore, in both cases, the solution (7) and has no special points other than poles.

The equation (8) is of no interest to be considered as Cauchy problem, since the solution is reduced to quadrature calculation. However, in the case of an odd $k \geqslant 1$, the solution (7) and equation (8) can be converted to the form

$$u(t) = \tan t (1 + \tan^2 t)^{(k-1)/2}, \tag{9}$$

$$du/dt = (1 + k \tan^2 t)(1 + \tan^2 t)^{(k-1)/2}. \tag{10}$$

Let us consider (9) as equation in $\tan t$ and express $\tan t$ in terms of u. Next, we substitute the obtained expression into (10) and obtain autonomous form of the equation.

Practically, explicit relations expressed in elementary functions can be derived only in two cases. The first one corresponding to $k = 1$ is trivial

$$u(t) = \tan t, \quad du/dt = 1 + u^2. \tag{11}$$

This example was used in [8] as a test for simple pole.

The second case with $k = 3$ is non-trivial

$$\begin{gathered} u(t) = \tan t + \tan^3 t, \\ du/dt = (1 + \xi(u)^2)(1 + 3\xi(u)^2), \\ \xi(u) = -2 \cdot 3^{-0.5} \operatorname{sign}(u) \sinh \varphi(u), \quad \varphi(u) = 3^{-1} \operatorname{arsinh}(0.5 \cdot 3^{1.5}|u|). \end{gathered} \tag{12}$$

This test is used in the present work.

2.4. Numerical example

Calculation of the test (12) was performed on the segment $0 \leqslant t \leqslant 15$ containing 5 poles of the third order. The calculation was performed on a sequence of uniform meshes using an explicit Runge–Kutta scheme of the fourth order of accuracy (ERK4). The first grid had a step of $\tau = 0.15$, the remaining grids were obtained by successive decreasing of all steps by the factor of 2 from mesh to mesh. Figure 1 compares the numerical solution on the first grid (markers) with the exact one (solid line). The vertical lines show the asymptotes of the exact solution. Even with such a large step, one can see good agreement between the numerical solution and the exact one.

Figure 2 shows the solution error in mean-squared analogue of the Hausdorff metrics [11] as function of the mesh step. The plot is given in double logarithmic scale. The calculated points lie on a straight line with a slope of -4. This corresponds to the power-law nature of convergence with the theoretical order of accuracy $p = 4$. One can see that the error reaches round-off errors $\sim 10^{-14}$ (which is only 100 times greater than the error of

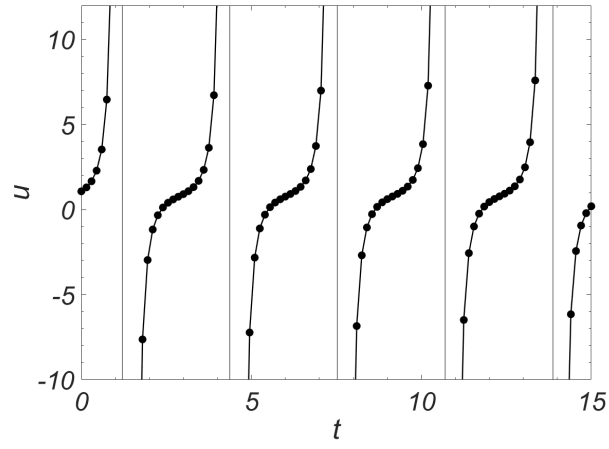


Figure 1. Calculation of the test (12) with step $\tau = 0.15$ using the ERK4 scheme. Comparison of the numerical solution on the first grid (markers) with the exact one (solid line)

a single rounding equal 10^{-16}) at $N \approx 10^5$ of grid nodes. This indicates high accuracy and reliability of the method. The position of the poles is determined by interpolation of w_n at two points to the right and left of zero. This procedure is described in [8]. The error of the fifth pole position is shown in figure 2 with triangles.

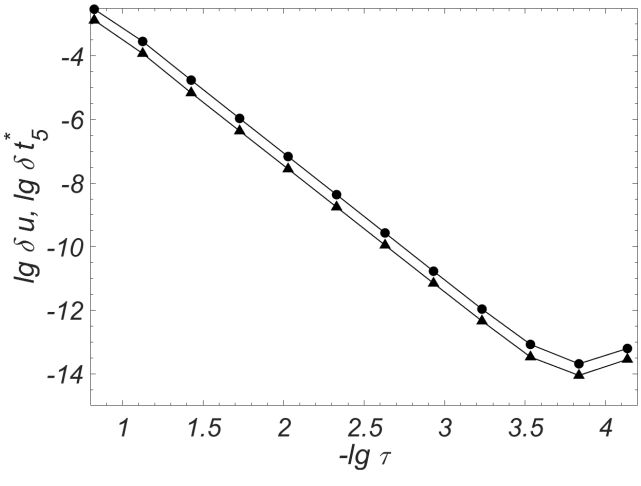


Figure 2. Dependence of the error of the solution and the fifth pole position (triangles) on the step size for the test (12)

3. Non-autonomous problems

3.1. Difficulties

For illustration, we used a test in which the differential equation was autonomous. However, in applied problems we also have to deal with non-autonomous equations. For such problems, the reliability of numerical methods deteriorates. In the vicinity of the pole, the calculated profile v_n may look like an alternating "saw", which does not allow to determine the position of zero. Let's explain the reason of this phenomenon.

Take, for example, the non-autonomous equation (8). The zero of the grid solution v_n, understood in the sense of different signs of this value in neighboring nodes, does not coincide with the exact pole. At the same time, the sign of the right side (8) that depends only on t is determined by the position of the exact pole. The value v_n changes sign when passing through the "mesh" pole, and the right part does so when passing through the exact pole. This lack of synchronization can lead to an unpredictable sign of increment of the value v_n at the next step. The higher is the pole multiplicity the stronger is this effect.

These effects usually reveal on insufficiently fine meshes. To overcome these difficulties, we recommend to choose fine enough mesh. Increasing digit capacity is also a helpful strategy.

3.2. Even multiplicity

For a pole of even multiplicity, the Cauchy problem can be non-autonomous only. In fact, near the pole $u \approx A(t-t^*)^{-k}$, and $du/dt \approx -kA(t^*-t)^{-k-1}$. For even k, du/dt has different signs on different sides of the pole. Therefore, it cannot be an unambiguous function of $f(u)$. Thus, any problems for an even k face all the difficulties that are typical for non-autonomous problems. The ways to overcome them are also indicated above.

3.3. Example

Consider the following non-autonomous problem

$$du/dt = \left(0.5 + \sqrt{0.25 + u^2} + 2u^2\right)\cos t, \quad u(0) = 0. \tag{13}$$

The exact solution is as follows:

$$u(t) = \sin t \cos^{-2} t. \tag{14}$$

It has poles of the order $k = 2$ at $t^*_m = \pi/2 + \pi m$.

Calculations were performed using the ERK4 scheme. Figure 3 shows the numerical solution for $\tau =$ and the exact solution (the notation corresponds to figure 1). One can see that the numerical calculation through 5 poles is successful, although the visual difference at the end of the calculation is somewhat greater than for the autonomous problem in figure 1.

Figure 4 shows the dependence of the error of the solution itself and the one of the fifth pole position on the mesh step. The notations correspond to

figure 2. One can see that the calculated points are slightly scattered around the average line. This is a manifestation of the difficulties associated with solving non-autonomous problems. However, the average slope of the straight line corresponds to the theoretical order of accuracy $p = 4$, and very high accuracy is achieved on moderate meshes, close to unit rounding errors.

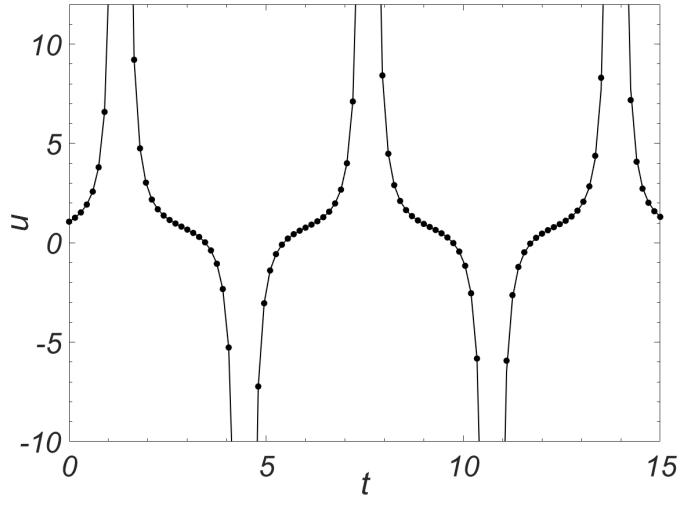


Figure 3. Calculation of the test (13) with step $\tau = 0.15$ using the ERK4 scheme. The notations correspond to figure 1

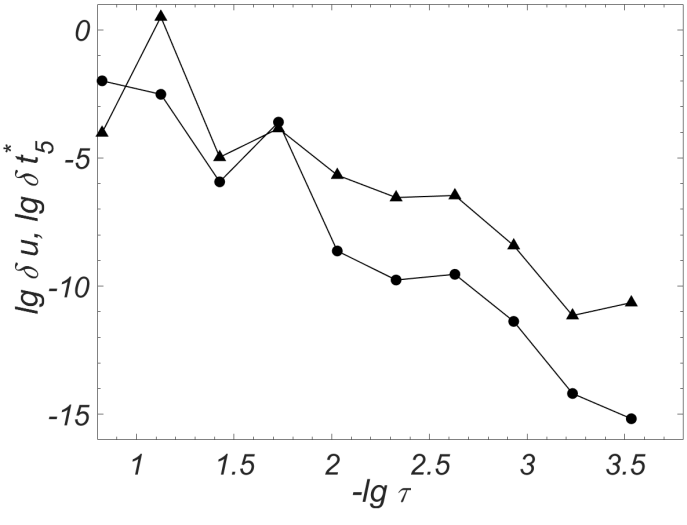


Figure 4. Dependence of the error of the solution and the fifth pole position on the step size for the test (13). The notations correspond to figure 2

3.4. Note

The same exact solution may correspond to different non-autonomous problem formulations. For example, the function (14) is the exact solution of a differential equation

$$du/dt = \cos^{-1} t + 2\sin^2 t \cos^{-3} t. \tag{15}$$

However, all attempts to calculate this equation using various quadrature formulas were unsuccessful due to "blow up" of the calculation.

Therefore, the tests (12) and (13) constructed here are of value themselves. Solutions have sequences of poles of the specified orders, special points of other types are absent, and the influence of non-autonomy is minimized. These problems are recommended for validation of other methods of calculation through poles.

Acknowledgments

This work was supported by grant MK-3630.2021.1.1.

References

[1] L. F. Janke E. Emde F., *Taffeln horere Functionen*. B.G. Teubbner Verlagsgesellschaft, Stuttgart, 1960.

[2] *NIST digital library of mathematical functions*, https://dlmf.nist.gov.

[3] C. F. Corliss, "Integrating ODE's in the complex plane – pole vaulting", *Mathematics of Computation*, vol. 35, pp. 1181–1189, 1980. DOI: 10.1090/S0025-5718-1980-0583495-8.

[4] B. Fornberg and J. A. C. Weideman, "A numerical methodology for the Painlevé equations", *Journal of Computational Physics*, vol. 230, pp. 5957–5973, 2011. DOI: 10.1016/j.jcp.2011.04.007.

[5] M. Fasondini, B. Fornberg, and J. A. C. Weideman, "Methods for the computation of the multivalued Painlevé transcendents on their Riemann surfaces", *Journal of Computational Physics*, vol. 344, pp. 36–50, 2017. DOI: 10.1016/j.jcp.2017.04.071.

[6] I. M. Willers, "A new integration algorithm for ordinary differential equations based on continued fraction approximations", *Communications of the ACM*, vol. 17, pp. 504–508, 1974. DOI: 10.1145/361147.361150.

[7] A. A. Abramov and L. F. Yukhno, "A method for calculating the Painleve transcendents", *Applied Numerical Mathematics*, vol. 93, pp. 262–267, 2015. DOI: 10.1016/j.apnum.2014.05.002.

[8] A. A. Belov and N. N. Kalitkin, "Reciprocal function method for Cauchy problems with first-order poles", *Doklady Mathematics*, vol. 101, no. 2, pp. 165–168, 2020. DOI: 10.1134/S1064562420020040.

[9] E. Hairer, G. Wanner, and S. P. Nørsett, "Solving ordinary differential equations: I. Nonstiff problems", in *Springer Series in Computational Mathematics*. Berlin: Springer, 1993, vol. 8. DOI: 10.1007/978-3-540-78862-1.

[10] E. Hairer and G. Wanner, "Solving ordinary differential equations: II. Stiff and differential-algebraic problems", in *Springer Series in Computational Mathematics*. Berlin: Springer, 1996, vol. 14. DOI: 10.1007/978-3-642-05221-7.

[11] A. A. Belov and N. N. Kalitkin, "Efficient numerical integration methods for the Cauchy problem for stiff systems of ordinary differential equations", *Differential equations*, vol. 55, no. 7, pp. 871–883, 2019. DOI: 10.1134/S0012266119070012.

For citation:

A. A. Belov, N. N. Kalitkin, Numerical solution of Cauchy problems with multiple poles of integer order, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 105–114. DOI: 10.22363/2658-4670-2022-30-2-105-114.

Information about the authors:

Belov, Aleksandr A. — Candidate of Physical and Mathematical Sciences, Associate Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); Researcher of Faculty of Physics, Lomonosov Moscow State University (e-mail: aa.belov@physics.msu.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-0918-9263, ResearcherID: Q-5064-2016, Scopus Author ID: 57191950560)

Kalitkin, Nikolay N. — Doctor of Physical and Mathematical Sciences, Professor, Corresponding member of the RAS, head of department, Keldysh Institute of Applied Mathematics RAS (e-mail: kalitkin@imamod.ru, phone: +7(499)2509726, ORCID: https://orcid.org/0000-0002-0861-1792)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-105-114

Численное решение задач Коши со множественными полюсами целого порядка

А. А. Белов[1,2], **Н. Н. Калиткин**[3]

[1] *Московский государственный университет им. М. В. Ломоносова, Ленинские горы, д. 1, стр. 2, Москва, 119991, Россия*
[2] *Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия*
[3] *Институт прикладной математики им. М. В. Келдыша РАН, Миусская пл., д. 4А, Москва, 125047, Россия*

Аннотация. Рассмотрена задачи Коши для обыкновенного дифференциального уравнения с решением, обладающим последовательностью кратных полюсов целого порядка. Предложен обобщённый метод обратной функции, который сводит вычисление кратного полюса к расчёту простого нуля соответственно выбранной функции. Преимущества такого подхода проиллюстрированы на численных примерах. Предложены сложные тестовые задачи, которые представляют интерес для проверки других численных методов для задач с полюсами.

Ключевые слова: задача Коши, сингулярности, продолжение за полюс, кратные полюсы

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 115–126
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-115-126

Optimization of an isotropic metasurface on a substrate

Zhanna O. Dombrovskaya

Lomonosov Moscow State University,
1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation

(received: March 12, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. Mathematical statement of one-wavelength antireflective coating based on two-dimensional metamaterial is formulated for the first time. The constraints on geometric parameters of the structure are found. We propose a penalty function, which ensures the applicability of physical model and provides the uniqueness of the desired minimum. As an example, we consider the optimization of metasurface composed of PbTe spheres located on germanium substrate. It is shown that the accuracy of the minimization with properly chosen penalty term is the same as for the objective function without it.

Key words and phrases: antireflective coating optimization, penalty function method, constrains on geometric parameters, all-dielectric metasurface on a substrate

1. Introduction

Last few years the designs of nanostructured coatings with the reflection coefficient close to zero attract a great attention. Such coatings are promising for solar cells and other photovoltaic elements which work both in the visible and in the infrared ranges. Nowadays, high refractive index all-dielectric meta-atoms are used [1], [2] instead of plasmonics [3], [4] in order to reduce Joule losses.

Commonly, the properties of substrated metasurfaces are calculated numerically. The computations are complicated due to big divergence of characteristic scales: resonator size can be 3–20 times smaller then the wavelength λ. Consequently, it is necessary to choose nonuniform grids with extra fine steps to describe all areas accurately. It makes computations ineffective for optimization problems. To increase the productivity, we propose to use analytical formulas from a combination of physical models [4]–[6]. However, each model has its applicability limitations. Moreover, there are restrictions on structure geometric parameters caused by fabrication limitations. They should be taken into account to obtain reasonable solutions. As a result, optimization parameters vary only in some ranges. The optimization problem should be stated as a nonlinear inverse problem of conditional minimization.

Due to resonant response of the particle array, there are numerous peaks and dips in the metasurface spectrum. Therefore, the result of objective function minimization strongly depends on the initial approximation. By performing calculations with several raffled off initial guesses, it is impossible to guarantee that the deepest of minima we found is global rather than a local one [7]. We cannot be sure that another deeper minimum does not exist. In this paper, basing on the idea of the penalty function method, we propose a well-posed statement of the inverse problem of one-wavelength antireflective coating based on isotropic two-dimensional metamaterial. The formulation allows to find global extremum, the location of which is approximately known from physical considerations. To solve the problem, we use the interior point method [8]. Its stability and accuracy are discussed.

2. Problem of one-wavelength antireflective coating

Depending on the specific formulation of the problem, it is required to minimize or maximize reflectance, transmittance, absorptance or their combination. Commonly, a list of materials used for fabrication of particles and a substrate is known in advance. The parameters to be determined are period p of the structure and radius r of the meta-atoms.

According to the Sveshnikov–Ilinskiy approach [9], the solution of the optimization problem is reduced to multiple solutions of the direct problem (in our case, calculations of electrodynamic characteristics of the substrated metasurface) with directionally modified optimization parameters. To simplify calculations, it is preferable to model the structure under study by combining numerical algorithm (for the objective function minimization) with simple analytical formulas (to solve the direct problem). Similar joint approach is often used for designing multilayer coatings with the given properties [10], [11].

2.1. Physical statement of the problem

To start with, consider square periodic array composed of spherical dielectric scatterers with refractive index n and radius r. The one-layer structure is located at "air-dielectric" interface with refractive index n_s of the dielectric substrate. Such isotropic metasurface (MS) with periodicity p is normally illuminated by an external plane electromagnetic wave (figure 1).

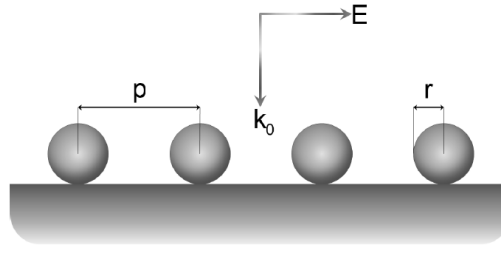


Figure 1. Schematic representation of a metasurface consisting of spherical particles on a semi-infinite substrate. The structure is normally irradiated by an electromagnetic wave

To describe electrodynamic properties of MS in air, A. B. Evlyukhin proposed the model of interacting induced dipoles [6]. According to the model, each sphere is replaced by a pair of electric and magnetic dipoles. To account for the interaction with other particles, Green's tensor of the medium is constructed. Such approach seems to be general since there is no homogenization of the structure [12]. In the case of normal incidence, the reflection R and transmission T Fresnel coefficients are

$$R = \frac{ik_0}{2p^2}\left(\alpha_{\mathrm{e}}^{\mathrm{eff}} - \alpha_{\mathrm{m}}^{\mathrm{eff}}\right), \quad T = 1 + \frac{ik_0}{2p^2}\left(\alpha_{\mathrm{e}}^{\mathrm{eff}} + \alpha_{\mathrm{m}}^{\mathrm{eff}}\right), \tag{1}$$

where $k_0 = 2\pi/\lambda$ is the free-space wave number, $\alpha_{\mathrm{e}}^{\mathrm{eff}}$ and $\alpha_{\mathrm{m}}^{\mathrm{eff}}$ are effective electric and magnetic polarizabilities that take into account interaction between the meta-atoms in the lattice. Here and after temporal dependence is assumed to be $e^{-i\omega t}$.

The presence of dielectric substrate influences on the field amplitude at electric (EDR) and magnetic (MDR) dipole resonances [13]. It was shown that for all-dielectric MSs, even if the refractive indexes n and n_s are high, the interaction between spherical particle and the substrate is weak enough [2]. For this reason, the MS located on the interface is modeled as imaginary sheet described with surface susceptibility electric χ_{e} and magnetic χ_{m} densities depending on R and T from (1). The reflection R_{s} and transmission T_{s} coefficients of substrated MS in the uncoupled-element model [4] are as follows:

$$\begin{aligned} R_{\mathrm{s}} &= \frac{(1+e)(1-\sqrt{\varepsilon}m) - (\sqrt{\varepsilon}-e)(1+m)}{(1-e)(1-\sqrt{\varepsilon}m) + (\varepsilon-e)(1-m)}, \\ T_{\mathrm{s}} &= \frac{(1+e)(1+m) + (1-e)(1+m)}{(1-e)(1-\sqrt{\varepsilon}m) + (\varepsilon-e)(1-m)}, \end{aligned} \tag{2}$$

where $\varepsilon = n_{\mathrm{s}}^2$ is a relative dielectric constant of the substrate, $e = ik_0\chi_{\mathrm{e}}/2$ and $m = ik_0\chi_{\mathrm{m}}/2$.

The above described approach gives a good qualitative description of the properties of isotropic all-dielectric MS on a substrate. Namely, it predicts the number of maxima and minima in the spectrum and dipole resonances positions [2]. This is quite enough to use it as a block for a direct problem solution. However, if more accurate model is proposed, formulas (2) will be easily replaced by the refined ones.

2.2. Constraints on geometric parameters

Limitations on structure periodicity and meta-atom size can be of several types. Firstly, there are conditions imposed from physical considerations. Obviously, the geometric parameters of the MS are positive quantities $p > 0$ and $r > 0$ and the particles do not touch each other $p > 2r$. Secondly, there are limitations associated with the fabrication process. Thus, radius of identical spherical particles manufactured by dielectric material is usually not less than 50 nm. They are not located on the substrate closely, but with the interval equals to the particle diameter or more, therefore, $p \geqslant 4r$. And, thirdly, it is necessary to take into account the conditions when the physical model works.

In our case, we should keep in mind that the Evlyukhin model gives correct results only when the dipole approximation is applicable. In [6], the condition for the minimal period is derived

$$p_{\min} = \sqrt{\frac{k_0}{2} \frac{|\alpha_{\mathrm{e}}^{\mathrm{eff}}|^2 + |\alpha_{\mathrm{m}}^{\mathrm{eff}}|^2}{\mathrm{Im}\,(\alpha_{\mathrm{e}}^{\mathrm{eff}}) + \mathrm{Im}\,(\alpha_{\mathrm{m}}^{\mathrm{eff}})}}. \tag{3}$$

The maximum radius for lossless materials can be found from the criterion which requires that the dipole contribution to the scattered radiation is greater than or equal to 95% [14], [15]:

$$r_{\max} \approx \frac{\lambda}{1.3n + 1}. \tag{4}$$

Some conditions listed in this subsection are overlapped. To find physical solutions, the strongest ones should be used. In addition, as upper limit on p, it seems reasonable to choose $p \leqslant \lambda$ in order to exclude far-located and, thus, weakly interacting meta-atoms.

2.3. Objective functions and mathematical statement of the problem

Consider the simplest formulation of the problem: the reflectance should be minimized at some fixed wavelength $\lambda = \lambda_*$. We introduce the vector $\mathbf{x} = \{p, r\}$ describing the optimization parameters. Denote the reflectivity of the structure $|R_{\mathrm{s}}(\mathbf{x}, \lambda)|^2$ as $f(\mathbf{x}, \lambda)$. Let E_2 be a two-dimensional vector space, C_2 is the closed convex set

$$C_2 = \{\mathbf{x} \in E_2 : 4r \leqslant p \leqslant \lambda_{\max},\ r_{\mathrm{fabric}} \leqslant r \leqslant r_{\max}\}. \tag{5}$$

Then our goal is to determine the vector $\mathbf{x}$ which minimizes the function

$$f(\mathbf{x}, \lambda) = \min, \quad \mathbf{x} \in C_2. \tag{6}$$

A preliminary analysis of the function $f(\mathbf{x}, \lambda)$ behavior shows that it strongly depends on the particle radius (figure 2). For small r, its values practically do not change. The presence of such a horizontal plateau, which is a local minimum, leads to computation looping and further breakdown. At resonant radii, there are deep "ravines". Imposing restrictions on r, we exclude needless ravines: only dipole resonance is located to the left of $r_{\max}$. However, such limitation does not eliminate the plateau. Therefore, the problem remains multi-extremal.

To make the desired minimum unique, we modify $f(\mathbf{x}, \lambda)$ by adding a term in the form $y(r) = (Ar + B)^\beta$, where β is an even natural number. Simple estimation for the radius r_0, corresponding to MDR at the wavelength λ_*, is known [16], [17]. The figure 3 shows a symmetric "gutter" centered at $r_0 \approx \lambda_*/(2n)$ with width equal to $(r_{\max} - r_0)$. Changing the value of β, it is possible to control the slope of walls and the flatness of its bottom.

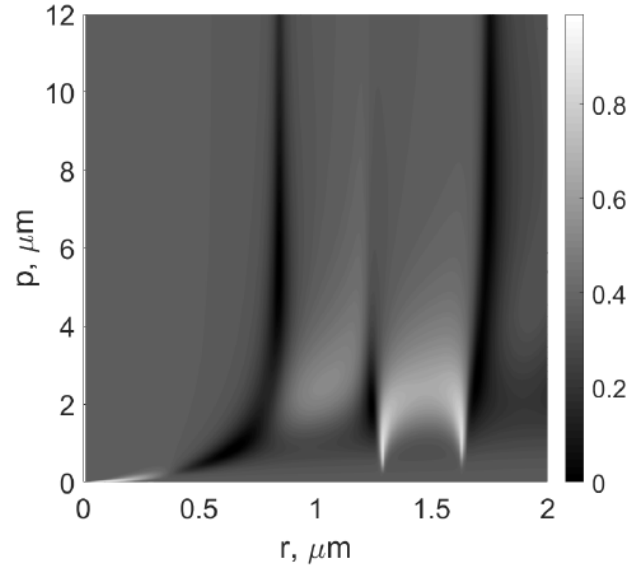


Figure 2. Dependence of $|R_s|^2$ on period p and radius r at $\lambda = 10$ μm

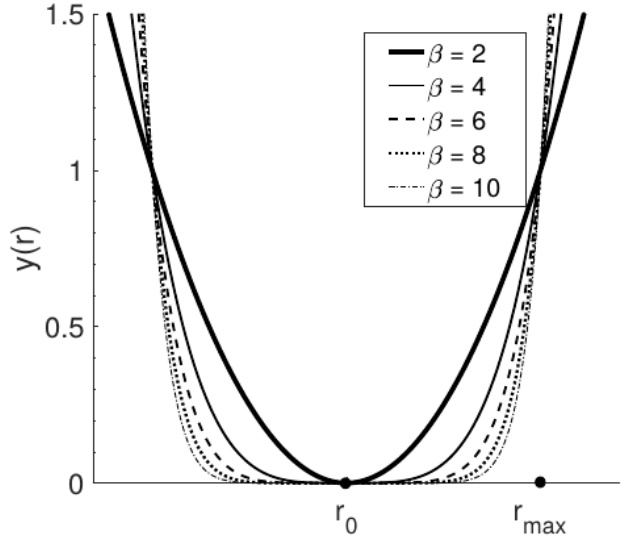


Figure 3. Penalty function $y(r)$ for different values of power β

Assume that $y(r_0) = 0$ and $y(r_{\max}) = 1$ on the wall of the gutter, then the proposed term is

$$y(r) = \left(\frac{r - r_0}{r_{\max} - r_0} \right)^{\beta}. \tag{7}$$

Due to $y(r)$ selection in the form (7), we discard minima at large radii (for which the dipole approximation does not work) and make the plateau at small radii non-horizontal, see the figure 4. Thus, $y(r)$ is a penalty function

that keeps one of optimization parameters within certain range. Finally, mathematical statement of the problem is the following

$$F^{\beta}[\mathbf{x}] = f(\mathbf{x}, \lambda) + y(r), \quad \mathbf{x} \in C_2. \tag{8}$$

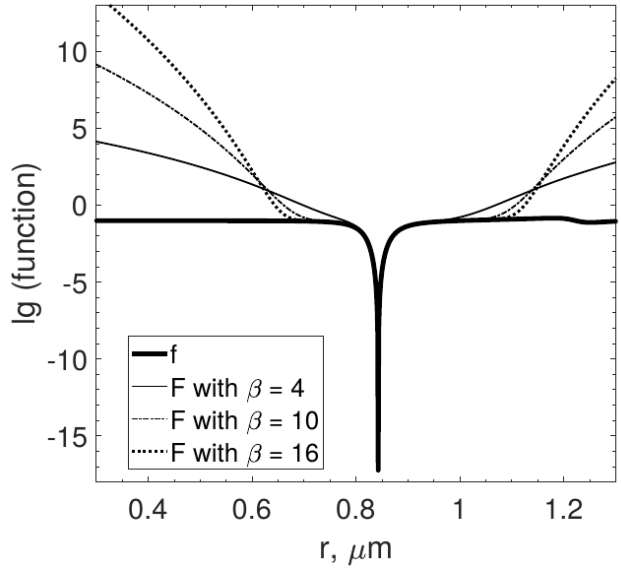


Figure 4. Dependencies of the objective functions (8) with different values of β (colored curves) and (6) (black curve) on radius r for fixed period $p = 5.4676$ μm

3. Optimization of the structure

Calculations were carried out for substrated MS (figure 1) with $n = 5$ (lead telluride PbTe) and $n_s = 4$ (germanium Ge) at the wavelength $\lambda_* = 10$ μm, which approximately corresponds to the human body temperature. The direct problem (i.e., one of the optimization algorithm blocks) is solved using analytical formulas (1)–(2). To find global minimum of $|R_{\text{s}}|^2$ for $p \in [4r, 12]$ μm and $r \in [0.05, 1.1983]$ μm, the standard function fmincon from MATLAB Optimization Toolbox was used. It solves minimization problem of a scalar nonlinear function of multiple variables with constraints using the interior point method.

3.1. Practical recommendations

For a prevailing part of software packages, the number of function evaluations is limited by default (i.e., there is the maximum number of iterations). For example, fmincon permits only 3000 evaluations. This measure prevents cycling. However, in the case of low gradient of the objective function, it stops the calculations before some minimum is found.

In the figure 5, the percentage of the initial approximations, for which numerical calculations converge to the minimum, is indicated near the points.

For small values of β, this value is 100%, and while β grows it decreases. The reason for this is as follows. In the objective function (8) with the penalty term, between the steep wall for large $(r - r_0)$ and the minimum there are quite flat areas (minimum "sides"), which become flatter with increasing of β (figure 4). These areas require more steps than available. Computations are interrupted and fmincon returns an error. In this case, it is recommended to take the last obtained values of p and r as new initial approximations and continue minimization. Since these sides are flat, but not horizontal, calculations converge to the minimum point.

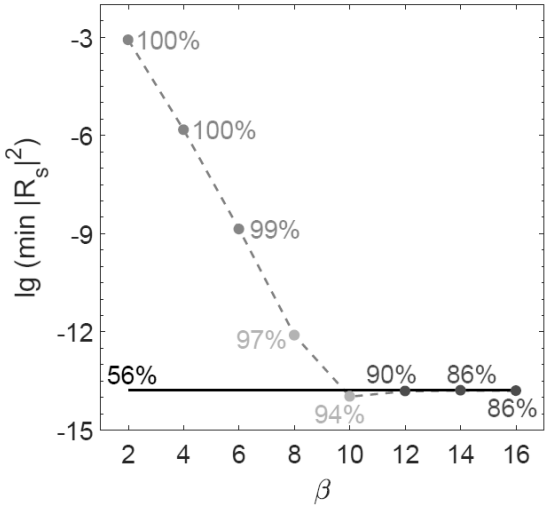


Figure 5. Minimal values of $|R_s|^2$ at logarithmic scale for computations with proposed penalty function (dotted curve) and without it (straight line). The percents of initial approximations, for which fmincon function converges to the minimum (see text), are indicated nearby

3.2. Comparison of the objective functions

To demonstrate the advantages of our approach, we compared the results of minimization with two objective functions (6) and (8). Initial approximations were chosen randomly: 10 computations were carried out with 100 points.

Their coordinates had Gaussian distribution, the average and the standard deviation were r_0 for meta-atom radius and $4r_0$ for structure periodicity. Each of these initial approximations was used for both objective functions.

For the minimum of $|R_s|^2$, the dependence of the depth on β is shown in the figure 5. For comparison, black line corresponding to the averaged value of minimum for the objective function (6) is added. It is clear that, using the penalty function with power $\beta \in [2, 6]$, we make the depth smaller because the center of the gutter r_0 does not exactly coincide with the coordinate of minimum point (figure 4). With the growth of β, bottom of the gutter becomes flatter, and the depth increases. Starting from $\beta = 10$, the minimum

depth is almost independent of β and does not differ from the value obtained without the penalty term (7).

3.3. Choice of the penalty function power

To choose the value of β, we were guided by the following considerations. The penalty function is introduced in order to eliminate all local minima that are not located near to MDR (approximately r_0) and EDR (close to $r_{\max}$) or between them. Therefore, the power β should satisfy the following conditions. On the one hand, the value $y(r)$ has to be be greater than 1 outside the specified range (all extra minima are automatically excluded from consideration). On the other hand, the penalty term should not distort the objective function (8) outside the range. These requirements are satisfied for $\beta = 10$ best of all.

Note that the usage of the penalty function (7) with power $\beta = 10$ practically does not affect the accuracy of obtained geometric parameters p and r, since it has a very flat bottom and does not distort the objective function (8). The figure 6 illustrates the accuracy δ of the obtained solutions using the interior point method versus the power β of the penalty term $y(r)$. The accuracy is the distance between the minimum points of the objective functions (6) and (8) under consideration $\delta = \sqrt{(p - p(\beta))^2 + (r - r(\beta))^2}$, where p and r are the coordinates of the best result of minimization without the penalty function ($|R_s|^2 \approx 8.0579 \cdot 10^{15}$). Here $p(\beta)$ and $r(\beta)$ denote minimum coordinates of (8).

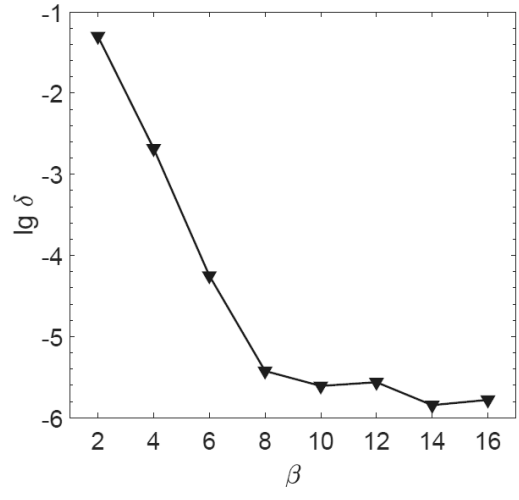


Figure 6. Accuracy of the minimization for different values of the penalty function power

Because of the presence of the penalty term $y(r)$ with insufficiently at bottom at the vicinity of the desired minimum, for small values of $\beta \in [2, 6]$, the accuracy, with which p and r are found, is not high enough (figure 6). Beginning with $\beta = 8$, the accuracy of the results of minimization δ coincides with the tolerance of fmincon function that is 10^{-6}.

3.4. Results of the minimization

The results of one of the computations with 100 random initial approximations for $\beta = 10$ are depicted on the graph of $|R_s(p, r)|^2|$ (figure 7). The domain of the arguments is shown by red lines. The results of minimization are marked with light dots for the objective function (8) and with dark ones for (6) without the penalty function. It is clearly seen that in the first case all 100 points converge to the same answer that is the global minimum. However, in the second case, 43 points "get stuck" on the plateau and 1 point on the horizontal area near the right boundary of the domain. They do not reach the desired minimum.

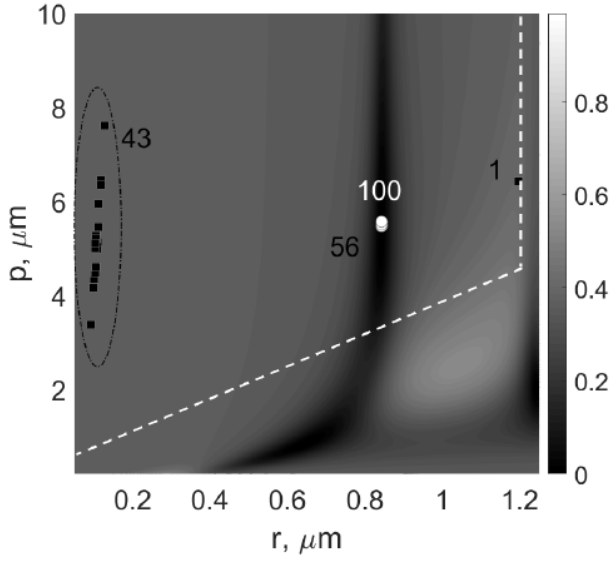


Figure 7. Results of the minimization of substrated metasurface at the wavelength $\lambda = 10$ μm. Found minima of the objective functions (8) and (6) are pointed out on the graph of the reflectance $|R_s(p, r)|^2$ with white and black markers, respectively. The number of points is indicated beside them. Red lines are the boundaries of C_2

To sum up, minimization of the first objective function is complicated and unstable (it depends on the choice of the initial approximation very strongly). Because of the existence of horizontal areas, in half of the cases the computations do not provide the correct answer for the position of narrow dip to be found. On the contrary, the objective function with power-law penalty term that we have constructed allows to find the desired global minimum without reference to the position of initial points. For default number of iterations, not more than 7–10 initial approximations are required.

4. Conclusions

The paper is devoted to the optimization of the geometric parameters of all-dielectric high refractive index isotropic metasurface placed on a semi-infinite dielectric substrate. To solve a direct problem, it is suggested to

use an analytical model combining several approaches of different authors. Constraints on period of the structure and radius of spherical meta-atoms are discussed. To construct the domain of geometric parameters, technological limitations and the conditions for physical model applicability were taken into account.

For the first time, the formulation of the problem of one-wavelength antireflective substrated metasurface is proposed, based on preliminary physical considerations about the location of narrow global minimum. Using the idea of the penalty functions, we suggest new objective function, which allows to cut off all minima except the desired one: a horizontal wide region at small radii and the local minima for large particles beyond the applicability of the dipole approximation. The results of minimization with power-law penalty term and without it are compared. The choice of the power for the penalty function providing the best result of optimization is described.

The developed technique is illustrated by the example of calculating a antireflective metasurface from PbTe on a Ge substrate for a wavelength of 10 μm when both materials are non-absorbent. The reflection spectrum of the structure under consideration is constructed in the range relevant for applications from 8 to 12 μm. It is shown that for non-absorbing materials, zero reflection occurs between the magnetic dipole resonance and the zero reflection region of the same metasurface, but located in the air.

References

[1] K. V. Baryshnikova, M. I. Petrov, V. E. Babicheva, and P. A. Belov, "Plasmonic and silicon nanoparticle anti-reflective coatings", *Scientific Reports*, vol. 6, p. 22 136, 2016. DOI: 10.1038/srep22136.

[2] V. E. Babicheva, M. I. Petrov, K. V. Baryshnikova, and P. A. Belov, "Reflection compensation mediated by electric and magnetic resonances of all-dielectric metasurfaces", *Journal of the Optical Society of America B*, vol. 34, no. 7, pp. D18–D28, 2017. DOI: 10.1364/JOSAB.34.000D18.

[3] H. A. Atwater and A. Polman, "Plasmonics for improved photovoltaic devices", *Nature Materials*, vol. 9, pp. 205–213, 2010. DOI: 10.1038/nmat2629.

[4] M. Albooyeh, D. Morits, and C. R. Simovski, "Electromagnetic characterization of substrated metasurfaces", *Metamaterials*, vol. 5, pp. 178–205, 2011. DOI: 10.1016/j.metmat.2011.08.002.

[5] Z. O. Dombrovskaya *et al.*, "Inverse problem for recovering of meta-atom characteristics by transmittance and reflectance of a metafilm", *Bulletin of the Russian Academy of Sciences: Physics*, vol. 79, no. 12, pp. 1496–1498, 2015. DOI: 10.3103/S1062873815120151.

[6] A. B. Evlyukhin *et al.*, "Optical response features of Si-nanoparticle arrays", *Physical Review B*, vol. 82, p. 045 404, 2010. DOI: 10.1103/PhysRevB.82.045404.

[7] N. N. Kalitkin and E. A. Al'shina, *Numerical Methods [Chislennye metody], book 1*, in Russian. Moscow: Akademiya, 2013.

[8] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming", *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 877–900, 1999. DOI: 10.1137/S1052623497325107.

[9] A. G. Sveshnikov and A. S. Ilinskiy, "Design problems in electrodynamics [Zadachi proyektirovaniya v elektrodinamike]", in Russian, *Proceedings of the USSR Academy of Sciences*, vol. 204, pp. 1077–1080, 1972.

[10] V. B. Glasko, A. N. Tikhonov, and A. V. Tikhonravov, "On the synthesis of multilayer coatings [O sinteze mnogosloynykh pokrytiy]", *USSR Computational Mathematics and Mathematical Physics*, vol. 14, p. 135, 1974, in Russian.

[11] A. V. Tikhonravov *et al.*, "Design and production of antireflection coating for the $8 - 10$ μm spectral region", *Optics Express*, vol. 22, pp. 32 174–32 179, 2014. DOI: 10.1364/OE.22.032174.

[12] Z. O. Dombrovskaya and A. V. Zhuravlev, "Investigation of the possibility of metafilm modeling as a conventional thin film", *Applied Physics A*, vol. 123, p. 27, 2017. DOI: 10.1007/s00339-016-0642-2.

[13] A. E. Miroshnichenko *et al.*, "Substrate-induced resonant magnetoelectric effects with dielectric nanoparticles", *ACS Photonics*, vol. 2, pp. 1423–1428, 2015. DOI: 10.1021/acsphotonics.5b00117.

[14] G. V. Belokopytov and A. V. Zhuravlev, "Dipole polarizability of spherical particles [Dipol'naya polyarizuyemost' sfericheskikh chastits]", in Russian, *Physics of Wave Processes and Radio Systems*, vol. 2, pp. 41–49, 2008.

[15] Z. O. Dombrovskaya *et al.*, "Phonon-polariton meta-atoms for far infrared range", *Physics of Wave Phenomena*, vol. 24, pp. 96–102, 2016. DOI: 10.3103/S1541308X16020023.

[16] A. I. Kuznetsov *et al.*, "Magnetic light", *Scientific Reports*, vol. 2, p. 492, 2012. DOI: 10.1038/srep00492.

[17] D. G. Baranov *et al.*, "All-dielectric nanophotonics: the quest for better materials and fabrication techniques", *Optica*, vol. 4, no. 7, pp. 814–825, 2017. DOI: 10.1364/OPTICA.4.000814.

For citation:

Z. O. Dombrovskaya, Optimization of an isotropic metasurface on a substrate, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 115–126. DOI: 10.22363/2658-4670-2022-30-2-115-126.

Information about the authors:

Dombrovskaya, Zhanna O. — Candidate of Physical and Mathematical Sciences, Junior Researcher of Faculty of Physics, Lomonosov Moscow State University (e-mail: dombrovskaya@physics.msu.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0003-0609-1065)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-115-126

Оптимизация изотропной метаповерхности на подложке

Ж. О. Домбровская

Московский государственный университет им. М. В. Ломоносова, Ленинские горы, д. 1, стр. 2, Москва, 119991, Россия

Аннотация. Впервые приведена математическая формулировка одноволнового безотражательного покрытия на основе двумерного метаматериала. Найдены ограничения на геометрические параметры конструкции. Предложена штрафная функция, которая обеспечивает применимость физической модели и обеспечивает единственность искомого минимума. В качестве примера рассмотрена оптимизация метаповерхности, состоящей из сфер PbTe, расположенных на германиевой подложке. Показано, что точность минимизации с правильно выбранным штрафным термином такая же, как и для целевой функции без него.

Ключевые слова: оптимизация безотражательного покрытия, метод штрафной функции, ограничения на геометрические параметры, диэлектрическая метаплёнка на подложке

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 127–138
http://journals.rudn.ru/miph

UDC 519.872:519.217
DOI: 10.22363/2658-4670-2022-30-2-127-138

Multistage pseudo-spectral method (method of collocations) for the approximate solution of an ordinary differential equation of the first order

Konstantin P. Lovetskiy[1],
Dmitry S. Kulyabov[1,2], **Ali Weddeye Hissein**[1]

[1] *Peoples' Friendship University of Russia (RUDN University),*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Joint Institute for Nuclear Research,*
6, Joliot-Curie St., Dubna, Moscow Region, 141980, Russian Federation

(received: February 4, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. The classical pseudospectral collocation method based on the expansion of the solution in a basis of Chebyshev polynomials is considered. A new approach to constructing systems of linear algebraic equations for solving ordinary differential equations with variable coefficients and with initial (and/or boundary) conditions makes possible a significant simplification of the structure of matrices, reducing it to a diagonal form. The solution of the system is reduced to multiplying the matrix of values of the Chebyshev polynomials on the selected collocation grid by the vector of values of the function describing the given derivative at the collocation points. The subsequent multiplication of the obtained vector by the two-diagonal spectral matrix, 'inverse' with respect to the Chebyshev differentiation matrix, yields all the expansion coefficients of the sought solution except for the first one. This first coefficient is determined at the second stage based on a given initial (and/or boundary) condition. The novelty of the approach is to first select a class (set) of functions that satisfy the differential equation, using a stable and computationally simple method of interpolation (collocation) of the derivative of the future solution. Then the coefficients (except for the first one) of the expansion of the future solution are determined in terms of the calculated expansion coefficients of the derivative using the integration matrix. Finally, from this set of solutions only those that correspond to the given initial conditions are selected.

Key words and phrases: initial value problems, pseudo spectral collocation method, Chebyshev polynomials, Gauss–Lobatto sets, numerical stability

1. Introduction

Spectral methods are a class of methods used in applied mathematics and scientific computing to solve many differential equations numerically [1]–[4]. The main idea of the method is to represent the desired solution of a differential equation as a sum of certain 'basis functions' [5] (e.g., as an expansion into a sum in power functions — a Taylor series, or a sum of

sinusoids, which is a Fourier series), and then calculate the coefficients in the sum to satisfy the differential equation in the best possible way.

Spectral and finite element methods are closely related and are based on the same ideas. The main difference between them is that spectral methods use nonzero basis functions over the entire domain, while finite element methods use nonzero basis functions only on small subdomains. In other words, spectral methods use a global approach, while finite element methods use a local approach. It is for this reason that spectral methods provide excellent convergence, their 'exponential convergence' being the fastest possible when the solution is smooth.

Spectral methods for the numerical solution of ordinary differential equations with given initial conditions are often reduced to solving a system of linear algebraic equations (SLAE), which includes both the initial conditions and conditions that ensure the fulfillment of differential relations [6]. However, a priori embedding of the initial (boundary) conditions into the system of linear equations leads to a significant increase in the filling of the matrices and, consequently, to the complication of the algorithm and method for solving the problem [7].

A more interesting approach is to select a basis that automatically takes into account the boundary conditions [1], [5], [6]. This is a frequently used trick when formulating the SLAE of the initial problem, and it reduces to taking into account the required initial/boundary conditions when creating the basis (a set of good basis functions-orthogonal, etc.) in a natural way, i.e., a basis is selected in which each basis function satisfies the initial conditions. The solution obtained using this approach is automatically sought in the class of functions satisfying the initial conditions. However, in this case it becomes much more difficult to work with new basis functions.

The novelty of the approach proposed by the authors is that first, a class (set) of functions that satisfy the differential equation is selected using a stable and computationally simple method of interpolation (collocation) of the derivative of the future solution. Then the coefficients (except for some) of the expansion of the future solution are determined in terms of the calculated expansion coefficients of the derivative using the integration matrix. Only after that, from this set of solutions those that correspond to the given initial conditions are selected.

Here we propose to divide the main problem into independent subproblems and to calculate the solution components in parts — separately those that determine the behavior of the derivative of the solution, and separately those that are determined by the boundary conditions. Thus, the problem is divided into two independent subproblems, each allowing stable and simple solution. The solution of the first problem in the simplest case is reduced to multiplying the vector of the right-hand side by the matrix of the Chebyshev functions values on the Gauss–Lobatto grid. At the next step, we solve the SLAE with a diagonal positive definite matrix and, multiplying the resulting vector on the left by the two-diagonal matrix, inverse (anti–derivative) with respect to the spectral Chebyshev matrix of differentiation, we obtain all the expansion coefficients of the desired solution, except for the first one. At the second, 'most difficult' stage, we determine the first coefficient of the expansion of the solution in terms of basis polynomials, solving a linear algebraic equation of the first order with respect to this coefficient.

2. Numerical solution of ordinary differential equations

Exact solution of a trivial ordinary differential equation for a given initial (boundary) condition

$$y' = f(x), \quad x \geqslant x_0, \quad y(x) = y_0, \tag{1}$$

the right–hand side of which is independent of y, can be presented in the form $y_0 + \int_{t_0}^{t} f(\tau)d\tau$.

Since the numerical methods for integrating functions are well developed from theoretical and practical points of view, it seems natural to apply them to the numerical solution of ordinary differential equations of general form

$$y'(x) = f(x, y(x)), \quad x \geqslant x_0, \quad y(x_0) = y_0, \tag{2}$$

and this is exactly the fact that naturally explains the development and wide use of the methods of the Runge–Kutta type.

Usually, the method implies obtaining the solution in the interval $[x_0, x_0 + c_k h]$. The coefficients $0 \leqslant c_1 < c_2 < ... < c_n \leqslant 1$ are chosen. Then, using the method of polynomial collocation, the solution is approximated by a polynomial p of the degree n, which satisfies two types of conditions

— the initial condition: $p(x_0) = y_0$, and
— the differential equation, $p'(x_k) = f(x_k, p(x_k))$, at all the *collocation* points $[x_k = x_0 + c_k h]$, $k = 1, ..., n$.

Satisfying these $(n + 1)$ conditions allows calculating $(n + 1)$ coefficients of the expansion of the sought polynomial p of the degree n.

Thus, the collocation methods are actually implicit Runge–Kutta methods [8].

It is important to note that to solve the approximation problem, it is not necessary to try solving Eq. (1) with simultaneous satisfaction of both the initial condition and the differential equation at the collocation points. In some cases, a fast and stable result can be achieved in two stages. First, to find those coefficients of the sought solution expansion that satisfy the differential equation at the collocation points. Then, to determine the deficient coefficients of the sought function expansion using the initial (final or intermediate) value.

3. Approximation of derivative. Cauchy problem

First, consider the problem of determining (recovering) a function from its derivative and some (one) additional condition. In this formulation, the problem naturally splits into two sub-problems:

— polynomial interpolation of the derivative (calculating the coefficients of the expansion of the derivative into a finite series in basis functions) and
— calculation of the coefficients of the required function by the initial (boundary, etc.) condition and the coefficients of the derivative expansion.

Without loss of generality, we assume that the domain of definition of the solution is the interval $[-1, 1]$.

Most often, the approximation of continuous functions is obtained by discarding the terms of the Chebyshev series, the magnitude of which is small [9], [10]. In contrast to the approximations obtained using other power series, the approximation in Chebyshev polynomials (having the property of being almost optimal) minimizes the number of terms required to approximate a function by polynomials with a given accuracy. This is also related to the property that the approximation based on the Chebyshev series turns out to be close to the best uniform approximation (among polynomials of the same degree), but easier to find. In addition, it allows avoiding the Gibbs effect with a reasonable choice of interpolation points.

Let us consider in more detail the problem of finding the derivative of the desired function, or rather the approximating polynomial $p(x)$, satisfying condition (1) at a given number of points in the interval $[-1, 1]$. The pseudospectral (collocation) method [11] for solving the problem consists in representing the desired approximating function in the form of an expansion in a finite series in Chebyshev polynomials

$$p(x) = \sum_{k=0}^{n} c_k T_k(x), \quad x \in [-1, 1] \tag{3}$$

using the basis of Chebyshev polynomials of the first kind $\{T_k(x)\}_{k=0}^{\infty}$, defined in the Hilbert space of functions on the segment $[-1, 1]$.

Let us differentiate the function (3). The derivative is expressed as

$$p'(x) = \frac{d}{dx}\left(\sum_{k=0}^{n} c_k T_k(x)\right) = \sum_{k=0}^{n} c_k T_k'(x) = \sum_{k=0}^{n} b_k T_k(x), \ x \in [-1, 1]. \tag{4}$$

Using the recurrence relations, which are satisfied by the Chebyshev polynomials of the first kind and their derivatives [3], [12] and equating the coefficients at the same polynomials in (4), we come [3] to the following dependence of the coefficients c_k on b_k:

$$\begin{bmatrix} 1 & 0 & -1/2 & 0 & \dots & 0 & 0 \\ 0 & 1/4 & 0 & -1/4 & \dots & 0 & 0 \\ 0 & 0 & 1/6 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1/8 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dfrac{1}{2(n-1)} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1/(2n) \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix}. \tag{5}$$

That is, the vector calculation of the coefficients $\{c_1, c_2, \dots, c_n\}$ is the result of multiplying a simple tridiagonal matrix by a vector and it can be implemented by the following explicit formulas

$$\begin{cases} c_1 = b_0 - b_2/2, \quad k = 1, \\ c_k = \left(b_{k-1} - b_{k+1}\right)/2k, \quad k > 1, \quad k < n-1, \\ c_k = b_{k-1}/2k, \quad k = n-1, n. \end{cases} \tag{6}$$

Thus, known the expansion coefficients b_k of the function $f(x)$ of problem (1) in Chebyshev polynomials of the first kind, we can recover the last N expansion coefficients of the sought function in the same basis functions by formulas (2.1.3) from [3].

Therefore, the first part of the problem is to calculate the coefficients $\{b_0, b_1, \dots, b_n\}$ of the representation of the right–hand side of (1) on the interval $[-1, 1]$

$$\sum_{k=0}^{n-1} b_k T_k(x) = f(x).$$

The collocation method consists in the selection of such coefficients $\{b_0, b_1, \dots, b_n\}$ of the expansion of the interpolation polynomial $p'(x)$ that the following equalities are satisfied for the desired coefficients b_k, $k = 0, 1, \dots, n-1$.

$$\sum_{k=0}^{n-1} b_k T_k\left(x_j\right) = f\left(x_j\right), \quad j = 0, \dots, n-1 \tag{7}$$

at the collocation points $\{x_0, x_1, \dots, x_n\}$.

The last statement is equivalent to the fact that the coefficients b_k, $k = 0, \dots, n$ must be a solution to the system of linear algebraic equations (7) of the collocation method. In matrix form, this can be written as

$$Tb = f. \tag{8}$$

The choice of collocation points should ensure the nondegeneracy of the system of Eqs. (7); for this it is sufficient that all grid points are different, and otherwise their choice is arbitrary, that is, the solution of system (7) on an arbitrary grid of the interval $[-1, 1]$ determines the required approximation. For an arbitrary grid, the matrix T is completely filled and the solution of such a system is rather laborious. To simplify the form of the matrix and speed up the search for the vector b, we use the discrete orthogonality property of the Chebyshev matrix T on the Gauss–Lobatto grid. Consider the set $x_j = \cos(j/n)$, $j = 0, \dots, n$ as collocation points. To further improve the properties of the system of linear equations, the solution of which will be the vector $\{b_0, b_1, \dots, b_n\}$, we multiply the first and last equations (7) by the factor $1/\sqrt{2}$. We obtain an equivalent 'modified' system with a new matrix $\tilde{T}$ (instead of T) and a vector $\tilde{f}$ instead of f. The good thing about the new system is that it has the property of discrete 'orthogonality' and multiplying it on the left by the transposed $\tilde{T}^T$ gives a diagonal matrix:

$$\tilde{T}^T\tilde{T} = \begin{bmatrix} n & 0 & 0 & ... & 0 \\ 0 & n/2 & 0 & ... & 0 \\ 0 & 0 & n/2 & ... & 0 \\ ... & ... & ... & \ddots & ... \\ 0 & 0 & 0 & ... & n \end{bmatrix}.$$

We transform system (8), multiplying it on the left by the transposed matrix $\tilde{T}^T$. As a result, we obtain a simple matrix equation with a diagonal matrix to determine the required expansion coefficients $\{b_0, b_1, ..., b_n\}$:

$$\begin{bmatrix} n & 0 & 0 & ... & 0 \\ 0 & n/2 & 0 & ... & 0 \\ 0 & 0 & n/2 & ... & 0 \\ ... & ... & ... & \ddots & ... \\ 0 & 0 & 0 & ... & n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ ... \\ b_n \end{bmatrix} = \tilde{T}^T \begin{bmatrix} f_0/\sqrt{2} \\ f_1 \\ f_2 \\ ... \\ f_n/\sqrt{2} \end{bmatrix}. \tag{9}$$

Denoting by $\left(\tilde{f}_0, \tilde{f}_1, ..., \tilde{f}_{n-1}, \tilde{f}_n\right)^T$ the product of matrix $\tilde{T}^T$ by vector $\left(f_0/\sqrt{2}, f_1, ..., f_{n-1}, f_n/\sqrt{2}\right)^T$ in the right–hand side of equation (9), we write down the required expansion coefficients of the derivative of the solution – the function $f(x)$ — in the explicit form

$$\begin{cases} b_0 = \dfrac{\tilde{f}_0}{n}, \\ b_1 = \dfrac{2\tilde{f}_1}{n}, \\ b_2 = \dfrac{2\tilde{f}_2}{n}, \\ ... \\ b_n = \dfrac{\tilde{f}_n}{n}. \end{cases} \tag{10}$$

Consequently, relations (10), (6) uniquely determine the last **n** coefficients $\{c_1, c_2, ..., c_n\}$ of the expansion of the sought function $p(x)$, and to determine one more coefficient c_0 it is necessary to involve at least one more additional condition. This can be both a boundary condition at the left or right end of the interval of consideration of a function, or a condition for the passage of the desired function through any given point within the interval of specifying the function.

That is, the considered method makes it possible to solve, depending on the type of the additional condition, both the Cauchy problem with initial conditions and problems with boundary conditions of a general form, requiring, for example, the use of the iterative shooting method [4].

In the case when the boundary condition is specified at the left end of the integration interval, the zero coefficient is determined from the equation

$$c_0 + \sum_{k=1}^{n} c_k T_k(-1) = y_0 \tag{11}$$

by the formula (taking into account that $T_k(-1) = (-1)^k$) for any Chebyshev polynomial

$$c_0 = y_0 - \sum_{k=1}^{n} c_k T_k(-1) = y_0 - \sum_{k=1}^{n} c_k (-1)^k. \tag{12}$$

If the additional 'boundary' condition is specified at an arbitrary point of the integration interval, $y_b = y(x_b)$, $x_b \in [-1, 1]$, then the coefficient c_0 is determined by the formula

$$c_0 = y_b - \sum_{k=1}^{n} c_k T_k(x_b). \tag{13}$$

At the right–hand end of the integration interval $y_r = y(1)$, $x_r = 1$, the Chebyshev polynomials of any order take the value equal to 1 ($T_k(1) = 1$). Therefore, the coefficient c_0 is determined by the formula

$$c_0 = y_r - \sum_{k=1}^{n} c_k T_k(x_r) = y_r - \sum_{k=1}^{n} c_k. \tag{14}$$

4. Examples with simplest differential equations

Reconstructing a function from its derivative and a boundary condition. Comparison with the Runge–Kutta–Fehlberg method [13]

$$\frac{dy}{dx} = f(x), \quad y(0) = y_0, \quad x \in [a, b].$$

Let us compare the solutions obtained by the Runge–Kutta method (subroutine RKF45) and the solutions obtained as previously described.

Let us specify a grid in the interval $[a, b]$, calculated by the formula

$$x_j = \frac{b-a}{2} \cos\left(\frac{j}{N-1}\right) + \frac{b+a}{2}, \quad j = 0, 1, \ldots, N-1,$$

and related to the chosen Gauss–Lobatto grid in the interval $[-1, 1]$. The number of grid points equals N, i.e., to recover the function from the given derivative and additional condition by our method, only N calculations of the function (the right–hand side) are needed, and the recalculation of these values into the expansion coefficients in Chebyshev polynomials will require only $2N$ divisions and $2N$ additions.

To solve the Cauchy problem by the Runge–Kutta–Fehlberg method, we applied the RKF45 algorithm on each subinterval of the grid calculated above on $[a, b]$.

Algorithms are compared when looking for a solution to the simplest problem

$$\frac{dy}{dx} = \cos(x), \quad y(0) = 0, \quad x \in [-\pi, \pi].$$

The calculation carried out by the Runge–Kutta method with automatic control of accuracy (not worse than 10^{-9}) required about 800 calculations of the function values over the entire interval.

For the two–stage method of separation of unknowns, the results of the deviation of the calculated values from the exact ones at the grid points are given in the table 1.

Table 1

Deviation of the calculated values from the exact ones

Number of grid points	11	13	15	30
Maximum deviation	$< 4 \cdot 10^{-7}$	$< 5 \cdot 10^{-9}$	$< 2 \cdot 10^{-13}$	$< 10^{-19}$

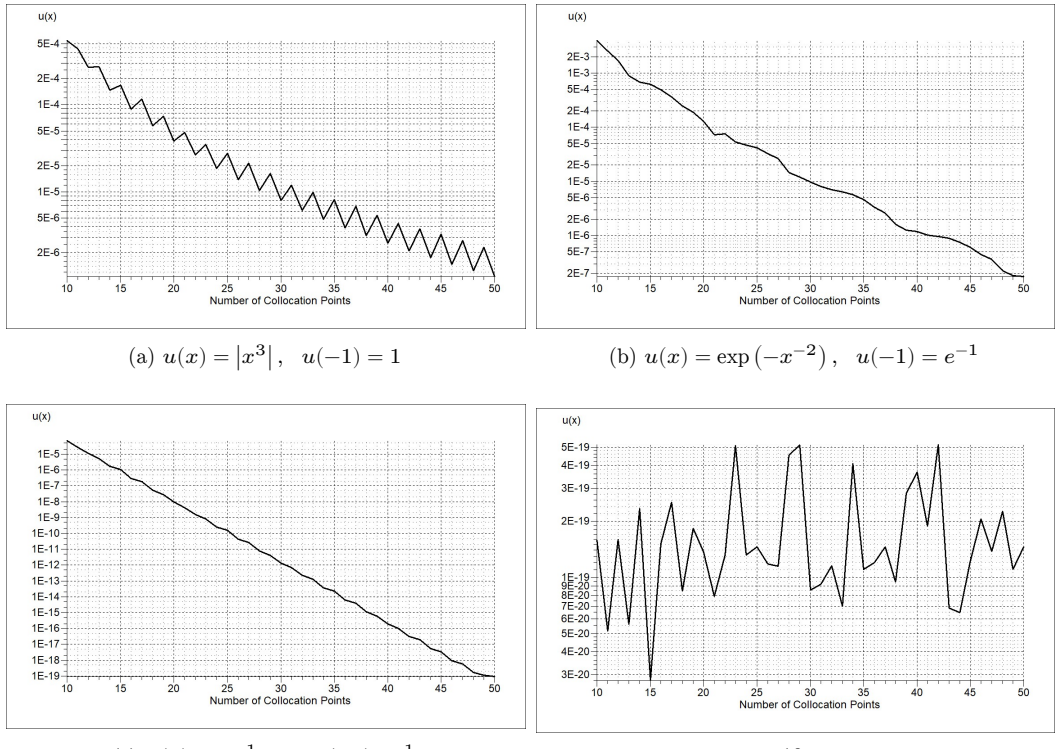
Consider a few more model examples of solving the Cauchy problem, i.e., recovering functions from given derivatives and an initial condition. Functions from [14], in which the accuracy of calculating derivatives with the help of Chebyshev matrices of differentiation in physical space, were studied as model ones. The selected examples systematically illustrate the accuracy of calculating derivatives as a function of the number of approximation points (see the figure 1).

Four functions characterized by different smoothness are considered: $|x^3|$, $\exp(-x^{-2})$, $1/(1+x^2)$, and x^{10}. The first function has the third derivative of bounded variation, the second function is smooth, but not analytical, the third one is analytical in the vicinity of $[-1, 1]$, and the fourth function is a polynomial. The accuracy of solutions obtained by us is by 1.5–3 orders of magnitude better than Trefethen's solutions [14] when using a similar number of collocation points.

5. Discussion and conclusion

There are methods based mainly on the local approximation of the solution, which primarily use the initial approximation (boundary conditions) when solving differential equations. These are methods like Euler, Runge–Kutta method, etc. Other methods based on the approximation of the solution using global functions [global collocation methods — Mason, Boyd, Fornberg, Iserles, Townsend] are based on the construction of such systems of equations that simultaneously include both initial (boundary) conditions and conditions specifying the behavior of the derivatives of the desired solution.

In our study (within the framework of the pseudospectral collocation method), the problem is divided into two independent subproblems. The first is to select a set of solutions that satisfies the differential equation. However, it does not necessarily satisfy the initial (boundary) conditions. The choice of suitable bases for representing the solution in the form of an expansion



(a) $u(x) = |x^3|, \quad u(-1) = 1$

(b) $u(x) = \exp\left(-x^{-2}\right), \quad u(-1) = e^{-1}$

(c) $u(x) = \frac{1}{1+x^2}, \quad u(-1) = \frac{1}{2}$

(d) $u(x) = x^{10}, \quad u(-1) = 1$

Figure 1. The accuracy of derivative recovering for four functions with increasing smoothness depending on the number of approximation points

in polynomials, e.g., Jacobi ones, and grids taking into account the discrete orthogonality of the considered bases, makes it possible to use highly efficient algorithms.

Perhaps, the most promising from the point of view of the application of numerical methods is the use of a particular case of Jacobi polynomials — Chebyshev polynomials, as specific basis functions [15]. The Chebyshev polynomials provide an almost optimal approximation of the ODE solution, the ease of calculating the Gauss–Lobatto grid for using the discrete orthogonality condition, and three-term relations determining the ease of constructing differentiation and integration matrices of the sought solutions [16].

The initial (boundary) conditions are considered at the second stage of solving the original problem, which is actually reduced to solving a linear equation with one unknown coefficient.

The solution of the first problem is reduced to multiplying the matrix of values of the Chebyshev functions on the Gauss–Lobatto grid by the vector of values of the function that defines the right-hand side of the original differential equation to determine the expansion coefficients of the solution derivative. Further, the multiplication of the two-diagonal integration matrix [3] by the vector of coefficients yields all the coefficients of the desired solution, except for the first one. At the second stage, the use of the initial (boundary) condition makes it possible to determine the first coefficient of the solution expansion.

The approach based on dividing the problem of solving first–order ODEs into two subproblems seems to be very promising. The authors will continue to develop approaches and algorithms for the application of the multistage pseudospectral method for solving initial and boundary value problems with differential equations of higher orders.

Acknowledgments

This paper has been supported by the RUDN University Strategic Academic Leadership Program.

References

[1] J. P. Boyd, *Chebyshev and Fourier spectral methods*, 2nd ed. Dover Books on Mathematics, 2013.

[2] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. London: Chapman and Hall/CRC Press, 2002.

[3] B. Fornberg, *A practical guide to pseudospectral methods*. Cambridge, England: Cambridge University Press, 1996. DOI: 10.1017/cbo9780511626357.

[4] M. Planitz *et al.*, *Numerical recipes: the art of scientific computing*, 3rd. New York: Cambridge University Press, 2007.

[5] J. Shen, T. Tang, and L.-L. Wang, *Spectral methods*. Berlin, Heidelberg: Springer, 2011, vol. 41.

[6] S. Olver and A. Townsend, "A fast and well-conditioned spectral method", *SIAM Rev.*, vol. 55, no. 3, pp. 462–489, 2013. DOI: 10.1137/120865458.

[7] S. Chandrasekaran and M. Gu, "Fast and stable algorithms for banded plus semiseparable systems of linear equations", *SIAM Journal on Matrix Analysis and Applications*, vol. 25, no. 2, pp. 373–384, 2003. DOI: 10.1137/S0895479899353373.

[8] A. Iserles, *A first course in the numerical analysis of differential equations*, 2nd edition. Cambridge: Cambridge University Press, 2008. DOI: 10.1017/CBO9780511995569.

[9] X. Zhang and J. P. Boyd, "Asymptotic coefficients and errors for Chebyshev polynomial approximations with weak endpoint singularities: effects of different bases", Online. https://arxiv.org/pdf/2103.11841.pdf, 2021.

[10] J. P. Boyd and D. H. Gally, "Numerical experiments on the accuracy of the Chebyshev–Frobenius companion matrix method for finding the zeros of a truncated series of Chebyshev polynomials", *Journal of Computational and Applied Mathematics*, vol. 205, no. 1, pp. 281–295, 2007. DOI: 10.1016/j.cam.2006.05.006.

[11] D. Dutykh, "A brief introduction to pseudo-spectral methods: application to diffusion problems", Online. https://arxiv.org/pdf/1606.05432.pdf, 2016.

[12] A. Amiraslani, R. M. Corless, and M. Gunasingam, "Differentiation matrices for univariate polynomials", *Numer. Algorithms*, vol. 83, no. 1, pp. 1–31, 2020. DOI: 10.1007/s11075-019-00668-z.

[13] E. Hairer, G. Wanner, and S. P. Nørsett, *Solving ordinary differential equations I*. Berlin: Springer, 1993. DOI: 10.1007/978-3-540-78862-1.

[14] L. N. Trefethen, *Spectral methods in MATLAB*. Philadelphia: SIAM, 2000.

[15] K. P. Lovetskiy, L. A. Sevastianov, D. S. Kulyabov, and N. E. Nikolaev, "Regularized computation of oscillatory integrals with stationary points", *Journal of Computational Science*, vol. 26, pp. 22–27, 2018. DOI: 10.1016/j.jocs.2018.03.001.

[16] L. A. Sevastianov, K. P. Lovetskiy, and D. S. Kulyabov, "An effective stable numerical method for integrating highly oscillating functions with a linear phase", *Lecture Notes in Computer Science*, vol. 12138, pp. 29–43, 2020. DOI: 10.1007/978-3-030-50417-5_3.

For citation:

K. P. Lovetskiy, D. S. Kulyabov, A. W. Hissein, Multistage pseudo-spectral method (method of collocations) for the approximate solution of an ordinary differential equation of the first order, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 127–138. DOI: 10.22363/2658-4670-2022-30-2-127-138.

Information about the authors:

Lovetskiy, Konstantin P. — Candidate of Physical and Mathematical Sciences, Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: lovetskiy-kp@rudn.ru, phone: +7(495)9522572, ORCID: https://orcid.org/0000-0002-3645-1060 , ResearcherID: A-5725-2017, Scopus Author ID: 18634692900)

Kulyabov, Dmitry S. — Docent, Doctor of Sciences in Physics and Mathematics, Professor at the Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: kulyabov-ds@rudn.ru, phone: +7(495)9520250, ORCID: https://orcid.org/0000-0002-0877-7063, ResearcherID: I-3183-2013, Scopus Author ID: 35194130800)

Hissein, Ali Weddeye — student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: 1032209306@rudn.ru, phone: +7(977)4294833, ORCID: https://orcid.org/0000-0003-1100-4966, ResearcherID: AAD-7566-2022)

УДК 519.872:519.217
DOI: 10.22363/2658-4670-2022-30-2-127-138

Многостадийный псевдоспектральный метод (метод коллокаций) приближенного решения обыкновенного дифференциального уравнения первого порядка

К. П. Ловецкий[1], **Д. С. Кулябов**[1,2], **Али Уэддей Хиссен**[1]

[1] *Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия*
[2] *Объединённый институт ядерных исследований, ул. Жолио-Кюри, д. 6, Дубна, Московская обл., 141980, Россия*

Аннотация. Рассмотрен классический псевдоспектральный метод коллокации, основанный на разложении решения по базису из полиномов Чебышева. Новый подход к формированию систем линейных алгебраических уравнений для решения обыкновенных дифференциальных уравнений с переменными коэффициентами и с начальными (и/или граничными) условиями позволяет значительно упростить структуру матриц, приводя её к диагональной форме. Решение системы сводится к умножению матрицы значений полиномов Чебышева на выбранной сетке коллокации на вектор значений функции, описывающей заданную производную в точках коллокации. Следующее за этой операцией умножение полученного вектора на двухдиагональную спектральную «обратную» по отношению к матрице дифференцирования Чебышева приводит к получению всех коэффициентов разложения искомого решения за исключением первого. Этот первый коэффициент определяется на втором этапе исходя из заданного начального (и/или граничного) условия. Новизна подхода заключается в том, чтобы сначала выделить класс (множество) функций, удовлетворяющих дифференциальному уравнению, с помощью устойчивого и простого с вычислительной точки зрения метода интерполяции (коллокации) производной будущего решения. Затем рассчитать коэффициенты (кроме первого) разложения будущего решения по вычисленным коэффициентам разложения производной с помощью матрицы интегрирования. И лишь после этого выделять из этого множества решений те, которые соответствуют заданным начальным условиям.

Ключевые слова: начальные задачи, метод псевдоспектральных коллокаций, многочлены Чебышева, множества Гаусса–Лобатто, численная устойчивость

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 139–148
http://journals.rudn.ru/miph

Research article

UDC 19.62:530.145:519.614
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-139-148

Complex eigenvalues in Kuryshkin–Wodkiewicz quantum mechanics

Alexander V. Zorin[1],
Mikhail D. Malykh[1, 2], **Leonid A. Sevastianov**[1, 2]

[1] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Joint Institute for Nuclear Research*
6, Joliot-Curie St., Dubna, Moscow Region, 141980, Russian Federation

(received: February 11, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. One of the possible versions of quantum mechanics, known as Kuryshkin–Wodkiewicz quantum mechanics, is considered. In this version, the quantum distribution function is positive, but, as a retribution for this, the von Neumann quantization rule is replaced by a more complicated rule, in which an observed value A is associated with a pseudodifferential operator $\hat{O}(A)$. This version is an example of a dissipative quantum system and, therefore, it was expected that the eigenvalues of the Hamiltonian should have imaginary parts. However, the discrete spectrum of the Hamiltonian of a hydrogen-like atom in this theory turned out to be real-valued. In this paper, we propose the following explanation for this paradox. It is traditionally assumed that in some state ψ the quantity A is equal to λ if ψ is an eigenfunction of the operator $\hat{O}(A)$. In this case, the variance $\hat{O}((A-\lambda)^2)\psi$ is zero in the standard version of quantum mechanics, but nonzero in Kuryshkin's mechanics. Therefore, it is possible to consider such a range of values and states corresponding to them for which the variance $\hat{O}((A-\lambda)^2)$ is zero. The spectrum of the quadratic pencil $\hat{O}(A^2) - 2\hat{O}(A)\lambda + \lambda^2\hat{E}$ is studied by the methods of perturbation theory under the assumption of small variance $\hat{D}(A) = \hat{O}(A^2) - \hat{O}(A)^2$ of the observable A. It is shown that in the neighborhood of the real eigenvalue λ of the operator $\hat{O}(A)$, there are two eigenvalues of the operator pencil, which differ in the first order of perturbation theory by $\pm i\sqrt{\langle\hat{D}\rangle}$.

Key words and phrases: models of quantum measurements, perturbation of discrete spectrum, complex eigenvalues, operator pencils

1. Introduction

The Kuryshkin–Wodkiewicz quantum mechanics [1] is an example of a dissipative quantum system. The quantum part of the measuring device is the

'environment of an open quantum system'. In the process of quantum measurement, an open quantum system interacts with its 'environment'. We study the result of this interaction [2]–[12]. Therefore, wave vectors must have a finite lifetime, inversely proportional to the imaginary part of eigenvalues.

In this version of quantum mechanics, the von Neumann quantization rule was abandoned and observable quantities are assigned to pseudo-differential operators, not necessarily self-adjoint. Therefore, the appearance of the imaginary part of the eigenvalues is not surprising. However, our studies of hydrogen-like atoms have shown that the operator corresponding to the Hamiltonian is essentially self-adjoint, so its discrete spectrum turned out to be real [13], [14].

This is quite surprising, since the von Neumann rule can be derived from general considerations, if we assume that the relation between the quantities $A = g(B)$ is inherited by their operators $\widehat{A} = g(\widehat{B})$ [15, P. 36]. Violation of this property inevitably means that the Kuryshkin–Wodkiewicz theory must be interpreted within the framework of the measurement theory and imaginary eigenvalues must appear. In this paper, we point out a spectral problem that has properties that are correct from this point of view.

2. Quantization in Kuryshkin–Wodkiewicz mechanics

Consider a Hamiltonian system with coordinates $q \in \mathbb{R}^n$, momenta $q \in \mathbb{R}^n$, and Hamiltonian H. We will assume that the Hamiltonian and all observables considered below belong to a commutative ring $\mathcal{M}$, for example, to the polynomial ring $\mathbb{R}[p, q]$ or the ring $C^\infty(\mathbb{R}^n)[p]$.

In classical statistical mechanics, the state of the system is described by the distribution function f, in quantum mechanics by the wave function $\psi \in L^2(\mathbb{R}^n)$. In statistical mechanics, the mean value of the observable quantity $A \in \mathcal{M}$ is given by

$$\langle A \rangle = \iint\limits_{\mathbb{R}^{2n}} A(p, q) f dp dq,$$

and in quantum mechanics by the expression

$$\langle A \rangle = \int\limits_{\mathbb{R}^n} \psi^*(q) \widehat{A} \psi(q) dq,$$

where $\widehat{A}$ is the operator corresponding to the observable A. In 1966, Cohen [16] proved that these two equalities for the mean cannot be combined in one theory, if it is assumed that the density takes strictly positive values, and the transition from mechanical quantities to operators is carried out according to the von Neumann rule.

However, if this rule of 'quantization' of mechanical quantities is abandoned, then it is possible to construct a version of quantum mechanics in which the average can be calculated by both formulas and the density takes positive values. Instead of the von Neumann rule, this theory uses a more complicated

mapping of the commutative ring $\mathcal{M}$ into the ring of linear operators on the space $L^2(\mathbb{R}^n)$: $\hat{O} : \mathcal{M} \to L(L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n))$.

This correspondence does not satisfy the Neumann rule, i.e., generally speaking, $\hat{O}(A) \neq A(\hat{p}, \hat{q})$, but it is linear, namely: for any $A, B \in \mathcal{M}$ and any $k \in \mathbb{C}$

1. $\hat{O}(A + B) = \hat{O}(A) + \hat{O}(B)$,
2. $\hat{O}(kA) = k\hat{O}(A)$,
3. $\hat{O}(0) = 0$,
4. $\hat{O}(1) = \hat{E}$.

In the early 1970s, V. V. Kuryshkin [1] not only proved the existence of such mappings, but also proposed an explicit construction for them. In doing so, it was necessary to extend the class of operators, in which the mapping $\hat{O}$ takes value, from the class of self-adjoint differential operators to a non-commutative ring of non-self-adjoint pseudo-differential operators. The resulting new version of quantum mechanics was called Kuryshkin–Wodkiewicz mechanics.

It turned out that 'perturbed operators' satisfy a certain condition for the proximity of the new quantization rule to the von Neumann rule:

$$\hat{O}(A) = A(\hat{p}, \hat{q}) + \hat{V},$$

where the addition of $\hat{V}$ to the standard quantization rule is an operator compact in the sense of Jorgens [17]. Therefore, the lower bounds of the essential spectra of the operators $\hat{O}(A)$ and $A(\hat{p}, \hat{q})$, as well as the points of the discrete spectra of these operators, may not coincide, but the structure of the spectrum is preserved: the points of the discrete spectrum lie below the continuous spectrum [18].

For what follows, the explicit form of the mapping $\hat{O}$ is not important. For hydrogen-like atoms, we explicitly computed $\hat{O}(p_i)$ and $\hat{O}(g)$ for any function g of coordinates q [14]. It turned out that in all these cases self-adjoint operators are obtained. This implies, in particular, that the spectrum of the operator $\hat{O}(H)$ consists of a continuous part, which coincides with the spectrum $H(\hat{p}, \hat{q})$, found in standard quantum mechanics, below the lower boundary of which lie the discrete spectrum points, which are slightly different from the points of the discrete spectrum of the operator $H(\hat{p}, \hat{q})$. However, all these points are real due to the self-adjointness of the operator $\hat{O}(H)$.

3. Spectral problem for a quadratic pencil

Let $A \in \mathcal{M}$ be an arbitrary observable. For brevity, we take

$$\hat{O}(A) = \hat{A}, \quad \hat{O}(A^2) = \hat{A}^2 + \hat{D}.$$

If the von Neumann rule is not satisfied, then two eigenvalue problems arise here:

1) classic problem

$$\hat{O}(A-\lambda)\psi = 0$$

or

$$\hat{A}\psi = \lambda\psi;$$

2) eigenvalue problem for a quadratic operator pencil

$$\hat{O}((A-\lambda)^2)\psi = 0$$

or

$$(\hat{A}-\lambda\hat{E})^2\psi + \hat{D}\psi = 0.$$

In standard quantum mechanics, $\hat{D} = 0$ and these problems are indistinguishable. The meaning of the first one has been discussed many times, but the second problem has a clear meaning. Expression

$$(\psi, O((A-\lambda)^2)\psi) = \langle (A-\lambda)^2 \rangle$$

is the mean square deviation of the observable value A from the value λ for the system in the ψ state. In mechanics with a positive distribution function, which is the Kuryshkin–Wodkiewicz mechanics, this value coincides with

$$\langle (A-\lambda)^2 \rangle = \iint\limits_{\mathbb{R}^{2n}} (A(p,q)-\lambda)^2 f dp dq$$

and therefore is non-negative. The same is true in standard quantum mechanics, but for a different reason:

$$\langle (A-\lambda)^2 \rangle = \int\limits_{\mathbb{R}} (\mu-\lambda)^2 d(\psi, \hat{\sigma}_\mu \psi) \geqslant 0.$$

If we assume that $\hat{D}$ is small, then the eigenvalues of these spectral problems are close to each other. Let us study this circumstance in more detail.

But first, we note that in [19] we were looking for the ψ states that provide a minimum to this expression for fixed values of the parameter λ, for which we took the eigenvalues of the operator $\hat{A}$. It turned out that the minimum values are nonzero, that is, there is some nonzero variance. However, the problem can be formulated differently: to find the values of λ and the states ψ, at which the mean square deviation of the observed value A from λ is minimal. On the eigenfunctions of the pencil $\hat{O}((A-\lambda)^2)$, this standard deviation is zero, therefore, on the pencil eigenfunctions, the mean square deviation of the observable A from the eigenvalue λ has a minimum, i.e., zero value. Thus, we can observe the value A in 'pure' states corresponding to the eigenfunctions of the operator $\hat{A}$, or in 'pure' states that provide zero root-mean-square deviation A from some value other than $\langle A \rangle$.

We have already used perturbation theory [19] to find states with minimal variance, but now we will apply it to finding eigenfunctions of a quadratic pencil.

4. Spectrum of a quadratic pencil

Let us introduce a small parameter ε and consider the problem

$$(\hat{A} - \lambda\hat{E})^2\psi + \varepsilon\hat{D}\psi = 0. \tag{1}$$

Let λ_0 be a single eigenvalue of the operator $\hat{A}$, and ψ_0 be the normalized eigenfunction corresponding to it. Let us study the eigenvalues of a quadratic pencil lying in a small neighborhood of this eigenvalue.

If the space under consideration is finite-dimensional, then all eigenvalues are roots of the determinant

$$\det\Big((\hat{A} - \lambda\hat{E})^2 + \varepsilon\hat{D}\Big) = 0.$$

In a neighborhood of the point $(\lambda, \varepsilon) = (\lambda_0, 0)$ the determinant

$$\det\Big(\hat{A} - \lambda\hat{E}\Big)^2 = \det\Big(\hat{A} - \lambda\hat{E}\Big)^2$$

has a zero of multiplicity 2, so

$$\det\Big((\hat{A} - \lambda\hat{E})^2 + \varepsilon\hat{D}\Big) = a(\lambda_0 - \lambda)^2 + \varepsilon b + \dots .$$

As is known from the theory of uniformization of curves [20], the curve

$$a(\lambda_0 - \lambda)^2 + \varepsilon b + \dots = 0$$

in the plane $\lambda\varepsilon$ has a node at the point $(\lambda_0, 0)$ through which two arcs of this curve pass, which can be represented as two Puiseux series:

$$\lambda - \lambda_0 = \pm c\sqrt{\varepsilon} + \dots .$$

Thus, in the vicinity of a single eigenvalue of the operator $\hat{A}$ there are two eigenvalues of the quadratic pencil:

$$\lambda = \lambda_0 \pm \lambda_1\sqrt{\varepsilon} + \dots . \tag{2}$$

This can be justified in the case of infinite-dimensional spaces, for completely continuous operators $\hat{A}$, $\hat{D}$ this can be done using the well-known results of the perturbation theory of multiple eigenvalues [21]. Of course, in quantum theory, the operator $\hat{A}$ is pseudo-differential, and the question requires additional study. For the time being, we assume without further justification that the formally developed perturbation theory can be justified for this class of operators as well.

To find the first coefficient in the expansion, as in regular perturbation theory, we multiply (1) by ψ_0:

$$(\psi_0, (\hat{A} - \lambda\hat{E})^2\psi) = -\varepsilon(\psi_0, \hat{D}\psi). \tag{3}$$

Since the operator A is self-adjoint, we have

$$(\psi_0, (\hat{A} - \lambda\hat{E})^2\psi) = ((\hat{A} - \lambda^*\hat{E})^2\psi_0, \psi) = (\lambda_0 - \lambda)^2(\psi_0, \psi) = \varepsilon\lambda_1^2(\psi_0, \psi).$$

Substituting this expression into (3) and reducing by ε, we get

$$\lambda_1^2(\psi_0, \psi) = -(\psi_0, \hat{D}\psi).$$

Hence, in the limit $\varepsilon \to 0$, we have $\lambda_1^2 = -(\psi_0, \hat{D}\psi_0)$.

Substituting this expression into (2) and setting $\varepsilon = 1$, we have: in the neighborhood of eigenvalue λ_0 of the operator $\hat{A}$ there are two eigenvalues of the quadratic pencil $\hat{O}((A - \lambda)^2)$, namely $\lambda = \lambda_0 \pm i\sqrt{(\psi_0, \hat{D}\psi_0)} + ...$ where $\hat{D} = \hat{O}(A^2) - \hat{A}^2$.

5. Conclusion and discussion

Let us now discuss the physical meaning of the resulting splitting of the eigenvalue of the operator $\hat{O}(A)$. The standard deviation of the observed value A from the value λ for a system in the ψ state is given by

$$\left(\psi, \hat{O}((A - \lambda)^2)\psi\right) = \langle(A - \lambda)^2\rangle.$$

This expression is non-negative both in standard quantum mechanics and in Kuryshkin–Wodkiewicz mechanics. It reaches zero on the eigenvectors of the quadratic pencil $\hat{O}((A - \lambda)^2)$.

In standard quantum mechanics

$$\hat{O}((A - \lambda)^2) = (\hat{A} - \lambda)^2$$

and therefore the eigenvectors of the pencil coincide with the eigenvectors of the operator $\hat{A}$. Therefore, the minimum standard deviation will be on those values of λ that are eigenvalues of the operator $\hat{A}$. They are traditionally considered as observed values of A.

In the mechanics of Kuryshkin–Wodkiewicz

$$\hat{O}((A - \lambda)^2) = (\hat{A} - \lambda)^2 + \hat{D}$$

and, as we just found out, the minimum standard deviation will be at those values of λ that differ from the eigenvalues λ_n of the operator $\hat{A}$ by $\pm i\sqrt{(\psi_n, \hat{D}\psi_n)}$.

Thus, the observed values of A will slightly differ from the eigenvalues of the operator $\hat{A}$. If $\langle\hat{D}\rangle > 0$, then this difference will manifest itself in the appearance of an imaginary additive, as one would expect in a dissipative quantum system. From this, two conclusions can be drawn.

Firstly, the transition to the root-mean-square deviation makes it possible to remove the difficulty with the reality of the spectrum of self-adjoint operators and obtain the expected dissipation in the Kuryshkin–Wodkiewicz mechanics.

Secondly, one of the two eigenvalues into which the eigenvalue $\hat{A}$ splits has the sign of the imaginary part corresponding to dissipation, and the second inevitably has a sign indicating antidissipation. We have already encountered a similar circumstance in the development of perturbation theory in the mathematical theory of waveguides [22], [23]: the spectral parameter λ should be considered as a point on the Riemann surface, only one sheet of which is physical, to which attention has been first drawn by V.P. Shestopalov [24]. In the case of Kuryshkin–Wodkiewicz mechanics, the eigenvalues of the operator $\hat{O}(A)$ are branch points on the Riemann surface, one of whose sheets describes a dissipative quantum system.

Acknowledgments

This work is supported by the Russian Science Foundation (grant no. 20-11-20257).

References

[1] V. V. Kuryshkin, "La mécanique quantique avec une fonction non-négative de distribution dans l'espace des phases", *Annales Henri Poincaré. Physique théorique*, vol. 17, no. 1, pp. 81–95, 1972.

[2] U. Weiss, *Quantum dissipative systems*, 4th ed. World Scientific, 2012. DOI: 10.1142/8334.

[3] H.-P. Breuer and F. Petruccione, *The theory of open quantum systems*. Oxford: Oxford University Press, 2002.

[4] V. E. Tarasov, *Quantum mechanics of non-Hamiltonian and dissipative systems*. Elsevier, 2008.

[5] M. Ahmadi, D. Jennings, and T. Rudolph, "The Wigner–Araki–Yanase theorem and the quantum resource theory of asymmetry", *New Journal of Physics*, vol. 15, no. 1, p. 013 057, 2013. DOI: 10.1088/1367-2630/15/1/013057.

[6] M. Ozawa, "Uncertainty relations for noise and disturbance in generalized quantum measurements", *Annals of Physics*, vol. 311, no. 2, pp. 350–416, 2004. DOI: 10.1016/j.aop.2003.12.012.

[7] M. Ozawa, "Universally valid reformulation of the Heisenberg uncertainty principle on noise and disturbance in measurement", *Physical Review A*, vol. 67, no. 4, p. 042 105, 2003. DOI: 10.1103/PhysRevA.67.042105.

[8] P. Busch and P. J. Lahti, "The standard model of quantum measurement theory: history and applications", *Foundations of Physics*, vol. 26, pp. 875–893, 1996. DOI: 10.1007/BF02148831.

[9] A. S. Holevo, *Statistical structure of quantum theory*. Springer, 2001.

[10] J. A. Wheeler and W. H. Zurek, *Quantum theory and measurement*. Princeton University Press, 1983.

[11] K. Jacobs, *Quantum measurement theory and its applications*. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139179027.

[12] W. H. Zurek, "Quantum theory and measurement in complexity", in *Complexity, Entropy And The Physics Of Information*, W. H. Zurek, Ed., CRC Press, 1990, ch. 6.

[13] A. V. Zorin and L. A. Sevastianov, "Hydrogen-like atom with nonnegative quantum distribution function", *Physics of Atomic Nuclei*, vol. 70, no. 4, pp. 792–799, 2007. DOI: 10.1134/S1063778807040229.

[14] A. V. Zorin, "Kuryshkin–Wodkiewicz quantum measurement model for alkaline metal atoms", *Discrete and Continuous Models and Applied Computational Science*, vol. 28, no. 3, pp. 274–288, 2020. DOI: 10.22363/2658-4670-2020-28-3-274-288.

[15] J. P. Rybakov and J. P. Terleckij, *Quantum mechanics [Kvantovaja mehanika]*. Moscow: RUDN, 1991, in Russian.

[16] L. Cohen, "Can quantum mechanics be formulated as a classical probability theory?", *Philosophy of Science*, vol. 33, no. 4, pp. 317–322, 1966. DOI: 10.1086/288104.

[17] K. Jorgens and J. Weidmann, *Spectral properties of Hamiltonian operators*. Berlin, Heidelberg, New York: Springer, 1973.

[18] A. V. Zorin and L. A. Sevastianov, "Bottom estimation method for the eigenvalues of the Hamilton differential operator in Kuryshkin quantum mechanics [Metod ocenok snizu dlja sobstvennyh znachenij differencial'nogo operatora Gamil'tona v kvantovoj mehanike Kuryshkina]", *RUDN Journal. Seria "Prikladnaja i komp'juternaja matematika"*, vol. 1, no. 1, pp. 134–144, 2002, in Russian.

[19] A. V. Zorin, L. A. Sevastianov, and N. P. Tretyakov, "States with Minimum Dispersion of Observables in Kuryshkin-Wodkiewicz Quantum Mechanics", *Lecture Notes in Computer Science*, vol. 11965, no. 4, pp. 508–519, 2019. DOI: 10.1007/978-3-030-36614-8_39.

[20] R. H. Nevanlinna, *Uniformisierung*. Berlin: Springer, 1953.

[21] T. Kato, *Perturbation theory for linear operators*. Berlin: Springer, 1966.

[22] A. N. Bogolyubov, M. D. Malykh, and A. G. Sveshnikov, "Instability of eigenvalues embedded in the waveguide's continuous spectrum with respect to perturbations of its filling", *Proceedings of the USSR Academy of Sciences*, vol. 385, no. 6, pp. 744–746, 2002.

[23] A. N. Bogolyubov and M. D. Malykh, "On the perturbation theory of spectral characteristics of waveguides", *Computational Mathematics and Mathematical Physics*, vol. 43, no. 7, pp. 1004–1015, 2003.

[24] V. P. Shestopalov, *Spectral theory and excitation of open structures [Spektral'naja teorija i vozbuzhdenie otkrytyh struktur]*. Moscow: Nauka, 1987, in Russian.

For citation:

A. V. Zorin, M. D. Malykh, L. A. Sevastianov, Complex eigenvalues in Kuryshkin–Wodkiewicz quantum mechanics, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 139–148. DOI: 10.22363/2658-4670-2022-30-2-139-148.

Information about the authors:

Zorin, Alexander V. — Candidate of Physical and Mathematical Sciences, Assistant Professor of Department of AppliedProbability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: `zorin-av@rudn.ru`, phone: +7(495)9550927, ORCID: https://orcid.org/00-0002-5721-4558, ResearcherID: AH-4011-2019, Scopus Author ID: 7193219091)

Malykh, Mikhail D. — Doctor of Physical and Mathematical Sciences, Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); (e-mail: `malykh-md@rudn.ru`, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0001-6541-6603, ResearcherID: P-8123-2016, Scopus Author ID: 6602318510)

Sevastianov, Leonid A. — Doctor of Physical and Mathematical Sciences, Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: `sevastianov-la@rudn.ru`, phone: +7(495)9522572, ORCID: https://orcid.org/0000-0002-1856-4643, ResearcherID: B-8497-2016, Scopus Author ID: 8783969400)

УДК 19.62:530.145:519.614
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-139-148

Комплексные собственные значения в квантовой механике Курышкина–Вудкевича

А. В. Зорин [1], **М. Д. Малых**[1, 2], **Л. А. Севастьянов**[1, 2]

[1] *Кафедра прикладной информатики и теории вероятностей*
Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Объединённый институт ядерных исследований*
ул. Жолио-Кюри, д. 6, Дубна, Московская обл., Россия, 141980

Аннотация. Рассматривается одна из возможных версий квантовой механики, известная как квантовая механика Курышкина–Вудкевича. В этой версии существует положительная квантовая функция распределения, но, в расплату за это, правило квантования фон Неймана заменено более сложным правилом, при котором наблюдаемой величине A ставится в соответствие псевдодифференциальный оператор $\hat{O}(A)$. Эта версия представляет собой пример диссипативной квантовой системы и поэтому ожидалось, что собственные значения гамильтониана должны иметь мнимые части. Однако точечный спектр гамильтониана водородоподобного атома в этой теории оказался вещественным. В настоящей статье мы предлагаем следующее объяснение этого парадокса. Традиционно принимают, что в некотором состоянии ψ величина A равна λ, если ψ — собственная функция оператора $\hat{O}(A)$. При этом дисперсия $\hat{O}((A-\lambda)^2)\psi$ равна нулю в стандартной версии квантовой механике, но не равна нулю в механике Курышкина. Поэтому можно рассмотреть такой спектр значений и соответствующих им состояний, при которых дисперсия $\hat{O}((A-\lambda)^2)$ равна нулю. В статье исследован спектр квадратичного пучка $\hat{O}(A^2) - 2\hat{O}(A)\lambda + \lambda^2\hat{E}$ методами теории возмущений в предположении малости дисперсии $\hat{D}(A) = \hat{O}(A^2) - \hat{O}(A)^2$ наблюдаемой A. Показано, что в окрестности вещественного собственного значения λ оператора $\hat{O}(A)$, имеется два собственных значения операторного пучка, которые в первом порядке теории возмущений различаются на величину $\pm i\sqrt{\langle\hat{D}\rangle}$.

Ключевые слова: модели квантовых измерений, возмущение дискретного спектра, комплексные собственные значения, пучки операторов

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 149–159
http://journals.rudn.ru/miph

UDC 535:535.3:681.7
DOI: 10.22363/2658-4670-2022-30-2-149-159

Investigation of adiabatic waveguide modes model for smoothly irregular integrated optical waveguides

Anton L. Sevastyanov

Higher School of Economics,
11, Pokrovsky Bulvar, Moscow, 109028, Russian Federation

(received: March 18, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. The model of adiabatic waveguide modes (AWMs) in a smoothly irregular integrated optical waveguide is studied. The model explicitly takes into account the dependence on the rapidly varying transverse coordinate and on the slowly varying horizontal coordinates. Equations are formulated for the strengths of the AWM fields in the approximations of zero and first order of smallness. The contributions of the first order of smallness introduce depolarization and complex values characteristic of leaky modes into the expressions of the AWM electromagnetic fields. A stable method is proposed for calculating the vertical distribution of the electromagnetic field of guided modes in regular multilayer waveguides, including those with a variable number of layers. A stable method for solving a nonlinear equation in partial derivatives of the first order (dispersion equation) for the thickness profile of a smoothly irregular integrated optical waveguide in models of adiabatic waveguide modes of zero and first orders of smallness is described. Stable regularized methods for calculating the AWM field strengths depending on vertical and horizontal coordinates are described. Within the framework of the listed matrix models, the same methods and algorithms for the approximate solution of problems arising in these models are used. Verification of approximate solutions of models of adiabatic waveguide modes of the first and zero orders is proposed; we compare them with the results obtained by other authors in the study of more crude models.

Key words and phrases: smoothly irregular thin–film dielectric waveguides, adiabatic waveguide modes, regularized methods for calculating field strengths

1. Introduction

The adiabatic waveguide propagation of optical radiation was previously described in optical fibers using the method of cross sections in the papers by B. Z. Katsenelenbaum [1], V. V. Shevchenko [2], M. V. Fedoruk [3], and in integrated optical waveguides using the method of adiabatic waveguide modes — in the papers by A. A. Egorov, L. A. Sevastyanov and their co-authors [4]–[6]. In the papers by A. L. Sevastyanov [7], [8], the model of adiabatic waveguide modes was substantiated.

It should be noted that in the last decade there has been an interest in the adiabatic waveguide propagation of electromagnetic radiation for the study

of coherent quantum effects in atomic, molecular or condensed matter systems. These effects are difficult to investigate because of dephasing effects or fast temporal dynamics. Optical Bloch oscillations [9], quantum-mechanical analogy of dynamic mode stabilization and radiation loss suppression [10], quantum enhancement and suppression of tunneling in directional optical couplers [11], [12], as well as Landau–Zener tunneling in coupled waveguides [13] can serve as optical models of coherent quantum effects. An interesting example is the three-level system with stimulated Raman adiabatic passage (STIRAP), which vividly illustrates counterintuitive quantum effects [14]–[19].

2. Model of adiabatic waveguide modes in a multilayer waveguide

Let us specify the class of integrated optical waveguides to be considered and the electromagnetic radiation propagating through them.

1. Electromagnetic radiation is polarized, monochromatic with a given wavelength $\lambda \in [\,380; 780\,]$, nm.
2. The thickness of the guiding layer of the base thin-film waveguide is comparable to the wavelength of the propagating monochromatic electromagnetic radiation $d \sim \lambda$.
3. The surface of the additional guiding layer ($x = h(y, z)$) satisfies the following restrictions: $\left|\dfrac{\partial h}{\partial y}, \dfrac{\partial h}{\partial z}\right| \ll \dfrac{hk_0}{2\pi}$, $\left|\dfrac{\Delta\varphi}{\nabla\varphi}\right| \ll \dfrac{k_0}{2\pi}$.
4. The integrated optical waveguide is a material medium consisting of dielectric subregions, which together fill the entire three-dimensional space.
5. The permittivities of the subregions are different and real-valued, and the permeability is everywhere equal to that of vacuum.
6. There are no external currents and charges. Therefore, in the absence of foreign currents and charges, the induced currents and charges are zero.
7. The Cartesian coordinate system is introduced as follows: the interfaces between the dielectric media of the basic three-layer waveguide are parallel to the yOz plane. The subdomains of the space corresponding to the cover and substrate layers are infinite; the additional guiding layers are asymptotically parallel to the yOz plane. Therefore, $\varepsilon = \varepsilon(x)$.

In Cartesian coordinates associated with the geometry of the substrate (or a three-layer planar dielectric waveguide underlying a smoothly irregular integrated optical waveguide), with the introduced restrictions taken into account, the Maxwell equations have the form

$$\begin{aligned}
&\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = \frac{\varepsilon}{c}\frac{\partial E_x}{\partial t}, \quad \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -\frac{\mu}{c}\frac{\partial H_x}{\partial t},\\
&\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = \frac{\varepsilon}{c}\frac{\partial E_y}{\partial t}, \quad \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -\frac{\mu}{c}\frac{\partial H_y}{\partial t},\\
&\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = \frac{\varepsilon}{c}\frac{\partial E_z}{\partial t}, \quad \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\frac{\mu}{c}\frac{\partial H_z}{\partial t}.
\end{aligned} \tag{1}$$

Note that variable x is fast, and variables y, z are slow with respect to the small dimensioned parameter $1/\omega$. The approximate solutions to the Maxwell equations (1) within the asymptotic method [20], [21], with the separation of slow and fast variables taken into account are sought in the form

$$\vec{E}(x,y,z,t) = \sum_{s=0}^{\infty} \frac{\vec{E}_s(x;y,z)}{(-i\omega)^{\gamma+s}} \exp\left\{i\omega t - ik_0\varphi(y,z)\right\}, \tag{2}$$

$$\vec{H}(x,y,z,t) = \sum_{s=0}^{\infty} \frac{\vec{H}_s(x;y,z)}{(-i\omega)^{\gamma+s}} \exp\left\{i\omega t - ik_0\varphi(y,z)\right\}. \tag{3}$$

Keeping in the solution (2), (3) the terms of the zero and first order of smallness leads to the model of adiabatic waveguide modes (AWMs) that describes the guided-wave propagation of a polarized optical radiation through irregular segments of smoothly irregular (multilayer) optical waveguides. In regular parts, the adiabatic waveguide modes become normal modes of a regular planar optical waveguide.

In the notation $\vec{E}_s(x;y,z)$, $\vec{H}_s(x;y,z)$, the separation by a semicolon means the following assumptions:

$$\left\|\frac{\partial\vec{E}_s\,(x;y,z)}{\partial y}\right\|, \left\|\frac{\partial\vec{E}_s(x;y,z)}{\partial z}\right\| \sim \frac{1}{\omega}\left\|\frac{\partial\vec{E}_s\,(x;y,z)}{\partial x}\right\| \tag{4}$$

and

$$\left\|\frac{\partial\vec{H}_s(x;y,z)}{\partial y}\right\|, \left\|\frac{\partial\vec{H}_s\,(x;y,z)}{\partial z}\right\| \sim \frac{1}{\omega}\left\|\frac{\partial\vec{H}_s(x;y,z)}{\partial x}\right\| \tag{5}$$

for each s, where $\|\|$ is the Hilbert norm of functions of x, and ω is the circular frequency of the propagating monochromatic electromagnetic radiation.

2.1. AWM model equations in the zero-order approximation

In Ref. [7] it was shown that the zero-order approximation (within the asymptotic approach) of the waveguide solution to the Maxwell equations is given by the following relations:

$$\begin{Bmatrix}\vec{E}(x,y,z,t)\\ \vec{H}(x,y,z,t)\end{Bmatrix} = \begin{Bmatrix}\vec{E}_0(x;y,z)\\ \vec{H}_0(x;y,z)\end{Bmatrix} \exp\left\{i\omega t - i\varphi(y,z)\right\}, \tag{6}$$

with

$$\varepsilon\frac{\partial E_0^y}{\partial x} = -ik_0\left(\frac{\partial\varphi}{\partial y}\right)\left(\frac{\partial\varphi}{\partial z}\right)H_0^y - ik_0\left(\varepsilon\mu - \left(\frac{\partial\varphi}{\partial y}\right)^2\right)H_0^z, \tag{7}$$

$$\varepsilon\frac{\partial E_0^z}{\partial x} = ik_0\left(\varepsilon\mu - \left(\frac{\partial\varphi}{\partial z}\right)^2\right)H_0^y + ik_0\left(\frac{\partial\varphi}{\partial z}\right)\left(\frac{\partial\varphi}{\partial y}\right)H_0^z, \tag{8}$$

$$\mu\frac{\partial H_0^y}{\partial x} = ik_0\left(\frac{\partial\varphi}{\partial y}\right)\left(\frac{\partial\varphi}{\partial z}\right)E_0^y + ik_0\left(\varepsilon\mu - \left(\frac{\partial\varphi}{\partial y}\right)^2\right)E_0^z, \tag{9}$$

$$\mu\frac{\partial H_0^z}{\partial x} = -ik_0\left(\varepsilon\mu - \left(\frac{\partial\varphi}{\partial z}\right)^2\right)E_0^y - ik_0\left(\frac{\partial\varphi}{\partial z}\right)\left(\frac{\partial\varphi}{\partial y}\right)E_0^z \tag{10}$$

and

$$E_0^x = -\frac{\partial\varphi}{\partial y}\frac{1}{\varepsilon}H_0^z + \frac{\partial\varphi}{\partial z}\frac{1}{\varepsilon}H_0^y, \tag{11}$$

$$H_0^x = \frac{\partial\varphi}{\partial y}\frac{1}{\mu}E_0^z - \frac{\partial\varphi}{\partial z}\frac{1}{\mu}E_0^y, \tag{12}$$

as well as

$$\left(\frac{\partial\varphi}{\partial y}(y,z)\right)^2 + \left(\frac{\partial\varphi}{\partial z}(y,z)\right)^2 = n_{\text{eff}}^2(y,z). \tag{13}$$

For a thin-film multilayer waveguide consisting of optically homogeneous layers, the conditions for matching the electromagnetic field at the interfaces between the media are valid, namely

$$\vec{n}\times\vec{E}^- + \vec{n}\times\vec{E}^+ = 0, \tag{14}$$

$$\vec{n}\times\vec{H}^- + \vec{n}\times\vec{H}^+ = 0. \tag{15}$$

In addition, the asymptotic conditions

$$E_y^0,\ E_z^0,\ H_y^0,\ H_z^0 \xrightarrow[x\to\pm\infty]{} 0 \tag{16}$$

are fulfilled.

The system of Eqs. (7)–(10), (16) for any fixed (y,z) defines the problem of finding eigenvalues $\left(\vec{\nabla}\varphi\right)_j^2(y,z)$ and eigenfunctions $\left(E_y^j, E_z^j, H_y^j, H_z^j\right)^T(y,z)$, normalized to unity:

$$\int\limits_{-\infty}^{\infty}\left|E_y^j\right|^2 dx = 1, \quad \int\limits_{-\infty}^{\infty}\left|H_y^j\right|^2 dx = 1. \tag{17}$$

2.2. AWM model equations in the first approximation

We continue to apply the approach based on the small parameter expansion and arrive at the system of equations in the first approximation of the method:

$$-\frac{\partial E_1^z}{\partial x} + \frac{ik_0}{\varepsilon}\frac{\partial\varphi}{\partial z}\left(\frac{\partial\varphi}{\partial y}H_1^z - \frac{\partial\varphi}{\partial z}H_1^y\right) + ik_0\mu H_1^y = \\ = i\omega\frac{\partial E_0^x}{\partial z} + \frac{i\omega}{\varepsilon}\frac{\partial\varphi}{\partial z}\left(\frac{\partial H_0^y}{\partial z} - \frac{\partial H_0^z}{\partial y}\right), \tag{18}$$

$$\frac{\partial E_1^y}{\partial x} - \frac{ik_0}{\varepsilon}\frac{\partial \varphi}{\partial y}\left(\frac{\partial \varphi}{\partial y}H_1^z - \frac{\partial \varphi}{\partial z}H_1^y\right) + ik_0\mu H_1^z = \\ = -i\omega\frac{\partial E_0^x}{\partial y} - \frac{i\omega}{\varepsilon}\frac{\partial \varphi}{\partial y}\left(\frac{\partial H_0^y}{\partial z} - \frac{\partial H_0^z}{\partial y}\right), \quad (19)$$

$$-\frac{\partial H_1^z}{\partial x} + \frac{ik_0}{\mu}\frac{\partial \varphi}{\partial z}\left(\frac{\partial \varphi}{\partial z}E_1^y - \frac{\partial \varphi}{\partial y}E_1^z\right) - ik_0\varepsilon E_1^y = \\ = i\omega\frac{\partial H_0^x}{\partial z} - \frac{i\omega}{\mu}\frac{\partial \varphi}{\partial z}\left(\frac{\partial E_0^y}{\partial z} - \frac{\partial E_0^z}{\partial y}\right), \quad (20)$$

$$\frac{\partial H_1^y}{\partial x} - \frac{ik_0}{\mu}\frac{\partial \varphi}{\partial y}\left(\frac{\partial \varphi}{\partial z}E_1^y - \frac{\partial \varphi}{\partial y}E_1^z\right) - ik_0\varepsilon E_1^z = \\ = -i\omega\frac{\partial H_0^x}{\partial y} + \frac{i\omega}{\mu}\frac{\partial \varphi}{\partial y}\left(\frac{\partial E_0^y}{\partial z} - \frac{\partial E_0^z}{\partial y}\right), \quad (21)$$

$$E_1^x + \frac{1}{\varepsilon}\left(\frac{\partial \varphi}{\partial y}H_1^z - \frac{\partial \varphi}{\partial z}H_1^y\right) = \frac{1}{\varepsilon}\frac{\omega}{k_0}\left(\frac{\partial H_0^y}{\partial z} - \frac{\partial H_0^z}{\partial y}\right), \quad (22)$$

$$H_1^x + \frac{1}{\mu}\left(\frac{\partial \varphi}{\partial z}E_1^y - \frac{\partial \varphi}{\partial y}E_1^z\right) = -\frac{1}{\mu}\frac{\omega}{k_0}\left(\frac{\partial E_0^y}{\partial z} - \frac{\partial E_0^z}{\partial y}\right). \quad (23)$$

The system of zero order equations (7)–(12) coincides with the system of equations (18)–(23), if in the latter we put zero into the right-hand sides (the contributions with zero-order quantities).

Substituting the solutions of system (7)–(12) into the right-hand sides of equations (18)–(23) leads to the following form of expressions for electromagnetic fields in the first (plus zero) approximation

$$\vec{E}(x;y,z) = \vec{E}_0(x;y,z) + \frac{i}{\omega}\vec{E}_1(x;y,z),$$

$$\vec{H}(x;y,z) = \vec{H}_0(x;y,z) + \frac{i}{\omega}\vec{H}_1(x;y,z).$$

These fields are necessarily complex-valued. Thus, the contributions of the first order of smallness introduce into the expressions for the AWM electromagnetic fields the characteristic features of leaky modes.

3. Implementation of numerical experiment

In Ref. [22], an hierarchy of mathematical models for the adiabatic waveguide propagation of optical radiation in integrated optical waveguides was proposed. The AWM model consists in representing the electromagnetic field in the form (6). The dependences of the field strengths on the fast variable have the form (7)–(12) in the zero approximation and (18)–(23) in the

first approximation. Of course, the rigging conditions (13)–(17) of the AWM mathematical model are assumed to be fulfilled.

3.1. Algorithm for calculating the AWM electromagnetic field

A. Stage 1: reconstructing the dependence of the AWM electromagnetic field on the fast variable at fixed values of the slow variables

1. Solve the system (7)–(12) for $\vec{E}^0$, $\vec{H}^0$ describing the AWM model in the zero order of smallness in $1/\omega$, rigged with (6), (18)–(23) using the method, asymptotic with respect to δ, to obtain systems for contributions of different orders of smallness with respect to δ.
2. Solve the system (13)–(17) for $\vec{E}^1$, $\vec{H}^1$ describing the AWM model in the first order of smallness in $1/\omega$, rigged with (6), (18)–(23) using the method, asymptotic with respect to δ, to obtain systems for contributions of different orders of smallness with respect to δ.

B. Stage 2: reconstructing the dependence of the AWM electromagnetic field on the slow variables.

In Ref. [7] it is shown how the general solutions of the system of ODEs (7)–(12) and (13)–(17), represented in the form of expansion in the fundamental system of solutions with indefinite coefficients $\left(\vec{A}, \vec{B}\right)^T$, can be reduced to a homogeneous system of linear algebraic equations (SLAE) with respect to these indefinite coefficients using the conditions (14)–(16).

3. Implement stable methods of approximate solutions of the homogeneous SLAE

$$\hat{M}^0\left[(z,y), h(z,y), \varphi(z,y), \vec{\nabla}\varphi(z,y)\right]\left(\vec{A}^0(z,y), \vec{B}^0(z,y)\right)^T = \left(\vec{0}, \vec{0}\right)^T, \quad (24)$$

satisfying the conditions

$$\det\left\{\hat{M}^0\right\}\left[(z,y), h(z,y), \varphi(z,y), \vec{\nabla}\varphi(z,y)\right] = 0. \quad (25)$$

4. Implement stable methods of approximate solutions of the homogeneous SLAE

$$\hat{M}^1\left[(z,y), h(z,y), \varphi(z,y), \vec{\nabla}\varphi(z,y)\right]\left(\vec{A}^1(z,y), \vec{B}^1(z,y)\right)^T = \left(\vec{0}, \vec{0}\right)^T \quad (26)$$

satisfying the conditions

$$\det\left\{\hat{M}^1\right\}\left[(z,y), h(z,y), \varphi(z,y), \vec{\nabla}\varphi(z,y)\right] = 0. \quad (27)$$

In both cases, the solution for the field strengths depending on the fast variable x for a fixed value of the slow variables y, z makes it possible, using the rigging (6), (18)–(23), to find the dependence of the AWM electromagnetic field for all values of the slow variables (see, e.g., Ref. [8]).

Homogeneous systems of linear algebraic equations (24) and (26) are uniquely solvable under conditions (25) and (27). In both cases, these equations with respect to the derivative $\vec{\nabla}\varphi(z,y)$ are partial differential equations

of the form

$$F^0\left(\vec{\nabla}\varphi(z,y);\, h(z,y), \vec{\nabla}h(z,y)\right) = 0 \tag{28}$$

and

$$F^1\left(\vec{\nabla}\varphi(z,y);\, h(z,y), \vec{\nabla}h(z,y)\right) = 0. \tag{29}$$

5. Solve Eqs. (28) and (29) numeric-symbolically using the Cauchy method (see, e.g. [23], [24]).
6. For each $\vec{\nabla}\varphi(z,y)$ calculate $\left(\vec{A}^0\left(z,y,\vec{\nabla}\varphi(z,y)\right), \vec{B}^0\left(z,y,\vec{\nabla}\varphi(z,y)\right)\right)^T$ using the Tikhonov regularization method, which consists in minimizing the Nelder–Mead functional:

$$F^0(\beta) = \left\|\hat{M}^0\left[(z,y), h(z,y), \varphi(z,y), \vec{\nabla}\varphi(z,y)\right]\left(\vec{A}^0(z,y), \vec{B}^0(z,y)\right)^T\right\|^2 +$$

$$+\alpha\left\|\left(\left(\vec{A}^0(z,y) - \vec{A}_0(z-\Delta z, y-\Delta y)\right), \left(\vec{B}^0(z,y) - \vec{B}_0(z-\Delta z, y-\Delta y)\right)\right)^T\right\|^2.$$

C. Stage 3: verifying the obtained numerical results and AWM models of the first and zero orders of smallness.

The validation of the asymptotic method of constructing AWM models is carried out by comparing solutions $\vec{E}^1,\ \vec{H}^1$ and $\vec{E}^0,\ \vec{H}^0$.

The formulation of the third condition from the set of conditions 1–7 implicitly implies the presence of the second small parameter $\delta \equiv \max\limits_{y,z} \dfrac{|\Delta\varphi|}{k_0\left|\vec{\nabla}\varphi\right|} \ll 1$ (see the beginning of the first section).

To verify the obtained approximate solutions of the zero-order model of adiabatic modes, we compare them with the results obtained by other authors using more crude models:

— matrix model of adiabatic modes in the approximation of horizontal boundary conditions (a stepped set of plates for a Luneburg thin-film generalized waveguide lens)

Such configurations are impossible in optical fibers and can be implemented in the case of adiabatic waveguide propagation of a nonparallel (converging or diverging) 2D beam of rays, normal to a nonplanar (2D) wave front.

— matrix model of comparison waveguides (passing to the horizontal boundary conditions + replacement $\beta_y \to 0$, $\beta_z \to \beta$).

Thus, three levels of making the AWM model cruder were used.

4. Discussion and conclusion

In the paper, we consider three levels of making the AWM model cruder:

— replacing the first-order AWM model with the zero-order one;
— replacing the tangential boundary conditions with the horizontal ones — the matrix model still having no name;
— replacing the tangential boundary conditions with the horizontal ones and $\beta_y \to 0$, $\beta_z \to \beta$ — the matrix model of comparison waveguides.

Two latter approximations have been used by other authors.

Within the listed matrix models, similar methods and algorithms are used for the approximate solution of problems, arising in the models. The method of studying the matrix model of adiabatic waveguide modes in the zero and first approximation of a smoothly irregular multilayer integrated optical waveguide is proposed for the first time. It allows to grade the crudeness of the approximate models used by other authors and approximate solutions in the adiabatic mode models of different order of smallness.

References

[1] B. Z. Katsenelenbaum, *Theory of irregular waveguides with slowly varying parameters [Teoriya neregulyarnyh volnovodov s medlenno menyayushchimisya parametrami]*. Moscow: Akad. Nauk SSSR, 1961, in Russian.

[2] V. V. Shevchenko, *Continuous transitions in open waveguides [Plavnyye perekhody v otkrytykh volnovodakh]*. Moscow: Nauka, 1969, in Russian.

[3] M. V. Fedoryuk, "Justification of the method of cross-sections for an acoustic waveguide with inhomogeneous filling", *USSR Computational Mathematics and Mathematical Physics*, vol. 13, no. 1, pp. 162–173, 1973. DOI: 10.1016/0041-5553(74)90012-3.

[4] A. A. Egorov and L. A. Sevast'yanov, "Structure of modes of a smoothly irregular integrated optical four–layer three–dimensional waveguide", *Quantum Electronics*, vol. 39, no. 6, pp. 566–574, 2009. DOI: 10.1070/QE2009v039n06ABEH013966.

[5] A. A. Egorov *et al.*, "Simulation of guided modes (eigenmodes) and synthesis of a thin-film generalised waveguide Luneburg lens in the zero-order vector approximation", *Quantum Electronics*, vol. 40, no. 9, pp. 830–836, 2010. DOI: 10.1070/QE2010v040n09ABEH014332.

[6] A. A. Egorov, L. A. Sevastianov, and A. L. Sevastianov, "Method of adiabatic modes in research of smoothly irregular integrated optical waveguides: zero approximation", *Quantum Electronics*, vol. 44, no. 2, pp. 167–173, 2014. DOI: 10.1070/QE2014v044n02ABEH015303.

[7] A. L. Sevastianov, "Asymptotic method for constructing a model of adiabatic guided modes of smoothly irregular integrated optical waveguides", *Discrete and Continuous Models and Applied Computational Science*, vol. 20, no. 3, pp. 252–273, 2020. DOI: 10.22363/2658-4670-2020-28-3-252-273.

[8] A. L. Sevastianov, "Single–mode propagation of adiabatic guided modes in smoothly irregular integral optical waveguides", *Discrete and Continuous Models and Applied Computational Science*, vol. 28, no. 4, pp. 361–377, 2020. DOI: 10.22363/2658-4670-2020-28-4-361-377.

[9] G. Lenz, I. Talanina, and C. M. de Sterke, "Bloch oscillations in an array of curved optical waveguides", *Physical Review Letters*, vol. 83, no. 5, pp. 963–966, 1999. DOI: 10.1103/PhysRevLett.83.963.

[10] S. Longhi, D. Janner, M. Marano, and P. Laporta, "Quantum–mechanical analogy of beam propagation in waveguides with a bent axis: dynamic–mode stabilization and radiation–loss suppression", *Physical Review E*, vol. 67, no. 3, p. 036 601, 2003. DOI: 10.1103/PhysRevE.67.036601.

[11] I. Vorobeichik *et al.*, "Electromagnetic realization of orders-of-magnitude tunneling enhancement in a double well system", *Physical Review Letters*, vol. 90, p. 176 806, 17 2003. DOI: 10.1103/PhysRevLett.90.176806.

[12] S. Longhi, "Coherent destruction of tunneling in waveguide directional couplers", *Physical Review A*, vol. 71, p. 065 801, 6 2005. DOI: 10.1103/PhysRevA.71.065801.

[13] R. Khomeriki and S. Ruffo, "Nonadiabatic Landau–Zener tunneling in waveguide arrays with a step in the refractive index", *Physical Review Letters*, vol. 94, p. 113 904, 11 2005. DOI: 10.1103/PhysRevLett.94.113904.

[14] K. Bergmann, H. Theuer, and B. W. Shore, "Coherent population transfer among quantum states of atoms and molecules", *Reviews of Modern Physics*, vol. 70, pp. 1003–1025, 3 1998. DOI: 10.1103/RevModPhys.70.1003.

[15] F. T. Hioe and J. H. Eberly, "N–Level coherence vector and higher conservation laws in quantum optics and quantum mechanics", *Physical Review Letters*, vol. 47, pp. 838–841, 12 1981. DOI: 10.1103/PhysRevLett.47.838.

[16] J. Oreg, F. T. Hioe, and J. H. Eberly, "Adiabatic following in multilevel systems", *Physical Review A*, vol. 29, pp. 690–697, 2 1984. DOI: 10.1103/PhysRevA.29.690.

[17] C. E. Carroll and F. T. Hioe, "Three–state systems driven by resonant optical pulses of different shapes", *Journal of the Optical Society of America B: Optical Physics*, vol. 5, no. 6, pp. 1335–1340, 1988. DOI: 10.1364/JOSAB.5.001335.

[18] J. Oreg, K. Bergmann, B. W. Shore, and S. Rosenwaks, "Population transfer with delayed pulses in four-state systems", *Physical Review A*, vol. 45, pp. 4888–4896, 7 1992. DOI: 10.1103/PhysRevA.45.4888.

[19] N. V. Vitanov and S. Stenholm, "Analytic properties and effective two-level problems in stimulated Raman adiabatic passage", *Physical Review A*, vol. 55, pp. 648–660, 1 1997. DOI: 10.1103/PhysRevA.55.648.

[20] V. M. Babich and V. S. Buldyrev, *Asymptotic methods in short–wavelength diffraction theory (Alpha Science Series on Wave Phenomena)*, English. Harrow, UK: Alpha Science International, 2009.

[21] Y. A. Kravtsov and Y. I. Orlov, *Geometrical optics of inhomogeneous media*. Berlin: Springer–Verlag, 1990.

[22] A. L. Sevastyanov, "Single–mode waveguide spread of light in a smooth irregular integral optical waveguide [Komp'yuternoe modelirovanie polej napravlyaemyh mod tonkoplenochnoj obobshchennoj volnovodnoj linzy Lyuneberga]", in Russian, Ph.D. dissertation, Peoples' Friendship University of Russia, Moscow, 2010.

[23] M. D. Malykh, "On integration of the first order differential equations in a finite terms", *Journal of Physics: Conference Series*, vol. 788, p. 012 026, 2017. DOI: 10.1088/1742-6596/788/1/012026.

[24] A. D. Polyanin and V. E. Nazaikinskii, *Handbook of linear partial differential equations for engineers and scientists*, 2nd ed. Boca Raton, London: CRC Press, 2016. DOI: 10.1201/b19056.

For citation:

A. L. Sevastyanov, Investigation of adiabatic waveguide modes model for smoothly irregular integrated optical waveguides, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 149–159. DOI: 10.22363/2658-4670-2022-30-2-149-159.

Information about the authors:

Sevastyanov, Anton L. — PhD in Physical and Mathematical Sciences, Deputy head of department: Department of Digitalization of Education (e-mail: alsevastyanov@gmail.com, phone: +7(495)772-95-90 (28571), ORCID: https://orcid.org/0000-0002-0280-485X)

УДК 535:535.3:681.7
DOI: 10.22363/2658-4670-2022-30-2-149-159

Исследование модели адиабатических волноводных мод для плавно-нерегулярных интегрально-оптических волноводов

А. Л. Севастьянов

Национальный исследовательский университет «Высшая школа экономики», Покровский бульвар, д. 11, Москва, 109028, Россия

Аннотация. Проведено исследование модели адиабатических волноводных мод плавно-нерегулярного интегрально-оптического волновода. В модели явно учтена зависимость от быстропеременной поперечной координаты и от медленно-переменных горизонтальных координат. Сформулированы уравнения для напряженностей полей АВМ в приближениях нулевого и первого порядка малости. Вклады первого порядка малости вносят в выражения электромагнитных полей АВМ деполяризацию и комлекснозначность, т.е. характерные черты вытекающих мод. Предложен устойчивый метод вычисления вертикального распределения электромагнитного поля направляемых мод регулярных многослойных волноводов, в том числе с переменным числом слоев. Описан устойчивый метод решения нелинейного уравнения в частных производных первого порядка (дисперсионного уравнения) для профиля толщины плавно-нерегулярного интегрально-оптического волновода в моделях адиабатических волноводных мод нулевого и первого порядков малости. Описаны устойчивые регуляризованные методы вычисления напряженностей полей АВМ в зависимости от вертикальных и горизонтальных координат. В рамках перечисленных матричных моделей используются одинаковые методы и алгоритмы приближенного решения задач, возникающих в этих моделях. Предложена верификация приближенных решений моделей адиабатических волноводных мод первого и нулевого порядков; проведено сравнение их с результатами других авторов, полученных при исследовании более грубых моделей.

Ключевые слова: модели квантовых измерений, возмущение дискретного спектра, комплексные собственные значения, пучки операторов

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2022, 30 (2) 160–182
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-160-182

Analysis of queuing systems with threshold renovation mechanism and inverse service discipline

Ivan S. Zaryadov[1,2], **Hilquias C. C. Viana**[1], **Tatiana A. Milovanova**[1]

[1] *Peoples' Friendship University of Russia (RUDN University), 6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation*
[2] *Institute of Informatics Problems, FRC CSC RAS, 44-2, Vavilova St., Moscow 119333, Russian Federation*

(received: February 22, 2022; revised: April 18, 2022; accepted: April 19, 2022)

Abstract. The paper presents a study of three queuing systems with a threshold renovation mechanism and an inverse service discipline. In the model of the first type, the threshold value is only responsible for activating the renovation mechanism (the mechanism for probabilistic reset of claims). In the second model, the threshold value not only turns on the renovation mechanism, but also determines the boundaries of the area in the queue from which claims that have entered the system cannot be dropped. In the model of the third type (generalizing the previous two models), two threshold values are used: one to activate the mechanism for dropping requests, the second — to set a safe zone in the queue. Based on the results obtained earlier, the main time-probabilistic characteristics of these models are presented. With the help of simulation modeling, the analysis and comparison of the behavior of the considered models were carried out.

Key words and phrases: queuing system, active queue management, renovation mechanism, threshold, time-probabilistic characteristics, GPSS modelling

1. Introduction

According to [1] the problem of congestion avoidance for communication networks does not have a satisfying solution, so the development and the analysis of new active queue management (AQM) algorithms appears to be the actual task for researches [2]–[13] and practitioners [14]–[24].

In this paper we will consider queuing systems with probabilistic renovation mechanism, which allows to adjust the number of packets in the system by dropping (resetting) them from the queue depending on the ratio of a certain control parameter with specified thresholds [25], [26] at the moment of the end of service on the device (server) [27]–[29] in contrast to standard RED algorithm, when a possible reset occurs at the time of the next packet arrival and the control parameter is an exponentially weighted average queue

length [30]–[34]. In our models the renovation mechanism uses one or two thresholds (which determine as the place in the buffer from which the packets are dropped, but also the place to which the reset of packets occurs).

The previous works devoted to the analysis of queuing systems with threshold based renovation are [35]–[38]. In [35], [36] some aspects of using the renovation mechanism (different types of renovation, definitions and brief overview were also given) with one or several thresholds as the mathematical models of active queue management mechanisms were considered. Some results of comparing classic RED algorithm with renovation mechanism were presented. In [37] two queuing models with threshold based renovation mechanism were presented: in the first model the threshold value is only responsible for activating the renovation mechanism (the mechanism for probabilistic reset of claims), in the second model the threshold value not only turns on the renovation mechanism, but also determines the boundaries of the area in the queue from which claims that have entered the system cannot be dropped. In [38] the queuing system with two threshold values (one to activate the mechanism for dropping requests, the second — to set a safe zone in the queue) for renovation mechanism was investigated. All three queuing systems have been studied for the service discipline FCFS (First Come First Served), and in this article we will present some results for the discipline LCFS (Last Come First Served). The study will again be carried out using embedded Markov chains. We will not consider in detail the derivation of the stationary distribution of the number of customers (which does not depend on the service discipline and presented in [37], [38]) and will focus only on the service (reset) probabilities and on time characteristics.

The structure of the article is following. In the section 2 the results for the queuing model, where the threshold value is only responsible for activating the renovation mechanism, are presented; the section 3 is devoted to the queuing model, in which the threshold value not only turns on the renovation mechanism, but also determines the boundaries of the area in the queue from which claims that have entered the system cannot be dropped. In section 4 the characteristics for the queuing system with two threshold values (one to activate the mechanism for dropping requests, the second — to set a safe zone in the queue) for renovation mechanism are presented. In section 5 the results of GPSS simulation are considered. The last section 6 concludes the paper with the short discussion.

2. The first model

Consider the $GI/M/1/\infty$ queuing system, shown in the figure 1, with the implemented renovation mechanism and a threshold value Q_1, which determines the boundary in the queue, starting from which the dropping of customers begins. If the current number of packets in the system $i \leqslant Q_1 + 1$ (the threshold value Q_1 is not been overcome), then none of the packets will be dropped from the queue. If the current number of packets in the system $i \geqslant Q_1 + 1$, then with probability q the packet finishing the service can drop all packets from the queue and leave the system, or with probability $p = 1 - q$ the serviced packet simply leaves the system.

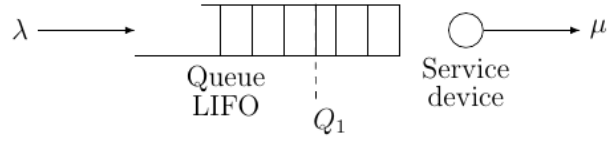


Figure 1. Queuing system model

2.1. The service probability and the loss probability for a received packet

Let $p^{(\text{loss})}$ be the probability that the packet received in the system will be dropped by renovation mechanism and let $p_i^{(\text{loss})}$ be the probability that a packet arriving and finding in the system exactly i packets will be dropped.

Let $p_i^{(\text{loss})}(x)$ be the probability that in a time less than x a packet that finds other i packets in the system will be dropped. Then:

$$p_i^{(\text{loss})} = \int_0^\infty p_{i,0}^{(\text{loss})}(x)dx,$$

where $p_{i,j}^{(\text{loss})}(x)$ is the probability that in time less than x the packet, before which there are i other packets in the queue and after which there are other j packets, will be dropped, $i, j \geqslant 0$.

Let $\tau_{i,j}^{(\text{loss})}(x)$ be the probability density functions and let $\rho_{i,j}^{(\text{loss})}(s)$ be the Laplace–Stieltjes transforms. Then:

$$\tau_{i,j}^{(\text{loss})}(x) = \left(p_{i,j}^{(\text{loss})}(x)\right)'_x, \quad \rho_{i,j}^{(\text{loss})}(s) = \int_0^\infty \tau_{i,j}^{(\text{loss})}(x)dx.$$

a) If $i + j + 1 \leqslant Q_1$ the threshold is not crossed, then:

$$\tau_{i,j}^{(\text{loss})}(x) = \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} dA(y) \tau_{i,j-k+1}^{(\text{loss})}(x-y).$$

b.1) If $i + j + 1 > Q_1$, $i + 1 \leqslant Q_1$, then:

$$\tau_{i,j}^{(\text{loss})}(x) = \sum_{k=1}^{\min(j,i+j+1-Q_1)} \frac{\mu^k x^{k-1}}{(k-1)!} e^{-\mu x} \cdot p^{k-1} \cdot q \cdot \overline{A}(x) + \\ + \int_0^y \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} \cdot p^{\min(k,i+1+j-Q_1)} dA(y) \tau_{i,j}^{(\text{loss})}(x-y).$$

b.2) If $i+j+1>Q_1$, $i+1>Q_1$, then:

$$\tau_{i,j}^{(\text{loss})}(x)=\sum_{k=1}^{j}\frac{\mu^k x^{k-1}}{(k-1)!}e^{-\mu x}\cdot p^{k-1}\cdot q\cdot\overline{A}(x)+ \\ +\int_0^y\sum_{k=0}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^k dA(y)\tau_{i,j-k+1}^{(\text{loss})}(x-y).$$

Then for the Laplace–Stieltjes transforms $\rho_{i,j}^{(\text{loss})}(s)$ we have:
a) If $i+j+1\leqslant Q_1$, then:

$$\rho_{i,j}^{(\text{loss})}(s)=\sum_{k=0}^{j}\frac{(-1)^k\mu^k}{k!}\alpha^{(k)}(\mu+s)\cdot\rho_{i,j-k+1}^{(\text{loss})}(s).$$

b.1) If $i+j+1>Q_1$, $i+1\leqslant Q_1$, then:

$$\rho_{i,j}^{(\text{loss})}(s)=\sum_{k=1}^{\min(j,i+1+j-Q_1)}\frac{(-1)^{k-1}\mu^k}{(k-1)!}\overline{\alpha}^{(k-1)}(\mu+s)\cdot p^{k-1}\cdot q+ \\ +\sum_{k=0}^{j}\frac{(-1)^k\mu^k}{k!}p^{\min(k,i+j+1-Q_1)}\alpha^{(k)}(\mu+s)\cdot\rho_{i,j-k+1}^{(\text{loss})}(s).$$

b.2) If $i+j+1>Q_1, i+1>Q_1$, then:

$$\rho_{i,j}^{(\text{loss})}(s)=\sum_{k=1}^{j}\frac{(-1)^{k-1}\mu^k}{(k-1)!}\overline{\alpha}^{(k-1)}(\mu+s)\cdot p^{k-1}\cdot q+ \\ +\sum_{k=0}^{j}\frac{(-1)^k\mu^k}{k!}p^k\alpha^{(k)}(\mu+s)\cdot\rho_{i,j-k+1}^{(\text{loss})}(s).$$

2.2. Time characteristics of the system

Let $W^{(\text{serv})}(x)$ and $W^{(\text{loss})}(x)$ be the distribution functions of the time spent in the queue by the served and dropped packets.

2.2.1. Time characteristics for a served packet

$W_{i,j}^{(\text{serv})}(x)$ — the intermediary distribution function of the time spent by the served packet in the queue, if there are i other packets in the queue before the considered one and there are j others after it. Then

$$W^{(\text{serv})}(x)=\left(\sum_{i=0}^{\infty}\pi_i W_{i,0}^{(\text{serv})}(x)\right)\cdot\frac{1}{p^{(\text{serv})}},$$

where steady-state probabilities π_i $(i \geqslant 0f)$ are defined in [37], [38]. For densities $w_{i,j}^{(\text{serv})}(x) = \left(W_{i,j}^{(\text{serv})}(x)\right)'$, we will consider several cases.

a) Consider the case when $i + j + 1 > Q_1$, $0 \leqslant i < Q_1$

$$w_{i,j}^{(\text{serv})}(x) = \frac{\mu^{j+1}x^j}{j!}e^{-\mu x}p_{i+1,j}^{(\text{serv})}\overline{A}(x) + \\ + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!}e^{-\mu y}p^{\min(k,j+i+1-Q_1)}dA(y)w_{i,j-k+1}^{(\text{serv})}(x-y),$$

$$p^{\min(k,j+i+1-Q_1)} = \begin{cases} p^k, k \leqslant j+i+1-Q_1, \\ p^{j+1+i-Q_1}, k > j+i-Q_1. \end{cases}$$

b) Let's move on to the case when $i \geqslant Q_1$

$$w_{i,j}^{(\text{serv})}(x) = \frac{\mu^{j+1}x^j}{j!}e^{-\mu x}p^j\overline{A}(x) + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!}e^{-\mu y}p^k dA(y)w_{i,j}^{(\text{serv})}(x-y).$$

If $i + j + 1 \geqslant Q_1$ the threshold is not crossed, then:

$$w_{i,j}^{(\text{serv})}(x) = \frac{\mu^{j+1}x^j}{j!}e^{-\mu x}\overline{A}(x) + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!}e^{-\mu y}dA(y)w_{i,j}^{(\text{serv})}(x-y).$$

The Laplase–Stieltjes transforms for derived densities.

If $i + j + 1 \leqslant Q_1$, then:

$$\omega_{i,j}^{(\text{serv})}(s) = \frac{(-1)^j\mu^{j+1}}{j!}\overline{\alpha}^{(j)}(\mu+s) + \sum_{k=0}^{j}\frac{(-1)^k\mu^k}{k!}\alpha^{(k)}(\mu+s)\omega_{i,j-k+1}^{(\text{serv})}(s),$$

$$\omega_{i,j}^{(\text{serv})}(s) = \int_0^\infty w_{i,j}^{(\text{serv})}(x)e^{-sx}dx \text{ — Laplace–Stieltjes transform.}$$

If $0 \leqslant i < Q_1$, but $i + j + 1 > Q_1$, then:

$$\omega_{i,j}^{(\text{serv})}(s) = \frac{(-1)^j\mu^{j+1}}{j!}\overline{\alpha}^{(j)}(\mu+s)\cdot p^{j+i+1-Q_1} + \\ + \sum_{k=0}^{j}\frac{(-1)^k\mu^k}{k!}\alpha^{(k)}(\mu+s)\cdot p^{\min(k,j+i+1-Q_1)}\cdot\omega_{i,j-k+1}^{(\text{serv})}(s).$$

If $i \geqslant Q_1$, then:

$$\omega_{i,j}^{(\text{serv})}(s) = \frac{(-1)^j \mu^{j+1}}{j!} \overline{\alpha}^{(j)}(\mu+s) \cdot p^j + \sum_{k=0}^{j} \frac{(-1)^k \mu^k}{k!} \alpha^{(k)}(\mu+s) \cdot p^k \cdot \omega_{i,j-k+1}^{(\text{serv})}(s).$$

2.2.2. Time characteristics for a dropped packet

$W_{i,j}^{(\text{loss})}(x)$ — the intermediary distribution function of the time spent by the dropped packet in the queue, if there are i other packets in the queue before the considered one and there are j others after it. Then

$$W^{(\text{loss})}(x) = \left(\sum_{i=0}^{\infty} \pi_i W_{i,0}^{(\text{loss})}(x) \right) \cdot \frac{1}{p^{(\text{loss})}}.$$

For densities $w_{i,j}^{(\text{loss})}(x) = \left(W_{i,j}^{(\text{loss})}(x) \right)'$, we also will consider several cases.

a) The first case is when $i+1+j \leqslant Q_1$, so the selected packet can be dropped only due to the reception of new packets in the system and overcoming the threshold value

$$w_{i,j}^{(\text{loss})}(x) = \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} dA(y) w_{i,j-k+1}^{(\text{loss})}(x-y).$$

b) for the second case, when $i + 1 + j > Q_1$, $(i + 1 \leqslant Q_1)$, several subcases should be considered:

b.1)

$$w_{i,j}^{(\text{loss})}(x) = \sum_{k=1}^{\min(i, i+1+j-Q_1)} \frac{\mu^k x^{k-1}}{(k-1)!} e^{-\mu x} \cdot p^{k-1} \cdot q \cdot \overline{A}(x) + \\ + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} \cdot p^{\min(k, i+1+j-Q_1)} dA(y) w_{i,j-k+1}^{(\text{loss})}(x-y).$$

b.2) If $i + 1 > Q_1$, then:

$$w_{i,j}^{(\text{loss})}(x) = \sum_{k=1}^{j} \frac{\mu^k x^{k-1}}{(k-1)!} e^{-\mu x} \cdot p^{k-1} \cdot q \cdot \overline{A}(x) + \\ + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} \cdot p^k dA(y) w_{i,j-k+1}^{(\text{loss})}(x-y).$$

The Laplase–Stieltjes transforms for derived densities.

a) For the case when $i + j + 1 \leqslant Q_1$ we have

$$\omega_{i,j}^{(\text{loss})}(s) = \sum_{k=0}^{j} \frac{(-1)^k \mu^k}{k!} \alpha^{(k)}(\mu + s) \cdot \omega_{i,j-k+1}^{(\text{loss})}(s).$$

b) For the case when $i + j + 1 > Q_1$, $i + 1 \leqslant Q_1$ we obtain:
b.1)

$$\omega_{i,j}^{(\text{loss})}(s) = \sum_{k=1}^{\min(j,i+1+j-Q_1)} \frac{(-1)^{k-1} \mu^k}{(k-1)!} \overline{\alpha}^{(k-1)}(\mu + s) \cdot p^{k-1} \cdot q + \\ + \sum_{k=0}^{j} \frac{(-1)^k \mu^k}{k!} p^{\min(k,i+j+1-Q_1)} \alpha^{(k)}(\mu + s) \cdot \omega_{i,j-k+1}^{(\text{loss})}(s).$$

b.2)

$$\omega_{i,j}^{(\text{loss})}(s) = \sum_{k=1}^{j} \frac{(-1)^{k-1} \mu^k}{(k-1)!} \overline{\alpha}^{(k-1)}(\mu + s) \cdot p^{k-1} \cdot q + \\ + \sum_{k=0}^{j} \frac{(-1)^k \mu^k}{k!} p^k \alpha^{(k)}(\mu + s) \cdot \omega_{i,j-k+1}^{(\text{loss})}(s).$$

3. The second model

The second queuing model is also $GI/M/1/\infty$ queuing system, shown in the figure 2, with the implemented renovation mechanism, but the threshold value Q_1 determines the boundary in the queue, starting from which the dropping of customers begins and also determines the safe zone from where packets cannot be dropped.

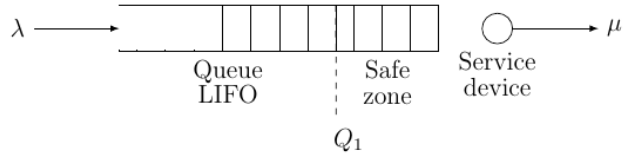


Figure 2. Queuing system model 2

If the current number of packets in the system i is less or equal to $Q_1 + 1$ (the threshold value Q_1 has not been overcome), then none of the packets will be dropped from the queue. If the current number of packets in the system i is greater then $Q_1 + 1$, then with probability q the packet, finishing the service and leaving the system, will drop all packets from the queue (outside the safe zone), or with probability $p = 1 - q$ the serviced packet simply leaves the system.

Let π_i be the steady-state probability distribution of the embedded Markov chain that the packet comming into the system will find in it i other packets ($i \geqslant 0$) [37], [38].

Let $p^{(\text{loss})}$ and $p^{(\text{serv})}$ be the probability that the received packet in the system will be dropped from the queue or will be transferred to service device.

The $p_i^{(\text{serv})}$ is the auxiliary probability that the packet will be served if it finds other i packets in the system.

$$p^{(\text{serv})} = \sum_{i=0}^{\infty} p_i^{(\text{serv})} \cdot \pi_i = 1 - \pi_{Q_1+1} \cdot \frac{q}{(1-g)(1-pg)}.$$

$$p^{(\text{loss})} = 1 - p^{(\text{serv})} = 1 - \left(1 - \pi_{Q_1+1} \cdot \frac{q}{(1-g)(1-pg)}\right),$$

$$p^{(\text{loss})} = \pi_{Q_1+1} \cdot \frac{q}{(1-g)(1-pg)}.$$

3.1. Time characteristics of the system

3.1.1. Time characteristics for serviced packets

$W^{(\text{serv})}(x)$ is the cumulative waiting time distribution function for the accepted into the system packet, $W_i^{(\text{serv})}(x)$ is the cumulative waiting time distribution function for the accepted into the system packet, if at the moment of its arrival there were i other packets in the system. Then:

$$W^{(\text{serv})}(x) = \frac{1}{p^{(\text{serv})}} \sum_{i=0}^{\infty} W_i^{(\text{serv})}(x) \cdot \pi_i,$$

$$\text{w}_i^{(\text{serv})}(x) = \left(W_i^{(\text{serv})}(x)\right)'$$

— probability density function.

The auxiliary functions $W_{i,j}^{(\text{serv})}(x)$ and $\text{w}_{i,j}^{(\text{serv})}(x) = \left(W_{i,j}^{(\text{serv})}(x)\right)'$ $(i, j \geqslant 0)$ are the distribution functions and the densities of distribution functions of the time spent by the served packet in the queue, if there were i other packets in the queue before the considered one and j others after it.

a) If $i = 0$, then the cumulative distribution functions $W_i^{(\text{serv})}(x) = 1, (x = 0)$. **b)** If $0 < i \leqslant Q_1$ — (the safe zone is not completely filled) then the received in the system packet will be in the safe zone (cannot be dropped). Then

$$\text{w}_i^{(\text{serv})}(x) = \mu e^{-\mu x} \cdot \overline{A}(x) + \int_0^x e^{-\mu y} d(y) \cdot \text{w}_{i,1}^{(\text{serv})}(x-y).$$

b.1) $0 < i + j \leqslant Q_1$, $j > 0$ (taking into account the packets that came after ours), the threshold value Q_1 has not been overcome in the queue, that is,

the renovation mechanism has not turned on. Then

$$\mathrm{w}_{ij}^{(\mathrm{serv})}(x) = \frac{\mu^{j+1}x^j}{j!}e^{-\mu x}\cdot\overline{A}(x) + \int\limits_0^x \sum_{k=0}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y).$$

b.2) $Q_1 < j+1$ $(j>0)$ the renovation mechanism was activated, but our packet is in a safe zone. Then

$$\begin{aligned}\mathrm{w}_{ij}^{(\mathrm{serv})}(x) = {} & \frac{\mu^{j+1}x^j}{j!}p^{j-(Q_1-i)+1}\cdot\overline{A}(x) + \frac{\mu^{Q_1-i+1}x^{Q_1-i}}{(Q_1-i)!}\cdot qe^{-\mu x}\cdot\overline{A}(x)+\\ & + \sum_{k=1}^{1+(j-(Q_1-i)-1)}\tilde{\pi}_k(j-(Q_1-i)-k)\cdot\frac{\mu^{k+Q_1-i}x^{k+Q_1-i-1}}{(k+Q_1-i-1)!}e^{-\mu x}\cdot\overline{A}(x)+\\ & + \int\limits_0^x e^{-\mu y}dA(y)\cdot\mathrm{w}_{i,j+1}^{(\mathrm{serv})}(x-y)+\\ & + \int\limits_0^x \sum_{k=1}^{j-(Q_1-i)-1}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^k dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y)+\\ & + \int\limits_0^x \sum_{k=1-(Q_1-i)}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^{i-Q_1-i}dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y),\end{aligned}$$

$$\begin{aligned}\mathrm{w}_{ij}^{(\mathrm{serv})}(x) = {} & \sum_{k=1}^{j-(Q_1-1)}\tilde{\pi}_k(j-(Q_1-i)-k)\cdot\frac{\mu^{k+Q_1-i}x^{k+Q_1-i-1}}{(k+Q_1-i-1)!}e^{-\mu x}\cdot\overline{A}(x)\\ & + \int\limits_0^x \sum_{k=1}^{j-(Q_1-i)-1}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^k dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y)+\\ & + \int\limits_0^x \sum_{k=1-(Q_1-i)}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^{i-Q_1-i}dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y).\end{aligned}$$

c) $i \geqslant Q_1+1$ — at the time of receipt of our packet, the safe zone is filled and there are packets outside the safe zone — the renovation mechanism is enabled. Then

$$\mathrm{w}_{i,0}^{(\mathrm{serv})}(x) = \mu e^{-\mu x}p\cdot\overline{A}(x) + \int\limits_0^x e^{-\mu y}dA(y)\cdot\mathrm{w}_{i,1}^{(\mathrm{serv})}(x-y),$$

$$\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = \frac{\mu^{j+1}x^j}{j!}e^{-\mu x}p^{j+1}\overline{A}(x) + \int\limits_0^x \sum_{k=0}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}\cdot p^k dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y).$$

3.1.2. Time characteristics for dropped packets

Let $W^{(\text{loss})}(x)$ be the cumulative distribution functions of the time spent by the packet in the queue before dropping.

$$W^{(\text{loss})}(x) = \frac{1}{p^{(\text{loss})}} \cdot \sum_{i=0}^{\infty} W_i^{(\text{loss})}(x)\pi_i.$$

$W_i^{(\text{loss})}(x)$ is the conditional probability that in a time less than x the packet that has found exactly i of other packets in the system will be dropped from the queue. The auxiliary functions $W_{i,j}^{(\text{loss})}(x)$ and $\mathrm{w}_{i,j}^{(\text{loss})}(x) = \left(W_{i,j}^{(\text{loss})}(x)\right)'$ $(i, j \geqslant 0)$ are the distribution functions and the densities of distribution functions of the time spent by the dropped packet in the system, if there were i other packets in the queue before the considered one and j others after it.

a) $0 \leqslant i \leqslant Q_1$ (that is, the system was either empty, or at least there was one free space in the safe zone)

$$W_i^{(\text{loss})}(x) = 0.$$

b) $Q_1 < i$ $(i \geqslant Q_1 + 1)$

$$\mathrm{w}_{i,0}^{(\text{loss})}(x) = \mu e^{-\mu x} q \cdot \overline{A}(x) + \int_0^x e^{-\mu y} dA(y) \cdot \mathrm{w}_{i,1}^{(\text{loss})}(x-y),$$

$$\mathrm{w}_{i,j}^{(\text{loss})}(x) \sum_{k=1}^{j+1} \frac{\mu^k x^{k-1}}{(k-1)!} e^{-\mu x} \cdot \tilde{\pi}_k(j+i-Q_1-k)\overline{A}(x) + \\ + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} \cdot \sum_{l=0}^{j-k} \pi_k(l) dA(y) \cdot \mathrm{w}_{i,j-k-l+1}^{(\text{loss})}(x).$$

4. The third model

Consider the $GI/M/1/\infty$ queuing system, shown in the figure 3.

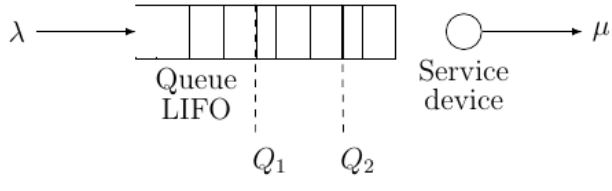


Figure 3. Queuing system model 3

In this section, a single-server queueing system with an infinite queue capacity and two threshold values is considered. Threshold values:

— Q_1 — the threshold value in the queue, when overcoming which by the queue length packets (from $Q_1 + 1$) will be dropped from the queue with a probability q.
— Q_2 — the threshold value in the queue to which packets are dropped (i.e. packets standing in the queue up to the Q_2 threshold are not dropped).

4.1. The service probability and loss probability of the received packet

Let's introduce the probability $p^{(\text{serv})}$ that the packet, entering the system, will be served, auxiliary probabilities $p_i^{(\text{serv})}$ $(i \geqslant 0)$ of incoming packet to be served if there were other i $(i \geqslant 0)$ packets in the system, and auxiliary probabilities $p_{i,j}^{(\text{serv})}(x)$ that during the time x the packet, which found exactly i other packets in the system at the moment of arrival and behind which there are j more packets, will be served

$$p^{(\text{serv})} = \sum_{i=0}^{\infty} p_i^{(\text{serv})} \pi_i,$$

where π_i — the stationary probabilities [37], [38].

Let's consider several cases

a) The first one, when the system is empty: $p_0^{(\text{serv})} = 1$.

b) The second case is when $1 \leqslant i \leqslant Q_2$, so $p_i^{(\text{serv})} = 1$.

c) The third case $Q_2 < i \leqslant Q_1$ includes two subcases:

c.1) the first subcase, $Q_2 + 1 \leqslant i + 1 + j \leqslant Q_1 + 1$ — the Q_1 threshold in the queue has not been overcome (taking into account the packets after the considered one), that is, the renovation mechanism has not turned on

$$p_{i,j}^{(\text{serv})}(x) = \overline{A}(x) \cdot \frac{(\mu x)^{j+1}}{(j+1)!} e^{-\mu x} + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} dA(y) \cdot p_{i,j-k+1}^{(\text{serv})}(x-y).$$

c.2) the second subcase, $i + 1 + j > Q_1 + 1$ — the Q_1 threshold in the queue has been overcome, so the renovation mechanism has been activated

$$\begin{aligned} p_{i,j}^{(\text{serv})}(x) = \overline{A}(x) \cdot \frac{(\mu x)^{j+1}}{(j+1)!} e^{-\mu x} \cdot p^{i+j+1-(Q_1+1)} + \\ + \int_0^x \sum_{k=0}^{i+j-Q_1} \frac{(\mu y)^k}{k!} e^{-\mu y} p^k dA(y) \cdot p_{i,j-k+1}^{(\text{serv})}(x-y) + \\ + \int_0^x \sum_{k=i+j-Q_1+1}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} p^{i+j-Q_1} dA(y) \cdot p_{i,j-k+1}^{(\text{serv})}(x-y). \end{aligned}$$

d) the fourth case is when the Q_1 threshold in the queue has been overcome at the moment of the arrival of the considered packet, $(i > Q_1)$ so the renovation mechanism has been already activated

$$p_{i,j}^{(\text{serv})}(x) = \overline{A}(x) \cdot \frac{(\mu x)^{j+1}}{(j+1)!} e^{-\mu x} p^{j+1} + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} p^k dA(y) \cdot p_{i,j-k}^{(\text{serv})}(x-y),$$

$$p_i^{(\text{serv})} = \int_0^\infty p_{i,0}^{(\text{serv})}(x) dx.$$

Loss probability of the received packet

$$p^{(\text{loss})} = \sum_{i=0}^{\infty} p_i^{(\text{loss})} \pi_i,$$

where $p_i^{(\text{loss})}$ — the probability that the incoming packet will be dropped if at the moment of its arrival there were i, $i \geqslant 0$ other packets in the system, and $p_{i,j}^{(\text{loss})}(x)$ is the probability that in time less than x the packet, before which there are i other packets in the queue and after which there are other j packets, will be dropped, $i, j \geqslant 0$.

a) $p_1^{(\text{loss})} = 0$, $i = \overline{0, Q_2}$;

b) $Q_2 < i \leqslant Q_1$ the threshold value of Q_1 has not been reached at the time of receipt;

b.1) $i + 1 + j \leqslant Q_1 + 1$ — (the threshold has not been crossed even taking into account the application that came later)

$$p_{i,j}^{(\text{loss})}(x) = \int_0^y \sum_{k=0}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} dA(y) \cdot p_{i,j-k+1}^{(\text{loss})}(x-y).$$

b.2) $i+1+j > Q_1+1$ — (the Q_1 threshold was overcome due to applications after the incoming one)

$$\begin{aligned} p_{i,j}^{(\text{loss})}(x) = \overline{A}(x) & \sum_{k=1}^{i+j+1-(Q_1+1)} \frac{(\mu x)^k}{k!} e^{-\mu x} p^{k-1} q + \\ & + \int_0^x \sum_{k=0}^{i+j-Q_1} \frac{(\mu y)^k}{k!} e^{-\mu y} p^k dA(y) p_{i,j-k+1}^{(\text{loss})}(x-y) + \\ & + \int_0^x \sum_{k=i+j-Q_1+1}^{j} \frac{(\mu y)^k}{k!} e^{-\mu y} p^{i+j-Q_1} dA(y) p_{i,j-k+1}^{(\text{loss})}(x-y). \end{aligned}$$

c) $i > Q_1$

$$p_{i,j}^{(\text{loss})}(x) = \overline{A}(x)\sum_{k=1}^{j+1}\frac{(\mu x)^k}{k!}e^{-\mu x}p^{k-1}q+$$
$$+\int_0^x\sum_{k=0}^{j}\frac{(\mu y)^k}{k!}\cdot e^{-\mu y}p^k dA(y)p_{i,j-k+1}^{(\text{loss})}(x-y);$$
$$p_i^{(\text{loss})} = \int_0^\infty p_{i,0}^{(\text{loss})}(x)dx.$$

4.2. Time characteristics of the system

Let $W^{(\text{loss})}(x)$ and $W^{(\text{serv})}(x)$ be the cumulative distribution functions of the time spent in the system by the packet before being dropped or served. The auxiliary functions $W_{i,j}^{(\text{serv})}(x)$ and $\mathrm{w}_{i,j}^{(\text{serv})}(x) = \left(W_{i,j}^{(\text{serv})}(x)\right)'$, $W_{i,j}^{(\text{loss})}(x)$ and $\mathrm{w}_{i,j}^{(\text{loss})}(x) = \left(W_{i,j}^{(\text{serv})}(x)\right)'$ $(i, j \geqslant 0)$ are the distribution functions and the densities of distribution functions of the time spent by the served (lossed) packet in the queue, if there were i other packets in the queue before the considered one and j others after it. Then

$$W^{(\text{serv})}(x) = \frac{1}{p^{(\text{serv})}}\sum_{i=0}^{\infty}W_{i,j}^{(\text{serv})}(x)\cdot\pi_i,$$
$$W^{(\text{loss})}(x) = \frac{1}{p^{(\text{loss})}}\sum_{i=0}^{\infty}W_{i,j}^{(\text{loss})}(x)\cdot\pi_i.$$

a) If a packet enters the empty system ($i = 0$), it immediately starts to be served.

$$\mathrm{w}_{0,0}^{(\text{serv})}(x) = \begin{cases}0, & x < 0,\\ 1, & x \geqslant 0,\end{cases}$$
$$\omega_{0,0}^{(\text{serv})}(s) = \int_0^\infty e^{-sx}\mathrm{w}_{0,0}^{(\text{serv})}(x)d(x) = 1,$$
$$\mathrm{w}_{0,0}^{(\text{loss})}(x) = 0.$$

b) If the total number of packets in the system has not overcome the threshold Q_2 $(0 < i \leqslant Q_1,\ i + j + 1 \leqslant Q_1)$, then the considered packet will be in the safe area and the renovation mechanism is not enabled.

$$\mathrm{w}_{i,0}^{(\text{serv})}(x) = \overline{A}(x)\cdot\mu e^{-\mu x} + \int_0^x e^{-\mu y}dA(y)\cdot\mathrm{w}_{i,1}^{(\text{serv})}(x-y).$$

$$\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = \overline{A}(x)\frac{\mu^{j+1}x^j}{j!}e^{-\mu x} + \int_0^x \sum_{k=0}^{j} \frac{(\mu y)^k}{k!}e^{-\mu y}dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y),$$

$$\omega_{i,j}^{(\mathrm{serv})}(s) = \frac{(-1)^j\mu^{j+1}}{j!}\overline{\alpha}^{(j)}(s+\mu) + \sum_{k=0}^{j}\frac{(-\mu)^k}{k!}\times\alpha^{(k)}(s+\mu)\cdot\omega_{i,j-k+1}^{(\mathrm{serv})}(s),$$

$$\mathrm{w}_{i,j}^{(\mathrm{loss})}(x) = 0.$$

c) The case, when at the moment of arrival of the considered packet there were $0 < i < Q_2$ other packets in the system (our packet was in the safe area), but currently the total number of packets in the system is equal to $i + j + 1 > Q_1$ (so the renovation mechanism is enabled)

$$\begin{aligned}\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = {} & \frac{\mu^{i+j}x^j}{j!}e^{-\mu x}p^{i+j+1-Q_1}\overline{A}(x) + \\ & + \overline{A}(x)\int_0^x \sum_{k=1}^{i+j+1-Q_1}\frac{(\mu y)^k}{k!}e^{-\mu y}p^{k-1}q\mu dy\frac{(\mu(x-y))^{Q_2-i-1}}{(Q_2-i-1)}e^{-\mu(x-y)} + \\ & + \int_0^x \sum_{k=0}^{i+j+1-Q_1}\frac{(\mu y)^k}{k!}e^{-\mu x}p^{k-1}qdA(y)\mathrm{w}_{i,Q_2-i-1+1}^{(\mathrm{serv})}(x-y) + \\ & + \int_0^x \sum_{k=i+j+1-Q_1-1}^{j}\frac{(\mu y)^k}{k!}e^{\mu y}p^{i+j+1-Q_1}dA(y)\mathrm{w}_{i,j-k}^{(\mathrm{serv})}(x),\end{aligned}$$

$$\mathrm{w}_{i,j}^{(\mathrm{loss})}(x) = 0.$$

d) The case, when at the moment of arrival of the considered packet there were $Q_2 < i < Q_1$ other packets in the system (our packet was out of the safe area), includes several subcases.

d.1) The first subcase — currently the total number of packets in the system is $Q_2 < i + j + 1 \leqslant Q_1$ (the renovation mechanism is not enabled)

$$\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = \overline{A}(x)\frac{\mu^{j+1}x^j}{j!}e^{-\mu x} + \int_0^x \sum_{k=0}^{j}\frac{(\mu y)^k}{k!}\times e^{-\mu y}dA(y)\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y),$$

$$\mathrm{w}_{i,j}^{(\mathrm{loss})}(x) = \int_0^x \sum_{k=0}^{i+j+1-Q_2}\frac{\mu y}{k!}e^{-\mu y}dAy\cdot\mathrm{w}_{i,j-k+1}^{(\mathrm{loss})}(x-y).$$

d.2) The second subcase, when currently the total number of packets in the system has overcome the threshold Q_1 ($i + j + 1 > Q_1$), so the renovation mechanism is activated

$$\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = \overline{A}(x)\frac{\mu^{j+i}x^j}{j!}e^{-\mu y}\cdot p^{i+j+1-Q_1}+$$
$$+\int_0^x \sum_{k=0}^{i+j+1-Q_1}\frac{(\mu y)^k}{k!}e^{-\mu y}p^k dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y)+$$
$$+\int_0^x \sum_{k=i+j+1-Q_1+1}^{j}\frac{(\mu y)^k}{k!}\cdot p^{i+j+1-Q_1}e^{-\mu y}dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y),$$

$$\mathrm{w}_{i,j}^{(\mathrm{loss})}(x) = \overline{A}(x)\sum_{k=1}^{i+j+1-Q_1}\frac{\mu^k x^{k-1}}{(k-1)!}p^{k-1}qe^{-\mu x}+$$
$$+\int_0^x \sum_{k=0}^{i+j+1-Q_1}\frac{(\mu u)^k}{k!}p^k e^{-\mu y}dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{loss})}(x-y)+$$
$$+\int_0^x \sum_{k=i+j+1-Q_1+1}^{j}\frac{(\mu y)^k}{k!}\cdot p^{i+j+1-Q_1}e^{-\mu y}dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{loss})}(x-y).$$

e) The last case, when the threshold Q_1 was overcome ($i > Q_1$) at the moment of our packet arrival

$$\mathrm{w}_{i,j}^{(\mathrm{serv})}(x) = \overline{A}(x)\frac{\mu^{j+1}x^j}{j!}e^{-\mu x}p^{j+1}+\int_0^x\sum_{k=0}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}p^k dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{serv})}(x-y),$$

$$\mathrm{w}_{i,j}^{(\mathrm{loss})}(x) = \overline{A}(x)\sum_{k=1}^{j+1}\frac{\mu^k x^{k-1}}{(k-1)!}e^{-\mu x}p^{k-1}q+$$
$$+\int_0^x\sum_{k=0}^{j}\frac{(\mu y)^k}{k!}e^{-\mu y}p^k dA(y)\cdot \mathrm{w}_{i,j-k+1}^{(\mathrm{loss})}(x-y).$$

5. GPSS simulation results

Below (see table 1) is presented a table with GPSS simulation results that was performed with the following initial parameters: threshold value $Q_1 = 30$, arrival rate — 14 task per 1 unit of time, service rate — 16 task per 1 unit of time, and the simulation time is 100000 units of time) for different drop probabilities.

The table 2 shows the results of GPSS simulation that was performed with the following initial parameters: arrival rate — 14 task per 1 unit of time, service rate — 16 task per 1 unit of time, $q = 0.01$, and the simulation time

is 100000 units of time) for different threshold values. For the third model the threshold value $Q_2 = 10$.

Table 1

Simulation results for different drop probabilities

q **propability**		**0.0025**	**0.005**	**0.01**	**0.025**	**0.05**	**0.1**	**0.15**
Generated tasks	sys.1	1401525	1401566	1401134	1400127	1400915	1399127	1398795
	sys.2	1400992	1401374	1401547	1400816	1401421	1400971	1401135
	sys.3	1401647	1401379	1400564	1400333	1400889	1400251	1399581
Serviced tasks	sys.1	1400084	1398863	1396791	1394210	1393457	1389597	1389540
	sys.2	1400752	1400843	1400879	1399692	1399428	1399166	1399030
	sys.3	1400537	1399411	1397201	1395975	1395643	1393555	1393104
Serviced tasks without calling the renv. mech.	sys.1	1379233	1381969	1385859	1388162	1388647	1386899	1387651
	sys.2	1378347	1381669	1385318	1388493	1387780	1391338	1391897
	sys.3	1379887	1382616	1385828	1389605	1390628	1390814	1391166
Dropped tasks	sys.1	1436	2698	4332	5917	7456	9530	9249
	sys.2	240	527	663	1117	1984	1803	2104
	sys.3	1091	1967	3357	4357	5240	6696	6472
Service Probability	sys.1	0.9990	0.9981	0.9969	0.9958	0.9947	0.9932	0.9934
	sys.2	0.9998	0.9996	0.9995	0.9992	0.9986	0.9987	0.9985
	sys.3	0.9992	0.9986	0.9976	0.9969	0.9963	0.9952	0.9954
Drop Probability	sys.1	0.0010	0.0019	0.0031	0.0042	0.0053	0.0068	0.0066
	sys.2	0.0002	0.0004	0.0005	0.0008	0.0014	0.0013	0.0015
	sys.3	0.0008	0.0014	0.0024	0.0031	0.0037	0.0048	0.0046
Average queue length	sys.1	6.0930	5.9230	5.7090	5.5240	5.4820	5.3080	5.2360
	sys.2	6.1800	6.0780	6.0220	5.8580	5.9530	5.7980	5.8550
	sys.3	6.1230	5.9360	5.7330	5.5720	5.5560	5.4120	5.3290
Maximum queue length	sys.1	92	71	63	67	54	46	43
	sys.2	92	64	61	65	60	51	49
	sys.3	92	71	71	67	54	46	43
Average waiting time	sys.1	0.497	0.483	0.467	0.453	0.449	0.437	0.431
	sys.2	0.503	0.495	0.491	0.478	0.485	0.473	0.478
	sys.3	0.499	0.484	0.469	0.456	0.454	0.444	0.438

Table 2

Simulation results for different threshold values

Threshold value Q_1		**10**	**20**	**25**	**30**	**40**	**50**	**75**
Generated tasks	sys.1	1399202	1401573	1401188	1401134	1399645	1400335	1400451
	sys.2	1399603	1400523	1399393	1401547	1402003	1400032	1399596
	sys.3	1399603	1400753	1400647	1400564	1399680	1400321	1400448
Serviced tasks	sys.1	1368353	1389618	1393927	1396791	1398462	1399917	1400367
	sys.2	1387180	1397457	1397721	1400879	1401813	1399986	1399562
	sys.3	1387180	1393344	1395743	1397201	1398764	1399969	1400374
Serviced tasks without calling the renv. mech.	sys.1	1166280	1343186	1370099	1385859	1394747	1398969	1400319
	sys.2	1145456	1336931	1365038	1385318	1396545	1398819	1399341
	sys.3	1145456	1346681	1372422	1385828	1395050	1399021	1400326
Dropped tasks	sys.1	30833	11955	7261	4332	1176	407	83
	sys.2	12423	3065	1672	663	190	42	33
	sys.3	12423	7409	4902	3357	916	337	73
Service Probability	sys.1	0.9780	0.9915	0.9948	0.9969	0.9992	0.9997	0.9999
	sys.2	0.9911	0.9978	0.9988	0.9995	0.9999	1.0000	1.0000
	sys.3	0.9911	0.9947	0.9965	0.9976	0.9993	0.9997	0.9999
Drop Probability	sys.1	0.0220	0.0085	0.0052	0.0031	0.0008	0.0003	0.0001
	sys.2	0.0089	0.0022	0.0012	0.0005	0.0001	0.0000	0.0000
	sys.3	0.0089	0.0053	0.0035	0.0024	0.0007	0.0002	0.0001
Average queue length	sys.1	4.564	5.273	5.5330	5.7090	5.9110	5.934	6.158
	sys.2	5.069	5.7	5.8540	6.0220	6.0780	6.014	6.089
	sys.3	5.069	5.37	5.5630	5.7330	5.9210	5.933	6.158
Maximum queue length	sys.1	67	64	71	63	80	76	89
	sys.2	67	75	62	61	64	76	102
	sys.3	67	75	59	71	80	76	89
Average waiting time	sys.1	0.381	0.433	0.454	0.467	0.484	0.485	0.502
	sys.2	0.418	0.466	0.479	0.491	0.496	0.491	0.497
	sys.3	0.418	0.441	0.456	0.469	0.485	0.485	0.502

6. Conclusion

Based on the simulation results 1, the following conclusions can be drawn. The largest number of dropped packets, as expected, is observed in the first model, the smallest — in the second model (due to the safe zone). The third model shows an average result compared to the first and the second models. The largest number of serviced packets is in the second model, then — in the third model. The smallest number of serviced packets is in the first model.

The probability of a packet to be dropped is about five times greater for the first model than for the second model, and 20–30 percent more than for the third model.

The average waiting time for the second model is about 5–10 percent greater than the same characteristic for the first and third models.

As the value of the renovation probability q increases, the drop probability increases for all three models, and the service probability decreases accordingly. Also, with an increase of the renovation probability q, both the average and maximum queue lengths decrease, and the average waiting time also decreases.

Based on the simulation results 2, the following conclusions can be drawn. With an increase of the threshold value Q_1 responsible for switching on the renovation mechanism, the number of dropped packets decreases for all three models (the second model is characterized by the smallest number of dropped packets), the service probability increases to unity (the second model), and the drop probability decreases almost to zero. The average and maximum queue lengths increase, and the values for the first and third models become approximately the same. The average waiting time also increases, and again for the first and third models, the values become approximately the same.

The third model, which generalizes the first and the second models, shows average results compared to the above models, and is more preferable for use as a queue length management model.

Acknowledgments

The publication was funded by RFBR according to the research projects No. 20-07-00804.

References

[1] F. Baker and G. Fairhurst. "IETF Recommendations Regarding Active Queue Management. RFC 7567". (Jul. 2015), [Online]. Available: `https://tools.ietf.org/html/rfc7567`.

[2] K. Nichols and V. Jacobson, "Controlling queue delay", *Communications of the ACM*, vol. 55, no. 7, pp. 42–50, May 2012. DOI: `10.1145/2209249.2209264`.

[3] T. Hoeiland-Joergensen *et al.* "The flow queue codel packet scheduler and active queue management algorithm. RFC 8290". (2018), [Online]. Available: `https://www.rfc-editor.org/info/rfc8290`.

[4] S. Jung, J. Kim, and J.-H. Kim, "Intelligent active queue management for stabilized QoS guarantees in 5G mobile networks", *IEEE Systems Journal*, vol. 15, pp. 4293–4302, 2021. DOI: `10.1109/JSYST.2020.3014231`.

[5] W.-c. Feng, D. Kandlurz, D. Sahaz, and K. Shin, "BLUE: a new class of active queue management algorithms", University of Michigan, Tech. Rep., Sep. 2000.

[6] W.-c. Feng, D. Kandlur, and D. Saha, "The BLUE active queue management algorithms", *Networking, IEEE/ACM Transactions on*, vol. 10, pp. 513–528, Sep. 2002. DOI: 10.1109/TNET.2002.801399.

[7] C. Zhang, J. Yin, and Z. Cai, "RSFB: a resilient stochastic fair blue algorithm against spoofing DDoS attacks", in *9th International Symposium on Communications and Information Technology*, 2009, pp. 1566–1567. DOI: 10.1109/ISCIT.2009.5341033.

[8] T. Hoiland-Jorgensen, D. Taht, and J. Morton, "Piece of CAKE: a comprehensive queue management solution for home gateways", in *IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, Jun. 2018, pp. 37–42. DOI: 10.1109/LANMAN.2018.8475045.

[9] J. Palmei, S. Gupta, P. Imputato, J. Morton, M. Tahiliani, S. Avallone, and D. Taht, "Design and evaluation of COBALT queue discipline", in *IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, Jul. 2019, pp. 1–6. DOI: 10.1109/LANMAN.2019.8847054.

[10] A. Roy, J. L. Pachuau, and A. K. Saha, "An overview of queuing delay and various delay based algorithms in networks", *Computing*, vol. 103, pp. 2361–2399, 2021. DOI: 10.1007/s00607-021-00973-3.

[11] W. de Morais, C. E. M. Santos, and C. M. Pedroso, "Application of active queue management for real-time adaptive video streaming", *Telecommun Syst*, vol. 79, pp. 261–270, 2022. DOI: 10.1007/s11235-021-00848-0.

[12] J. George and R. Santhosh, "Congestion control mechanism for unresponsive flows in Internet through active queue management system (AQM)", *Lecture Notes on Data Engineering and Communications Technologies*, vol. 68, pp. 765–777, 2022. DOI: 10.1007/978-981-16-1866-6_58.

[13] S. Singha, B. Jana, N. K. Mandal, S. Jana, S. Bandyopadhyay, and S. Midya, "Application of dynamic weight with distance to reduce packet loss in RED based algorithm", *Lecture Notes in Networks and Systems*, vol. 292, pp. 530–543, 2022. DOI: 10.1007/978-981-16-4435-1_52.

[14] R. Adams, "Active queue management: a survey", *Communications Surveys & Tutorials, IEEE*, vol. 15, pp. 1425–1476, Jan. 2013. DOI: 10.1109/SURV.2012.082212.00018.

[15] M. Menth and S. Veith, "Active queue management based on congestion policing (CP-AQM)", in Jan. 2018, pp. 173–187. DOI: 10.1007/978-3-319-74947-1_12.

[16] A. Chydzinski and L. Chrost, "Analysis of AQM queues with queue size based packet dropping", *Applied Mathematics and Computer Science*, vol. 21, pp. 567–577, Sep. 2011. DOI: 10.2478/v10006-011-0045-7.

[17] A. Chydzinski and P. Mrozowski, "Queues with dropping functions and general arrival processes", *PloS one*, vol. 11, e0150702, Mar. 2016. DOI: 10.1371/journal.pone.0150702.

[18] M. Konovalov and R. Razumchik, "Numerical analysis of improved access restriction algorithms in a GI/G/1/N system", *Journal of Communications Technology and Electronics*, vol. 63, pp. 616–625, Jun. 2018. DOI: 10.1134/S1064226918060141.

[19] M. Konovalov and R. Razumchik, "Comparison of two active queue management schemes through the M/D/1/N queue", *Informatika i ee Primeneniya*, vol. 12, no. 4, pp. 9–15, 2018, in Russian. DOI: 10.14357/19922264180402.

[20] C. SSo-In, R. Jain, and J. Jiang, "Enhanced forward explicit congestion notification (E-FECN) scheme for datacenter Ethernet networks", in *International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, 2008, pp. 542–546.

[21] C. Gomez, X. Wang, and A. Shami, "Intelligent active queue management using explicit congestion notification", in *IEEE Global Communications Conference (GLOBECOM)*, Sep. 2019, pp. 1–6. DOI: 10.20944/preprints201909.0077.v1.

[22] S. Shahzad, E.-S. Jung, J. F. Chung M., and R. Kettimuthu, "Enhanced explicit congestion notification (EECN) in TCP with P4 programming", in *International Conference on Green and Human Information Technology (ICGHIT)*, Feb. 2020. DOI: 10.1109/ICGHIT49656.2020.00015.

[23] S. Wang, J. Zhang, T. Huang, T. Pan, J. Liu, and Y. Liu, "A-ECN minimizing queue length for datacenter networks", *IEEE Access*, vol. 8, pp. 49 100–49 111, 2020. DOI: 10.1109/ACCESS.2020.2979216.

[24] A. Bashir, E. Machnev, and E. Mokrov, "Queueing model of hysteretic congestion control for cloud wireless sensor networks", in *13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2021, pp. 104–108. DOI: 10.1109/ICUMT54235.2021.9631576.

[25] S. Li, Q. Xu, J. Gaber, Z. Dou, and J. Chen, "Congestion control mechanism based on dual threshold DI-RED for WSNs", *Wireless Personal Communications*, vol. 115, pp. 2171–2195, 2020. DOI: 10.1007/s11277-020-07676-6.

[26] S. Singha, B. Jana, S. Jana, and N. K. Mandal, "An innovative active queue management model through threshold adjustment using queue size", *Advances in Intelligent Systems and Computing*, vol. 1406, pp. 257–273, 2022. DOI: 10.1007/978-981-16-5207-3_23.

[27] A. Kreinin, "Queueing systems with renovation", *Journal of Applied Mathematics and Stochastic Analysis*, vol. 10, pp. 431–443, Jan. 1997. DOI: 10.1155/S1048953397000464.

[28] M. Konovalov and R. Razumchik, *Queueing systems with renovation vs. queues with red. supplementary material*, 2017. arXiv: 1709.01477.

[29] A. V. Gorbunova and A. V. Lebedev, "Queueing system with two input flows, preemptive priority, and stochastic dropping", *Automation and Remote Control*, vol. 81, no. 12, pp. 2230–2243, 2020. DOI: 10.1134/S0005117920120073.

[30] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", *IEEE/ACM Transactions on Networking*, vol. 1, pp. 397–413, Sep. 1993. DOI: 10.1109/90.251892.

[31] K. Ramakrishnan, S. Floyd, and D. Black. "RFC3168: The Addition of Explicit Congestion Notification (ECN) to IP". (2001), [Online]. Available: https://tools.ietf.org/html/rfc3168.

[32] S. Floyd, R. Gummadi, and S. Shenker, *Adaptive RED: an algorithm for increasing the robustness of RED's active queue management*, Sep. 2001.

[33] A. V. Korolkova, D. S. Kulyabov, and A. I. Chernoivanov, "On the classification of RED algorithms", *Bulletin of Peoples' Friendship University of Russia*, no. 3, pp. 34–46, 2009, in Russian.

[34] W.-C. Feng, "Improving Internet congestion control and queue management algorithms", The University of Michigan, Tech. Rep., 1999.

[35] H. C. C. Viana, I. Zaryadov, V. Tsurlukov, T. Milovanova, E. Bogdanova, A. Korolkova, and D. Kulyabov, "The general renovation as the active queue management mechanism. Some aspects and results", *Communications in Computer and Information Science*, vol. 1141, pp. 488–502, 2019. DOI: 10.1007/978-3-030-36625-4_39.

[36] H. C. C. Viana, I. S. Zaryadov, and T. A. Milovanova, "Queueing systems with different types of renovation mechanism and thresholds as the mathematical models of active queue management mechanism", *Discrete and Continuous Models and Applied Computational Science*, vol. 28, no. 4, pp. 305–318, 2020. DOI: 10.22363/2658-4670-2020-28-4-305-318.

[37] H. C. C. Viana, I. S. Zaryadov, and T. A. Milovanova, "Two types of single-server queueing systems with threshold-based renovation mechanism", *Lecture Notes in Computer Science*, vol. 13144, pp. 196–210, 2021. DOI: 10.1007/978-3-030-92507-9_17.

[38] H. C. C. Viana and I. S. Zaryadov, "Single-server queuing systems with exponential service times and threshold-based renovation", in *13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2021, pp. 91–97. DOI: 10.1109/ICUMT54235.2021.9631585.

For citation:

I. S. Zaryadov, H. C. C. Viana, T. A. Milovanova, Analysis of queuing systems with threshold renovation mechanism and inverse service discipline, Discrete and Continuous Models and Applied Computational Science 30 (2) (2022) 160–182. DOI: 10.22363/2658-4670-2022-30-2-160-182.

Information about the authors:

Zaryadov, Ivan S. — Candidate of Physical and Mathematical Sciences, Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); Senior Researcher of Institute of Informatics Problems of Federal Research Center "Computer Science and Control" Russian Academy of Sciences (e-mail: zaryadov-is@rudn.ru,

phone: +7(495)9550927, ORCID: https://orcid.org/0000-0002-7909-6396, ResearcherID: B-8154-2018, Scopus Author ID: 35294470000)

Viana, Hilquias C. C. — PHD student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: hilvianamat1@gmail.com, phone: +7(495)9550927, Scopus Author ID: 57212930802)

Milovanova, Tatiana A. — Candidate of Physical and Mathematical Sciences, Lecturer of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: milovanova-ta@rudn.ru, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0002-9388-9499, Scopus Author ID: 26641495400)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2022-30-2-160-182

Анализ систем массового обслуживания с пороговым механизмом обновления и инверсионной дисциплиной обслуживания

И. С. Зарядов[1,2], Илкиаш К. К. Виана[1], Т. А. Милованова[1]

[1] *Российский университет дружбы народов,*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Институт проблем информатики,*
Федеральный исследовательский центр «Информатика и управление» РАН,
ул. Вавилова, д. 44, кор. 2, Москва, 119333, Россия

Аннотация. В работе представлено исследование трёх систем массового обслуживания с пороговым механизмом обновления и инверсионной дисциплиной обслуживания. В модели первого типа пороговое значение отвечает только за активацию механизма обновления — механизма вероятностного сброса заявок. Во второй модели пороговое значение не только включает механизм обновления, но и определяет в накопителе границы области, из которой поступившие в систему заявки не могут быть сброшены. В модели третьего типа, обобщающей предыдущие две модели, используются два пороговых значения: одно для активации механизма сброса заявок, второе — для задания безопасной зоны в накопителе. На основе полученных ранее результатов представлены основные вероятностно-временные характеристики рассмотренных моделей. С помощью имитационного моделирования проведён анализ и сравнение поведения изученных моделей.

Ключевые слова: система массового обслуживания, активное управление очередью, механизм обновления, пороговое значение, временные характеристики, GPSS