# DISCRETE AND CONTINUOUS MODELS AND APPLIED COMPUTATIONAL SCIENCE


## Volume 31 Number 2 (2023)

**Founded in 1993**

**Founder: PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA NAMED AFTER PATRICE LUMUMBA**

**DOI: 10.22363/2658-4670-2023-31-2**

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 named after Patrice Lumumba (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), 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:** *Anna V. Korolkova*, *Dmitry 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 7.10. Printing run 500 copies. Open price. The order 664.
PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA NAMED AFTER PATRICE LUMUMBA
6 Miklukho-Maklaya St., 117198 Moscow, Russian Federaion
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)
2023, 31 (2) 104
http://journals.rudn.ru/miph

# Contents

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

2023, 31 (2) 105–119
http://journals.rudn.ru/miph

*Research article*

UDC 519.872
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-105-119
EDN: VUBLKP

# Heterogeneous queueing system with Markov renewal arrivals and service times dependent on states of arrival process

**Evgeny P. Polin**[1, 2], **Svetlana P. Moiseeva**[1], **Alexander N. Moiseev**[1]

[1] *National Research Tomsk State University,*
*36, Lenin Avenue, Tomsk, 634050, Russian Federation*
[2] *National Research Tomsk Polytechnic University,*
*30, Lenin Avenue, Tomsk, 634050, Russian Federation*

(received: April 14, 2023; revised: April 25, 2023; accepted: June 26, 2023)

**Abstract.** In the proposed work, we consider a heterogeneous queueing system with a Markov renewal process and an unlimited number of servers. The service time for requests on the servers is a positive random variable with an exponential probability distribution. The service parameters depend on the state of the Markov chain nested over the renewal moments. It should be noted that these parameters do not change their values until the end of maintenance. Thus, the devices in the system under consideration are heterogeneous. The object of the study is a multidimensional random process — the number of servers of each type being served with different intensities in the stationary regime. The method of asymptotic analysis under the condition of equivalent growing of service times in the units of servers is applied for the study. The method of asymptotic analysis is implemented in the construction of a sequence of asymptotic of increasing order, in which the asymptotic of the first order determines the asymptotic mean value of the number of occupied servers. The second-order asymptotic allows one to construct a Gaussian approximation of the probability distribution of the number of occupied servers in the system. It is shown that this approximation coincides with the Gaussian distribution.

**Key words and phrases:** queuing system, random environment, Markov renewal process, asymptotic analysis method

## 1. Introduction

Queueing theory is a field of applied mathematics that deals with the study and analysis of processes in various service, production, management, and communication systems in which homogeneous events are repeated many times. Examples of such systems include consumer services; systems for

receiving, processing, and transmitting information, automatic production lines, telecommunication systems, and others [1].

The independence of processes in queueing systems is generally assumed when developing queueing models. However, real systems often involve several process dependencies, and failure to consider these can lead to a serious under errors in the estimation of the performance measures. Semi-Markov processes are used in modeling stochastic control problems arising in Markovian dynamic systems where the sojourn time in each state is a general continuous random variable. They are powerful, natural tools for the optimization of queues, production scheduling, reliability/maintenance [2, 3]. For example, in a machine replacement problem with deteriorating performance over time, a decision-maker, after observing the current state of the machine, decides whether to continue its usage, initiate maintenance (preventive or corrective) repair or replace the machine.

Semi-Markov Processes include renewal processes and continuous-time Markov chains as special cases. In a semi-Markov process similar to Markov chains, state changes occur according to the Markov property, i.e., states in the future do not depend on the states in the past given the present. However, the sojourn time in a state is a continuous random variable with distribution depending on that state and the next state. A renewal process is a generalization of a Poisson process that allows arbitrary holding times. Its applications include such as planning for replacing worn-out machinery in a factory. A Markov renewal process is a generalization of a renewal process that the sequence of holding times is not independent and identically distributed. Their distributions depend on the states in a Markov chain. The Markov renewal processes were studied by Pyke in the 1960s [4, 5].

In the proposed work, we consider a heterogeneous queueing system (QS) with a Markov renewal process (MRP) for the process of its arrival and an unlimited number of servers. The service time for requests have an exponential probability distribution. Parameter of the service depends on the state of the Markov chain nested over the renewal moments. It should be noted that these parameters do not change their values until the end of maintenance. Thus, the devices in the system under consideration are heterogeneous.This problem for the Queueing System $M|M|\infty$ in a Markov Random Enviroment was addressed in [6–8].

The objects of the study are the number of servers of each type being served in the stationary regime. Such a QS can be attributed to the class of non-homogeneous QS operating in a random environment.

Currently, a significant part of the information, telecommunication, and other systems operate in a changing environment. The impact of a random environment can be expressed, for example, in a change in the parameters of the functioning of the system. In this regard, questions arise about the stability of such systems to external influences. Therefore, the study of systems operating in a random environment is an urgent task. In various works devoted to the study of systems in Markov and semi-Markov random environments, various variants of the system's response to a change in the state of the external environment were considered in [9–11].

In this paper, we consider the case assuming that the service mode of claims does not change until they leave the system. The method of asymptotic analysis under the condition of equivalent growing of service times in the

units of servers is applied for the study. This asymptotic condition means proportional growth of the average service times in both service units and it is taken from practice. The method of asymptotic analysis is implemented in the construction of a sequence of asymptotic of increasing order, in which the asymptotic of the first order determines the asymptotic mean value of the number of occupied servers. The second-order asymptotic allows to construct an approximation of the probability distribution of the number of occupied servers in the system. It is shown that this approximation coincides with the Gaussian distribution.

## 2. Markov renewal process

A renewal process is a generalization of a Poisson process that allows arbitrary waiting time between events. Its applications include such as planning for replacing worn-out machinery in a factory. A Markov renewal process is a generalization of a renewal process that the sequence of holding times is not independent and identically distributed. Their distributions depend on the states in a Markov chain. The Markov renewal processes were studied by Pyke [4, 5] in 1960s.

### 2.1. Mathematical model of the Markov renewal process

Consider a two-dimensional homogeneous Markov random process $\{\xi(n), \tau(n)\}$ with discrete time $n = 1, 2, 3, ...$, where $\xi(n)$ takes values from some discrete set $\xi(n) = k = 1, 2, 3, ...$ and $\tau(n)$ takes on non-negative values.

We denote

$$F(k_2, x; k_1, y) = P\{\xi(n+1) = k_2,\ \tau(n+1) < x | \xi(n) = k_1,\ \tau(n) = y\} = \\ = F(k_2, x; k_1) = P_{k_1 k_2} A_{k_2}(x).$$

A random stream of homogeneous events $t_1 < ... < t_n < t_{n+1} < ...$ will be called the Markov renewal flow or MR-flow given by the matrix of transition probabilities $\mathbf{P}$ and functions $A_k(x)$ distribution of interval lengths $\tau_{n+1} = t_{n+1} - t_n$, for which the equalities $\tau_{n+1} = \tau_n$ hold.

To study the MR-flow, we define the process $z(t)$ as the length of the interval from the time $t$ to the time $t_{n+1}$ of the next event in the considered flow and the process

$$k(t) = \xi(n), \quad t_n \leqslant t < t_{n+1},$$

that is, the process $k(t)$ on the interval $t_n \leqslant t < t_{n+1}$ retains the value that it received at the beginning of this interval and which coincides with the value $\xi(n)$ of the embedded Markov chain.

For a Markov renewal flow, the three-dimensional process $\{k(t), z(t), m(t)\}$ is Markov, therefore, for its probability distribution

$$P(k, z, m, t) = P\{k(t) = k, z(t) < z, m(t) = m\}$$

by the formula of total probability we obtain the equality

$$P(k, z-\Delta t, m, t+\Delta t) = \\ = P(k, z, m, t) - P(k, \Delta t, m, t) + \sum_{\nu} P(\nu, \Delta t, m-1, t) P_{\nu k} A_k(z) + o(\Delta t)$$

from which it follows that the probability distribution $P(k, z, m, t)$ is a solution to the Kolmogorov equations

$$\frac{\partial P(k, z, m, t)}{\partial t} = \\ = \frac{\partial P(k, z, m, t)}{\partial z} - \frac{\partial P(k, 0, m, t)}{\partial z} + \sum_{\nu} \frac{\partial P(\nu, 0, m-1, t)}{\partial z} P_{\nu k} A_k(z). \quad (1)$$

By defining the functions

$$H(k, z, u, t) = \sum_{m=0}^{\infty} e^{jum} P(k, z, m, t),$$

the equations (1) can be rewritten as

$$\frac{\partial H(k, z, u, t)}{\partial t} = \frac{\partial H(k, z, u, t)}{\partial z} - \frac{\partial H(k, 0, u, t)}{\partial z} + \sum_{\nu} \frac{\partial H(\nu, 0, u, t)}{\partial z} e^{ju} P_{\nu k} A_k(z).$$

The basic equation for a semi-Markov flow has the form

$$\frac{\partial \mathbf{h}(z, u, t)}{\partial t} = \frac{\partial \mathbf{h}(z, u, t)}{\partial z} + \frac{\partial \mathbf{h}(0, u, t)}{\partial z} \left( e^{ju} \mathbf{P} \mathbf{A}(z) - \mathbf{I} \right), \quad (2)$$

where $\mathbf{P}$ is the matrix of transition probabilities, $\mathbf{A}(z) = \operatorname{diag}[A_k(z)]$, $\mathbf{I}$ is identity diagonal matrix. To find its particular solution, we define the initial condition in the form

$$\mathbf{h}(z, u, 0) = \mathbf{r}(z),$$

where $\mathbf{r}(z)$ — stationary probability distribution of the values of a two-dimensional random process $\{k(t), z(t)\}$.

### 2.2. Finding the distribution $\mathbf{r}(z)$

Vector $\mathbf{r}(z)$ is a solution to the equation obtained from (2)

$$\frac{\partial \mathbf{r}(z)}{\partial z} + \frac{\partial \mathbf{r}(0)}{\partial z} \left( \mathbf{P} \mathbf{A}(z) - \mathbf{I} \right) = 0,$$

therefore it can be written as

$$\mathbf{r}(z) = \int_0^z \frac{\partial \mathbf{r}(0)}{\partial z} \left( \mathbf{I} - \mathbf{P} \mathbf{A}(x) \right) dx. \quad (3)$$

Since $r(k,z) = P\{k(t) = k,\ z(t) < z\}$ then $\mathbf{r} = \mathbf{r}(\infty)$. Therefore, we obtain

$$\mathbf{r} = \int_0^\infty \frac{\partial \mathbf{r}(0)}{\partial z} \left(\mathbf{I} - \mathbf{P}\mathbf{A}(x)\right) dx. \tag{4}$$

By virtue of the necessary condition for the convergence of the improper integral, we can write down the equality to zero of the integrand at $x \to \infty$, we obtain the system of equations

$$\frac{\partial \mathbf{r}(0)}{\partial z} \left(\mathbf{I} - \mathbf{P}\right) = 0 \tag{5}$$

for $\dfrac{\partial \mathbf{r}(0)}{\partial z}$, where $\mathbf{P} = \mathbf{A}(\infty)$.

Since the system (4) coincides with the system of Kolmogorov equations for the stationary probability distribution $\mathbf{r}$ of values of the embedded Markov chain, then

$$\frac{\partial \mathbf{r}(0)}{\partial z} = \lambda \mathbf{r}, \tag{6}$$

where $\lambda$ is some multiplicative constant, the value of which is found as follows.

Substituting (6) into (4), we obtain

$$\mathbf{r} = \lambda \int_0^\infty \mathbf{r} \left(\mathbf{P} - \mathbf{A}(x)\right) dx.$$

Since $\mathbf{re} = 1$ then

$$\lambda = \frac{1}{\int_0^\infty \mathbf{r} \left(\mathbf{P} - \mathbf{A}(x)\right) \mathbf{e} dx} = \frac{1}{\int_0^\infty \left(1 - F(x)\right) dx}. \tag{7}$$

Equalities (7), (6) and (3) solve the problem of finding the probability distribution $\mathbf{r}(z)$.

## 3. Mathematical model

Consider a queueing system $MRP|M|\infty$ with an unlimited number of servers of different types, operating in a semi-Markov random environment (see the figure 1). Arrivals are determined as Markov renewal process, interarrival periods have cumulative distribution functions $A_1(x), A_2(x), \ldots A_K(x)$ and the matrix of transition probabilities $\mathbf{P} = [p_{ij}]$, $i, j = 1, 2, \ldots, K$ — embedded in the moments of occurrence of events Markov chains with a finite number of states $k(t) = 1, 2, \ldots, K$. The service discipline is defined as follows: if the embedded Markov chain is in the state $k(t) = i$, then the incoming customer will be serviced on the $i$-th type server during a random time, exponentially distributed $F_i(x) = 1 - e^{-\mu_i x}$.

The problem is to study a multidimensional random process — numbers occupied servers of different types in the system at time $t$, which is denoted by $\mathbf{i}(t) = [i_1(t), i_2(t), \ldots, i_K(t)]$. The process $\mathbf{i}(t)$ is not Markov. For clarity,

consider the case when the external environment takes only 2 different states. We define a four-dimensional Markov random process $\{k(t), z(t), i_1(t), i_2(t)\}$, where $z(t)$ is the length of the interval from the time $t$ to the time of the next event in the stream Markov renewal, $k(t)$ is a Markov chain embedded with respect to renewal times.

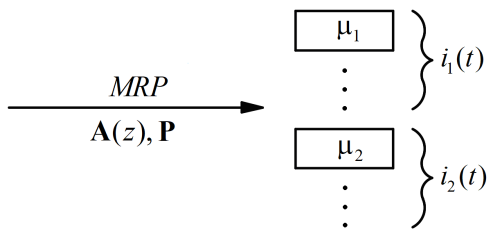


Figure 1. Queueing system $MRP|M|\infty$ in a semi-Markov random environment

For research, we will obtain some characteristics for the number of events occurring in the MR stream.

For the probability distribution

$$P(k, z, i_1, i_2, t) = P\{k(t) = k,\ z(t) < z,\ i_1(t) = i_1, i_2(t) = i_2\}$$

we write down the Kolmogorov system of differential equations:

$$\begin{aligned}\frac{\partial P(1, z, i_1, i_2, t)}{\partial t} &= \frac{\partial P(1, z, i_1, i_2, t)}{\partial z} - \frac{\partial P(1, 0, i_1, i_2, t)}{\partial z} - \\ &- (i_1\mu_1 + i_2\mu_2)P(1, z, i_1, i_2, t) + \frac{\partial P(1, 0, i_1 - 1, i_2, t)}{\partial z} p_{11}A_1(z) + \\ &+ \frac{\partial P(2, 0, i_1 - 1, i_2, t)}{\partial z} p_{21}A_1(z) + P(1, z, i_1 + 1, i_2, t)(i_1 + 1)\mu_1 + \\ &+ P(1, z, i_1, i_2 + 1, t)(i_2 + 1)\mu_2,\end{aligned}$$

$$\begin{aligned}\frac{\partial P(2, z, i_1, i_2, t)}{\partial t} &= \frac{\partial P(2, z, i_1, i_2, t)}{\partial z} - \frac{\partial P(2, 0, i_1, i_2, t)}{\partial z} - \\ &- (i_1\mu_1 + i_2\mu_2)P(2, z, i_1, i_2, t) + \frac{\partial P(2, 0, i_1, i_2 - 1, t)}{\partial z} p_{22}A_2(z) + \\ &+ \frac{\partial P(1, 0, i_1, i_2 - 1, t)}{\partial z} p_{12}A_2(z) + P(2, z, i_1 + 1, i_2, t)(i_1 + 1)\mu_1 + \\ &+ P(2, z, i_1, i_2 + 1, t)(i_2 + 1)\mu_2.\end{aligned}$$

For a stationary probability distribution, we write this system in the form

$$\begin{aligned}&\frac{\partial P(1, z, i_1, i_2)}{\partial z} - \frac{\partial P(1, 0, i_1, i_2)}{\partial z} - \\ &- (i_1\mu_1 + i_2\mu_2)P(1, z, i_1, i_2) + \frac{\partial P(1, 0, i_1 - 1, i_2)}{\partial z} p_{11}A_1(z) +\end{aligned}$$

$$+\frac{\partial P(2,0,i_1-1,i_2)}{\partial z}p_{21}A_1(z)+P(1,z,i_1+1,i_2)(i_1+1)\mu_1+$$
$$+P(1,z,i_1,i_2+1)(i_2+1)\mu_2=0,$$

$$\frac{\partial P(2,z,i_1,i_2)}{\partial z}-\frac{\partial P(2,0,i_1,i_2)}{\partial z}-$$
$$-(i_1\mu_1+i_2\mu_2)P(2,z,i_1,i_2)+\frac{\partial P(2,0,i_1,i_2-1)}{\partial z}p_{22}A_2(z)+$$
$$+\frac{\partial P(1,0,i_1,i_2-1)}{\partial z}p_{12}A_2(z)+P(2,z,i_1+1,i_2)(i_1+1)\mu_1+$$
$$+P(2,z,i_1,i_2+1)(i_2+1)\mu_2=0.$$

We introduce partial characteristic functions of the form

$$H(k,z,u_1,u_2)=\sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty}e^{ju_1i_1}e^{ju_2i_2}P(k,z,i_1,i_2),\ \text{where } j=\sqrt{-1}.$$

Let us write the system of differential equations for the partial characteristic functions

$$\frac{\partial H(1,z,u_1,u_2)}{\partial z}-\frac{\partial H(1,0,u_1,u_2)}{\partial z}+$$
$$+j\mu_1\left(1-e^{-ju_1}\right)\frac{\partial H(1,z,u_1,u_2)}{\partial u_1}+j\mu_2\left(1-e^{-ju_2}\right)\frac{\partial H(1,z,u_1,u_2)}{\partial u_2}+$$
$$+\frac{\partial H(1,0,u_1,u_2)}{\partial z}e^{ju_1}p_{11}A_1(z)+\frac{\partial H(2,0,u_1,u_2)}{\partial z}e^{ju_1}p_{21}A_1(z)=0,$$

$$\frac{\partial H(2,z,u_1,u_2)}{\partial z}-\frac{\partial H(2,0,u_1,u_2)}{\partial z}+$$
$$+j\mu_1\left(1-e^{-ju_1}\right)\frac{\partial H(2,z,u_1,u_2)}{\partial u_1}+j\mu_2\left(1-e^{-ju_2}\right)\frac{\partial H(2,z,u_1,u_2)}{\partial u_2}+$$
$$+\frac{\partial H(1,0,u_1,u_2)}{\partial z}e^{ju_2}p_{12}A_2(z)+\frac{\partial H(2,0,u_1,u_2)}{\partial z}e^{ju_2}p_{22}A_2(z)=0$$

with initial conditions

$$H(k,z,0,0)=r(k,z).$$

In vector-matrix form, this system will take the form

$$\frac{\partial \mathbf{h}(z,u_1,u_2)}{\partial z}+\frac{\partial \mathbf{h}(0,u_1,u_2)}{\partial z}\left(\mathbf{P}\mathbf{A}(z)\mathbf{B}(\mathbf{u})-\mathbf{I}\right)+$$
$$+j\mu_1\left(1-e^{-ju_1}\right)\frac{\partial \mathbf{h}(z,u_1,u_2)}{\partial u_1}+j\mu_2\left(1-e^{-ju_2}\right)\frac{\partial \mathbf{h}(z,u_1,u_2)}{\partial u_2}=0, \quad (8)$$

with initial conditions

$$\mathbf{h}(z,0,0)=\mathbf{r}(z),$$

where

$$\mathbf{h}(z,u_1,u_2)=\left[H(1,z,u_1,u_2),H(2,z,u_1,u_2)\right],$$

$$\mathbf{B}(\mathbf{u})=\begin{bmatrix}e^{ju_1} & 0\\ 0 & e^{ju_2}\end{bmatrix},\quad \mathbf{I}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}.$$

The resulting system of equations (8) is the main one for further research. Since it is not possible to find an explicit form of a solution to the problem (8), we will seek the solution under the asymptotic condition of equivalent growing of service times in the units of servers. This asymptotic condition means proportional growth of the average service times in both service units and it is taken from practice.

## 4. Asymptotic analysis of the first order

We denote $\mu_1=\epsilon$, $\mu_2=q\epsilon$, $q=\text{const}$ ($\epsilon$ is an infinitesimal quantity). Then we can write the asymptotic condition of equivalent growing of service times in the units of servers in the form $\mu_1,\mu_2\to 0$. In (8) we perform the replacements

$$\mathbf{h}(z,u_1,u_2)=\mathbf{f}(z,w_1,w_2,\epsilon),\ u_1=\epsilon w_1,\ u_2=\epsilon w_2,$$

we obtain the matrix equation for $\mathbf{f}(z,w_1,w_2,\epsilon)$

$$\begin{aligned}&\frac{\partial\mathbf{f}(z,w_1,w_2,\epsilon)}{\partial z}+\frac{\partial\mathbf{f}(0,w_1,w_2,\epsilon)}{\partial z}\left(\mathbf{PA}(z)\mathbf{B}(\mathbf{u},\epsilon)-\mathbf{I}\right)+\\ &+j\left(1-e^{-j\epsilon w_1}\right)\frac{\partial\mathbf{f}(z,w_1,w_2,\epsilon)}{\partial w_1}+jq\left(1-e^{-j\epsilon w_2}\right)\frac{\partial\mathbf{f}(z,w_1,w_2,\epsilon)}{\partial w_2}=0,\end{aligned}\tag{9}$$

**Theorem 1.** *The limiting solution for $\epsilon\to 0$ to the equation* (9) $\mathbf{f}(z,w_1,w_2,\epsilon)$ *has the form*

$$\mathbf{f}(z,w_1,w_2,\epsilon)=\mathbf{r}(z)\exp\left\{j\lambda\left(r_1w_1+\frac{r_2w_2}{q}\right)\right\},\tag{10}$$

*where $\mathbf{r}(z)=[r_1(z),r_2(z)]$ is the vector of the probability distribution of the values of the embedded Markov chain, $\mathbf{r}=[r_1,r_2]$ is vector of stationary probability distribution of the values of the embedded Markov chain.*

**Proof.** In the equation (9) we carry out the passage to the limit for $\epsilon\to 0$, we obtain that $\mathbf{f}(z,w_1,w_2)$ is a solution to the equation

$$\frac{\partial\mathbf{f}(z,w_1,w_2)}{\partial z}+\frac{\partial\mathbf{f}(0,w_1,w_2)}{\partial z}\left(\mathbf{PA}(z)-\mathbf{I}\right)=0,$$

which defines the vector function $\mathbf{r}(z)$, therefore we will seek the function $\mathbf{f}(z, w_1, w_2, \epsilon)$ in the form of the expansion

$$\mathbf{f}(z, w_1, w_2, \epsilon) = \mathbf{r}(z)\Phi(w_1, w_2) + o(\epsilon). \tag{11}$$

In the equation (9) we carry out the passage to the limit as $z \to \infty$, multiply this equation by the unit column vector $\mathbf{e}$, expand the exponents in a Maclaurin series up to the first order. In the resulting expression, we substitute the expansion (11), divide by $\epsilon$ and carry out the passage to the limit at $\epsilon \to 0$, we obtain the equation for the function $\Phi(w_1, w_2)$

$$w_1 \frac{\partial \Phi(w_1, w_2)}{\partial w_1} + q w_2 \frac{\partial \Phi(w_1, w_2)}{\partial w_2} = j \frac{\partial \mathbf{r}(0)}{\partial z} \mathbf{PWe} \Phi(w_1, w_2),$$

where $\dfrac{\partial \mathbf{r}(0)}{\partial z} = \lambda \mathbf{r}$, $\mathbf{rP} = \mathbf{r}$, $\mathbf{re} = 1$, $\lambda = \dfrac{1}{\int_0^\infty (1 - \mathbf{rA}(x)\mathbf{e})\, dx}$, $\mathbf{W} = \begin{bmatrix} w_1 & 0 \\ 0 & w_2 \end{bmatrix}$.
The solution will have the following form

$$\Phi(w_1, w_2) = \exp\left\{ j\lambda \left( r_1 w_1 + \frac{r_2 w_2}{q} \right) \right\}$$

Substituting the obtained solution into (11), we get (10).

The theorem is proved. □

By substitution and equality (3), we write down the approximate (asymptotic) equality

$$\mathbf{h}(z, u_1, u_2) \approx \mathbf{f}(z, w_1, w_2) = \mathbf{r}(z) \exp\left\{ j\lambda \left( r_1 w_1 + \frac{r_2 w_2}{q} \right) \right\} = \\ = \mathbf{r}(z) \exp\left\{ j\lambda \left( \frac{r_1 u_1}{\mu_1} + \frac{r_2 u_2}{\mu_2} \right) \right\}.$$

Let us define the characteristic of the process $\{i_1(t), i_2(t)\}$ in the stationary mode

$$h(u_1, u_2) = \exp\left\{ j\lambda \left( \frac{r_1 u_1}{\mu_1} + \frac{r_2 u_2}{\mu_2} \right) \right\},$$

which we will call the first-order asymptotics of the characteristic functions of the number of occupied servers in the system.

## 5. Asymptotic analysis of the second order

In the equation (8) we replace

$$\mathbf{h}(z, u_1, u_2) = \mathbf{h}_2(z, u_1, u_2) \exp\left\{ j\lambda \left( \frac{r_1 u_1}{\mu_1} + \frac{r_2 u_2}{\mu_2} \right) \right\},$$

we obtain the equation for $\mathbf{h}_2(z,u_1,u_2)$

$$\begin{aligned}&\frac{\partial \mathbf{h}_2(z,u_1,u_2)}{\partial z}+\frac{\partial \mathbf{h}_2(0,u_1,u_2)}{\partial z}\left(\mathbf{PA}(z)\mathbf{B}(\mathbf{u})-\mathbf{I}\right)+\\&+j\mu_1\left(1-e^{-ju_1}\right)\frac{\partial \mathbf{h}_2(z,u_1,u_2)}{\partial u_1}+j\mu_2\left(1-e^{-ju_2}\right)\frac{\partial \mathbf{h}_2(z,u_1,u_2)}{\partial u_2}-\\&-\lambda r_1\left(1-e^{-ju_1}\right)\mathbf{h}_2(z,u_1,u_2)-\lambda r_2\left(1-e^{-ju_2}\right)\mathbf{h}_2(z,u_1,u_2)=0.\end{aligned}\tag{12}$$

We denote $\mu_1=\epsilon^2$, $\mu_2=q\epsilon^2$, in (12) we replace

$$\mathbf{h}_2(z,u_1,u_2)=\mathbf{f}_2(z,w_1,w_2,\epsilon),\ u_1=\epsilon w_1,\ u_2=\epsilon w_2,$$

we obtain the equation for $\mathbf{f}_2(z,w_1,w_2,\epsilon)$

$$\begin{aligned}&\frac{\partial \mathbf{f}_2(z,w_1,w_2,\epsilon)}{\partial z}+\frac{\partial \mathbf{f}_2(0,w_1,w_2,\epsilon)}{\partial z}\left(\mathbf{PA}(z)\mathbf{B}(\mathbf{w},\epsilon)-\mathbf{I}\right)+\\&+j\epsilon\left(1-e^{-j\epsilon w_1}\right)\frac{\partial \mathbf{f}_2(z,w_1,w_2,\epsilon)}{\partial w_1}+j\epsilon q\left(1-e^{-j\epsilon w_2}\right)\frac{\partial \mathbf{f}_2(z,w_1,w_2,\epsilon)}{\partial w_2}-\\&-\lambda r_1\left(1-e^{-j\epsilon w_1}\right)\mathbf{f}_2(z,w_1,w_2,\epsilon)-\lambda r_2\left(1-e^{-j\epsilon w_2}\right)\mathbf{f}_2(z,w_1,w_2,\epsilon)=0.\end{aligned}\tag{13}$$

**Theorem 2.** *The limiting solution for $\epsilon\to 0$ to the equation* (13) $\mathbf{f}_2(z,w_1,w_2)$ *has the form*

$$\begin{aligned}\mathbf{f}_2(z,w_1,w_2)=\mathbf{r}(z)\exp\Bigg\{\frac{j^2}{2}\Bigg(&\lambda\left(r_1w_1^2+r_2\frac{w_2^2}{q}\right)+\\&+\kappa\left(r_1^2w_1^2+4r_1r_2\frac{w_1w_2}{q+1}+r_2^2\frac{w_2^2}{q}\right)\Bigg)\Bigg\},\end{aligned}\tag{14}$$

*where* $\kappa=\lambda^2\displaystyle\int_0^\infty\left(\mathbf{rA}(x)-\mathbf{r}(x)\right)\mathbf{e}dx$.

**Proof.** We will obtain the solution of the equation (14) in the following form

$$\mathbf{f}_2(z,w_1,w_2,\epsilon)=\Phi(w_1,w_2)\left(\mathbf{r}(z)+j\epsilon(r_1w_1+r_2w_2)\mathbf{f}_2(z)\right)+o^2(\epsilon),\tag{15}$$

where $\mathbf{f}_2(z)$ satisfies the condition $\mathbf{f}_2(\infty)\mathbf{e}=0$. Substitute (15) into (13) and expand the exponents in a series up to the first order. Considering that

$$\frac{\partial \mathbf{r}(z)}{\partial z}+\frac{\partial \mathbf{r}(0)}{\partial z}\left(\mathbf{PA}(z)-\mathbf{I}\right)=0,$$

we obtain the equation for finding the function $\mathbf{f}_2(z)$

$$\mathbf{e}\frac{\partial \mathbf{f}_2(z)}{\partial z}-\lambda\mathbf{er}(z)+\mathbf{e}\frac{\partial \mathbf{f}_2(0)}{\partial z}\left(\mathbf{PA}(z)-\mathbf{I}\right)+\lambda\mathbf{A}(z)=0.\tag{16}$$

From the equation (16) we find that

$$\frac{\partial \mathbf{f}_2(0)}{\partial z} = \kappa \mathbf{r}, \quad \text{where} \quad \kappa = \lambda^2 \int_0^\infty (\mathbf{r}\mathbf{A}(x) - \mathbf{r}(x))\,\mathbf{e}dx.$$

Substitute (15) into (13) and expand the exponents in a series up to the second order. Multiply by $\mathbf{e}$ and perform the passage to the limit $z \to \infty$, we obtain the equation for finding the function $\Phi(w_1, w_2)$

$$w_1 \frac{\partial \Phi(w_1, w_2)}{\partial w_1} + w_2 q \frac{\partial \Phi(w_1, w_2)}{\partial w_2} = \\ = \Phi(w_1, w_2)\left(-\lambda\left(r_1 w_1^2 + r_2 w_2^2\right) - \kappa\left(r_1 w_1 + r_2 w_2\right)^2\right). \quad (17)$$

The solution of the equation (17) has the form

$$\Phi(w_1, w_2) = \\ = \exp\left\{\frac{j^2}{2}\left(\lambda\left(r_1 w_1^2 + r_2 \frac{w_2^2}{q}\right) + \kappa\left(r_1^2 w_1^2 + 4 r_1 r_2 \frac{w_1 w_2}{q+1} + r_2^2 \frac{w_2^2}{q}\right)\right)\right\} \quad (18)$$

Substituting the solution (18) into (15) and performing the passage to the limit $\epsilon \to 0$, we obtain (14).

The theorem is proved. □

Due to the change, as well as the equality (14) for the function $\mathbf{h}_2(z, u_1, u_2)$ we can write down the approximate (asymptotic) equality

$$\mathbf{h}_2(z, u_1, u_2) \approx \mathbf{f}_2(z, w_1, w_2) = \\ = \mathbf{r}(z) \exp\left\{\frac{j^2}{2}\left(\lambda\left(r_1 \frac{u_1^2}{\mu_1} + r_2 \frac{u_2^2}{\mu_2}\right) + \right.\right. \\ \left.\left. + \kappa\left(r_1^2 \frac{u_1^2}{\mu_1} + 4 r_1 r_2 \frac{u_1 u_2}{\mu_1 + \mu_2} + r_2^2 \frac{u_2^2}{\mu_2}\right)\right)\right\}.$$

Thus, the characteristic function of the number of occupied servers in the system under consideration has the form

$$h_2(u_1, u_2) = \exp\left\{j\lambda\left(\frac{r_1 u_1}{\mu_1} + \frac{r_2 u_2}{\mu_2}\right) + \frac{j^2}{2}\left[\lambda\left(r_1 \frac{u_1^2}{\mu_1} + r_2 \frac{u_2^2}{\mu_2}\right) + \right.\right. \\ \left.\left. + \kappa\left(r_1^2 \frac{u_1^2}{\mu_1} + 4 r_1 r_2 \frac{u_1 u_2}{\mu_1 + \mu_2} + r_2^2 \frac{u_2^2}{\mu_2}\right)\right]\right\}. \quad (19)$$

## 6. Numerical example

Let us consider a numerical example where we can illustrate the accuracy of approximating formula (19). Consider queueing system with MRP arrivals,

where the Markov renewal process is given by matrices

$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{bmatrix}, \quad \mathbf{A}(x) = \operatorname{diag}\{A_1(x), A_2(x)\}.$$

Here $A_1(x)$ and $A_2(x)$ are gamma distribution cdf-s with shape and rate parameters $\alpha$ and $\beta$ that have the following values:

$$\alpha_1 = 0.5,\ \beta_1 = 0.25, \quad \alpha_2 = 1.5,\ \beta_2 = 1.5.$$

The service times are exponentially distributed with service rates

$$\mu_1 = 1 \cdot \varepsilon, \quad \mu_2 = 2 \cdot \varepsilon$$

for the the first and the second types of arrivals respectively. Here parameter $\varepsilon$ will be varied to establish the accuracy of approximation (19) accordingly to the asymptotic condition $\varepsilon \to 0$.

To establish the accuracy of the approximation, we use its comparison with the results of simulation modeling of the corresponding system. For the error estimation (difference between the results), we use the Kolmogorov distance

$$\Delta = \max_{i_1, i_2 \in [0,\infty)} \left| F_{\text{approx}}(i_1, i_2) - F_{\text{sim}}(i_1, i_2) \right|,$$

where $F_{\text{approx}}(i_1, i_2)$ is a cdf of Gaussian distribution (19) and $F_{\text{sim}}(i_1, i_2)$ is a cdf built basing on the results of the simulation. The results of the comparison is presented in the table 1. We see that the Kolmogorov distance decreases with decreasing of parameter $\varepsilon$, so, approximation (19) becomes more accurate for small values of this parameter.

Table 1

Kolmogorov distance $\Delta$ between the approximation and distribution based on the simulation results for various values of asymptotic parameter $\varepsilon$

| $\varepsilon$ | 0.1 | 0.05 | 0.01 | 0.005 | 0.001 | 0.0005 | 0.0001 |
|---|---|---|---|---|---|---|---|
| $\Delta$ | 0.1137 | 0.0501 | 0.0371 | 0.0323 | 0.0253 | 0.0226 | 0.0197 |

For example, if we suppose that error $\Delta \leqslant 0.05$ means that the approximation is accurate enough then we can conclude that for the considered example, Gaussian approximation (19) is applicable for values $\varepsilon < 0.05$.

## 7. Conclusions

In this paper, the method of asymptotic analysis is used to study a mathematical model of the $MR|M|\infty$ system functioning under the condition of a changing environment. The case is considered when a semi-Markov random environment has 2 different states. It is proved that the asymptotic characteristic function of the number of occupied servers of each type in the considered

system is Gaussian with the vector of mathematical expectations

$$\mathbf{a} = \left[\lambda\frac{r_1}{\mu_1}, \lambda\frac{r_2}{\mu_2}\right]$$

and the covariance matrix

$$\mathbf{K} = \begin{bmatrix} \lambda\dfrac{r_1}{\mu_1} + \kappa\dfrac{r_1^2}{\mu_1} & 4\kappa\dfrac{r_1 r_2}{\mu_1+\mu_2} \\ 4\kappa\dfrac{r_1 r_2}{\mu_1+\mu_2} & \lambda\dfrac{r_2}{\mu_2} + \kappa\dfrac{r_2^2}{\mu_2} \end{bmatrix}.$$

## References


[1] A. Dudin, V. Klimenok, and V. Vishnevsky, *The Theory of Queuing Systems with Correlated Flow*. Springer Nature, 2020. DOI: 10.1007/978-3-030-32072-0.

[2] V. K. Malinovskii, "Asymptotic expansions in the central limit theorem for recurrent Markov renewal processes," *Theory of Probability & Its Applications*, vol. 51, no. 3, pp. 523–526, 1987. DOI: 10.1137/1131073.

[3] Y. Lim, S. Hur, and J. Seung, "Departure process of a single server queueing system with Markov renewal input and general service time distribution," *Computers & Industrial Engineering*, vol. 51, no. 3, pp. 519–525, 2006. DOI: 10.1016/j.cie.2006.08.011.

[4] R. Pyke, "Markov renewal processes: definitions and preliminary properties," *Ann. Math. Statist.*, vol. 32, pp. 1231–1242, 1961. DOI: 10.1214/aoms/1177704863.

[5] R. Pyke and R. Schaufele, "Stationary measures for Markov renewal processes," *Ann. Math. Statist.*, vol. 37, pp. 1439–1462, 1966. DOI: 10.1214/aoms/1177699138.

[6] J. Sztrik and D. Kouvatsos, "Asymptotic analysis of a heterogeneous multiprocessor system in a randomly changing environment," *IEEE Transactions on Software Engineering*, vol. 17, no. 10, pp. 1069–1075, 1991. DOI: 10.1109/32.99194.

[7] E. P. Polin, S. P. Moiseeva, and S. V. Rozhkova, "Asymptotic analysis of heterogeneous queueing system $M|M|\infty$ in a Markov random enviroment [Asimptoticheskiy analiz neodnorodnoy sistemy massovogo obsluzhivaniya $M|M|\infty$ v markovskoy sluchaynoy srede]," *Tomsk State University Jounal of Control and Computer Science [Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika]*, vol. 47, pp. 75–83, 2019, in Russian. DOI: 10.17223/19988605/47/9.

[8] E. P. Polin, S. P. Moiseeva, and A. N. Moiseev, "Heterogeneous queueing system $MR(S)|M(s)|\infty$ with service parameters depending on the state of the underlying Markov chain [Analiz veroyatnostnykh kharakteristik geterogennoy SMO vida $MR(S)/M(S)/\infty$ s parametrami obsluzhivaniya, zavisyashchimi ot sostoyaniya vlozhennoy tsepi Markova]," *Saratov University News. New Series. Series Mathematics. Mechanics. Informatics [Izv. Saratov Univ. (N. S.), Ser. Math. Mech. Inform.]*, vol. 20, no. 3, pp. 388–399, 2020, in Russian. DOI: 10.18500/1816-9791-2020-20-3-388-399.

[9] B. D'Auria, "$M|M|\infty$ queues in semi-Markovian random environment," *Queueing Systems*, vol. 58, pp. 221–237, 2008. DOI: 10.1007/s11134-008-9068-7.

[10] H. M. Jansen, "A large deviations principle for infinite-server queues in a random environment," *Queueing Systems*, vol. 82, pp. 199–235, 2016. DOI: 10.1007/s11134-015-9470-x.

[11] J. Blom, M. Mandjes, and H. Thorsdottir, "Time-scaling limits for Markov-modulated infinite-server queues," *Stochastic Models*, vol. 29, pp. 112–127, 2012. DOI: 10.1080/15326349.2013.750536.

**For citation:**

E. P. Polin, S. P. Moiseeva, A. N. Moiseev, Heterogeneous queueing system with Markov renewal arrivals and service times dependent on states of arrival process, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 105–119. DOI: 10.22363/2658-4670-2023-31-2-105-119.

**Information about the authors**:

**Polin, Evgeny P.** — Assistant of Department of Probability Theory and Mathematical Statistics, National Research Tomsk State University (e-mail: polin_evgeny@mail.ru, phone: +7(923)4480077, ORCID: https://orcid.org/0000-0002-0250-2368)

**Moiseeva, Svetlana P.** — Doctor in Physics and Mathematics, Professor at Department of Probability Theory and Mathematical Statistics, National Research Tomsk State University (e-mail: smoiseeva@mail.ru, ORCID: https://orcid.org/0000-0001-9285-1555, Scopus Author ID: 56436490300)

**Moiseev, Alexander N.** — Doctor in Physics and Mathematics, Head of the Department of Software Engineering, National Research Tomsk State University (e-mail: moiseev.tsu@gmail.com, ORCID: https://orcid.org/0000-0003-2369-452X, ResearcherID: N-7189-2014, Scopus Author ID: 55646953800)

УДК 519.872
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-105-119

# Гетерогенная система массового обслуживания с входящим потоком марковского восстановления и временем обслуживания, зависящими от состояний вложенной цепи Маркова

**Е. П. Полин**[1,2], **С. П. Моисеева**[1], **А. Н. Моисеев**[1]

[1] *Национальный исследовательский Томский государственный университет, пр. Ленина, д. 36, Томск, 634050, Россия*
[2] *Национальный исследовательский Томский политехнический университет, пр. Ленина, д. 30, Томск, 634050, Россия*

**Аннотация.** В работе рассматривается гетерогенная система массового обслуживания с входящим потоком марковского восстановления и неограниченным числом серверов. Время обслуживания запросов на серверах является положительной случайной величиной с экспоненциальным распределением вероятностей. Параметры обслуживания зависят от состояния цепи Маркова в моменты восстановления. Следует отметить, что эти параметры не меняют своих значений до окончания обслуживания. Таким образом, устройства в рассматриваемой системе являются неоднородными (гетерогенными). Объектом исследования становится многомерный случайный процесс — количество серверов каждого типа, обслуживаемых с разной интенсивностью в стационарном режиме. Для исследования применён метод асимптотического анализа при условии эквивалентно долгого времени обслуживания. Метод асимптотического анализа реализуется при построении последовательности асимптотик возрастающего порядка, в которой асимптотика первого порядка определяет асимптотическое среднее значение числа занятых серверов. Асимптотика второго порядка позволяет построить гауссовскую аппроксимацию распределения вероятностей числа занятых серверов в системе.

**Ключевые слова:** система массового обслуживания, случайная среда, поток марковского восстановления, метод асимптотического анализа

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

2023, 31 (2) 120–127
http://journals.rudn.ru/miph

*Research article*

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-120-127
EDN: WIMGRX

# Convergence of the grid method for the Fredholm equation of the first kind with Tikhonov regularization

**Aleksandr A. Belov**[1, 2]

[1] *M. V. Lomonosov Moscow State University,*
*1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation*
[2] *RUDN University,*
*6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation*

(received: April 25, 2023; revised: May 5, 2023; accepted: June 26, 2023)

**Abstract.** The paper describes a grid method for solving an ill-posed problem for the Fredholm equation of the first kind using the A. N. Tikhonov regularizer. The convergence theorem for this method was formulated and proved. A procedure for thickening grids with a simultaneous increase in digit capacity of calculations is proposed.

**Key words and phrases:** ill-posed problems, grid method, regularization

## 1. Introduction

A large number of applied tasks are ill-posed. A number of methods have been developed to solve them. Firstly, these are parametric methods in which the solution is represented as a decomposition over some basis, and the regularized equation is reduced to the problem of optimizing the coefficients of the decomposition (see, for example [1–3]). The success of this approach strongly depends on the successful choice of the basis. Such methods are difficult to study; finding estimates of accuracy and conditionality in calculations with finite digit numbers is particularly difficult. Most of the proofs are carried out for exact calculations with infinite digit capacity, i.e., without round-off errors.

Secondly, iterative methods with simple or implicit iterations [4, 5] are often used to obtain an approximate analytical solution. The number of iterations is also a regularizing parameter [6]. This looks tempting, since there is no need to introduce additional stabilizing terms and thereby increase the discrepancy. On the other hand, in the general case, iterations have to be implemented numerically. The finite-difference approximation of the corresponding quadratures introduces some systematic error in the operator

and the right part. To reduce it, it is necessary to perform calculations on thickening grids.

The third approach is represented by various grid methods (finite-difference or finite-element), in which the solution is calculated in a set of discrete grid nodes, that is, essentially replaced by a piecewise constant function. In this approach, the initial problem is reduced to a system of algebraic equations that can be solved by any direct or iterative method [7, 8]. Yu. L. Gaponenko showed that finite-difference approximation makes the problem correct, i.e., self-regulation takes place [9, 10]. The study of finite element approximations (for specific applied problems) was carried out, for example, in [11, 12]. However, the proofs and convergence estimates are valid for calculations with infinite digit capacity, since they do not take into account rounding errors.

The central point of all regularizing algorithms is the justification of convergence and the evaluation of the actual accuracy, that is, the difference between the exact solution and the approximate one found. A review of the literature on this issue is given in [13]. Known a posteriori estimates are majorant and often greatly overestimate the error (up to 10 times or more). Quite often, they require specific information and solutions that are not easy to obtain in complex application tasks [14].

Another important issue is the choice of the regularization parameter. This problem is not trivial, since in most applied calculations the error level is fixed and does not tend to zero [15]. The best known solution to this question is the well-known generalized residual principle [16].

In the present paper, we describe a grid method for solving an ill-posed problem for the Fredholm equation of the first kind using the Tikhonov regularizer of the zeroth order. For this method, we formulate and prove convergence theorem which takes into account finite digit capacity of calculations. For its practical implementation, we propose procedure of simultaneous grid thickening and increase of digit capacity.

## 2. Method

We consider the Fredholm equation of the first kind

$$Au = f, \quad Au = \int_a^b K(y,x)u(x)dx, \quad y \in [c,d]. \tag{1}$$

A well-known technique of regularization is to add the simplest Tikhonov stabilizer to the residual [8]. This leads to the following optimization problem

$$\|Au - f\|_{L2}^2 + \alpha\|u\|_{L2}^2 \to \min. \tag{2}$$

Here, $\alpha > 0$ is a regularization parameter.

Minimizing (2) by $u$ leads to the Euler equation. In the case of a non-self-adjoint operator $A$, it has the form

$$\begin{gathered} \int\limits_a^b Q(z,x)u(x)dx + \alpha u = F(z), \quad z \in [a,b], \\ Q(z,x) = Q(x,z) = \int\limits_c^d K(y,x)K(y,z)dy, \quad F(z) = \int\limits_c^d K(y,z)f(y)dy. \end{gathered} \tag{3}$$

To solve (2), let us use convenient mesh method [17]. We introduce meshes on $x \in [a,b]$ and $y \in [c,d]$. For simplicity, they are supposed to be uniform and to have the same number of steps $N$. The grid steps of $x$ and $y$ are denoted by $h = (b-a)/N$ and $\tau = (d-c)/N$, respectively. Let us replace all integrals in (1) by quadrature rules (for definiteness, using trapezoid rule). This leads to the difference problem

$$\begin{gathered} \sum_{n=0}^{N} \left[(A^*A)_{k,n} + \alpha E_{k,n}\right] u_n = F_k, \quad 0 \leqslant k \leqslant N, \\ (A^*A)_{k,n} = \tau h g_n \sum_{m=0}^{M} g_m K_{m,k} K_{m,n}, \quad F_k = \tau \sum_{m=0}^{M} g_m K_{m,k} f_m. \end{gathered} \tag{4}$$

Here, $g$ are the weights of the trapezoid formula, $E_{k,n}$ is the unit matrix. The system of equations (4) is solved by some direct method.

## 3. Convergence

Let us formulate a few preliminary considerations.

***1***$^o$ When replacing integrals with grid approximations, we introduce some error. It can be considered as systematic. This error can be estimated using the Richardson method. This method is rigorously substantiated in [18]. Recall the essence of this approach.

In sequential twofold mesh thickening, even nodes of the current mesh coincide exactly with the nodes of the previous one. In these nodes, one can directly compute the difference of solutions on the sequential grids $\delta = u_{\text{fine}} - u_{\text{coarse}}$. The error estimation takes the form

$$r = \frac{\delta}{(2^p - 1)}, \tag{5}$$

where $p$ is the accuracy order of the scheme. We emphasize that this approach does not require any information on the derivatives of the exact solution and provides asymptotically precise (i.e. unimprovable) error value instead of majorant one.

The described procedure can be controlled by graphs of $\lg \|r\|_{l2}$ versus $\lg N$. If $N$ is too small, the plot behavior is irregular. For "moderate" $N$, the plot is a straight line with slope $-p$. On this section of the plot, Richardson method is applicable. For excessively large $N$, the plot sharply passes to a horizontal line. This means that the calculation has reached round-off error background caused by finite digit capacity. Here, Richardson method is inapplicable, and one should terminate the calculations.

*2°* The matrix of a linear system (4) is ill-conditioned. Calculations with finite digit capacity lead to a random error associated with round-off errors. With a sufficiently small step, the calculation error becomes comparable with round-off errors and ceases to decrease with further thickening of the grids. To reduce the impact of rounding errors, one needs to increase the digit capacity of calculations. Apparently, Richtmyer was the first to point this out in the 1950s [19]. He noted that any difference scheme is incorrect in the sense that when the grid step tends to zero, it is necessary to increase the digit capacity of calculations.

At the same time, theorems on regularizing properties are usually proved for exact calculations (i.e., with infinite digit capacity). However, real calculations are carried out on finite round-off errors. It often turns out that in ill-conditioned problems, computer round-off errors can become predominant.

*3°* The use of a regularizer improves the conditionality of the linear system matrix. Therefore, increasing $\alpha$ reduces the random error (for calculations with fixed bit depth). However, the regularizer itself introduces a systematic error in the problem, which increases with increasing $\alpha$.

Based on these suggestive considerations, we formulate the convergence theorem of the grid method (4). As far as we know, it is new.

**Theorem 1.** *For any precision $\varepsilon > 0$, there exist $\alpha_0 > 0$, step $h_0$ and digit capacity $K_0$ such that for $h < h_0$, $K > K_0$ and $\alpha = \alpha_0$ the error is less than $\varepsilon$.*

**Proof.** The proof consists of 3 stages. We write down the regularized Fredholm equation. For it, according to Tikhonov's fundamental theorem, there is a required value of $\alpha_0$.

By virtue of the Ryabenky–Fillipov theorems, there is such a $h_0$ that provides a systematic approximation error that does not exceed the required one for calculations with infinite digit capacity.

Since, for the selected grid step $h$, the conditionality of the linear system is known, then there is such a digit capacity that provides the required smallness of the random error. The theorem is proved. □

## 4. Calculation procedure

For the practical implementation of this theorem, the following algorithm is proposed. Let us set some $K$ and $\alpha$ and perform the calculation with grid thickening. On each grid, we calculate the error estimate using the Richardson method. We thicken the grids until this estimate stops decreasing. Denote the last solution obtained as the ***limiting*** one.

Let us perform such calculations for a wide range of $\alpha$ values. The dependence of the true error of the limiting solution (i.e., the difference between numerical and exact solutions) on $\alpha$ has the following qualitative form. For $\alpha = 0$, the error is very large due to poor conditionality of the matrix $A^*A$. For small $\alpha$, the random error is predominant, and the systematic error is negligible. As $\alpha$ increases, the random error decreases, and the systematic error, on the contrary, increases due to the term $\sim \alpha\|u\|^2$ in the regularizer. With some $\alpha$, the random and systematic errors become equal. This $\alpha$ corresponds to the best achievable accuracy at the selected bit depth.

The value of the random error is estimated as the product of the unit rounding error $\delta_0$ by the condition number $\kappa$ of a linear system (4). For calculations with 64-bit numbers, we have $\delta_0 = 10^{-16.2}$. To estimate $\kappa$, it is advisable to use the angular conditionality number [20]. As noted above, the systematic error consists of the grid approximation error (which is calculated using the Richardson method) and the regularizer contribution. In the zeroth approximation, these contributions can be considered independent. Therefore, according to the rules of statistics, the total value of the systematic error can be estimated as $\sqrt{\|r\|^2 + \alpha^2\|u\|^2}$.

As the final one, we choose such a $\alpha$, in which the estimates of random and systematic error are equal. If the obtained accuracy is unsatisfactory, one should increase the digit capacity and repeat the described calculations.

As far as we know, such calculation procedures with simultaneous thickening of grids and increasing digit capacity have not been proposed before.

## 5. Conclusion

Let us discuss possible generalizations of the proposed approaches. Firstly, the convergence theorem admits generalization to the case when the difference scheme is compiled not for a regularized problem, but for an initial ill-posed one. The absence of a regularizer reduces the systematic error. However, obviously, a significantly larger number of digits is required, which increases the complexity of the calculation.

Secondly, it is also advisable to use the procedure of thickening grids with a simultaneous increase in digit capacity for the numerical solution of formally correct, but ill-conditioned problems. Examples are stiff Cauchy problems with contrast structures. It is easy to construct a problem in which, when calculating 64-bit numbers, there is not a single correct sign in the answer [17]. Note that ill-conditionality and round-off errors are one of the important factors limiting the applicability of grid methods. Therefore, the relaxation of this restriction is of great practical interest.

## Acknowledgments

This work did not receive specific funding.

## References

[1] W. Jun-Gang, L. Yan, and R. Yu-Hong, "Convergence of Chebyshev type regularization method under Morozov discrepancy principle," *Applied Mathematics Letters*, vol. 74, pp. 174–180, 2017. DOI: 10.1016/j.aml.2017.06.004.

[2] A. A. Belov and N. N. Kalitkin, "Processing of Experimental Curves by Applying a Regularized Double Period Method," *Doklady Mathematics*, vol. 94, no. 2, pp. 539–543, 2016. DOI: 10.1134/S1064562416050100.

[3] A. A. Belov and N. N. Kalitkin, "Regularization of the double period method for experimental data processing," *Computational Mathematics and Mathematical Physics*, vol. 57, no. 11, pp. 1741–1750, 2017. DOI: 10.1134/S0965542517110033.

[4] A. B. Bakushinsky and A. Smirnova, "Irregular operator equations by iterative methods with undetermined reverse connection," *Journal of Inverse and Ill-posed Problems*, vol. 18, pp. 147–165, 2010. DOI: 10.1515/jiip.2010.005.

[5] A. B. Bakushinsky and A. Smirnova, "Discrepancy principle for generalized GN iterations combined with the reverse connection control," *Journal of Inverse and Ill-posed Problems*, vol. 18, pp. 421–431, 2010. DOI: 10.1515/jiip.2010.019.

[6] T. Jian-guo, "An implicit method for linear ill-posed problems with perturbed operators," *Mathematical Methods in the Applied Sciences*, vol. 18, pp. 1327–1338, 2006. DOI: 10.1002/mma.729.

[7] A. S. Leonov, *Solving ill-posed inverse problems: essay on theory, practical algorithms and Matlab demonstrations [Resheniye nekorrektno postavlennykh obratnykh zadach. Ocherk teorii, prakticheskiye algoritmy i demonstratsii v Matlab]*. Moscow: Librokom, 2010, in Russian.

[8] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. New York: Halsted, 1977.

[9] Y. L. Gaponenko, "On the degree of decidability and the accuracy of the solution of an ill-posed problem for a fixed level of error," *USSR Computational Mathematics and Mathematical Physics*, vol. 24, pp. 96–101, 1984. DOI: 10.1016/0041-5553(84)90092-2.

[10] Y. L. Gaponenko, "The accuracy of the solution of a non-linear ill-posed problem for a finite error level," *USSR Computational Mathematics and Mathematical Physics*, vol. 25, pp. 81–85, 1985. DOI: 10.1016/0041-5553(85)90076-X.

[11] Y. Hon and T. Wei, "Numerical computation of an inverse contact problem in elasticity," *Journal of Inverse and Ill-posed Problems*, vol. 14, pp. 651–664, 2006. DOI: 10.1515/156939406779802004.

[12] H. Ben Ameur and B. Kaltenbacher, "Regularization of parameter estimation by adaptive discretization using refinement and coarsening indicators," *Journal of Inverse and Ill-posed Problems*, vol. 10, pp. 561–583, 2002. DOI: 10.1515/jiip.2002.10.6.561.

[13] A. B. Bakushinsky and A. S. Leonov, "New a posteriori error estimates for approximate solutions to irregular operator equations [Novyye aposteriornyye otsenki pogreshnosti priblizhennykh resheniy neregulyarnykh operatornykh uravneniy]," *Vychisl. Metody Programm.*, vol. 15, pp. 359–369, 2014, in Russian.

[14] A. B. Bakushinsky, A. Smirnova, and L. Hui, "A posteriori error analysis for unstable models," *Journal of Inverse and Ill-posed Problems*, vol. 20, pp. 411–428, 2012. DOI: 10.1515/jip-2012-0006.

[15] M. V. Klibanov, A. B. Bakushinsky, and L. Beilina, "Why a minimizer of the Tikhonov functional is closer to the exact solution than the first guess," *Journal of Inverse and Ill-posed Problems*, vol. 19, pp. 83–105, 2011. DOI: 10.1515/jiip.2011.024.

[16] A. V. Goncharskii, A. S. Leonov, and A. G. Yagola, "A generalized discrepancy principle," *USSR Computational Mathematics and Mathematical Physics*, vol. 13, no. 2, pp. 25–37, 1973. DOI: 10.1016/0041-5553(73)90128-6.

[17] A. A. Belov and N. N. Kalitkin, "Solution of the Fredholm Equation of the First Kind by the Mesh Method with Tikhonov Regularization," *Mathematical Models and Computer Simulations*, vol. 11, pp. 287–300, 2018. DOI: 10.1134/S2070048219020042.

[18] V. S. Ryabenkii and A. F. Fillipov, *On stability of difference equations [Ob ustoychivosti raznostnykh uravneniy]*. Moscow: Gos. Izdat. Tekh.-Teor. Liter., 1956, in Russian.

[19] R. D. Richtmyer and K. W. Morton, *Difference methods for initial-value problems*. New York: Interscience publishers, 1967.

[20] N. N. Kalitkin, L. F. Yuhno, and L. V. Kuzmina, "Quantitative criterion of conditioning for systems of linear algebraic equations," *Mathematical Models and Computer Simulations*, vol. 3, pp. 541–556, 2011. DOI: 10.1134/S2070048211050097.

**For citation:**

A. A. Belov, Convergence of the grid method for the Fredholm equation of the first kind with Tikhonov regularization, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 120–127. DOI: 10.22363/2658-4670-2023-31-2-120-127.

**Information about the authors**:

**Belov, Aleksandr A.** — Candidate of Physical and Mathematical Sciences, Researcher of Faculty of Physics, M. V. Lomonosov Moscow State University; Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN 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)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-120-127
EDN: WIMGRX

# Сходимость сеточного метода для уравнения Фредгольма первого рода с регуляризацией по Тихонову

**А. А. Белов**[1, 2]

[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*

**Аннотация.** В статье описан сеточный метод решения некорректной задачи для уравнения Фредгольма первого рода с использованием регуляризатора А. Н. Тихонова. Сформулирована и доказана теорема о сходимости этого метода. Для её практической реализации предложена процедура сгущения сеток с одновременным увеличением разрядности вычислений.

**Ключевые слова:** некорректные задачи, сеточный метод, регуляризация

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

2023, 31 (2) 128–138
http://journals.rudn.ru/miph

*Research article*

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-128-138
EDN: XAUSJA

# Quadratures with super power convergence

**Aleksandr A. Belov**[1,2], **Maxim A. Tintul**[1], **Valentin S. Khokhlachev**[1]

[1] *M. V. Lomonosov Moscow State University,*
*1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation*
[2] *RUDN University,*
*6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation*

(received: April 25, 2023; revised: May 5, 2023; accepted: June 26, 2023)

**Abstract.** The calculation of quadratures arises in many physical and technical applications. The replacement of integration variables is proposed, which dramatically increases the accuracy of the formula of averages. For infinitely smooth integrand functions, the convergence law becomes super power. It is significantly faster than the power law and is close to exponential one. For integrals with bounded smoothness, power convergence is realized with the maximum achievable order of accuracy.

**Key words and phrases:** trapezoid rule, exponential convergence, error estimate, asymptotically sharp estimates

## 1. Introduction

**Applied tasks.** In many physical problems, it is required to approximate integrals that are not taken in elementary functions. Here are some examples:

1. Calculation of special functions of mathematical physics: Fermi–Dirac functions equal to the moments of the Fermi distribution, gamma function, cylindrical functions and a number of others.
2. Calculation of Fourier coefficients of a given function, Fourier and Laplace transforms.
3. Numerical solution of integral equations, both correctly posed and incorrect.
4. Solving boundary value problems for partial differential equations (including eigenvalue problems) written in integral form, etc.

Such integrals must be calculated with high accuracy up to computer round-off errors.

**Calculation of quadratures.** Commonly, trapezoid, mean and Simpson methods on a uniform grid are used for grid calculation of quadratures. The majorant error estimation is well known for these methods. For trapezoid and mean formulas it is $O(h^2)$, for Simpson's formula it is $O(h^4)$, where $h$ is the grid step. There are ways to improve accuracy: calculation on a set of thickening grids and extrapolation refinement by the Richardson method, refinement by the Euler–Maclaurin formula, etc. [1, 2]. All these methods give a power dependence of the error on the grid step $O(h^m)$.

If the integrand is periodic and the integral is calculated over the full period, then the dependence of the error on the step becomes exponential instead of power-law $\sim \exp(-1/h)$ [3–5]. This means that when the step is halved, the number of correct characters in the answer approximately doubles. This convergence rate is much faster than the power one. However, the corresponding class of integrand functions is rather narrow. Attempts have been made in the literature to expand this class [6–9], but they were considered unsuccessful [7].

**In the present paper,** an approach is proposed that dramatically accelerates the convergence of the mean rule. It is based on a special substitution of integration variables. The integrand function may be non-periodic. If it is infinitely smooth, then the proposed replacement provides super power convergence of the quadrature. This convergence rate is significantly faster than the power-law one and is close to the exponential one.

If the integrand has bounded smoothness, then the proposed method gives a power convergence with the maximum achievable order of accuracy.

The proposed approach does not require a priori information about the nature of the integrand function and is uniformly applicable to a wide range of tasks. The class of integrand functions, for which the super power convergence of quadratures is realized, is significantly expanded.

## 2. Change of integration variables

Consider the integral

$$I = \int_0^1 f(x)dx. \tag{1}$$

Let us perform the variable change in two stages. First, using fractional polynomial transformation $t(x)$, we map the segment $x \in (0, 1)$ to the straight line $t \in (-\infty, +\infty)$. Then we map this line to the segment $\xi \in (0, 1)$ using the transformation $t(\xi)$, whose derivatives tend to zero near $\xi = 0$ and $\xi = 1$ faster than any degree $\xi^m$.

Such substitutions can be made in various ways. In this paper, the following transformation was considered

$$t(\xi) = \frac{A(\xi - 0.5)}{\xi^\alpha (1-\xi)^\alpha}, \quad x(t) = \frac{1}{2} + \frac{1}{2}\mathrm{th}(Bt), \tag{2}$$

where $A, B, \alpha$ are constants. The mapping (2) is shown in figure 1 as $x(\xi)$ dependence. It is almost linear in the middle of the segment, but at its ends,

the derivatives of $x_\xi$ quickly tend to zero. It is also possible to implement the replacement (2), in which the error function $\Phi(Bt)$ is taken instead of the hyperbolic tangent.

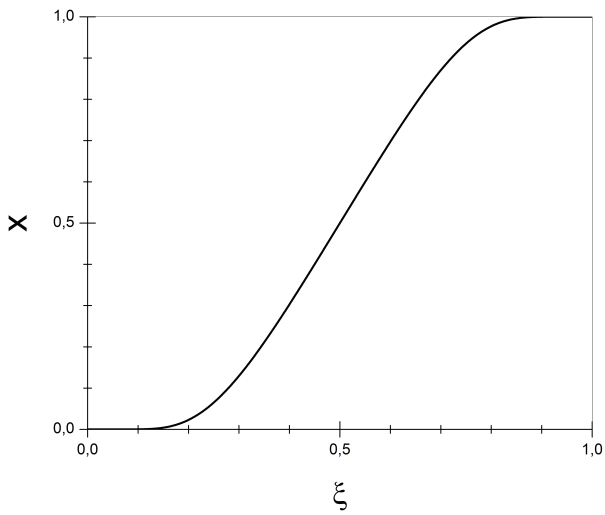


Figure 1. Variable transformation (2). Parameters $A$, $B$, $\alpha$ are equal to unity

After mapping (2), the integral takes the form

$$I = \int_0^1 \tilde{f}(\xi)d\xi, \quad \tilde{f}(\xi) = f\{x[t(\xi)]\}x_t[t(\xi)]t_\xi(\xi). \tag{3}$$

**Periodic continuation.** Let us show that the new integrand $\tilde{f}(\xi)$ admits an infinitely smooth periodic continuation beyond the boundaries of the segment $\xi \in (0, 1)$.

The expression $t_\xi(\xi) \sim \xi^{-\alpha-1}(1-\xi)^{-\alpha-1}$ has poles at the ends of the segment $\xi = 0$ and $\xi = 1$. However, for $\xi \to 0+0$ and $\xi \to 1-0$, the derivative $x_t \sim \exp(-\xi^{-\alpha}(1-\xi)^{-\alpha})$ tends to zero significantly faster. As a result, $x_t t_\xi \to 0$ when striving for points $\xi = 0$ and $\xi = 1$ from inside the segment. Therefore, $\tilde{f}(\xi)$ vanishes at the boundaries of the segment.

Similarly, it can be shown that all derivatives of this function tend to zero at $\xi \to 0+0$ and $\xi \to 1-0$. For example, the first derivative has the form

$$\tilde{f}_\xi = f_x x_t^2 t_\xi^2 + f x_{tt} t_\xi^2 + f x_t t_{\xi\xi}. \tag{4}$$

All derivatives of $dt^m/d\xi^m \sim \xi^{-\alpha-m}(1-\xi)^{-\alpha-m}$ at the boundaries of the segment have poles that are multiplied by the expression $\sim \exp(-\xi^{-\alpha}(1-\xi)^{-\alpha})$ in various degrees. Therefore, for $\xi \to 0+0$ and $\xi \to 1-0$ we have $\tilde{f}_\xi \to 0$.

The same is true for higher derivatives $d\tilde{f}_{\xi}^{m} \Big/ d\xi^{m}$. Thus, the integrand function $\tilde{f}$ can be periodically continued infinitely smoothly beyond the boundaries of the segment $\xi \in (0,1)$.

## 3. Mean rule convergence

On the segment $\xi \in (0,1)$, we introduce a uniform grid with a step $h = 1/N$. Half-integer nodes are denoted by $\xi_{n+1/2} = (n-1/2)h$, $n = 1, ..., N$. We write the mean rule quadrature

$$I_N = \sum_{n=1}^{N} h\tilde{f}(\xi_{n+1/2}). \tag{5}$$

The following statement holds.

**Theorem 1.**
*A) If $f(x)$ is infinitely smooth on the segment $x \in (0,1)$, then the quadrature* (5) *has super power convergence.*

*B) If $f(x)$ has $j$ continuous derivatives on $x \in (0,1)$, the $(j+1)$-th derivative has a discontinuity at the point $x = a \in (0,1)$, and this point is a grid node, then the quadrature* (5) *has power convergence. The order of accuracy is $j+2$ if $j$ is even, and $j+3$ if $j$ is odd. This order of accuracy is maximal for a given smoothness of the integrand function.*

**Proof.** Let us prove the statement A). The power part of the mean rule error is described by the Euler–Maclaurin formula [1]. It contains the differences of odd derivatives at the ends of the integration segment

$$\delta = \sum_{k=1}^{\infty} b_k h^{2k} \left( \tilde{f}^{(2k-1)}(1) - \tilde{f}^{(2k-1)}(0) \right), \quad b_k = \text{const}. \tag{6}$$

As noted above, due to the variable transform (2), the derivatives $\tilde{f}^{(k)}(\xi) \to 0$ for $\xi \to 0+0$ and $\xi \to 1-0$. All summands in the sum of (6) vanish. Therefore, there are no power terms left in the error of the mean rule, and the convergence turns out to be super power one.

Let us prove the statement B). Under these assumptions, the power-law contribution to the error of the mean formula has the form

$$\delta = \sum_{k=1}^{\infty} b_k h^{2k} \left( \tilde{f}^{(2k-1)}(1) - \tilde{f}^{(2k-1)}(0) \right) + \\ + \sum_{k=1}^{K} b_k h^{2k} \left( \tilde{f}^{(2k-1)}(a-0) - \tilde{f}^{(2k-1)}(a+0) \right). \tag{7}$$

The first sum in (7) is similar to (6). After the variable transform (2), it turns to zero.

The second sum is the error resulting from the singularity at the point $a$. If $2k-1 \leqslant j$, then by virtue of continuity, the right and left limit values of

derivatives of the order of $2k-1$ are the same $\tilde{f}^{(2k-1)}(a-0) = \tilde{f}^{(2k-1)}(a+0)$. What is the limit of summation of $K$? Since $\tilde{f}^{(j+1)}$ is discontinuous at the point $a$, and only odd derivatives are included in (7), two cases are possible. If $j$ is odd, then $2K-1 = j+2$. Then $\delta = O(h^{j+3})$. If $j$ is even, then $2K-1 = j+1$, and $\delta = O(h^{j+2})$. Obviously, this order of accuracy is the maximum, i.e. it cannot be improved. The theorem is proved. □

**Note.** The literature describes [6–9] variable substitutions similar to (2). In these works, trapezoid and Simpson formulas were used, in which one needs to calculate the integrand function at the boundary points. However, after variable change (2), the integral function $\tilde{f}(\xi)$ has essentially singular points within the boundaries of the segment $\xi = 0$ and $\xi = 1$. Therefore, calculating $\tilde{f}(0)$ and $\tilde{f}(1)$ presents a problem; in particular, computer numbers overflow occurs.

To avoid this, in [6] it was proposed to cut the integration segment, i.e. instead of $\xi \in (0,1)$, consider $\xi \in (\varepsilon, 1-\varepsilon)$, where $\varepsilon$ is some small number. Such a cutting introduced a significant error, and it was not possible to realize superstellar convergence. The authors of [7] conducted numerical experiments and found that this approach is inferior in quantitative accuracy to Simpson's formula without replacing variables. Therefore, this approach was considered unpromising [7].

We use the mean rule that does not require calculating $\tilde{f}(0)$ and $\tilde{f}(1)$. Therefore, the described difficulty does not arise, and super power convergence is realized.

## 4. Method validation

**Infinitely smooth integrand.** As an example, consider a test integral with a known exact value

$$I = \int_0^1 e^x/(e-1)dx = 1. \tag{8}$$

The integrand is infinitely smooth.

The calculation was carried out on a set of grids with different $N = 2, 4, 8, ...$ On each grid, the mean rule quadrature and its error $\Delta = |I - I_N|$, equal to the difference between the numerical and exact integrals, were calculated. Figure 2 shows a graph of the error $\Delta$ depending on the number of grid steps $N$. The scale of the graph is semi-logarithmic. At this scale, exponential convergence corresponds to a straight line, and a power-law curve corresponds to a logarithmic curve.

Dark circles correspond to the calculation with the replacement of variables (2), light circles correspond to the calculation without it. One can see that the proposed replacement of variables dramatically increases accuracy: already at $N \sim 100$, the error is $\Delta \sim 10^{-14}$, which is comparable to rounding errors. The gain in accuracy compared to the calculation without replacing variables reaches 10 orders of magnitude. The convergence rate is somewhat inferior to the exponential one, but cardinally exceeds the power one.

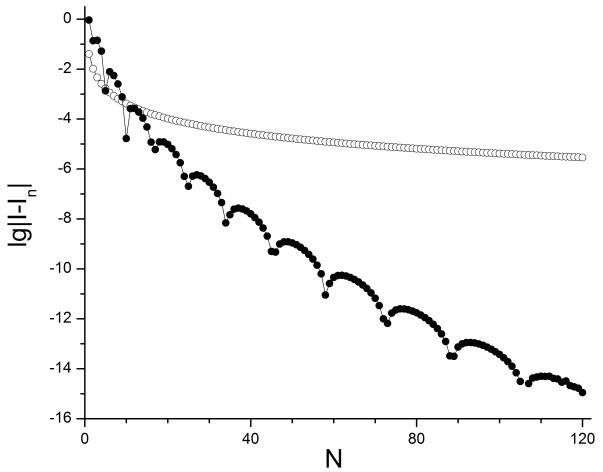


Figure 2. The error of the mean rule quadrature in the test (8). Notation are explained in the text

Due to the presence of essentially singular points of $\tilde{f}$ at the boundaries of the segment, the dependence of the error on the number of steps is non-monotonic and alternating [10, 11]. In this graph, this can be seen by the non-monotonic behavior of the curve. Local minima correspond to the change of the error sign.

Therefore, the proposed replacement dramatically increases the accuracy of the mean rule quadrature. We recommend it for wide application in practical computing.

**Integrand function with bounded smoothness.** Often in applications, it is necessary to calculate integrals from piecewise given spline approximations and interpolants. They have limited smoothness. So, the simplest linear interpolation is continuous, but has discontinuities of the first derivative. The cubic spline is continuous along with the second derivative, and the third derivative experiences a discontinuity.

As an example, consider the integral of the function

$$f(x) = \begin{cases} 1, & x < 0.5, \\ 1 + (2x-1)^m, & x \geqslant 0.5, \end{cases} \tag{9}$$

for integers $1 \leqslant m \leqslant 5$. The function (9) has a $m-1$ continuous derivative, and the $m$-th derivative experiences a discontinuity. The exact values of the integral $I$ are known, they are listed in table 1.

The calculation was carried out on several thickening grids. They were chosen so that the feature $x = 0.5$ was a node. For example, it is enough to take only even $N$ for this. The resulting errors depending on the number of steps are shown in figure 3. The scale of the graph is double logarithmic.

Table 1

Test (9)

| $m$ | $I$ | $q$ |
|---|---|---|
| 1 | $1 + 2e^{0.5} - e$ | 2 |
| 2 | $1 - 8e^{0.5} + 5e$ | 4 |
| 3 | $1 + 48e^{0.5} - 29e$ | 4 |
| 4 | $1 - 384e^{0.5} + 233e$ | 6 |
| 5 | $1 + 3840e^{0.5} - 2329e$ | 6 |

Therefore, the power convergence corresponds to a straight line whose slope is equal to the order of accuracy. The numbers near the lines are the values of $m$.

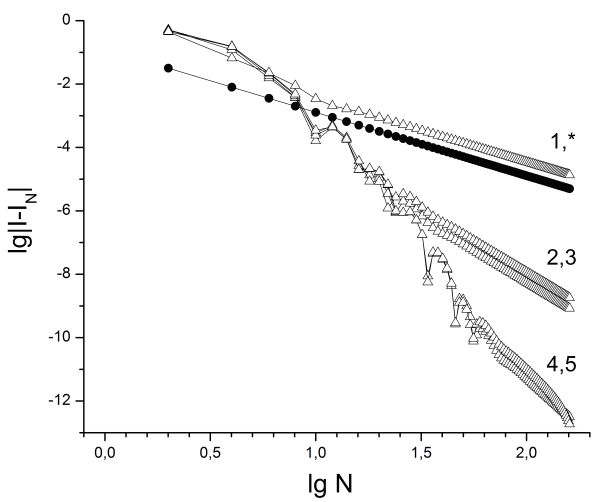


Figure 3. The error of the mean rule quadrature in the test 9. Notation are explained in the text

It can be seen that on sufficiently detailed grids, the curves for each $m$ tend to straight lines, i.e. power convergence is realized. The corresponding orders of accuracy of $q$ are given in the table 1. They are completely consistent with Theorem 1.

On coarse meshes, the behavior of curves is irregular. The error depends on $N$ nonmonotonically, changes sign and decreases significantly faster than the power law. Apparently, the law of convergence on coarse meshes corresponds to the super power one (similar to the figure 2). Justification of this consideration is beyond the scope of the present work.

For comparison, the figure 3 shows the calculation error according to the mean rule without variable change. For all the $m$ considered, the errors were approximately the same, so we showed them with one line. It is indicated by an asterisk (*). This line corresponds to a power convergence with a second order of accuracy. It can be seen that for $m \geqslant 2$ (i.e., if there is at least one continuous derivative), the proposed replacement increases the order of accuracy and sharply reduces the quantitative error. In Fig. 3, the accuracy gain was from 3 to 8 orders of magnitude.

We performed a similar calculation using grids in which the singularity $x = 0.5$ did not get into the node. To do this, it is enough to take only odd $N$. This case does not fall under Theorem 1. Nevertheless, the theorem turned out to be true for it as well. The resulting errors were similar to the figure 3. In particular, the convergence rate for the considered $m$ turned out to be the same. The quantitative accuracy was somewhat worse than figure 3. This was most noticeable for $m = 1$. For other $m$, the accuracy decreasing turned out to be insignificant.

## 5. Conclusion

In this paper, a special transformation of the integration variable is proposed, which dramatically increases the accuracy of the mean rule quadrature. For infinitely smooth integrand functions, convergence becomes super power one. For functions of bounded smoothness, the convergence law remains power-law, but the maximum achievable order of accuracy is realized.

Let us conduct a qualitative comparison of the proposed approach with other methods for improving the accuracy of quadratures listed in the introduction.

None of them provides super power convergence. Therefore, for infinitely smooth functions, the proposed approach provides obviously higher accuracy.

The use of Euler–Maclaurin corrections requires a large amount of a priori information about the integrand. It is necessary to accurately calculate high derivatives and a priori set the number of corrections to be taken into account. Therefore, the maximum order of accuracy is realized if the smoothness class of the integrand function is known.

On the contrary, the proposed approach is uniformly applicable to integrals both infinitely smooth and having bounded smoothness. One does not need to know the smoothness class in advance.

It is possible to increase the order of accuracy using Richardson extrapolation only on sufficiently detailed grids on which theoretical convergence is already being implemented, but rounding errors have not yet been achieved. On coarse grids, the use of extrapolation can even degrade accuracy.

In the proposed method, even on coarse grids, convergence is observed, and quite fast. A quantitative comparison of the Richardson extrapolation and the proposed method for bounded smoothness functions is beyond the scope of this paper.

## Acknowledgments

This work did not receive specific funding.

## References

[1] N. N. Kalitkin and E. A. Alshina, *Numerical Methods. Vol. 1: Numerical Analysis [Chislennye Metody. T. 1: Chislennyi analiz]*. Moscow: Akademiya, 2013, in Russian.

[2] N. N. Kalitkin, A. B. Alshin, E. A. Alshina, and V. B. Rogov, *Computations with Quasi-Uniform Grids [Vychisleniya na kvaziravnomernykh setkakh]*. Moscow: Fizmatlit, 2005, in Russian.

[3] L. N. Trefethen and J. A. C. Weideman, "The exponentially convergent trapezoidal rule," *SIAM Review*, vol. 56, no. 3, pp. 385–458, 2014. DOI: 10.1137/130932132.

[4] N. N. Kalitkin and S. A. Kolganov, "Quadrature formulas with exponential convergence and calculation of the Fermi–Dirac integrals," *Doklady Mathematics*, vol. 95, no. 2, pp. 157–160, 2017. DOI: 10.1134/S1064562417020156.

[5] N. N. Kalitkin and S. A. Kolganov, "Computing the Fermi–Dirac functions by exponentially convergent quadratures," *Mathematical Models and Computer Simulations*, vol. 10, no. 4, pp. 472–482, 2018. DOI: 10.1134/S2070048218040063.

[6] T. Sag and G. Szekeres, "Numerical evaluation of high-dimensional integrals," *Math. Comp.*, vol. 18, pp. 245–253, 1964. DOI: 10.1090/S0025-5718-1964-0165689-X.

[7] A. Sidi, "Numerical evaluation of high-dimensional integrals," *International Series Numer. Math.*, vol. 112, pp. 359–373, 1993. DOI: 10.1007/978-3-0348-6338-4_27.

[8] M. Iri, S. Moriguti, and Y. Takasawa, "On a certain quadrature formula," *International Series Numer. Math.*, vol. 17, pp. 3–20, 1987. DOI: 10.1016/0377-0427(87)90034-3.

[9] M. Mori, "An IMT-Type Double Exponential Formula for Numerical Integration," *Publ. Res. Inst . Math. Sci. Kyoto Univ.*, vol. 14, no. 3, pp. 713–729, 1978. DOI: 10.2977/prims/1195188835.

[10] A. A. Belov, N. N. Kalitkin, and V. S. Khokhlachev, "Improved error estimates for an exponentially convergent quadratures [Uluchshennyye otsenki pogreshnosti dlya eksponentsial'no skhodyashchikhsya kvadratur]," *Preprints of IPM im. M. V. Keldysh*, no. 75, 2020, in Russian. DOI: 10.20948/prepr-2020-75.

[11] V. S. Khokhlachev, A. A. Belov, and N. N. Kalitkin, "Improvement of error estimates for exponentially convergent quadratures [Uluchsheniye otsenok pogreshnosti dlya eksponentsial'no skhodyashchikhsya kvadratur]," *Izv. RAN. Ser. fiz.*, vol. 85, no. 2, pp. 282–288, 2021, in Russian. DOI: 10.31857/S0367676521010166.

**For citation:**

A. A. Belov, M. A. Tintul, V. S. Khokhlachev, Quadratures with super power convergence, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 128–138. DOI: 10.22363/2658-4670-2023-31-2-128-138.

**Information about the authors**:

**Belov, Aleksandr A.** — Candidate of Physical and Mathematical Sciences, Researcher of Faculty of Physics, M. V. Lomonosov Moscow State University; Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN 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)

**Tintul, Maxim A.** — Master's Degree Student of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: maksim.tintul@mail.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-5466-1221)

**Khoklachev, Valentin S.** — Master's Degree Student of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: valentin.mycroft@yandex.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-6590-5914)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-128-138
EDN: XAUSJA

# Квадратуры со сверхстепенной сходимостью

**А. А. Белов**[1, 2], **М. А. Тинтул**[1], **В. С. Хохлачев**[1]

[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*

**Аннотация.** Вычисление квадратур возникает во многих физических и технических приложениях. В статье предложена замена переменных интегрирования, кардинально повышающая точность формулы средних. Для бесконечно гладких подынтегральных функций закон сходимости становится сверхстепенным. Он существенно быстрее степенного и близок к экспоненциальному. Для подынтегральных функций с ограниченной гладкостью реализуется степенная сходимость с максимально достижимым порядком точности.

**Ключевые слова:** формула трапеций, экспоненциальная сходимость, оценки точности, асимптотически точные оценки

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

2023, 31 (2) 139–149
http://journals.rudn.ru/miph

*Research article*

UDC 004.92:004.928:519.67
DOI: 10.22363/2658-4670-2023-31-2-139-149
EDN: XKNIYV

# Asymptote-based scientific animation

**Migran N. Gevorkyan**[1],
**Anna V. Korolkova**[1], **Dmitry S. Kulyabov**[1, 2]

[1] *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: April 5, 2023; revised: May 5, 2023; accepted: June 26, 2023)

**Abstract.** This article discusses a universal way to create animation using Asymptote the language for vector graphics. The Asymptote language itself has a built-in library for creating animations, but its practical use is complicated by an extremely brief description in the official documentation and unstable execution of existing examples. The purpose of this article is to eliminate this gap. The method we describe is based on creating a PDF-file with frames using Asymptote, with further converting it into a set of PNG images and merging them into a video using FFmpeg. All stages are described in detail, which allows the reader to use the described method without being familiar with the used utilities.

**Key words and phrases:** vector graphics, TeX, asymptote, scientific graphics

## 1. Introduction

In this paper we study the creation of animation animation using the vector graphics language Asymptote [1–4].

Asymptote is an interpreted language, that is a translator into the PostScript vector graphics language. Designed to create vector images for mathematical publications. It is closely integrated with the TEX system and is an integral part of the TEX Live [5] distribution. It has a `C`-like syntax, supports the creation of functions, custom data structures, and comes with an extensive set of modules for various tasks. Unlike PGF/TikZ [6], Asymptote is more imperative, so it is easier to implement complex program logic on it.

In the official documentation of this language, only a few paragraphs are devoted to the animation creation process and the user is referred to the source code examples located in the `animations` directory. Asymptote creates animation in two steps. At the first step, a multi-page PDF-file is created containing images that will become frames of future animation. Then, using the external utility ImageMagick [7] (command `convert`), this PDF-file is

converted into a GIF image. If the ImageMagick utility is not installed on the user's system, all examples will stop at creating a multi-page PDF-file with a set of images and a GIF image with animation will not be received.

In this article, we are considering a universal way to create animation in video format using the ffmpeg [8, 9] and Ghostscript [10] utilities. All external programs will be called explicitly from the command line. With the help of Asymptote, only a multi-page PDF-file with frames for the future video will be created.

The reader should be familiar with the basic capabilities of the Asymptote language. For an introduction to the basics of the language, we recommend the manual [11]. The information from it will be enough to understand this work. As an example, we chose the animation of the process of constructing epitrochoids and hypotrochoids. In the first part of the work, we will recall the definitions of these curves, some of their properties and reduce their construction to a composition of two rotations. In the second part of the article, we will describe in detail the implementation of their construction using Asymptote. And in the third part we will focus on the technical side of the issue and describe the process of creating a multi-page PDF-file, converting it into PNG images using Ghostscript and converting these images into video using ffmpeg.

## 2. Task description

Consider the task of animating the process of constructing cycloidal curves, namely hypotrochoids and epitrochoids. We will not use the parametric equation of these curves, but will reduce everything to the composition of two rotations applied to the starting point of the curve. This will better illustrate the capabilities of the Asymptote language.

### 2.1. Definition of epitrochoids

*Epitrochoid* is defined as a trajectory plotted by a fixed point $P$ lying on a radial line of circle with radius $r$, which rolls along the *outer side* of the circle with radius $R$ (figure 1). The parametric equation of the curve has the following form:

$$\begin{cases} x(t) = (R+r)\cos(\varphi) - d\cos\left(\dfrac{R+r}{r}\varphi\right), \\ y(t) = (R+r)\sin(\varphi) - d\sin\left(\dfrac{R+r}{r}\varphi\right), \end{cases}$$

where $d$ is the distance from the center of the rolling circle to the point of the curve, $\varphi$ is the angle of rotation of the rolling circle relative to the axis $Ox$.

Let us introduce the coefficient $k = r/R$, then it will possible to change the parameterization and the equation will take the form:

$$\begin{cases} x(t) = R(k+1)\cos(kt) - d\cos((k+1)t), \\ y(t) = R(k+1)\sin(kt) - d\sin((k+1)t), \end{cases}$$

where the parameters $t$ and $\varphi$ are related as $\varphi = kt$.

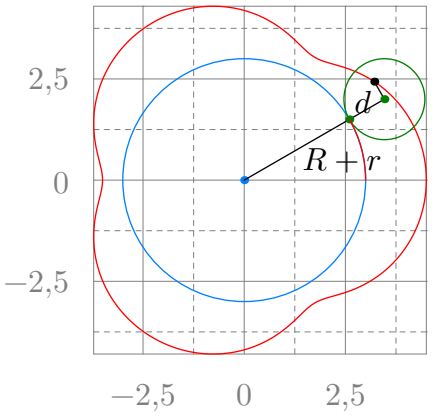


Figure 1. $R = 3, r = 1, d = 1/2$

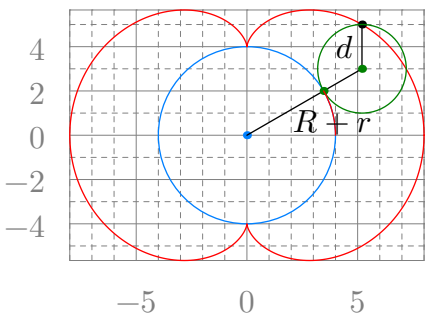


Figure 2. $R = 4, r = 2, d = 2$

Some special cases of epitrochoids have proper names. So for $r = R$, *Pascal's snail* is obtained, for $d = R + r$ — *rosy curve* or *rose*, and for $d = r$ — *epicycloid* (figure 2).

### 2.2. Definition of a hypotrochoid

*Hypotrochoid* is the trajectory described by a fixed point $P$ on a radial straight circle of radius $r$, which rolls along the *inner* side of the circle of radius $R$ (figure 3). The parametric equation of the curve has the following form:

$$\begin{cases} x(t) = (R - r)\cos(\varphi) + d\cos\left(\dfrac{R - r}{r}\varphi\right), \\ y(t) = (R - r)\sin(\varphi) - d\sin\left(\dfrac{R - r}{r}\varphi\right), \end{cases}$$

where, as in the case of the epitrochoid, $d$ is the distance from the center of the rolling circle to the point $P$. In particular, for $d = r$, *hypocycloid* is obtained (figure 4).

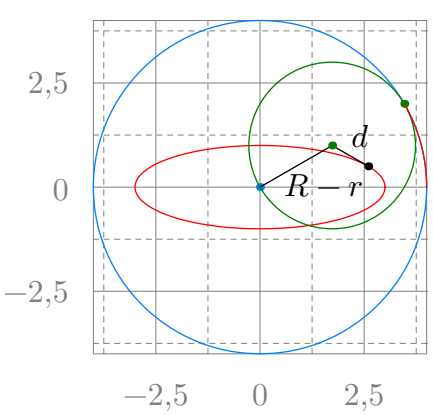


Figure 3. $R = 4, r = 2, d = 1$

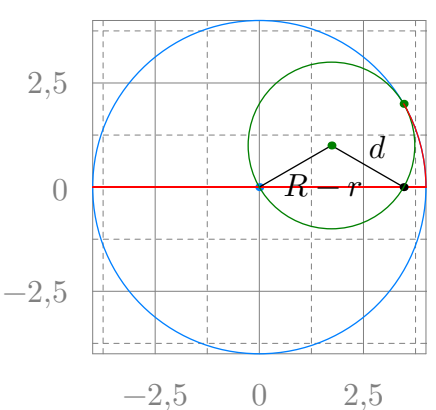


Figure 4. $R = 4, r = 2, d = 2$

It is also possible to parameterize $\varphi = kt$, where $k = r/R$, then the equation will take the form:

$$\begin{cases} x(t) = R(1-k)\cos(kt) + d\cos((1-k)t), \\ y(t) = R(1-k)\sin(kt) - d\sin((1-k)t). \end{cases}$$

### 2.3. Reducing the problem to a composition of turns

The construction of cycloidal curves begins by specifying two circles: a fixed circle of radius $R$ centered at point $O_R$ and a moving circle of radius $r$ centered at point $O_r$.

A fixed circle will be conventionally called "large", and a moving one — "small", since usually $R > r$. On the radial line of a small circle, we fix the point of the curve $P_0$.

From the definition of hypotrochoids and epitrochoids, it follows that a motion $T(\varphi)$ is performed over the point $P_0$, consisting of a composition of two turns (figures 5–10):

1. $T_1(\varphi)$ — rotation around the point $O_R$ by the angle $\varphi$, at which the point $O_r$ turns into $O'_r$, and the point $P_0$ into the point $P_{1/2}$;
2. $T_2(\theta(\varphi))$ — rotation around the point $O'_r$ by the angle $\theta$, at which $P_{1/2}$ turns into $P_1$.

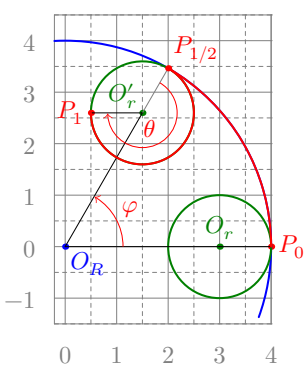


Figure 5. Hypocycloid $d = r$

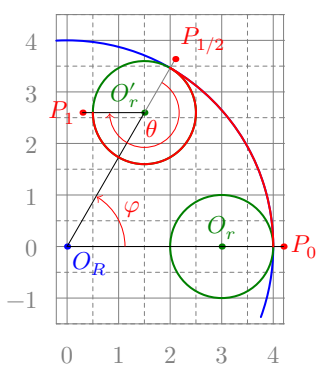


Figure 6. Hypotrochoid $d > r$

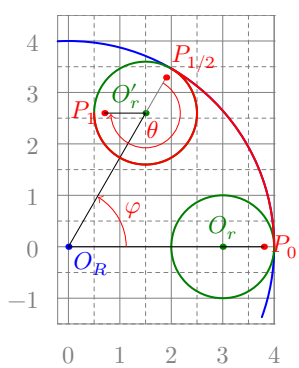


Figure 7. Hypotrochoid $d < r$

The rotation angle $\theta$ is related to the angle $\varphi$. A small circle must travel a distance equal to the length of the arc $PP_{1/2}$, which means that the lengths of the arcs $PP_{1/2}$ and $P_{1/2}P_1$ are equal.

$$|PP_{1/2}| = R\varphi = |P_{1/2}P_1| = \theta r \Rightarrow \theta = \frac{R\varphi}{r} = \frac{\varphi}{k}, \quad k = r/R.$$

Thus, to construct a curve, it is enough to set the parameters $R$, $r$ and $d$, the initial positions of the circles and the points $P_0$. It is usually assumed that the center of $O_R$ coincides with the origin, and the center of $O_r$ lies on the $Ox$ axis.

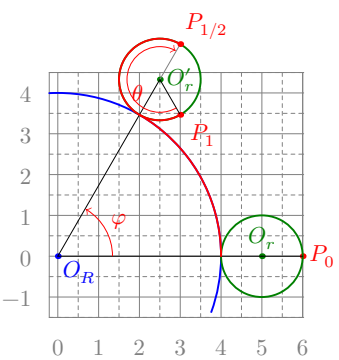


Figure 8. Epicycloid $d = r$

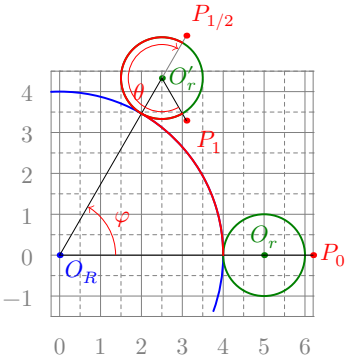


Figure 9. Epitrochoid $d > r$

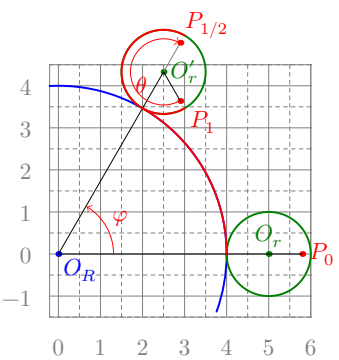


Figure 10. Epitrochoid $d < r$

Then the coordinates of the center $O_r$ are calculated as:

$$\mathbf{OO}_r = \mathbf{OO}_R + (R + s \cdot r, 0)^T, \quad s = \begin{cases} +1, \text{ if epitrochoid,} \\ -1, \text{ if hypotrochoid.} \end{cases} \tag{1}$$

And the coordinates of the curve point are $P_0$: $\mathbf{OP}_0 = \mathbf{OO}_r + (d, 0)^T$.

Now, to find any point of the curve, it is enough to act on $P_0$ by converting $T(\varphi) = T_2(\varphi/k) \circ T_1(\varphi)$ setting the required value to $\varphi$. If it is necessary to construct a set of points, then taking a sufficiently small step $\varphi$, one can consistently act on the point $P_0$ by transformations $T(i\varphi), i = 1, 2, \ldots, n$:

$$P_0 \overset{T(\varphi)}{\longrightarrow} P_1,\ P_0 \overset{T(2\varphi)}{\longrightarrow} P_2,\ P_0 \overset{T(3\varphi)}{\longrightarrow} P_3,\ P_0 \overset{T(4\varphi)}{\longrightarrow} P_4,\ , \ldots, P_0 \overset{T(n\varphi)}{\longrightarrow} P_n,\ .$$

## 3. Implementation on Asymptote

Below we present the source code of the program in the Asymptote language and comment on its key points:

```
include "config.asy";

import animation;
import graph;

unitsize(1cm);
size(10cm, 10cm);

string ssign(int d) {
  return d > 0 ? "+" : "-";
}

transform T1(real phi, pair O_R) { // (2)
  return rotate(angle=phi, z=O_R);
}

transform T2(real phi, int sign, real k, pair O_r) { // (4)
```

```
  return rotate(angle=sign*phi/k, z=O_r);
}

pen BigCircle = blue;
pen littleCircle = deepgreen;
pen curve = red;

int sign = -1; // (6)
real R = 4.0;
real r = 2.0;
real d = 2.0;
real k = r/R;
int N = 100; // (8)
pair O_R = (0, 0);
int turns = 1; // (10)
usersetting(); // (12)

pair O_r = O_R + polar(R + sign * r, 0); // (14)
pair P = O_r + polar(d, 0);  // (16)
pair Q = O_r - sign * polar(r, 0);  // (18)

guide xcycloid;
transform T;

animation A; // (20)
A.global = true;

draw(circle(c=O_R, r=R), p=BigCircle); // (22)
dot(O_R, p=BigCircle);

for(real phi: uniform(0, 360turns, N)) {
  save(); // (24)
  T = T1(phi, O_R) * T2(phi, sign, k, O_r); // (26)
  xcycloid = xcycloid -- T*P; // (28)
  draw(xcycloid, p=1bp+curve); // (30)
  dot(T*P, L=Label("P", align=NW)); // (32)
  draw(O_R -- T*O_r, L=Label("R"+ssign(sign)+"r")); // (34)
  draw(T*O_r -- T*P, L=Label("d"));
  draw(circle(c=T*O_r, r=r), p=littleCircle); // (36)
  dot(T*O_r, p=littleCircle);
  dot(T1(phi, O_R)*Q, p=littleCircle); // (38)

  include "axes.asy"; // (40)

  A.add(); // (42)
  restore(); // (44)
}

A.movie(); // (46)
currentpicture.erase();
```

This program creates a multi-page PDF-file, each page of which is a future frame of the video. The main work on calculating the points of the curve is performed by the functions `T1` (2) and `T2` (4). These functions are defined for convenience, so that the code reflects the above formulas as much as possible. All the work is done by the built-in function `rotate`, which allows you to determine rotation around an arbitrary point (argument `z`) by an arbitrary angle value in degrees (argument `angle`).

Next, we set a set of variables-parameters (6). The variable `sign` is $s$ from the formula (1), and the rest correspond to their mathematical notation. The variable `N` (8) sets the number of calculated points and, as a result, frames in the future video. The variable `turns` (10) sets the number of complete turns around the center of $O_R$. Calling the built-in function `usersetting` (12) to override the value of any variable specified above via the command line argument `-u`.

Then, based on the above-defined parameters, the coordinates of the initial position of the center of the moving circle $O_r$ (14), the points of the curve $P$ (16) and the point of tangency $Q$ of the moving circle with the stationary (18) are calculated.

Next, an object `A` (20) is created, into which animation frames will be recorded (objects of the type `picture` or `frame`). `A` has several fields, in particular the `global` field of the `bool` type allows you to enable and disable saving the created images as an array in RAM and writing them as files to disk only after they are all built.

The curve points are calculated in a loop, but before that, a fixed circle (22) and its center are drawn. Then, at the beginning of each iteration of the cycle, all the current stationary elements of the image are saved (object `picture`) (24), all movable elements are built, the resulting image is added to the structure `A` (42) and the image state is reset (44) to the one that was at the time of (24). The process continues until all frames are drawn and saved to `A`.

As the cycle progresses, the angle $\varphi$ changes from 0 to $2\pi n$ (in degrees). At each step, the rotation transformation $T(\varphi)$ is calculated (28), applied to the starting point of $P$ and added to the path (`guide`) `xcycloid` (28). With each iteration of the loop, new points are added to the path `xcycloid` and the curve grows.

The following drawing commands follow:

— of the already calculated part of the curve (30);
— of the new point position $P$ (32);
— of a segment of length $R + s \cdot r$ (34) connecting the center of $O_R$ to the new position of the center of $O_r$, as well as a segment of length $d$ connecting the new center of $O_r$ to the point $P$ of the curve;
— directly the moving circle itself in its new position (36) and its center;
— touch point $Q$ (38);
— coordinate grid, the settings of which are placed in a separate file (40).

Finally, after working out the loop, all created frames are recorded in a PDF-file. To do this, Asymptote sequentially creates separate PDF-files for each frame, then adds text processed by LaTeX (in our case LuaLaTeX) to them.

It is this procedure that takes the main time of the program, the calculations themselves practically do not take up time in comparison with this.

We also note the peculiarity of the Asymptote syntax, which allows omitting the `*` operator when multiplying numeric literal constants and variables, for example `360turn` (24).

# 4. Creating a video clip

## 4.1. Launching Asymptote

To run the program discussed above, run the following command

```
asy -noV -nobatchView -f pdf -globalwrite -u
  ↪ 'R=3;r=1;d=1;N=100' xcycloid.asy -o video/xcycloid.pdf
```

The source code file `xcycloid.asy` is started for execution and as a result the file `xcycloid.pdf` will be created. Consider the options used:

— Options `-noV` and `-nobatchView` prevent the newly created image from opening automatically. The `-noV` option disables this function when executed from the command line, and `-nobatchView` when executing the script (as in our case).

— Option `-f pdf` indicates that you should immediately create a PDF-file, bypassing the postscript-file stage.

— Option `-globalwrite` makes it possible to save the file `xcycloid.pdf` to any directory (in our case `video`), and not only to the one where the source file `xcycloid.asy` is located.

— Option `-u` allows you to interact with the `usersetting()` function and pass the values of variables inside the program. So we pass the values `R=3`, `r=1`, `d=1` and `N=100`. This feature allows you to use a single source code file to build multiple images, flexibly adjusting any parameters. Note that this parameter takes exactly a text string, which is then processed by the `usersetting()` function, so the passed parameters must be taken in quotation marks.

## 4.2. Converting to PNG using GhostScript

To convert the resulting multipage file into a video format, it is necessary to convert its pages into bitmaps. To do this, we suggest using the GhostScript [10] program. It is available for both Windows and Unix systems (GNU/Linux, macOS). It also comes with the TEXLive [5] distribution, as does Asymptote.

To convert a PDF-file, run the command

```
gs -sDEVICE=png16m -r600 -o video/xcycloid-%04d.png
  ↪ video/xcycloid.pdf
```

In the case of using GhostScript from the TEXLive distribution, you should call `gs` using the `rungs` script, which is located:

— in the directory `texlive\2023\bin\win32` in the case of Windows OS,

— in the `texlive/2023/bin/x86_64-linux` in the case of GNU/Linux.

The `2023` directory corresponds to the version of the TEXLive distribution and may differ.

### 4.3. Creating a video using FFmpeg

The process of gluing the resulting bitmap images into one video clip is carried out using FFmpeg [9]. This program is a command-line utility and has extensive functionality and, as a result, a huge number of options and settings. Let's give an example of creating a video clip from the PNG images generated in the previous step and give an explanation of the parameters used:

```
ffmpeg -r 30 -f image2 -start_number 1 -i
↪ video/xcycloid-%04d.png -c:v libx264 -vf
↪ "pad=ceil(iw/2)*2:ceil(ih/2)*2" video/xcycloid.mp4
```

— Parameter `-r` sets the frame rate.
— Parameter `-f` sets the format of the input file.
— Since a lot of files are submitted to the input, you should specify the format of their names. The same notation is used as in the case of `gs`. The `-start_number` parameter sets the starting number.
— Parameter `-c:v` allows you to specify the video encoder used. In our case `libx264`, but many other formats are supported.
— The important parameter `-vf` sets the filter that is applied to the processed frame. In our case, we round the width and height of the frame in pixels to an even number. After converting to PNG, the width and height of the image may be odd, which is unacceptable for the vast majority of encoders. The specified filter allows you to avoid this error and rescale the frame by ffmpeg.

At the output we will get a video packed in a container `mp4`. The x264 format we have chosen is widespread and can be played by any browser.

## 5. Conclusion

We have analyzed in detail the way to create vector graphics animation on a plane using the Asymptote language. This aspect of this language is poorly covered in the official manual and, in our opinion, this article fills this gap. Although the result is a video clip containing bitmaps, but thanks to the vector source (PDF), you can increase the resolution of the video almost limitlessly. It should also be noted that this method of creating animation is universal, since almost any data visualization tool can be used to create a set of image frames. FFmpeg does all the work on creating a video file.

## Acknowledgments

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

## References

[1] O. Shardt and J. C. Bowman, "Surface parameterization of nonsimply connected planar Bézier regions," *Computer-Aided Design*, vol. 44, no. 5, 484.e1–484.e10, May 2012. DOI: 10.1016/j.cad.2011.05.010.

[2] J. C. Bowman, "Asymptote: Interactive TEX-aware 3D vector graphics," *TUGboat*, vol. 31, no. 2, pp. 203–205, 2010.

[3] J. C. Bowman and A. Hammerlindl, "Asymptote: A vector graphics language," *TUGboat*, vol. 29, no. 2, pp. 288–294, 2008.

[4] J. C. Bowman. "Asymptote: The Vector Graphics Language." (May 2023), [Online]. Available: `https://asymptote.sourceforge.io/`.

[5] "TeX Live." (2023), [Online]. Available: `https://www.tug.org/texlive/`.

[6] T. Tantau and H. Menke. "PGF/TikZ." (2023), [Online]. Available: `https://ctan.org/pkg/pgf`.

[7] "ImageMagick." (Jun. 12, 2020), [Online]. Available: `https://imagemagick.org`.

[8] S. Tomar, "Converting video formats with FFmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[9] "FFmpeg Website." (2023), [Online]. Available: `https://ffmpeg.org/`.

[10] "Ghostscript Website." (2023), [Online]. Available: `https://www.ghostscript.com/`.

[11] C. I. Staats. "An Asymptote tutorial." (2015), [Online]. Available: `https://math.uchicago.edu/~cstaats/Charles_Staats_III/Notes_and_papers_files/asymptote_tutorial.pdf`.

**For citation:**
M. N. Gevorkyan, A. V. Korolkova, D. S. Kulyabov, Asymptote-based scientific animation, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 139–149. DOI: 10.22363/2658-4670-2023-31-2-139-149.

**Information about the authors**:
**Migran N. Gevorkyan** — Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: `gevorkyan-mn@rudn.ru`, phone: +7 (495) 955-09-27, ORCID: https://orcid.org/0000-0002-4834-4895)

**Anna V. Korolkova** — Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: `korolkova-av@rudn.ru`, phone: +7(495) 952-02-50, ORCID: https://orcid.org/0000-0001-7141-7610)

**Dmitry S. Kulyabov** — Doctor of Sciences in Physics and Mathematics, Professor of the Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University); Senior Researcher of Laboratory of Information Technologies, Joint Institute for Nuclear Research (e-mail: `kulyabov-ds@rudn.ru`, phone: +7 (495) 952-02-50, ORCID: https://orcid.org/0000-0002-0877-7063)

УДК 004.92:004.928:519.67
DOI: 10.22363/2658-4670-2023-31-2-139-149
EDN: XKNIYV

# Научная анимация на основе Asymptote

**М. Н. Геворкян**[1], **А. В. Королькова**[1], **Д. С. Кулябов**[1,2]

[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 область, 141980, Россия*

**Аннотация.** В статье рассматривается универсальный способ создания анимации с помощью языка для создания векторной графики Asymptote. В сам язык Asymptote встроена библиотека для создания анимации, однако практическое её использование осложнено крайне кратким описанием в официальной документации и нестабильной работой существующих примеров. Целью статьи является устранение данного пробела. Излагаемый нами способ основывается на создании PDF-файла с кадрами с помощью Asymptote с дальнейшей конвертацией его в набор PNG-изображений и склейкой их в видео с помощью FFmpeg. Все этапы подробно описываются, что даёт возможность читателю использовать изложенный метод, не будучи знакомым с используемыми утилитами.

**Ключевые слова:** векторная графика, TeX, asymptote, научная графика

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

2023, 31 (2) 150–163
http://journals.rudn.ru/miph

*Research article*

UDC 519.6:004.94
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-150-163
EDN: WFZCIO

# Chebyshev collocation method for solving second order ODEs using integration matrices

**Konstantin P. Lovetskiy**[1], **Dmitry S. Kulyabov**[1,2], **Leonid A. Sevastianov**[1,2], **Stepan V. Sergeev**[1]

[1] *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: April 5, 2023; revised: May 5, 2023; accepted: June 26, 2023)

**Abstract.** The spectral collocation method for solving two-point boundary value problems for second order differential equations is implemented, based on representing the solution as an expansion in Chebyshev polynomials. The approach allows a stable calculation of both the spectral representation of the solution and its pointwise representation on any required grid in the definition domain of the equation and additional conditions of the multipoint problem. For the effective construction of SLAE, the solution of which gives the desired coefficients, the Chebyshev matrices of spectral integration are actively used. The proposed algorithms have a high accuracy for moderate-dimension systems of linear algebraic equations. The matrix of the system remains well-conditioned and, with an increase in the number of collocation points, allows finding solutions with ever-increasing accuracy.

**Key words and phrases:** ordinary differential equation, spectral methods, two-point boundary value problems

## 1. Introduction

Ordinary differential equations (ODEs) and systems of ODEs of the second order describe most problems in classical mechanics. Most oscillatory processes are described by second order ODEs or systems of ODEs. Second order ODE systems describe a number of optical diffraction problems (see, for example, [1]). The model of adiabatic guided wave propagation of polarized light in integrated optical waveguides is also described by a system of two coupled oscillators [2–4].

There are many different methods for exact and approximate solution of initial/boundary value problems for different classes of second order ordinary

differential equations. Among them, the spectral methods of expansion in Chebyshev polynomials consistently occupy a well-deserved place.

In 1991, L. Greengard [5] formulated a method for solving a two-point boundary value problem for second order ODEs with constant coefficients, based on expanding the solution into a series of Chebyshev polynomials of the first kind. The method became stably referred to as the pseudo-spectral collocation method. In the same paper, mathematical constructions were introduced, which later received the names "differentiation matrix" and "integration matrix" (or "antidifferentiation matrix"). A detailed description of the properties of matrices that determine the relationship between the expansion coefficients in a series of approximated functions and the expansion coefficients of their derivatives and antiderivatives in the same set of basis functions is given in [6]. Greengard obtained estimates for the norms of these matrices and their condition numbers — large values for differentiation matrices and small values for integration (antidifferentiation) matrices.

Despite the poor conditionality of differentiation matrices, many authors used them to solve initial and boundary problems for ODEs of various orders. This is explained by the more familiar and therefore 'convenient' representation of physical models using the language of mathematical formulas.

The instability of widely used [7, 8] algorithms has been overcome by applying methods of preconditioning to the corresponding systems of linear algebraic equations. As a result of numerous studies, methods based on integration matrices in the physical space and in the spectral representation turned out to be the most preferable [9].

It is important to note that none of the applied methods for solving ODEs based on Chebyshev integration matrices [9, 10] allows obtaining systems of linear equations with sparse matrices [5]. The dense filling of matrices is a consequence of attempts to introduce boundary conditions into the system of linear algebraic equations along with differential relations [11]. The high sparseness of the matrices can be maintained by improving the algorithm by switching to the two-stage method. In this case, at the first stage, differential conditions are considered, which allow fixing the leading coefficients in the expansion of the solution into a series, thus defining the 'general solution'. The next step uses boundary/initial conditions to determine a pair (for second order equations) of missing coefficients. This makes it possible to obtain a complete set of expansion coefficients for the desired 'particular' solution.

The results of studies [5] demonstrate that the method of Chebyshev collocation that ensures the best accuracy in solving initial and boundary value problems is the method using Chebyshev integration matrices in the spectral space. This approach effectively relies on the use of operations with sparse matrices and its computational costs are quite comparable with the Fourier spectral discretization.

## 2. Setting of the problem

We consider an approximate solution to the two-point boundary value problem for the second-order differential equation having the form [12]

$$y''(x) + p(x)y'(x) + q(x)y(x) = r(x), \quad x \in (-1, 1), \tag{1}$$

where $p(x), q(x), r(x)$ are sufficiently regular functions. The uniqueness of the solution for any $\alpha$,$\beta$ is ensured by the boundary conditions

$$\alpha_0 y(-1) - \alpha_1 y'(-1) = \alpha, \quad \beta_0 y(1) + \beta_1 y'(1) = \beta, \tag{2}$$

the constants $\alpha_0$,$\alpha_1$,$\beta_0$,$\beta_1$ being nonnegative. For example, the condition of continuous $p(x)$ and $q(x)$, positive $q(x) > 0$, $x \in [-1, 1]$, and nonzero $\alpha_0 + \alpha_1 \neq 0$, $\alpha_0 + \beta_0 \neq 0$, $\beta_0 + \beta_1 \neq 0$ ensures the existence of the problem (1)–(2) [13].

## 3. Methods

The basic idea of spectral methods is to present the solution as a truncated series in known basis functions. The linear transformation (differentiation operator) that transforms the vector of coefficients $\mathbf{a} = \{a_k\}_{k\geqslant 0}$ of the function expansion $f(x) = \sum_{k\geqslant 0} a_k \phi_k(x)$ into the vector of coefficients $\mathbf{b} = \{b_k\}_{k\geqslant 0}$ of its derivative expansion $f'(x) = \sum_{k\geqslant 0} b_k \phi_k(x)$ into an analogous series in the same basis functions is known as the spectral differentiation matrix. The most widespread is the use of bases of Chebyshev functions of the first kind or Lagrange functions, which is due to high interpolative properties of these functions.

*Approximation by a finite series (on accuracy when discarding terms of the series with $n > N$).* The expansion of function $f \in C^n[-1, 1]$ ($n$ times differentiable function) in Chebyshev polynomials $T_k(x) : T_k(\cos\theta) = \cos(k\theta)$, is determined by the relation

$$g(x) = \frac{1}{2} a_0 T_0(x) + a_1 T_1(x) + \ldots + a_n T_n(x) + \ldots, \quad x \in [-1, 1], \tag{3}$$

where

$$a_k = \frac{2}{\pi} \int_{-1}^{1} f(x) T_k(x) (1 - x^2)^{-1/2} dx. \tag{4}$$

The residue of truncation of the series (3) to $N$ terms

$$g_N(x) = \frac{1}{2} a_0 T_0(x) + a_1 T_1(x) + \ldots + a_N T_N(x), \quad x \in [-1, 1], \tag{5}$$

has an order of $O\left(\frac{1}{N^{n-1}}\right)$ at $N \to \infty$ and at $f \in C^\infty[-1, 1]$ it tends to zero superalgebraically [14, 15].

**Remark 1.** According to Eq. (4), coefficients $a_k$ are the coefficients of Fourier cosine transformation, so that all $N$ coefficients $a_k$ can be obtained by the fast Fourier cosine transformation. And using the inverse Fourier cosine transformation, it is possible to simply calculate $g_n(\cos\theta_j)$ on a grid uniform in $\theta \in [0, \pi]$.

Most often, the approximation of continuous functions is restricted to a certain fixed number $n$ of the Chebyshev series, as a result of discarding

the components with such $T_k(x)$, $k > n$, the magnitude of which is small [16, 17]. In contrast to the approximations obtained using other power series, the approximation using the Chebyshev polynomials minimizes the number of terms necessary to approximate the function by polynomials with a given accuracy. Related to this is also the property that the approximation based on the Chebyshev series turns out to be quite close to the best uniform approximation (among polynomials of the same degree), but it is easier to calculate. In addition, it allows you to get rid of the Gibbs effect with a reasonable choice of interpolation points.

The differentiation matrices in the implicit or explicit form are presented in many publications related to the use of pseudospectral collocation methods [6–8]. The ODE solution using nondegenerate differentiation matrices in the $(N+1)$-dimensional physical and/or spectral space quite naturally led to poor conditioned systems of the linear algebraic equations to be solved. Refs. [5, 6, 18–20] formulate the specific features of the differentiation and integration matrices, considered on similar or mutually dependent grids. Using explicitly the differentiation matrices on the Chebyshev–Gauss–Lobatto to solve ODEs allows proposing stable and economic methods for solving ODEs. WE use the integration matrices $n$ Chebyshev–Gauss–Lobatto grids in the spectral representation. For more details on the form and properties of these matrices, see [6, 18–20].

### 3.1. The algorithm based on using integration matrices

Note first that the derivative of $T_k(x)$ can be explicitly written as an expansion in Chebyshev polynomials $T_0, T_1, \dots, T_{k-1}$ of lower order [6, 21] as a sum

$$\frac{dT_k(x)}{dx} = k\left(-[k\,odd]T_0(x) + 2\sum_{j=0}^{\lfloor (k-1)/2\rfloor} T_{k-1-2j}(x)\right), \quad x \in [-1,1], \tag{6}$$

where the notation $\lfloor x \rfloor$ means the largest integer less than $x$, and the expression $[k\,odd]$ takes the value equal to 1 when $k$ is odd, and 0 when $k$ is even.

We represent the desired function $y(x)$, the future approximate solution of equation (1), as an expansion of the form (3),(4),(5) in a finite set of Chebyshev polynomials $T_0, T_1, \dots, T_n$:

$$y(x) = \sum_{k=0}^{n} a_k T_k(x), \quad x \in [-1,1]. \tag{7}$$

By differentiating (7), it is possible to present the first derivative as a series:

$$y'(x) = \sum_{k=0}^{n} a_k T_k'(x), \quad x \in [-1,1]. \tag{8}$$

At the same time, the derivative $y'(x)$ as a polynomial of degree $n$ can be expanded is series with respect to the initial basis $T_0, T_1, \dots, T_n$ with coefficients $\mathbf{b} = \{b_0, b_1, \dots, b_n\}$:

$$y'(x) = \sum_{k=0}^{n} b_k T_k(x), \quad x \in [-1, 1], \tag{9}$$

the last expansion coefficient becoming zero, $b_n = 0$, in accordance with formula (6) of a transition to the expansion in lower-order polynomials.

Therefore, Eq. (6) describes the relation between the expansion coefficients $\mathbf{a} = \{a_0, a_1, \dots, a_n\}$ of a Chebyshev polynomial of the first kind and the expansion coefficients of its derivative. In matrix form, this relation can be represented using the differentiation matrix: $\mathbf{b} = \mathbf{D}\mathbf{a}$, where the infinite matrix of Chebyshev differentiation has the form:

$$\mathbf{D} \equiv \mathbf{D}_{\text{Chebyshev}} = \begin{bmatrix} 0 & 1 & 0 & 3 & 0 & 5 & 0 & 7 & \cdots \\ & 0 & 4 & 0 & 8 & 0 & 12 & 0 & \cdots \\ & & 0 & 6 & 0 & 10 & 0 & 14 & \cdots \\ & & & 0 & 8 & 0 & 12 & 0 & \cdots \\ & & & & 0 & 10 & 0 & 14 & \cdots \\ & & & & & 0 & 12 & 0 & \cdots \\ & & & & & & 0 & 14 & \ddots \\ & & & & & & & 0 & \ddots \\ & & & & & & & & \ddots \end{bmatrix}. \tag{10}$$

A similar transformation for the coefficients of the second derivative

$$y''(x) = \sum_{k=0}^{n} c_k T_k(x), \quad x \in [-1, 1] \tag{11}$$

allows using the formula $\mathbf{c} = \mathbf{D}\mathbf{D}\mathbf{a}$ to calculate the expansion coefficients $\mathbf{c} = \{c_0, c_1, \dots, c_n\}$ in the matrix form.

If an algorithm is needed to determine a part of the coefficients $\mathbf{a} = \{a_0, a_1, \dots, a_n\}$ of the expansion of function $y(x)$ from the known coefficients $\mathbf{b} = \{b_0, b_1, \dots, b_n\}$ of its derivative expansion, the appropriate matrix form for this operation is $\mathbf{a} = \mathbf{D}^{+}\mathbf{b}$, where the infinite tridiagonal matrix of integration (antidifferentiation) has the form [6, 18]:

$$\mathbf{D}^{+} \equiv \mathbf{D}^{+}_{\text{Chebyshev}} = \begin{bmatrix} 0 & 0 & & & \\ 1 & 0 & -\frac{1}{2} & & \\ & \frac{1}{4} & 0 & -\frac{1}{4} & \\ & & \frac{1}{6} & 0 & -\frac{1}{6} & \\ & & & \frac{1}{8} & 0 & \ddots \\ & & & & \frac{1}{10} & \ddots \end{bmatrix}. \tag{12}$$

For example, using the spectral integration matrices $\mathbf{D}^+$ to determine coefficients $\mathbf{a} = \{a_0, a_1, \dots, a_n\}$ of function $y(x)$ for the known coefficients $\mathbf{c} = \{c_0, c_1, \dots, c_n\}$ of the expansion of its second derivative $y''(x)$ allows calculating all coefficients of the function expansion by formula $\mathbf{a} = \mathbf{D}^+\mathbf{D}^+\mathbf{c}$, except the first two coefficients. This is because the first row of matrix $\mathbf{D}^+$ is zero.

Multiplying from the left the integration matrix $\mathbf{D}^+$ by vector $\mathbf{b} = \{b_0, b_1, \dots, b_n\}$ of the known coefficients of the derivative expansion allows revealing [6] the following dependence of coefficients $\mathbf{a} = \{a_0, a_1, \dots, a_n\}$ on $\mathbf{b} = \{b_0, b_1, \dots, b_n\}$, which can be written in the explicit form:

$$\mathbf{D}^+\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 \\ 1 & 0 & -\frac{1}{2} & 0 & 0 & \vdots & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & -\frac{1}{4} & 0 & \vdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & -\frac{1}{6} & \vdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{8} & 0 & \vdots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \ddots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 & \frac{-1/2}{(n-3)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & \frac{1/2}{(n-2)} & 0 & \frac{-1/2}{(n-2)} & 0 \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 & \frac{1/2}{(n-1)} & 0 & \frac{-1/2}{(n-1)} \\ 0 & 0 & 0 & 0 & 0 & \vdots & 0 & 0 & \frac{1/2}{n} & 0 \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ \vdots \\ b_{n-3} \\ b_{n-2} \\ b_{n-1} \\ b_n \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_{n-3} \\ a_{n-2} \\ a_{n-1} \\ a_n \end{bmatrix}$$

All the above, including relations (6) and (7),(9),(11) expressed through the representations $\mathbf{a} = \mathbf{D}^+\mathbf{D}^+\mathbf{c}$ and $\mathbf{b} = \mathbf{D}^+\mathbf{c}$, allows us to write equation (1) in spectral representation in the following matrix form:

$$\mathbf{Tc} + diag(\mathbf{p})\mathbf{TD}^+\mathbf{c} + diag(\mathbf{q})\mathbf{TD}^+\mathbf{D}^+\mathbf{c} = \mathbf{r}, \quad x \in (-1, 1). \tag{13}$$

Here $\mathbf{T}$ is the Chebyshev matrix of mapping a point (vector) from the space of coefficients to the space of function values [5]

$$\begin{bmatrix} T_{0,0} & T_{1,0} & T_{2,0} & \vdots & T_{n,0} \\ T_{0,1} & T_{1,1} & T_{2,1} & \vdots & T_{n,1} \\ T_{0,2} & T_{1,2} & T_{2,2} & \vdots & T_{n,2} \\ \dots & \dots & \dots & \ddots & \dots \\ T_{0,n} & T_{1,n} & T_{2,n} & \vdots & T_{n,n} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \dots \\ c_n \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \dots \\ p_n \end{bmatrix}, \tag{14}$$

so that $\mathbf{p} = \mathbf{Tc}$ is the vector of values of the desired function (also in the physical space). Here, to reduce formulas, we use the notation $T_{kj} = T_k(x_j)$, $k, j = 0, \dots, n$.

The system of linear algebraic equations (13) has a well-conditioned matrix [5] for any number of collocation points. We will use the Chebyshev–Gauss–Lobatto grid [7, 8], which has proven itself well in the Chebyshev

pseudospectral collocation method [9]. Since the matrix is not a symmetric real matrix, instead of the very convenient Cholesky method, we will use the widely used LU method to solve system (13).

The solution of the system of linear algebraic equations (13) is the vector of expansion coefficients $\{c_0, c_1, \dots, c_n\}$ in the $(n+1)$-dimensional space of the second derivative of the desired solution of equation (1). These components determine the set of 'general' solutions to the ordinary differential equation (1). To single out some specific 'particular' solution from this set, it is required to impose additional restrictions on the components $\{a_0, a_1\}$, which cannot be determined from the relation $\mathbf{a} = \mathbf{D}^+\mathbf{D}^+\mathbf{c}$.

The first two components that have not yet been found will have to be additionally determined (to obtain a 'particular' solution) from the boundary conditions (2). The remaining components of the vector $\mathbf{a}$ remain unchanged and allow satisfying equation (1) for any first expansion coefficients in terms of basis polynomials.

The solution of equation (13) gives us the vector of coefficients $\{c_0, c_1, \dots, c_n\}$ of the expansion of the second derivative of the solution of Eq. (1) in Chebyshev polynomials. Thus, the main problem is reduced to solving the simplest Poisson equation:

$$y''(x) = f(x), \quad -1 < x < 1, \tag{15}$$

where the function $f(x)$ is calculated at any point of the interval $x \in$ based on the known vector of coefficients $\{c_0, c_1, \dots, c_n\}$.

$$f(x) = \sum_{k=0}^{n} c_k T_k(x), \quad x \in [-1, 1]. \tag{16}$$

The method under consideration makes it possible to solve, depending on the type of additional conditions, both the Cauchy problem with initial conditions and the problem with boundary conditions of a general form, requiring, for example, the use of the iterative shooting method [22]. The boundary conditions of the original problem (2) allow extending the definition of the spectral coefficients of the solution. Let us consider some variants, such, e.g., as the Dirichlet conditions at both ends of the interval

$$y(-1) = \alpha, \quad y(1) = \beta. \tag{17}$$

Neumann–Dirichlet conditions

$$y'(-1) = \alpha, \quad y(1) = \beta \tag{18}$$

or Dirichlet–Neumann condition

$$y'(-1) = \alpha, \quad y'(1) = \beta. \tag{19}$$

The algorithm for finding a solution to the simplest Poisson equation (15) with one of the boundary conditions (17),(18),(19) consists of three stages:

— calculation of the coefficients of polynomial interpolation of the vector $f(x)$ in the right-hand side of Eq. (15) on the Gauss–Lobatto grid; an efficient method is presented in [23];

— calculation of those coefficients of the solution (except for the first two), which are determined by the differential conditions (15) of the problem (the solution must satisfy the differential conditions), by multiplying the transposed Chebyshev matrix by the vector of interpolation coefficients of the function $f(x)$;
— redefinition of solution coefficients based on boundary (or other independent additional) conditions (17),(18),(19).

In the case of Dirichlet boundary conditions (boundary conditions of the first kind): $p(-1) = \alpha$, $p(1) = \beta$, the determination of the still unknown coefficients $a_0$,$a_1$ is reduced to solving a system of two equations, which can be, e.g., the equations, which determine the behavior of the solution at the boundary points $x = \pm 1$:

$$\begin{aligned} a_0 + a_1 T_{1,0}(-1) + \sum_{k=2}^{n} a_k T_{k,0}(-1) = \alpha, \\ a_0 + a_1 T_{1,n}(1) + \sum_{k=2}^{n} a_k T_{k,n}(1) = \beta. \end{aligned} \tag{20}$$

If we additionally consider the fact that Chebyshev polynomials of the first kind take the values $T_{k,j}(\pm 1) = \pm 1$, $j, k = 0, 1, ...$ at the boundary of the interval, then the solution can be written explicitly

$$a_0 = \frac{1}{2}\left(\alpha + \beta - \sum_{k=2,k\,\text{even}}^{n} a_k\right), \quad a_1 = \frac{1}{2}\left(\beta - \alpha - \sum_{k=2,k\,\text{odd}}^{n} a_k\right). \tag{21}$$

In the case when the boundary conditions contain expressions of higher degrees of derivatives of the desired function, one can use the relation [7]

$$\left.\frac{d^p T_n}{dx^p}\right|_{x=\pm 1} = (\pm 1)^{n+p} \prod_{k=0}^{p-1} \frac{n^2 - k^2}{2k+1}. \tag{22}$$

For example, in the case of mixed Neumann–Dirichlet conditions (boundary conditions of the second and first kind): $p'(-1) = \alpha$, $p(1) = \beta$, the coefficients $c_0, c_1$ are determined by the formulas:

$$a_1 = \alpha - \sum_{k=2}^{n} (-1)^{k+1} k^2 a_k, \quad a_0 = \beta - a_1 - \sum_{k=2}^{n} a_k \tag{23}$$

and in the case of Dirichlet–Neumann conditions

$$a_1 = \beta - \sum_{k=2}^{n} k^2 a_k, \quad a_0 = \alpha - a_1 - \sum_{k=2}^{n} (-1)^{k+1} k^2 a_k. \tag{24}$$

## 4. Solution of model examples

To illustrate the capabilities of the proposed algorithm, consider as an example the solution of the following second order ODE with Dirichlet boundary conditions $y(-1) = \sin 1$, $y(1) = \sin 1$:

$$\begin{cases} y'' + xy' = (2 + x^2)\cos x, \quad x \in (-1, 1), \\ y(-1) = \sin 1, \\ y(1) = \sin 1. \end{cases} \tag{25}$$

The exact solution is $y(x) = x \sin x$.

The problem was solved by the collocation method using the integration matrices (see figures 1, 2).

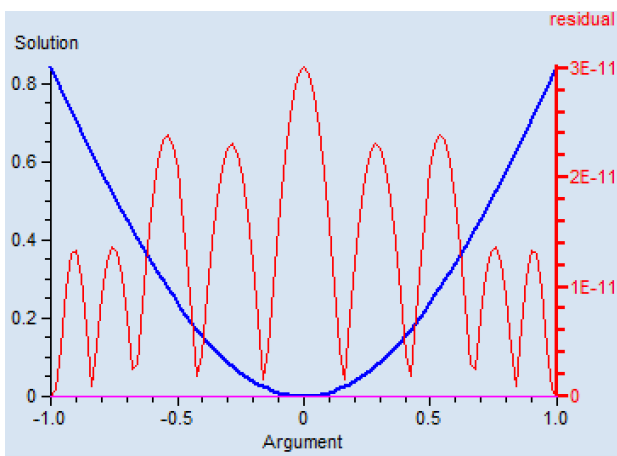


Figure 1. Ten collocation points. Solution is plotted in blue, residual – in red

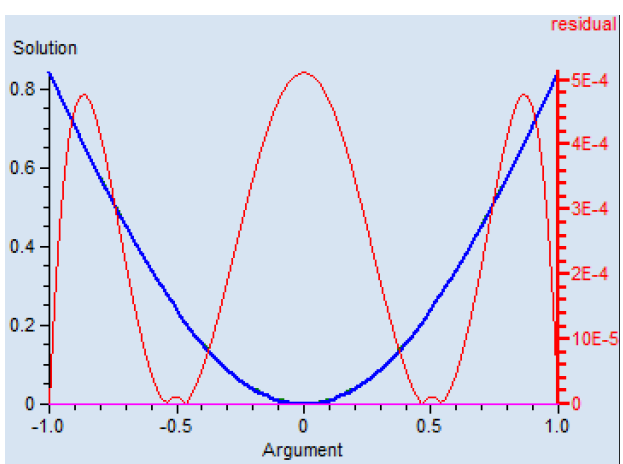


Figure 2. Five collocation points. Solution is plotted blue, residual – red

Comparison of the exact solution of the model equation with the numerical one is given in the table 1.

Table 1

Comparison of the exact solution of the model equation with the numerical one

| Number of collocation points | Mean deviation $abs(y_{\text{exact}}(x) - y_{\text{calc}}(x))/N$ | Maximum deviation of the calculated solution from the exact one |
|---|---|---|
| 6 | 1.82356474757341e-06 | 4.63901002387395e-06 |
| 7 | 1.50424878363523e-06 | 3.0061171892859e-06 |
| 9 | 5.23575446557936e-09 | 1.05369208025419e-08 |
| 11 | 1.19253244714073e-11 | 2.39715192140721e-11 |
| 13 | 1.91730425714347e-14 | 3.86046415624311e-14 |
| 14 | 4.4039495190215e-17 | 1.11022302462516e-16 |

The error was estimated numerically (see figure 3). The number of accuracy control points $N$ was taken equal to one hundred.

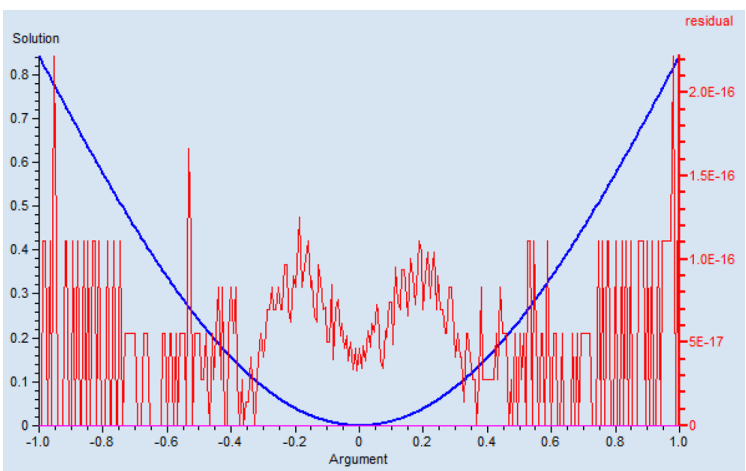


Figure 3. High solution accuracy with the average and maximum deviations of the numerical solution from the exact one $< 10^{-17}$ is achieved with a sufficiently small number of collocation points ($n > 13$)

As can be seen from the results, the accuracy of the solution depends significantly on the number of collocation points: with an increase in the number of collocation points, the algorithm, in contrast to the method using differentiation matrices, does not lose stability. Due to the inherent property of Chebyshev polynomials, when approximating smooth functions, the accuracy of the solution rapidly increases with a slight increase in the number of basis functions. In our experiment, the most accurate solution was obtained with the number of collocation points equal to 14. With a further

increase in the number of collocation points and, consequently, the number of approximation terms in the expansion series of the solution in Chebyshev polynomials, the accuracy does not increase.

## 5. Conclusion

In traditional algorithms, even in the most favorable cases when using differentiation matrices on arbitrary grids, the number of arithmetic operations for solving problems with acceptable accuracy turns out to be large. This fact is a consequence of the inclusion in the SLAE, obtained by passing from differential to algebraic relations, of additional equations that specify the initial and boundary conditions.

The algorithm presented in [23] uses a modified (improved) method of pseudo-spectral collocation, i.e., the solution of the problem in two stages. At the first stage, only the 'general' solution of the ODE is found, which is determined by the leading coefficients of the spectral expansion of the solution in the polynomial basis. This approach allows constructing an algorithm that uses only matrices of a simple structure to obtain the solution of the corresponding SLAE. The missing expansion coefficients are determined at the second stage based on the additional (initial or boundary) conditions, solving a simple system of two linear equations.

In this paper, we use an algorithm based on integration matrices. The matrix of the SLAE formed in this case turns out to be well conditioned even for large dimensions of the system. A high accuracy of the solution is achieved with a sufficiently small number of collocation points. The method based on integration matrices should be chosen in cases when there is a request for high and stable accuracy of solving the problem.

## Acknowledgments

The publication was supported by the RUDN University Strategic Academic Leadership Program.

## References

[1] V. A. Soifer, *Diffraction computer optics [Difraktsionnaya komp'yuternaya optika]*. M.: FIZMATLIT, 2007, in Russian.

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

[3] 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. 28, no. 3, pp. 252–273, 2020. DOI: 10.22363/2658-4670-2020-28-3-252-273.

[4] 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.

[5] L. Greengard, "Spectral integration and two-point boundary value problems," *SIAM Journal on Numerical Analysis*, vol. 28, no. 4, pp. 1071–1080, 1991. DOI: 10.1137/0728057.

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

[7] J. P. Boyd, *Chebyshev and Fourier spectral methods*, second. Dover Books on Mathematics, 2013.

[8] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. New York: Chapman and Hall/CRC Press, 2002. DOI: 10.1201/9781420036114.

[9] S. E. El-gendi, "Chebyshev solution of differential, integral and integro-differential equations," *The Computer Journal*, vol. 12, no. 3, pp. 282–287, Aug. 1969. DOI: 10.1093/comjnl/12.3.282.

[10] L. N. Trefethen, "Is Gauss quadrature better than Clenshaw–Curtis?" *SIAM Review*, vol. 50, no. 1, pp. 67–87, 2008. DOI: 10.1137/060659831.

[11] L. A. Sevastianov, K. P. Lovetskiy, and D. S. Kulyabov, "A new approach to the formation of systems of linear algebraic equations for solving ordinary differential equations by the collocation method [Novyy podkhod k formirovaniyu sistem lineynykh algebraicheskikh uravneniy dlya resheniya obyknovennykh differentsial'nykh uravneniy metodom kollokatsiy]," *Izvestiya of Saratov University. Mathematics. Mechanics. Informatics*, vol. 23, no. 1, pp. 36–47, 2023, in Russian. DOI: 10.18500/1816-9791-2023-23-1-36-47.

[12] N. Egidi and P. Maponi, "A spectral method for the solution of boundary value problems," *Applied Mathematics and Computation*, vol. 409, p. 125 812, 2021. DOI: 10.1016/j.amc.2020.125812.

[13] H. B. Keller, *Numerical methods for two-point boundary value problems*. Boston: Ginn-Blaisdell, 1968.

[14] D. Gottlieb and S. A. Orszag, *Numerical analysis of spectral methods*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1977.

[15] J. F. Epperson, *An introduction to numerical methods and analysis*, second. John Wiley & Sons, Inc, 2013.

[16] X. Zhang and J. P. Boyd, *Asymptotic coefficients and errors for Chebyshev polynomial approximations with weak endpoint singularities: effects of different bases*, 2022. DOI: 10.48550/arXiv.2103.11841.

[17] 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.

[18] B. Fornberg, *A practical guide to pseudospectral methods*. New York: Cambridge University Press, 1996.

[19] F. Rezaei, M. Hadizadeh, R. Corless, and A. Amiraslani, "Structural analysis of matrix integration operators in polynomial bases," *Banach Journal of Mathematical Analysis*, vol. 16, no. 1, p. 5, 2022. DOI: 10.1007/s43037-021-00156-4.

[20] L. C. Young, "Orthogonal collocation revisited," *Computer methods in Applied Mechanics and Engineering*, vol. 345, pp. 1033–1076, 2019. DOI: 10.1016/j.cma.2018.10.019.

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

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

[23] L. A. Sevastianov, K. P. Lovetskiy, and D. S. Kulyabov, "Multistage collocation pseudo-spectral method for the solution of the first order linear ODE," in *2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)*, 2022, pp. 1–6. DOI: 10.1109/ITNT55410.2022.9848731.

**For citation:**

K. P. Lovetskiy, D. S. Kulyabov, L. A. Sevastianov, S. V. Sergeev, Chebyshev collocation method for solving second order ODEs using integration matrices, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 150–163. DOI: 10.22363/2658-4670-2023-31-2-150-163.

**Information about the authors**:

**Lovetskiy, Konstantin P.** — Candidate of Sciences in Physics and Mathematics, Associate Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: lovetskiy-kp@rudn.ru, phone: +7(495)952-25-72, ORCID: https://orcid.org/0000-0002-3645-1060)

**Kulyabov, Dmitry S.** — Professor, Doctor of Sciences in Physics and Mathematics, Professor at the Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University); Senior Researcher of Laboratory of Information Technologies, Joint Institute for Nuclear Research (e-mail: kulyabov-ds@rudn.ru, phone: +7(495)952-02-50, ORCID: https://orcid.org/0000-0002-0877-7063)

**Sevastianov, Leonid A.** — Professor, Doctor of Sciences in Physics and Mathematics, Professor at the Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University), Leading Researcher of Bogoliubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research (e-mail: sevastianov-la@rudn.ru, phone: +7(495)952-25-72, ORCID: https://orcid.org/0000-0002-1856-4643)

**Sergeev, Stepan V.** — PhD student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: 1142220124@rudn.ru, ORCID: https://orcid.org/0009-0004-1159-4745)

УДК 519.6:004.94
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-150-163
EDN: WFZCIO

# Метод коллокации Чебышева для решения ОДУ второго порядка с использованием матриц интегрирования

**К. П. Ловецкий**[1], **Д. С. Кулябов**[1,2], **Л. А. Севастьянов**[1,2], **С. В. Сергеев**[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 область, 141980, Россия*

**Аннотация.** Реализован метод спектральной коллокации для решения двухточечных краевых задач для дифференциальных уравнений второго порядка, основанный на представлении решения в виде разложения по полиномам Чебышева. Подход позволяет устойчиво вычислять как спектральное представление решения, так и его поточечное представление на любой необходимой сетке в области определения уравнения и дополнительных условий многоточечной задачи. Для эффективного построения СЛАУ, решение которой дает искомые коэффициенты, активно используются матрицы Чебышева спектрального интегрирования. Предложенные алгоритмы обладают высокой точностью для систем линейных алгебраических уравнений средней размерности. Матрица системы остается хорошо обусловленной и с увеличением количества точек коллокации позволяет находить решения со все возрастающей точностью.

**Ключевые слова:** обыкновенное дифференциальное уравнение, спектральные методы, двухточечные краевые задачи

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

2023, 31 (2) 164–173
http://journals.rudn.ru/miph

*Research article*

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-164-173
EDN: XDVQBB

# Implementation of the Adams method for solving ordinary differential equations in the Sage computer algebra system

**Mikhail D. Malykh**[1, 2], **Polina S. Chusovitina**[1]

[1] *RUDN University,*
*6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation*
[2] *Meshcheryakov Laboratory of Information Technologies,*
*Joint Institute for Nuclear Research,*
*6, Joliot-Curie St., Dubna, Moscow Region, 141980, Russian Federation*

(received: April 25, 2023; revised: May 7, 2023; accepted: June 26, 2023)

**Abstract.** This work is devoted to the implementation and testing of the Adams method for solving ordinary differential equations in the Sage computer algebra system. The Sage computer algebra system has, to some extent, trivial means for numerical integration of ordinary differential equations, but at the same time, it is worth noting that this environment is convenient and practical for conducting computer experiments related to symbolic numerical calculations in it. The article presents the FDM package developed on the basis of the RUDN, which contains the developments of recent years, performed by M. D. Malykh and his students, for numerical integration of differential equations. In this package, attention is paid to the visualization of the calculation results, including the construction of various kinds of auxiliary diagrams, such as Richardson diagrams, as well as graphs of dependence, for example, the value of a function or step from a moment in time. The implementation of the Adams method will be considered from this package. In this article, this implementation of the Adams method will be tested on various examples of input data, and the method will also be compared with the Jacobi system. Exact and approximate values will be found and compared, and an estimate for the error will be obtained.

**Key words and phrases:** differential equations, Adams method, Sage, FDM package, Cauchy theorem, Taylor series, Richardson diagram

## 1. Introduction

To describe models in a variety of subject areas from mechanics to economics, ordinary differential equations are used [1]. These equations admit solutions in elementary functions only in some very special cases, therefore they are usually

investigated numerically. The finite difference method was proposed by Euler, the Runge–Kutta method of the 4th order is the most popular numerical method for solving initial problems for ordinary differential equations [2].

Old authors, including J. Scarborough [3, ch. XIII], mention numerical methods alternative to the Runge–Kutta method. The method that J. Scarborough has associated with the name of the English theoretical astronomer J. K. Adams, was forgotten for a long time, because it was very inconvenient to implement on a computer: before its use, a number of preparatory calculations had to be carried out on paper. However, with the development of computer computing, it became possible to perform these actions on a computer, which pushes us to study the possibility of implementing the Adams method in modern computer algebra systems.

Currently, RUDN University is developing an addition to Sage — the FDM package, which contains the achievements of recent years, made by M. D. Malykh and his students. The goal of the project is to create a convenient environment for numerical experiments with ODES in the Sage computer algebra system. This project is available to everyone on `https://github.com/malykhmd/fdm`. The general principles of the package are described in [4].

The purpose of this work is to test the implementation of the Adams method in FDM.

## 2. The Adams method and its implementation in FDM

Consider the initial problem

$$\frac{dx}{dt} = f(x,t), \quad x(0) = x_0. \tag{1}$$

Its solution exists by virtue of the Cauchy theorem. Decompose its solution into a Taylor series at $dt = 0$:

$$x(t+dt) = x(t) + \dot{x}(t)dt + \frac{1}{2}\ddot{x}(t)dt^2 + \dots. \tag{2}$$

Scarborough was forced to search numerically for the coefficients of the Taylor series, but now we can find them analytically, using the formulas specified by Cauchy himself. If $f(x,t)$ is known, then

$$\dot{x}(t) = f(x(t),t), \quad \ddot{x} = Df, \quad \dddot{x} = D^2 f,$$

where

$$Dg = f\frac{\partial g}{\partial x} + \frac{\partial g}{\partial t}.$$

To calculate the coefficients of the Taylor series using these formulas, it is required to differentiate the symbolic expression $f$ many times, which is naturally performed in Sage.

The Adams method in our interpretation is as follows. First, according to the given symbolic expression $f$, the Taylor polynomial is compiled

$$x + f(x,t)\cdot dt + \dots + \frac{1}{r!}D^{r-1}(f)\cdot dt^r \tag{3}$$

up to $r$-th order members. Its coefficients are calculated explicitly in symbolic form.

Then the segment $0 < t < T$ on which the initial problem is considered is divided into segments of width $dt$. At time $t = 0$, the solution is given to us. To find a solution at time $t = dt$, we substitute $t = 0$, $x = x_0$ into the Taylor polynomial (3) and thus obtain an approximate value of $x_1$ for $x(dt)$. To find a solution at time $t = 2 \cdot dt$, we substitute $t = dt$, $x = x_1$, etc. into the Taylor polynomial.

FDM implements two versions of the Adams method: with a constant step, and with a step that becomes smaller the larger the first discarded term

$$\frac{1}{(r+1)!} D^r(f) \cdot dt^{r+1}$$

in the Taylor formula (3).

## 3. Numerical experiments

We can test the implementation of the Adams method in FDM with a few examples. We will evaluate the error using the Richardson method [2].

**Example 1.** Consider the initial problem

$$\frac{dx}{dt} = t^2 + x, \quad x(0) = 0$$

on the segment $0 < t < 1$, its exact solution is known:

```
var("x,t")
pr1=Initial_problem(x,t^2+x,0,1)
P=adams_adaptive(pr1,  h=0.1, field=RealField(500))
```

We get the value of $x$ at the point $t = 0.8$:

```
P.value(x,0.8)
        0.21079360329767660164890230589662678539752960205078125
```

In figure 1 we see that the exact solution coincides with the approximate one. In figure 2 we see that the step depends on the moment of time $t$, in this case it decreases monotonically.

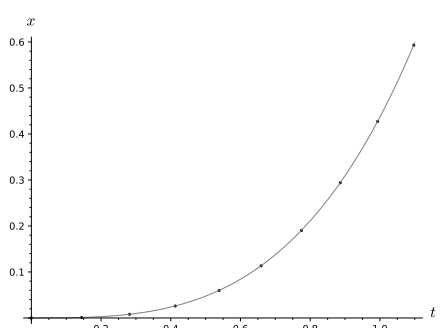


Figure 1. Dependence of $x$ on $t$

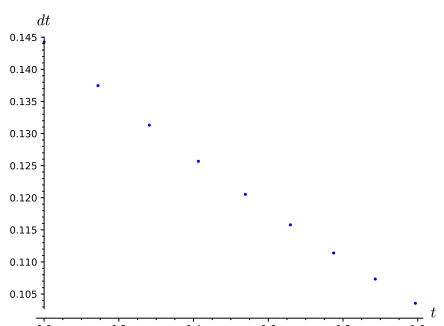


Figure 2. Dependence of the step on the moment of time $t$

Figure 3 shows the log–log plot for the dependence of the error $E$ on the value of the variable $x$ at time $t = 0.9$ on the parameter $h$ characterizing the step of the Adams method. Such a figure is known as Richardson diagram [2]. Theoretically, the error $E = ch^r + \mathcal{O}(h^{r+1})$.

In our experiments we use 3 terms in Taylor series, thus $r$ must be equal 3, so the error must be proportional to $h^3$. In our example (figure 3) there is a direct proportionality on the logarithmic scale and the slope is equal to 2.99, which is close to 3, so the error is proportional to $h^3$. To build a Richardson diagram [2] we used the standard tools of FDM [4]:

```
L=[adams_adaptive(pr1,  h=1/10/2^n, field=RealField(500))
        for n in range(10)]
richardson_plot(L,x,0.9)
```

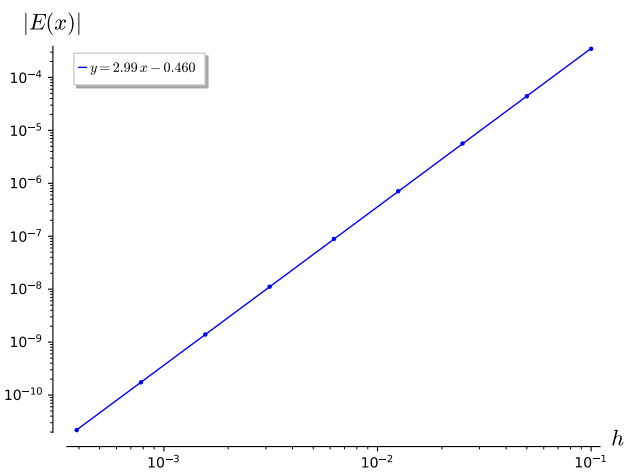


Figure 3. Richardson diagram

Each element of the list is an approximate solution with respect to the step $h = 0.1 \cdot (1/2^n)$, $n = 0, 1, \ldots, 9$.

Using standard tools of FDM, we can calculate the values of $x$ and the error and make sure that it is small. For example, at time $t = 0.8$ and $h = 1/20$:

```
richardson(L[0], L[1], x, 0.8)
        [0.21104359494629298943, -0.00003571309265949189 9]
```

If we take the exact solution and subtract the approximate one from it, we get an error value equal to $-0.0000382$:

```
x_exact=-t^2 - 2*t + 2*e^t - 2
L[1].value(x,0.8) - RR (x_exact.subs(t=0.8))
        -0.00003826203864254 47
```

It should be noted that the values are almost the same ($-0.0000357$ and $-0.0000382$), which means that Richardson's method gives an almost correct estimate for the error.

**Example 2.** Now we consider the system of two ODEs

$$\dot{x_1} = x_2, \quad \dot{x_2} = -x_1$$

with initial conditions $x_1(0) = 0$, $x_2(0) = 1$ at the interval $0 < t < 10$. Let's take the step $h = 1/4$ and solve the initial problem in FDM:

```
var("x1,x2,t")
pr2=Initial_problem([x1,x2],[x2,-x1],[0,1],10)
P=adams_adaptive(pr2,h=1/4)
```

In figure 4 we can see that our points lie on the sine wave. To plot graphs we use the line:

```
P.plot(t,x1,color='black') + plot(sin,(0,10),color='grey')
```

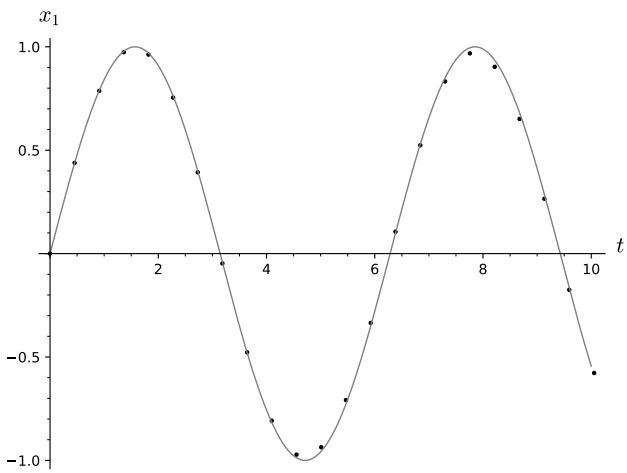


Figure 4. Dependence of $x$ on $t$

To plot a Richardson diagram (figure 5) we use the line:

```
L=[adams_adaptive(pr2,h=1/4/2^n) for n in range(10)]
```

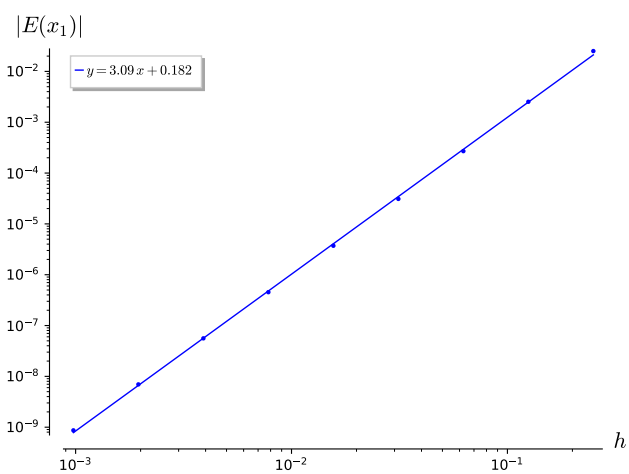


Figure 5. Richardson diagram

In this Richardson diagram we can see that not all points lie on a straight line, this is explained by the deviation of the slope value, which is not 3, but 3.09. But the error is also proportional to $h^3$.

Calculate the values of $x$ and the error:

```
richardson (L[-3],L[-2], x 1,9)
        [0.412118484375421, -8.69428719494993e-10]
```

We can compare the received error value by the Richardson method with the real error:

```
L[-2].value(x1,9)-sin(9).n()
        -8.66335503335591e-10
```

Note again that the values are almost identical.

**Example 3.** Now consider the Jacobi oscillator

$$\dot{p} = qr, \quad \dot{q} = -pr, \dot{r} = -k^2 pq$$

with initial conditions $p(0) = 0$, $q(0) = 1$, $r(0) = 1$ at the interval $0 < t < 100$. This example is interesting as a nonlinear oscillator. We described the problem by lines:

```
var("p,q,r,t")
k=0.99
pr3=Initial_problem([p,q,r], [q*r,-p*r,-k^2*p*q], [0,1,1], 100)
P=adams_adaptive(pr3,h=1/4)
```

and plot the graphs to compare approximate and exact solutions in Jacobi elliptic functions (figure 6):

```
P.plot(t,p,color='grey')+plot(jacobi('sn', t, k^2),(t,0,100),
       color='black')
```

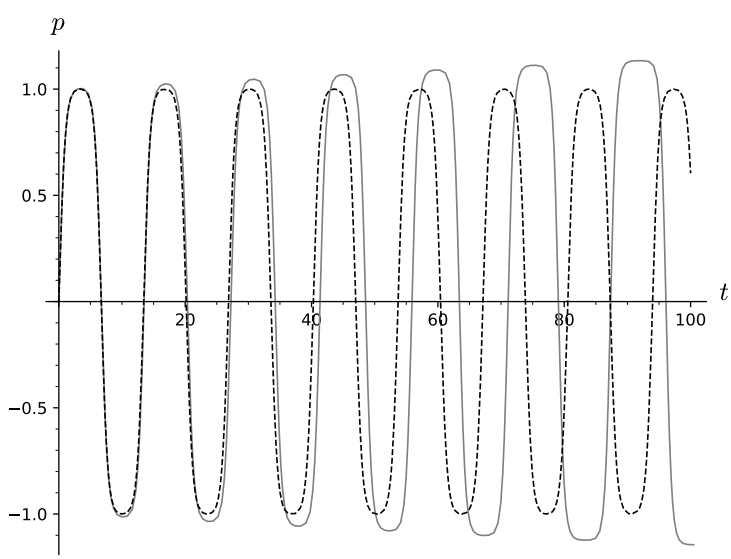


Figure 6. Comparison of values obtained by the Adams method and the Jacobi method

It is noticeable that the values diverge over a long-time interval, so that this does not happen, it is necessary to take a smaller step. In figure 7 we use a small-time interval, and the solutions coincide.

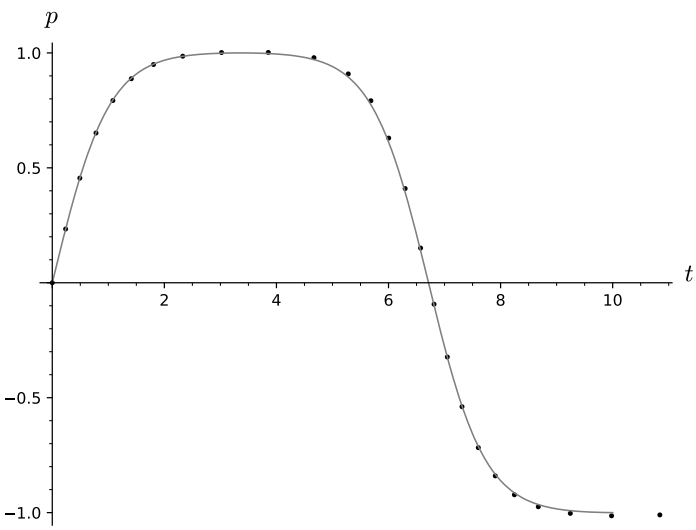


Figure 7. Comparison of exact and approximate solutions

We use adaptive Adams method what is important for nonlinear problems. In figure 8, we see the dependence of the step on the moment of time, while in places where the values change smoothly, the step is larger, and where the function changes quickly, the step is smaller, which indicates that the chosen method of adaptation according to the last term in the Taylor series is correct.

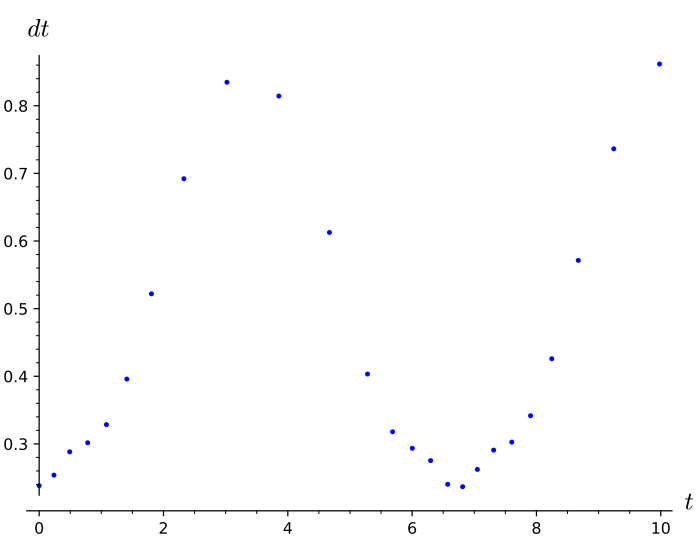


Figure 8. Dependence of the step on the moment of time $t$

At Richardson diagram (figure 9), we can see that the points lie on a straight line with a slope of 2.99. The error is also proportional to $h^3$. Also we calculate the value of $x$, the error by the Richardson method and the real error, and again we get identical error values:

```
richardson(L[3],L[4],p,9)
        [-0.983742403638134, -2.91251782518526e-6]
richardson(L[3],L[4],p,9)[0]-jacobi('sn',9,k^2)
        -2.91556109444091e-6
```

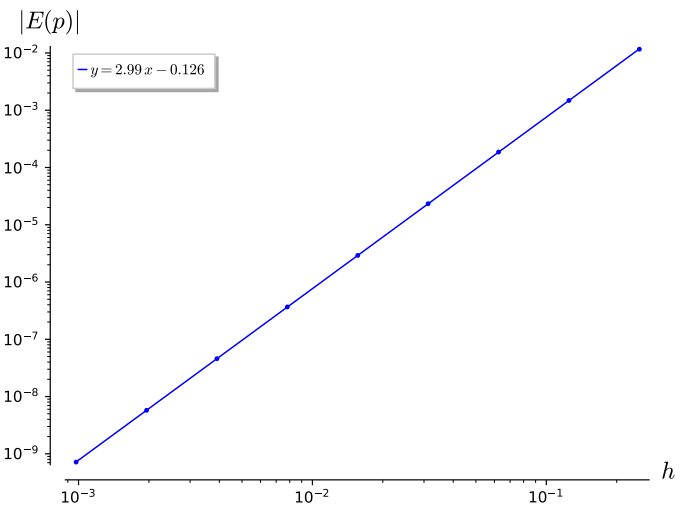


Figure 9. Richardson diagram

## 4. Conclusion

In this article, the Adams method for solving ordinary differential equations was tested on three examples. Also, shortcomings were identified in the fdm.sage package, after their correction, the execution of the examples became correct.

Thus, the computer experiments carried out confirm that the correct approximation method is indicated in the implementation of the Adams method in FDM, and this implementation itself allows to search for solutions with an accuracy close to the rounding error. Also, the achievement of the method is the natural adaptation of the step to changes in the function, where the function changes quickly, the step decreases. In this case, the method is symbolic-numerical, since the Taylor series is calculated symbolically once, and then used as a formula in which specific numeric values are substituted. At the beginning of the 20th century, this was considered the main drawback of the method proposed by Adams, since there were problems with symbolic computation – many iterations of differentiation led to large and complex expressions. And now, on the contrary, we can consider the advantage of the method that it can combine the powers of both symbolic and numerical methods, but, of course, the implementation of this method can still be refined and optimized.

## Acknowledgments

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

## References

[1] H. Gould and J. Tobochnik, *An introduction to computer simulation methods. Applications to physical systems*. Addison-Wesley Publishing Company, 1988.

[2] A. Baddour and M. D. Malykh, "Richardson–Kalitkin method in abstract description," *Discrete and Continuous Models and Applied Computational Science*, vol. 29, no. 3, pp. 271–284, 2021. DOI: 10.22363/2658-4670-2021-29-3-271-284.

[3] J. B. Scarborough, *Numerical methods of mathematical analysis*. Oxford book company, 1930.

[4] L. Gonzalez and M. D. Malykh, "On a new package for numerical solution of ordinary differential equations in Sage [O novom pakete dlya chislennogo resheniya obyknovennykh differentsial'nykh uravneniy v Sage]," in *Proceedings of ITTMM'22, Moscow, Russia*, in Russian, 2022, pp. 360–364.

**For citation:**

M. D. Malykh, P. S. Chusovitina, Implementation of the Adams method for solving ordinary differential equations in the Sage computer algebra system, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 164–173. DOI: 10.22363/2658-4670-2023-31-2-164-173.

**Information about the authors**:

**Malykh, Mikhail D.** — Doctor of Physical and Mathematical Sciences, Assistant Professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (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)

**Chusovitina, Polina S.** — Student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia named after Patrice Lumumba (RUDN University) (e-mail: 1032192941@rudn.ru, ORCID: https://orcid.org/0009-0006-4191-2454)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2023-31-2-164-173
EDN: XDVQBB

# Реализация метода Адамса для решения обычных дифференциальных уравнений в системе компьютерной алгебры Sage

**М. Д. Малых**[1,2], **П. С. Чусовитина**[1]

[1] *Российский университет дружбы народов,*
*ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия*
[2] *Лаборатория информационных технологий им. М. Г. Мещерякова,*
*Объединённый институт ядерных исследований,*
*ул. Жолио-Кюри, д. 6, Дубна, Московская область, 141980, Россия*

**Аннотация.** Работа посвящена реализации и тестированию метода Адамса для решения обыкновенных дифференциальных уравнений в системе компьютерной алгебры Sage. Система компьютерной алгебры Sage обладает в какой-то степени тривиальными средствами для численного интегрирования обыкновенных дифференциальных уравнений, но при этом, стоит заметить, что данная среда удобна и практична для проведения в ней компьютерных экспериментов, связанных с символьно-численными вычислениями. В работе представлен пакет FDM, разработанный на базе РУДН, в котором собраны наработки последних лет, выполненных М. Д. Малых и его учениками, для численного интегрирования дифференциальных уравнений. В данном пакете уделено внимание визуализации результатов вычисления, в том числе построению разного рода вспомогательных диаграмм, например диаграмм Ричардсона, а также графиков зависимости, например значения функции или шага от момента времени. В статье рассмотрена реализация метода Адамса, проведено её тестирование на различных примерах входных данных, а также выполнено сравнение метода с системой Якоби. Найдены и точные, и приближённые значения, проведено их сравнение, получена оценка для ошибки.

**Ключевые слова:** дифференциальные уравнения, метод Адамса, Sage, пакет FDM, теорема Коши, ряд Тейлора, диаграмма Ричардсона

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

2023, 31 (2) 174–188
http://journals.rudn.ru/miph

*Research article*

UDC 519.624.2:531.8
PACS 62.20
DOI: 10.22363/2658-4670-2023-31-2-174-188
EDN: XEAYRS

# Buckling in inelastic regime of a uniform console with symmetrical cross section: computer modeling using Maple 18

**Viktor V. Chistyakov, Sergey M. Soloviev**

*Laboratory of Rare Earth Semiconductor Physics,*
*Physical-Technical Institute named after A.F. Ioffe of RAS,*
*26, Politekhnicheskaya St., Saint Petersburg, 194021, Russian Federation*

(received: February 5, 2023; revised: May 7, 2023; accepted: June 26, 2023)

**Abstract.** The method of numerical integration of Euler problem of buckling of a homogeneous console with symmetrical cross section in regime of plastic deformation using Maple 18 is presented. The ordinary differential equation for a transversal coordinate $y$ was deduced which takes into consideration higher geometrical momenta of cross section area. As an argument in the equation a dimensionless console slope $p = \operatorname{tg}\theta$ is used which is linked in mutually unique manner with all other linear displacements. Real strain-stress diagram of metals (steel, titan) and PTFE polymers were modelled via the Maple nonlinear regression with cubic polynomial to provide a conditional yield point $(t,\sigma_f)$. The console parameters (free length $l_0$, $m$, cross section area $S$ and minimal gyration moment $J_x$) were chosen so that a critical buckling forces $F_{\text{cr}}$ corresponded to the stresses $\sigma$ close to the yield strength $\sigma_f$. To find the key dependence of the final slope $p_f$ vs load $F$ needed for the shape determination the equality for restored console length was applied. The dependences $p_f(F)$ and shapes $y(z)$, $z$ being a longitudinal coordinate, were determined within these three approaches: plastic regime with cubic strain-stress diagram, tangent modulus $E_{\text{tang}}$ approximations and Hook's law. It was found that critical buckling load $F_{\text{cr}}$ in plastic range nearly two times less of that for an ideal Hook's law. A quasi-identity of calculated console shapes was found for the same final slope $p_f$ within the three approaches especially for the metals.

**Key words and phrases:** Euler problem, plane cross-sections hypothesis, buckling, console, plastic deformation, strain-stress diagram, conditional yield point, critical buckling load, Maple programming, nonlinear estimation, Al/PTFE, steel

## 1. Introduction

The problem of stability loss in a beam under longitudinal load (buckling) in the range of inelastic strains is actual and important from many points of view such as sports (pole vaulting), civil engineering (bridges, truss constructions), aeronautics, robotics and elsewhere the requirements of a small weight and large strength are imposed on structural elements been designed [1]. Fatigue of materials, lowering the proportionality and elasticity limit due to the Bauschinger effect in periodically tensile and compressed elements, hysteresis etc. — all that results in falling of initially secure loads time into the zone of serious risk of buckling. Therefore, beginning from the pioneering work of F.R. Shanley [2] considered so called tangent and reduced moduli approaches [ibid], Euler's problem in inelastic range attracts more and more researchers — from engineers dealing with material strength to pure mechanicians and mathematicians dealing with bifurcations, nonlinear phenomena etc.

Of course, modern models of buckling are 2- or even 3-dimensional and they take into account not only bending shift component but a shear one too. To take all this into account the finite-element modeling (FEM) is widely used and it is implemented in the commercial software package ABAQUS (see e.g. [3–5]) and similar software. Many features and peculiarities both in thick so called Timoshenko beams [6] and in sandwich/fiber-composite/lattice/C-columns (see [7–9]) etc. are explained well in these multidimensional models.

The problem is studied in university courses of material sciences within a *plane cross-sections hypothesis* which leads to simple one-dimensional (1D) Euler ordinary differential equation (ODE) of the II-nd order. However, the attention is paid mainly to moment of arising of the phenomenon itself and its possible shapes for various ways of a beam fixation. Unfortunately, the linearized Euler ODE coupled with boundary condition (BC) on the beam ends looks like a classical eigenvalue problem with unstable higher modes corresponding to higher eigenvalues too. This ODE is similar to the Schröedinger equation for 1D particle in a potential well with infinitely high walls. This similarity misleads the students to the wrong conclusion that the non-zero solution of the ODE exists only for a set of "resonant" axial loads $F_n$, $n = 1, 2, ...$ just like in the aforementioned case of the well. And it is not clear whether for "non-resonant" forces from inside the intervals the ODE has purely compressive solution without any buckling or else power-like formula or something else. Or, may be, it shoots at some finite value at once just as the axial force $F$ reaches some critical value as we have seen from our own experience, compressing by the hands a steel ruler? In what way the non-linear and inelastic properties with yield point on strain-stress diagram of real materials influence the critical buckling load $F_{\rm cr}$ and the shape of column buckled?

This kind of questions inevitably arises by analytically thinking students which can't find the answers in many available textbooks where only simplified explanation of the phenomenon is presented. One of the reasons why it takes place is traditionally pure mathematical means to describe the buckling process not expressing in standard algebraic functions. Nowadays, at the time of rapid development of mathematical software, more and more new opportunities to study the buckling phenomenon both with practical, scientific and educational purposes are opening up.

As for the Maple itself it is permanently improving software package with simple programming language with commands close to English ones supplied with comprehensive parameters and easily read option Help. The undeniable advantage of the package in addition to the extremely broad coverage of the sciences from Bayesian statistics to Feynman diagram calculations is an extremely high computational accuracy due to so-called "long Arithmetics" (Matlab soft uses Maple calculations) and opportunity to choose an alternative computational method and compare the results to improve their reliability. All above makes this package most reliable means of numerical modeling compared to those packages where the control of the calculation process is reduced only to the choice of the "mouse" option from the menu.

The work is devoted to numerical modeling of the buckling phenomenon of uniform beam. The main purpose of this work is to present readers relatively simple and effective calculation algorithm and its realization with the Maple 2018 relying on which it is possible to learn in what way inelastic and plastic properties of the material in question influence the basic parameters of buckling. The versatile skills gained from this activity may be then applied in up to date theories and experiments in study of buckling of real constructive elements say within aforementioned FEM and others.

The application of the Maple not only solves many technical difficulties of mathematical nature but with the algorithm itself and with consequence of computational procedures it gives students better comprehension of the mechanism and nature of the phenomenon of buckling. Moreover, as part of university lessons and practices this kind of investigations may be joined in one collective interdisciplinary research project which results may be discussed, analyzed and then presented at student conference/contest.

## 2. Equation

We neglect in the paper the shear effects and stay within the classical Euler's 1D-model of buckling but with the axial compressive loads resulting in plastic deformations in the material.

Nevertheless, the fundamental properties of the phenomenon are described adequately both from qualitative and quantitative points of view in the frames. And the results of modelling with the use of software package Maple 18 fit well compressive tests for various real materials with non-linear strain-stress diagram.

We regard for simplicity the vertical column $AB$ of free length $l$ and uniform cross-section $(S)$ symmetrical with respect to the axis x of minimal gyration. The column is made of isotropic material with a typical for metals and polymers strain-stress diagram with conditional yield point (see later). The lower end $A$ of the column is pinned hardly while the upper one $B$ been exposed to the axial compressive load $F$, $N$ (figure 1). This vertical force provides normal stresses $\sigma_n$ beyond the elastic range on the diagram, and even greater than yield strength $\sigma_f$ above.

The choice of the pinned console solves problems with ambiguity of the bonds between the slope and the axis displacements. So, the transversal shift $y$ of the center of the $z$ cross-section and its vertical displacement $\Delta z$ are uniquely related to the current inclination $p = dy/dz$ of the console axis (figure 2).

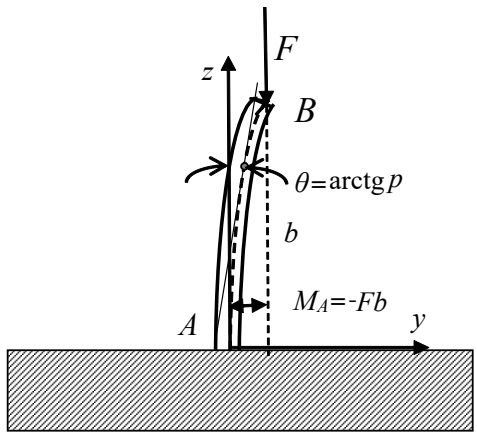


Figure 1. The uniform beam with pinned lower end $A$ and loaded with $F$ on the upper $B$

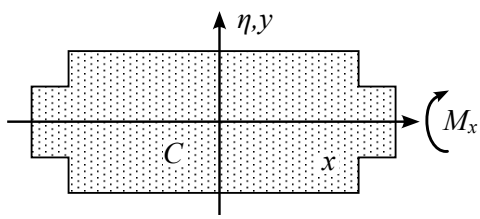


Figure 2. Symmetrical cross-section with zero odd momenta

Let's write down the fundamental relation between bending torque $M_x(z)$ and curvature radius $\rho$ $(z)$ in the same section $(z)$. First, within plane-section hypothesis we represent the normal strain in the layer with the local coordinate $\eta$ as $\varepsilon_\eta = \varepsilon_{ax} + \eta/\rho$ where $\varepsilon_{ax}$ is a compressive strain of the axis crossing the section $z$ in the center $C$.

Then regarding strain-stress diagram as obeying cubic law with respect to the access value we have

$$\sigma(\varepsilon) = \sigma(\varepsilon_{ax}) + \frac{d\sigma(\varepsilon_{ax})}{d\varepsilon}\frac{\eta}{\rho} + \frac{1}{2!}\frac{d^2\sigma(\varepsilon_{ax})}{d\varepsilon^2}\left(\frac{\eta}{\rho}\right)^2 + \frac{1}{3!}\frac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\left(\frac{\eta}{\rho}\right)^3.$$

Substituting this expression into the formula for the bending moment and taking into account the symmetry of the cross-section, we get

$$M_x(z) = \iint \sigma_z \eta dS = \frac{d\sigma(\varepsilon_{ax})}{d\varepsilon}\frac{J_x^{(II)}}{\rho} + \frac{1}{3!}\frac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\frac{J_x^{(IV)}}{\rho^3},$$

with $J_x^{(II)}$ and $J_x^{(IV)}$ being the 2-nd and 4-th momenta of inertia of the cross-section area. The 3-rd one drops due to the cross-section symmetry and we may generalize the concept of a cross-section symmetry in this way, i.e. $J_x^{(III)} = 0$.

Given that $1/\rho$ equals $y''_{zz}/(1+y'^2_z)^{3/2}$, we get the left-hand side and writing down the right-hand side the equation which determines the shape of the column buckled

$$\frac{d\sigma(\varepsilon_{ax})}{d\varepsilon}\frac{J_x^{(II)}y''_{zz}}{(1+y'^2_z)^{\frac{3}{2}}}+\frac{1}{6}\frac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\frac{J_x^{(IV)}(y''_{zz})^3}{(1+y'^2_z)^{\frac{9}{2}}}=-(Fy+M_A), \tag{1}$$

with $M_A=-Fb$ being a torque on hard seal $A$ (figure 1) and the value of $b$ as a transversal shift of the upper end $B$.

This equation is nonlinear on the senior second derivative but due to its autonomy, it can be lowered in its order and then solved within the framework of perturbative approach under the assumption that the second term with $J_x^{(IV)}$ is much smaller than the first one. Thus, making the substitute $v=y-b$ we get the equation with boundary condition

$$\begin{cases}\dfrac{J_x^{(II)}v''_{zz}}{(1+v'^2_z)^{\frac{3}{2}}}\left(\dfrac{d\sigma(\varepsilon_{ax})}{d\varepsilon}\right)+\dfrac{1}{6}\left(\dfrac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\right)\dfrac{J_x^{(IV)}v''^3_{zz}}{(1+v'^2_z)^{\frac{9}{2}}}=-Fv,\\ v(0)=-b,\quad v(z_B)=0,\quad v'_z(0)=0.\end{cases}$$

After substitution $v'=p$ and assigning the $p$ as an argument we get $v''_{zz}=p\cdot(dp)/(dv)$ and

$$\begin{cases}\dfrac{J_x^{(II)}p\frac{dp}{dv}}{(1+p^2)^{\frac{3}{2}}}\left(\dfrac{d\sigma(\varepsilon_{ax})}{d\varepsilon}\right)+\dfrac{1}{6}\left(\dfrac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\right)\dfrac{J_x^{(IV)}(p\frac{dp}{dv})^3}{(1+p^2)^{\frac{9}{2}}}=-Fv,\\ v(0)=-b,\quad v(p_f)=0,\end{cases}$$

where $p_f$ is a final slope at the end $B$.

After simple transformation we receive

$$\begin{cases}\dfrac{J_x^{(II)}dp^2}{(1+p^2)^{\frac{3}{2}}}\dfrac{d\sigma(\varepsilon_{ax})}{d\varepsilon}+\dfrac{1}{6}\dfrac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\dfrac{J_x^{(IV)}v^2dp^2}{(1+p^2)^{\frac{9}{2}}(\frac{dv^2}{dp^2})^2}=-Fdv^2,\\ v^2(0)=b^2,\quad v(p_f^2)=0,\end{cases}\quad\Leftrightarrow$$

$$\Leftrightarrow\left|\begin{matrix}v^2\equiv w\\ p^2\equiv s\end{matrix}\right|\begin{cases}\dfrac{dw}{ds}=-\dfrac{J_x^{(II)}}{F(1+s)^{\frac{3}{2}}}\dfrac{d\sigma(\varepsilon_{ax})}{d\varepsilon}-\dfrac{1}{6}\dfrac{d^3\sigma(\varepsilon_{ax})}{d\varepsilon^3}\dfrac{J_x^{(IV)}w}{F(1+s)^{\frac{9}{2}}(\frac{dw}{ds})^2},\\ w(0)=b^2,\quad w(p_f^2)=0.\end{cases} \tag{2}$$

In this equation the derivatives $(d\sigma(\varepsilon_{ax}))/(d\varepsilon)$ and $(d^3\sigma(\varepsilon_{ax}))/(d\varepsilon^3)$ depend on $p^2=s$ because both the strain $\varepsilon_{ax}$ of the axis and the normal stress $\sigma_{ax}$ at certain place caused by it depend the slope $p$ as

$$\sigma_{ax}(p)=\frac{F\cos\theta}{S}=\frac{F}{S(1+p^2)^{\frac{1}{2}}}=\frac{F}{S(1+s)^{\frac{1}{2}}}. \tag{3}$$

Also, the final slope $p_f$ is unknown and it should be found from some condition (see later). To solve (2) we should build and use a model strain-stress diagram both in direct and inverse type.

## 3. Modelling strain-stress diagram

We considered the diagrams which contain a) initial proportionality stage $\sigma(\varepsilon) = E\varepsilon$, the $E$ being Young's modulus, b) the yield stage containing *conditional yield point* $(\sigma_f, t)$, i.e. an inflection point with $(d^2\sigma(t))/(d\varepsilon^2) = 0$, c) and final *densification* stage with $(d^2\sigma(\varepsilon))/(d\varepsilon^2) > 0$.

The cubic formula meeting all the requirements above is as follows

$$\sigma(\varepsilon) = E\varepsilon - \frac{3E\mu}{2}\varepsilon^2 + \frac{E\mu}{2t}\varepsilon^2, \quad \mu = \frac{Et - \sigma_f}{Et^2}, \tag{4}$$

and it has the derivatives

$$\frac{d\sigma(\varepsilon)}{d\varepsilon} = E - 3E\mu\varepsilon + \frac{3E\mu}{2t}\varepsilon^2, \quad \frac{d^3\sigma(\varepsilon)}{d\varepsilon^3} = \frac{3E\mu}{t}. \tag{5}$$

(The parameter $\mu$ turns to zero at ideal linear diagram otherwise it describes in what extent the diagram is nonlinear. Namely, the greater $\mu$ the more non-linear $\sigma(\varepsilon)$-dependence and it manifests itself at smaller strains $\varepsilon$.)

For the equation (3) it corresponds to reverse approximate formula

$$\varepsilon(\sigma) = \frac{\sigma}{E} + \frac{3\mu\sigma^2}{2E^2} + \frac{\mu(9\mu t - 1)\sigma^3}{2E^3 t} \tag{6}$$

which gives identity with accuracy of $O(\varepsilon^4)$ when the stress value (3) is substituted into it.

Application of (4) for regression by Maple 2018 option "*Fit*" on experimental data received at students' practicum for low carbon steel compression test gives good match on the level of adjusted $R^2 = 0.999733$ of the data with the curve (3) (figure 3). An estimated Young's modulus $E$ lies in confidence (95%) interval (165; 175) GPa a little less of the handbook values of 180 ... 220 GPa. This is surely due to fatigue of the material as a result of numerous tests fulfilled by many generations of students in the workshop on material science at Yaroslavl branch of Moscow Institute of Transport Engineers.

The reversed formula (6) also fits well the data within yield stage though it doesn't contain a densification stage. Up to the beginning of the densification stage due to (3), the curves actually merge into a single line with discrepancies being of order of the residuals of estimation. Also, we see good quasi-linearity of the data in range of $\varepsilon s$ from about 0.007 to $\sim 0.017$ where the conditional yield point ($t = 0.0134$) is localized. This quasi-linearity justifies the use of the tangent modulus method in solving the Euler equation for buckled beam.

Not only for the steel but for other metals such as titan and wolfram the simple cubic formulas (4) and (6) fit well the experimental data. For the fluor polymers they hold too. Thus, for Al/PTFE (aluminum/polytetrafluoroethylene) the experimental data [10] fit well (4) (figure 4). Moreover, we see that

in the wide enough middle part of the diagram points fit well on a straight line corresponding to the tangent modulus $Et$ of about 50 MPa. Although the interpolating line does not emphasize this fact.

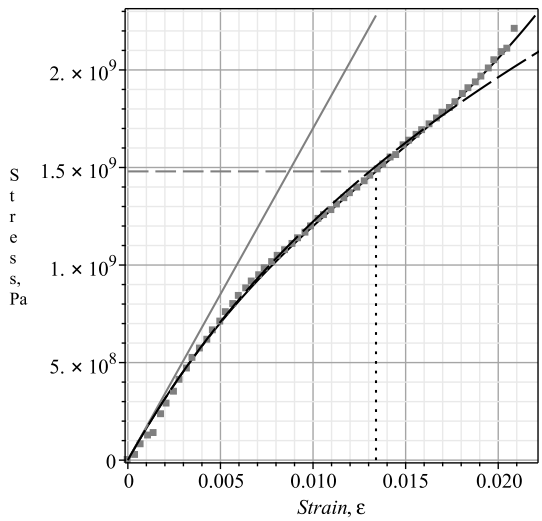


Figure 3. Cubic model direct (4) (grey solid) and reverse (6) (black long dash) diagrams built on experimental data (black diamonds) for low carbon steel. Hook's law (solid thin grey), yield strain $t = 0.0134$ (black dot), yield stress $\sigma_f = 1.48 \cdot 10^9$ Pa (grey dash)

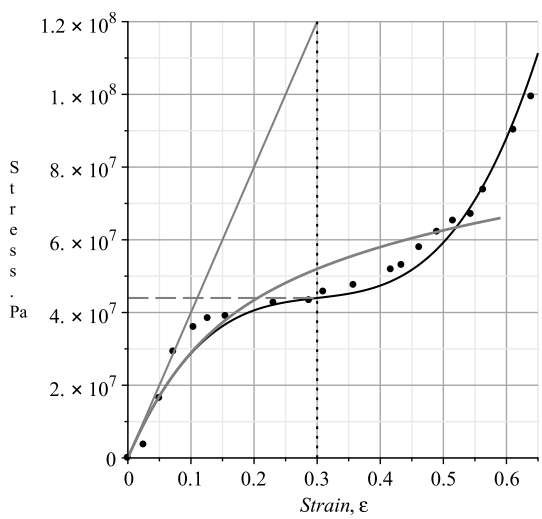


Figure 4. Al/PTFE strain-stress diagram: experiment [10] (solid circle); cubic model $\sigma(\varepsilon)$ (4) with the parameters $E = 400$ MPa, $\sigma_f = 44$ MPa, $t = 0.3$, $E_{\text{tang}} = 55.8$ MPa (solid black), Hook's law (solid thin grey), reversed diagram $\varepsilon(\sigma)$ (solid thick grey), yield strain $t = 0.3$ (black dot), yield strength (grey long dash)

All examples above confirm the effectiveness of a simple cubic formula for adequate describing the diagrams of many plastic materials, both metals and polymer composites, increasingly used in mechanical engineering, aeronautics and robotics.

## 4. Solving the equation

To write down the equation describing the buckling in inelastic regime we should substitute (6) in the formula (5) for the first derivative and limit the resulting expression to the first four members

$$\frac{d\sigma}{d\varepsilon} = E - 3\mu\sigma + \left(\frac{3\mu}{2Et} - \frac{9\mu^2}{2E}\right)\sigma^2 + \left(\frac{6\mu^2}{E^2t} - \frac{27}{2E^2}\right)\sigma^3. \tag{7}$$

Substituting the expression (3) for axial stress in (7) and then in (2) we receive the equation defining the dependence $w = v^2 \cdot vs \cdot s = p^2$.

$$\frac{dw}{ds} = -\frac{J_x^{(II)}}{F(1+s)^{\frac{3}{2}}} + \frac{3J_x^{(II)}\mu}{S(1+s)^2} - \frac{J_x^{(II)}\left(\frac{3\mu}{2Et} - \frac{9\mu^2}{2E}\right)}{S^2(1+s)^{\frac{5}{2}}}F - \\ - \frac{J_x^{(II)}\left(\frac{6\mu^2}{E^2t} - \frac{27}{2E^2}\right)}{S^3(1+s)^3}F^2 - \frac{\mu E}{2t}\frac{J_x^{(IV)}w}{F(1+s)^{\frac{9}{2}}\left(\frac{dw}{ds}\right)^2}. \tag{8}$$

Expressing the $w = w_0 + \delta w$ as a sum of the $w_0$ satisfying the equation (8) with $J_x^{(IV)} = 0$ and boundary conditions in (2), and a small additive $\delta w$ fitting zero boundary conditions at upper end $B$, we find formulas for the $w_0$ and $\delta w$:

$$w_0(s) = \frac{2J_x^{(II)}}{F(1+s)^{\frac{3}{2}}} - \frac{3J_x^{(II)}\mu}{S(1+s)^2} + \frac{2J_x^{(II)}\left(\frac{3\mu}{2Et} - \frac{9\mu^2}{2E}\right)}{3S^2(1+s)^{\frac{5}{2}}}F + \\ + \frac{J_x^{(II)}\left(\frac{6\mu^2}{E^2t} - \frac{27}{2E^2}\right)}{2S^3(1+s)^3}F^2 - b^2, \tag{9}$$

$$\delta w(s) = \frac{\mu E J_x^{(IV)}}{2Ft}\int_s^{p_f^2} \frac{w_0(s')ds'}{(1+s')^{\frac{9}{2}}\left(\frac{dw_0}{ds'}\right)^2}. \tag{9'}$$

From (8) a clear relationship between the transversal displacement $b$ of the upper end B of the console and its final slope $p_f$ follows

$$b = \left\{ \frac{2J_x^{(II)}}{F(1+p_f^2)^{1/2}} - \frac{3J_x^{(II)}}{S(1+p_f^2)^2} + \frac{2J_x^{(II)}\left(\frac{3\mu}{2Et} - \frac{9\mu^2}{2E}\right)}{3S^2(1+p_f^2)^{5/2}}F + \right.$$

$$+\frac{J_x^{(II)}\left(\frac{6\mu^2}{E^2t}-\frac{27}{2E^2}\right)}{2S^3(1+p_f^2)^3}F^2+\frac{\mu E J_x^{(IV)}}{2Ft}\int\limits_s^{p_f^2}\frac{w_0(s')ds'}{(1+s')^{\frac{9}{2}}\left(\frac{dw_0}{ds'}\right)^2}\Bigg\}^{1/2}. \quad (10)$$

Further calculations (see later) show that for real materials and standard cross sections (I,L-beam, channel, square, circle, etc.) the addition (9′) is at least 4 orders less than basic function (8). Thus, for a PTFE Teflon channel with a length of $l_0 = 0.75$ m and an area 64 times the area of channel No. 10 at a final slope of 0.5, the additive $\delta w$ was in maximum less than 0.01% of the basic function $w_0(s = p^2)$. And since they are added geometrically to form $v(p) = y(p) - b$, the contribution will be completely invisible (figure 5). So, we may easily neglect the integral amendment in (10).

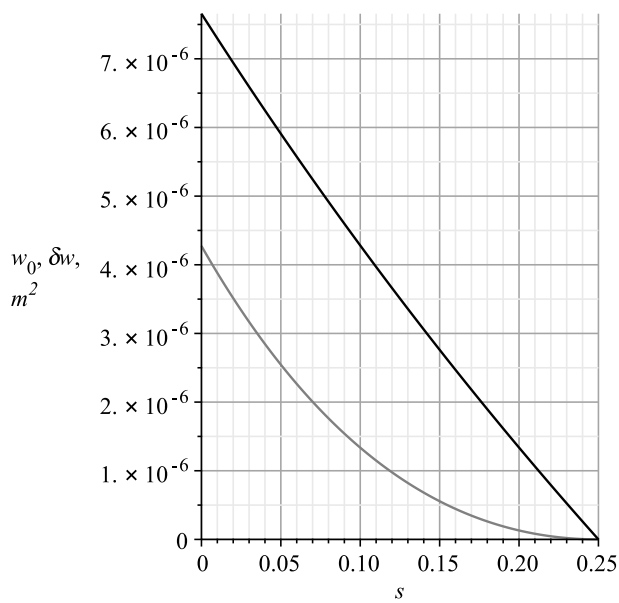


Figure 5. First order function $10^{-4}w_0(s = p^2)$ (black) (9) and additive $\delta w$ (grey) (9′) due to 4-th moment of inertia $J_x^{(IV)}$ for I-beam made of Teflon (PTFE), 0.75 m, 0.077 m$^2$, $J_x^{(II)} = 0.000065124$ m$^4$

## 5. Determining the final slope vs load dependence

Analyzing the expressions (8)–(10) it is easily seen that correct solution of ODE may be received only for known dependence of the load $F$ on the final slope $p_f$. To find it we are to compile so called characteristic equation on restored length of the console.

So, we have a transversal axis shifted coordinate $v_0(p) = y(p) - b = -\sqrt{w_0(p^2)}$.

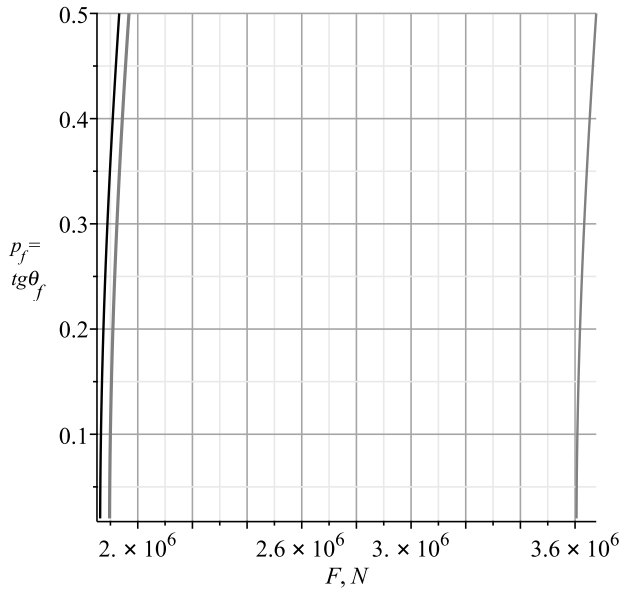


Figure 6. The $p_f(F)$ dependences for Al/PTFE I-console with $l_0 = 0.5$ m, $S = 0.077$ m$^2$, $J_x^{(II)} = 0.000065124$ m$^4$ within the three approaches: stress due (4) (black), tangent modulus (thick grey) and Hook's law (grey thin)

The longitudinal coordinate $z$ may be found as $z(p) = \int_0^p \frac{dv_0(p')}{p'}$ and elementary length of the axis as

$$dl(p) = \frac{dv_0(p)}{p}\sqrt{1+p^2} = -\frac{d\sqrt{w_0}}{ds}\cdot\frac{dw_0}{ds}\cdot\frac{ds}{dp}\cdot\frac{\sqrt{1+p^2}}{p} = -\frac{dw_0}{ds}\frac{\sqrt{1+p^2}}{\sqrt{w_0}}.$$

Being restored after the load is removed, this value becomes

$$dl_{\text{res}}(p) = \frac{dl(p)}{1-\varepsilon(p)} \approx \frac{dw_0}{dp^2}\cdot\frac{\sqrt{1+p^2}}{\sqrt{w_0}}\left(1+\varepsilon(p)+\varepsilon^2(p)+\varepsilon^3(p)\right),$$

where the strain $\varepsilon(p)$ taken positive, of the console axis element marked $p$ is received by substitution of the stress (3) in the reverse strain-stress diagram (6). By integrating by $p$ and equating the result with the free length of the console $l_0$, we obtain the function $p_f(F)$ specified implicitly:

$$-\int_0^{p_f}\frac{dw_0}{dp^2}\cdot\frac{\sqrt{1+p^2}}{\sqrt{w_0}}\left(1+\varepsilon(p)+\varepsilon^2(p)+\varepsilon^3(p)\right)dp = L(F,p_f) = l_0. \quad (11)$$

This key equation has non-empty solution if and only if the value of load $F$ exceeds some critical buckling force $F_{\text{cr}}$, The Maple package has successful

option *implicitplot* which builds precisely the graphs of implicitly specified functions.

To compare the results obtained at different approximations, dependencies $p_f(F)$ were also calculated for an ideal material with the same Young's modulus, as well as in the approximation of a tangent module when the first derivative in (7) was limited by the first two terms $\frac{d\sigma}{d\varepsilon}(\sigma) = E - 3\mu\sigma$.

Due to the availability of a quasilinear middle yield stage on the diagrams (figures 4) for Al/PTFE (aluminum/polytetrafluoroethylene) [ibid], the results for the cubic formula (4) obtained within the tangent modulus approach were very close, but hugely differing from the results within Hook's law (figure 5).

It is worth mentioning that the extremely large loads were chosen exclusively to reach the stresses close to yield strength $\sigma_f$. For the same reason, the geometric parameters of the console were chosen, so that its flexibility $\lambda$ varied from $\sim 5$ to $\sim 20$.

So, we see that regime of plastic deformations diminishes cardinally the classical critical load $F_{\text{cr}}$ predicted by Hook's law approach studied in most universities. Especially it takes place for the materials with low yield strength such as Teflon, polymers in general and composites based on them.

Also, we see that relatively simple tangent modulus approach gives the results extremely close to those received by modeling strain-stress diagram by cubic formula (4) with conditional yield point.

The cross-section symmetry in generalized meaning, i.e., $J_x^{(III)} = 0$ simplifies significantly the calculation due to the absence of a next-in-rank additive. And as for this for the 4-th gyration moment it occurs quite negligible so we may limit ourselves to only the terms containing the second moment of the cross-section. As for widely used non-symmetrical cross-sections such as L-beam the 3-rd moment doesn't equal exactly to zero but very close to it due to the area quasi-anti-symmetry. Thus, the method developed may be implemented for wide class of constructive profiles.

## 6. Buckling shape

The shape of the buckled console is easily calculated parametrically from the above formulas for the longitudinal $z(p)$ and transversal $y(p)$ coordinates. The shape was calculated in all three approximations: the plastic deformations of the axis due to a model cubic diagram with a conditional yield strength, in the approximation of the tangent modulus, and for Hook's law. The load $F$ for each case was taken to provide the same final slope $p_f = 0.5$ (figures 7, 8).

We see that the shapes are extremely close to each other though the compressed lengths differ significantly especially for ideal Hook's case with the load almost 2 times greater of those for the rest two approaches. As for the case of the low carbon steel with much greater Young's modulus the identity of the shapes was really ideal (figures 8) for different beam lengths. Due to approximate proportionality of the loads providing the same final slope within different approaches one may suggest not to solve the unwieldy case of cubic diagram (4) but to use a simple Hookean case with the subsequent recalculation of forces.

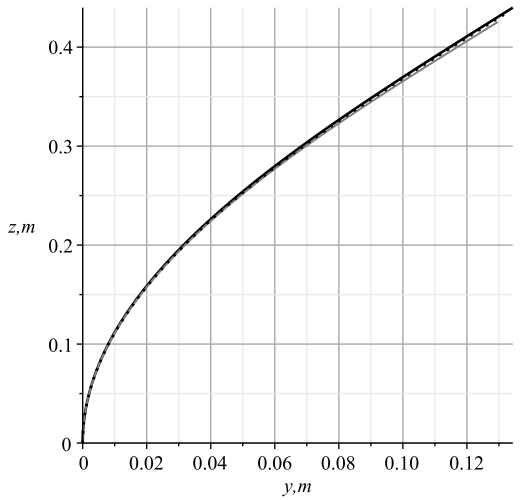


Figure 7. Quasi-identity of the shape of the Al/PTFE I-console, $l_0 = 0.5$ m. 0.077 m$^2$, $J_x^{(II)} = 0.000065124$ m$^4$ buckled under the loads $F(p_f = 0.5)$ within the three approaches: plastic strain (4) under $F = 1.932 \cdot 10^6 N$ (black solid), tangent modulus with $F = 1.97 \cdot 10^6 N$ (black dot) and Hook's law with $F = 3.68 \cdot 10^6 N$ (grey)

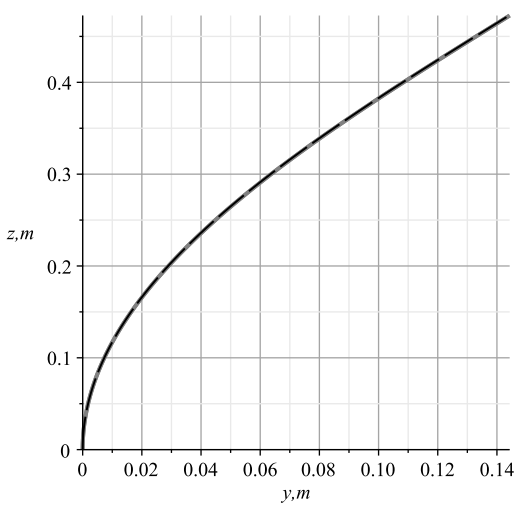


Figure 8. Really complete shape identity for the steel I-console No 10, $l_0 = 0.5$ m under loads resulting in equal $p_f$ of 0.5 due to ideal Hook's (black long dash) and cubic (solid grey) strain-stress diagram

## 7. Conclusion

So, we prove that the suggested numerical method of Euler problem solution within a plane section hypothesis using the Maple software is quite effective and implementable in the range of plastic strains. The software is versatile

useful for solving many related sub-problems such as bringing together similar terms, expansion expression into a series, curve fitting, nonlinear estimation of parameters from experimental data, plotting an implicitly specified function, 3D-plots, etc. The algebraic type of the functions involved which is provided by a lucky choice of integration variable facilitates the computational process and gives a gain in speed compared to analogous use of transcendental and moreover special functions. Therefore, the method may be further generalized on more complicate case of piecewise uniform beam.

Solving this kind problems, when any even minor invisible error can mislead student to qualitatively wrong results and conclusions, disciplines him and eventually makes him a specialist in mathematical modeling in a wide range of sciences. The specialist who is critical of "ready-made solutions" in the form of convenient commercial software products which may solve well one class of problems and occur useless and distractive for another one.

## References

[1] T. H. G. Megson, "Columns," in *Aircraft Structures for Engineering Students*, 6th. Elsevier Ltd., 2022, pp. 253–324.

[2] F. R. Shanley, "Inealstic Column Theory," *Journal of Aeronautical Sciences*, vol. 14, no. 5, pp. 261–280, 1947.

[3] A. Afroz and T. Fukui, "Numerical Analysis II: Branch Switching," in *Bifurcation and Buckling Structures*, 1st. CRC Press, 2021, p. 12.

[4] N. Shuang, J. R. Kim, and F. F. Rasmussen, "Local-Global Interaction Buckling of Stainless Steel I-Beams. II: Numerical Study and Design," *Journal of Structural Engineering*, vol. 141, no. 8, p. 04 014 195, 2014. DOI: 10.1061/(ASCE)ST.1943-541X.0001131.

[5] F. Shenggang, D. Daoyang, Z. Ting, *et al.*, "Experimental Study on Stainless Steel C-columns with Local-Global Interaction Buckling," *Journal of Constructional Steel Research*, vol. 198, no. 2, p. 107 516, 2022. DOI: 10.1016/j.jcsr.2022.107516.

[6] S. P. Timoshenko and J. M. Gere, *Theory of Elastic Stability*. NewYork, USA: McGraw-Hill, 1961.

[7] K. L. Nielsen and J. W. Hutchinson, "Plastic Buckling of Columns at the Micron Scale," *International Journal of Solids and Structures*, vol. 257, no. 5, p. 111 558, 2022. DOI: 10.1016/j.ijsolstr.2022.111558.

[8] A. Bedford and K. M. Liechti, "Buckling of Columns," in *Mechanics of Materials*, Springer, Cham., 2020. DOI: 10.1007/978-3-030-22082-2_10.

[9] Z. P. Bazant, "Shear buckling of sandwich, fiber-composite and lattice columns, bearings and helical springs: paradox resolved," *ASME Journal of Applied Mechanics*, vol. 70, pp. 75–83, 2003. DOI: 10.1115/1.1509486.

[10] C. Chuang, G. Zihan, and T. Enling, "Determination of Elastic Modulus, Stress Relaxation Time and Thermal Softening Index in ZWT Constitutive Model for Reinforced Al/PTFE," *Polymers*, vol. 15, p. 702, 2023. DOI: 10.3390/polym15030702.

**For citation:**

V. V. Chistyakov, S. M. Soloviev, Buckling in inelastic regime of a uniform console with symmetrical cross section: computer modeling using Maple 18, Discrete and Continuous Models and Applied Computational Science 31 (2) (2023) 174–188. DOI: 10.22363/2658-4670-2023-31-2-174-188.

**Information about the authors**:

**Chistyakov, Viktor Vladimirovich** (Russian Federation) — Candidate of Sciences in Physics and Mathematics, Senior Researcher of Laboratory of Physics of Rare Earth Semiconductors of Ioffe Physical-Technical Institute of the Russian Academy of Sciences (e-mail: `v.chistyakov@mail.ioffe.ru`, phone: +7(981) 815-74-95, ORCID: https://orcid.org/0000-0003-4574-0857, ResearcherID: F-9868-2016, Scopus Author ID: 44461256400)

**Soloviev, Sergey Mikhailovich** — Candidate of Sciences in Physics and Mathematics, Leading Researcher (Head of Laboratory) of Laboratory of Physics of Rare Earth Semiconductors of Ioffe Physical-Technical Institute of the Russian Academy of Sciences (e-mail: `serge.soloviev@mail.ioffe.ru`, phone: +7 (921) 439-62-13, ORCID: https://orcid.org/0000-0002-9019-7382, ResearcherID: D-5128-2015, Scopus Author ID: 7101661580)

УДК 519.624.2:531.8
PACS 62.20
DOI: 10.22363/2658-4670-2023-31-2-174-188
EDN: XEAYRS

# Продольный изгиб однородной консоли с симметричным сечением в режиме пластических деформаций: численное моделирование посредством Maple 18

**В. В. Чистяков, С. М. Соловьёв**

*Лаборатория физики редкоземельных полупроводников,*
*Физико-технический институт им. А. Ф. Иоффе РАН,*
*Политехническая ул., д. 26, Санкт-Петербург, 194021, Россия*

**Аннотация.** Представлен способ численного моделирования посредством Maple 2018 продольного изгиба однородной консоли с симметричным сечением в режиме пластических деформаций. Получено обыкновенное дифференциальное уравнение для поперечной координаты, учитывающее высшие моменты инерции сечения. В качестве аргумента в нём служил уникальный для каждого места безразмерный наклон консоли $p = \tan\theta$, взаимно однозначно связанный со всеми перемещениями. Диаграммы сжатия реальных материалов (сталь, титан, тефлон, алюминий-тефлон) моделировались в Maple при помощи нелинейной регрессии на экспериментальных и литературных данных с использованием полинома 3-го порядка, обеспечивающего условный предел текучести $(t,\sigma_f)$. Параметры консоли (длина $l_0$, площадь сечения $S$ и минимальный момент инерции $J_x$) подбирались так, чтобы изгибающая сила обеспечивала напряжение вблизи предела текучести $\sigma_f$. Для нахождения ключевой зависимости углового наклона свободного конца $p_f$ от закритической нагрузки $F > F_{\rm cr}$, что необходимо для определения формы прогиба, использовалось равенство проинтегрированной восстановленной элементарной длины её свободному значению $l_0$. Зависимости $p_f(F)$ и $y(z)$, $z$ — продольная координата, рассчитывались в рамках следующих трёх подходов: пластический характер деформаций согласно полиномиальной $(n = 3)$ диаграмме, приближение касательного модуля $E_{\rm tang}$ и приближение идеальной выполнимости закона Гука. Обнаружено, что в реальном случае пластических деформаций критическая нагрузка $F_{\rm cr}$ почти вдвое меньше, чем в идеальном случае. При этом наблюдается почти идентичность формы изгиба консоли в рамках этих трёх подходов при одинаковом конечном наклоне $p_f$, особенно для металлов.

**Ключевые слова:** проблема Эйлера, гипотеза плоских сечений, выгибание, консоль, пластические деформации, диаграмма сжатия, условный предел текучести, критическая выгибающая сила, программирование на Maple, нелинейная оценка, тефлон Al/PTFE, сталь