Київський національний університет імені Тараса Шевченка

Журнал обчислювальної та прикладної

2018 **МАТЕМАТИКИ** 2 (128)

Заснований у 1965 році

Адреса редакції: 03022 Київ, пр. Глушкова, 4 д
Київський національний університет імені Тараса Шевченка,
факультет кібернетики, кафедра обчислювальної математики,
тел.: (044) 259-04-36, E-mail: opmjournal@gmail.com
http://www.opmj.univ.kiev.ua

Адреса редакції серії: 79000 Львів, вул. Університетська, 1
Львівський національний університет імені Івана Франка,
Кафедра обчислювальної математики,
тел.: (032) 239-43-91, E-mail: kom@franko.lviv.ua
http://jnam.lnu.edu.ua

Затверджено Вченою радою факультету кібернетики
від 8 жовтня 2018 р. (протокол № 2)

Свідоцтво про державну реєстрацію КВ 4246 від 26.05.2000

Підписано до друку 8 жовтня 2018 р.

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 519.6

ON THE NUMERICAL SOLUTION OF A MIXED BOUNDARY VALUE PROBLEM FOR THE ELLIPTIC EQUATION WITH VARIABLE COEFFICIENTS IN DOUBLY CONNECTED PLANAR DOMAINS

A. V. Beshley

Резюме. Ми розглядаємо чисельне розв'язування мішаної задачі для еліптичного рівняння другого порядку зі змінними коефіцієнтами у двозв'язній області. Розв'язок задачі подається у вигляді суми потенціалів з невідомими густинами і функцією Леві у якості ядра. Підставляючи подання розв'язку в основне рівняння та дві крайові умови, ми отримуємо систему гранично-просторових інтегральних рівнянь. Заміна змінних приводить до параметризованої системи, яка трансформується у систему лінійних алгебричних рівнянь після застосування квадратур та колокації апроксимаційних рівнянь у відповідних вузлах. Наприкінці наведено деякі чисельні результати.

Abstract. We consider a numerical solution of a mixed boundary value problem for the second-order elliptic equation with variable coefficients in a doubly connected domain. A solution of the problem is represented as a sum of potentials with unknown densities and Levi function as a kernel. Substituting the solution representation in the main equation and two boundary conditions we obtain a system of boundary-domain integral equations. The change of variables leads to the parameterised system which is being transformed in a system of linear algebraic equations after quadratures application and collocation of the approximating equations at appropriate points. Some numerical results are provided at the end.

1. Introduction

The elliptic differential equations with variable coefficients are widely spread in many problems of mathematical physics. The coefficients presented in a differential operator mostly correspond to the specific material parameters (for instance, thermal, electrical or hydraulic conductivity) of a considered physical process.

There are well-known effective methods (the boundary element method, the boundary integral equation method) for solving problems defined in bounded or infinite domains. The main advantage of these approaches is decreasing of the dimension of the problem – the solution in a domain can be represented using specific expression only over the boundary. However, in this case, a fundamental solution for a general differential operator is required. Unfortunately, a fundamental solution, in general, is unknown for differential equations with

Key words. Elliptic equation with variable coefficients, mixed boundary value problem, parametrix, boundary-domain integral equations, quadrature formulas.

variable coefficients or its finding can be quite complicated (in contrast to equations with constant coefficients). Therefore, efficient methods to solve such kind of problems are welcomed.

One of the approaches that has been proposed for the numerical solution of so-called the generalized Laplace equation [9] (a second-order linear elliptic partial differential equation with variable coefficients) is described in [10]. The main idea is to transform the starting equation with variable coefficients into a constant-coefficient equation for which a fundamental solution is available and then any of mentioned above effective methods can be applied. The first step in the procedure is to avoid the first partial derivatives of the unknown function and next step is to approximate the transformed equation using constant coefficients.

It is not mandatory to obtain the constant-coefficient equation to solve the problem. As an example, in [1] for solving a two-dimensional mixed problem (where the Dirichlet condition prescribed on a part of the boundary and the Neumann condition prescribed on the remaining part of the domain boundary) with variable coefficients a special function (parametrix) has been used in the Green formula to reduce the initial boundary value problem to a boundary-domain integral equation or boundary-domain integro-differential-equation with the following discretisation of the domain and application of the collocation method. Another similar technique for solving this problem, but with using the radial integration method [5], has been proposed in [2]. The radial integration method was employed to convert domain integrals into equivalent boundary integrals.

In this paper, we consider the numerical solution of a mixed boundary value problem in a doubly connected domain where the Neumann condition is defined on the outer boundary, meanwhile as the Dirichlet condition prescribed on the inner boundary.

Let D_0 be a simple bounded domain in $\mathbb{R}^2$ with boundary $\Gamma_0 \in C^2$. Let D_{-1} be a domain bounded by curve $\Gamma_{-1} \in C^2$ and $\overline{D}_{-1} \subset D_0$. We define that $D = D_0 \setminus \overline{D}_{-1}$. We consider the following mixed boundary value problem in the doubly connected planar domain D for elliptic equation with variable coefficients: need to find function $u \in H^1(D)$ that satisfies the differential equation

$$Lu(x) = \operatorname{div}(\sigma(x) \operatorname{grad} u(x)) = 0, \quad x \in D, \tag{1}$$

the Dirichlet condition on Γ_{-1}

$$u = f_1 \ \text{ on } \ \Gamma_{-1} \tag{2}$$

and the Neumann condition on Γ_0

$$\sigma \frac{\partial u}{\partial \nu} = f_2 \ \text{ on } \ \Gamma_0. \tag{3}$$

Here, $\sigma \in C^\infty(\overline{D})$, $\sigma > 0$, f_1, f_2 are known functions and ν is the outward unit normal to the boundary.

This problem can be interpreted as a stationary heat transfer problem in an isotropic medium for a two-dimensional bounded body with prescribed temperature and heat flux on different boundaries. Since the main equation is homogeneous we assume that a heat source is not available. The function $\sigma(x)$, in this case, is a known thermal conductivity.

For the outline of the work, in Section 2, we reduce our differential problem to a system of boundary-domain integral equations, obtain an equivalent system in a parameterised form and split singularities from some kernels. A full discretisation of the system with applied quadratures and approximation formula of the solution in a domain are presented in Section 3. In Section 4, two numerical examples for different domain configurations are considered. Some conclusions are given in Section 5.

2. Reduction to a system of boundary-domain integral equations

As it was mentioned above, there is no ability to reduce the problem to a boundary integral equation as a fundamental solution is not available in the explicit form, in general case, for elliptic equations with variable coefficients. But, we can use a parametrix to work only with integrals instead of the differential equation and boundary conditions, however, it leads to domain integrals appearing. A parametrix (or Levi function) $P(x,y)$, $x,y \in \mathbb{R}^2$ should satisfy the following expression [8]

$$L_x P(x,y) = \delta(x-y) + R(x,y), \tag{4}$$

where δ is the Dirac function and the remainder function R has a weak singularity for $x = y$. In the two-dimensional case we can define the parametrix as the fundamental solution with frozen coefficients $a(x) = a(y)$ corresponding to the operator L, i.e., in the form

$$P(x,y) = \frac{\ln|x-y|}{2\pi\sigma(y)}, \quad x,y \in \mathbb{R}^2, \;\; x \neq y$$

with the remainder function

$$R(x,y) = \frac{(x-y)\cdot \operatorname{grad}\sigma(x)}{2\pi\sigma(y)|x-y|^2}, \quad x,y \in \mathbb{R}^2 \;\; x \neq y.$$

It is not difficult to verify that functions $P(x,y)$ and $R(x,y)$ satisfy (4). Should note that the parametrix function is not unique.

We seek the solution as a sum of potentials, but instead of the fundamental solution of the differential operator we use the Levi function

$$\begin{aligned} u(x) = &\int\limits_D \psi(y) P(x,y)\, dy + \int\limits_{\Gamma_{-1}} \psi_{-1}(y) P(x,y)\, ds(y) + \\ &+ \int\limits_{\Gamma_0} \psi_0(y) P(x,y)\, ds(y), \quad x \in D, \end{aligned} \tag{5}$$

where $\psi \in C(D)$, $\psi_{-1} \in C(\Gamma_{-1})$ and $\psi_0 \in C(\Gamma_0)$ are unknown densities. Substituting (5) in (1)-(3) we obtain the following system of a boundary-domain integral equations

$$
\begin{cases}
\psi(x) + \int\limits_D \psi(y) R(x,y)\, dy + \int\limits_{\Gamma_{-1}} \psi_{-1}(y) R(x,y)\, ds(y) + \\
\qquad + \int\limits_{\Gamma_0} \psi_0(y) R(x,y)\, ds(y) = 0,\ x \in D, \\
\\
\int\limits_D \psi(y) P(x,y)\, dy + \int\limits_{\Gamma_{-1}} \psi_{-1}(y) P(x,y)\, ds(y) + \\
\qquad + \int\limits_{\Gamma_0} \psi_0(y) P(x,y)\, ds(y) = f_1(x),\ x \in \Gamma_{-1}, \\
\\
-\frac{1}{2}\psi_0(x) + \int\limits_D \psi(y)\sigma(x) \frac{\partial P(x,y)}{\partial \nu(x)}\, dy + \\
\qquad + \int\limits_{\Gamma_{-1}} \psi_{-1}(y)\sigma(x) \frac{\partial P(x,y)}{\partial \nu(x)}\, ds(y) + \\
\qquad + \int\limits_{\Gamma_0} \psi_0(y)\sigma(x) \frac{\partial P(x,y)}{\partial \nu(x)}\, ds(y) = f_2(x), \quad x \in \Gamma_0.
\end{cases}
\tag{6}
$$

If $\sigma(x) = 1$ then the density $\psi(x)$ vanishes (together with domain integrals) and the system is being simplified to a system of boundary integral equations that correspond to the Laplace equation. The similar system for this case can be found in [4].

Let D is symmetric relative to the origin and assume that the closed boundary curves Γ_0, Γ_{-1} are homothetic with factor ξ_{-1} and have the following representations

$$
\begin{aligned}
\Gamma_0 &= \{x(t) = (x_1(t), x_2(t)),\ t \in [0, 2\pi)\}, \\
\Gamma_{-1} &= \{x_{-1}(t) = (\xi_{-1} x_1(t), \xi_{-1} x_2(t)),\ t \in [0, 2\pi)\},
\end{aligned}
\tag{7}
$$

where ξ_{-1} is a fixed parameter and $0 < \xi_{-1} < 1$. To obtain the system in the parametrized form we use the change of variables in the integrals over domain in (6)

$$
y_1 = p_1(\xi, \tau) = \xi x_1(\tau),
$$

$$
y_2 = p_2(\xi, \tau) = \xi x_2(\tau),
$$

where $(\xi, \tau) \in \Pi = (\xi_{-1}, 1) \times [0, 2\pi)$ and Jacobian $J(\xi, \tau) = \xi(x_1(\tau) x_2'(\tau) - x_2(\tau) x_1'(\tau))$. The notation $p = (p_1, p_2)$ is used for the function mapping into Π.

This yields the following system

$$
\begin{cases}
\varphi(\eta,t)+\dfrac{1}{2\pi}\displaystyle\int_{\Pi}\varphi(\xi,\tau)\widetilde{R}(\eta,t;\xi,\tau)\,d\tau d\xi+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{-1}(\xi_{-1},\tau)\widetilde{R}_{-1}(\eta,t;\xi_{-1},\tau)\,d\tau+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{0}(\tau)\widetilde{R}_{0}(\eta,t;\tau)\,d\tau=0,\ (\eta,t)\in\Pi, \\
\dfrac{1}{2\pi}\displaystyle\int_{\Pi}\varphi(\xi,\tau)\check{P}(\xi_{-1},t;\xi,\tau)\,d\tau d\xi+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{-1}(\xi_{-1},\tau)\check{P}_{-1}(\xi_{-1},t;\xi_{-1},\tau)\,d\tau+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{0}(\tau)\check{P}_{0}(\xi_{-1},t;\tau)\,d\tau=\widetilde{f}_1(\xi_{-1},t),\ t\in[0,2\pi), \\
-\dfrac{1}{2}\varphi_0(t)+\dfrac{1}{2\pi}\displaystyle\int_{\Pi}\varphi(\xi,\tau)\widehat{P}(t;\xi,\tau)\,d\tau d\xi+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{-1}(\xi_{-1},\tau)\widehat{P}_{-1}(t;\xi_{-1},\tau)\,d\tau+ \\
\qquad\qquad +\dfrac{1}{2\pi}\displaystyle\int_{0}^{2\pi}\varphi_{0}(\tau)\widehat{P}_{0}(t;\tau)\,d\tau=\widetilde{f}_2(t),\ t\in[0,2\pi),
\end{cases}
\tag{8}
$$

with the functions $\varphi(\eta,t)=\psi(p(\eta,t))$, $\varphi_{-1}(t)=\psi_{-1}(x(t))$, $\varphi_0(t)=\psi_0(x(t))$, $\widetilde{f}_1(t)=f_1(x_{-1}(t))$, $\widetilde{f}_2(t)=f_2(x(t))$ and kernels

$$\widetilde{R}(\eta,t;\xi,\tau)=2\pi R(p(\eta,t),p(\xi,\tau))J(\xi,\tau),$$

$$\widetilde{R}_0(\eta,t;\tau)=2\pi R(p(\eta,t),x(\tau))|x'(\tau)|;$$

$$\check{P}(\xi_{-1},t;\xi,\tau)=2\pi P(\xi_{-1}x(t),p(\xi,\tau))J(\xi,\tau),$$

$$\check{P}_0(\xi_{-1},t;\tau)=2\pi P(\xi_{-1}x(t),x(\tau))|x'(\tau)|;$$

$$\widehat{P}(t;\xi,\tau)=2\pi\sigma(x(t))\frac{\partial P(x(t),\xi x(\tau))}{\partial\nu(x(t))}J(\xi,\tau),$$

$$\widehat{P}_0(t;\tau)=2\pi\sigma(x(t))\frac{\partial P(x(t),x(\tau))}{\partial\nu(x(t))}|x'(\tau)|;$$

$$\widetilde{R}_{-1}(\eta,t;\xi_{-1},\tau)=2\pi R(p(\eta,t),\xi_{-1}x(\tau))\xi_{-1}|x'(\tau)|;$$

$$\check{P}_{-1}(\xi_{-1},t;\xi_{-1},\tau)=2\pi P(\xi_{-1}x(t),\xi_{-1}x(\tau))\xi_{-1}|x'(\tau)|;$$

$$\widehat{P}_{-1}(t;\xi_{-1},\tau) = 2\pi\sigma(x(t))\frac{\partial P(x(t),\xi_{-1}x(\tau))}{\partial\nu(x(t))}\xi_{-1}|x'(\tau)|.$$

Exploring the kernels it is easy to see that the kernels $\widetilde{R}$ and $\check{P}_{-1}$ have different singularities. The strong singularity in $\widetilde{R}$ can be handled by applying the ideas from [7] (for more details see [3]). The logarithmic singularity in the kernel $\check{P}_{-1}$ can be split [6] as follows

$$\check{P}_{-1}(\xi_{-1},t;\xi_{-1}\tau) = \check{P}^{(1)}_{-1}(\xi_{-1},\tau)\ln\frac{4}{e}\sin^2\frac{t-\tau}{2} + \check{P}^{(2)}_{-1}(\xi_{-1},t;\xi_{-1}\tau) \tag{9}$$

with

$$\check{P}^{(1)}_{-1}(t,\tau) = \frac{1}{2}\frac{\xi_{-1}|x'(\tau)|}{\sigma(\xi_{-1}x(\tau))},$$

and

$$\check{P}^{(2)}_{-1}(t,\tau) = \frac{\xi_{-1}|x'(\tau)|}{\sigma(\xi_{-1}x(\tau))}\begin{cases} \dfrac{1}{2}\ln\dfrac{|\xi_{-1}x(t)-\xi_{-1}x(\tau)|^2}{\frac{4}{e}\sin^2\frac{t-\tau}{2}} & \text{for} \quad t\neq\tau, \\ \dfrac{1}{2}\ln\left(e|\xi_{-1}x'(t)|^2\right) & \text{for} \quad t=\tau. \end{cases}$$

3. Full discretisation and numerical solution of the system

For solving the system (8) we use the interpolation quadrature rules for continuous integrands and integrands with weight function that corresponds to the specific singularity. For continuous integrands we use

$$\frac{1}{2\pi}\int\limits_{\Pi} g(\xi,\tau)d\tau d\xi \approx \frac{1}{2n}\sum_{k=1}^{N}\sum_{i=0}^{2n-1}\alpha_k g(\eta_k,t_i), \tag{10}$$

$$\frac{1}{2\pi}\int_0^{2\pi} f(\tau)\,d\tau \approx \frac{1}{2n}\sum_{k=0}^{2n-1} f(t_k). \tag{11}$$

The following quadratures are used for integrals with strong and logarithmic singularities

$$\frac{1}{2\pi}\int\limits_{\Pi} g(\xi,\tau)\cot\frac{\tau-t}{2}d\tau d\xi \approx \sum_{k=1}^{N}\sum_{i=0}^{2n-1}\alpha_k g(\eta_k,t_i)T_i(t), \tag{12}$$

$$\frac{1}{2\pi}\int_0^{2\pi} f(\tau)\ln\left(\frac{4}{e}\sin^2\frac{t-\tau}{2}\right)d\tau \approx \sum_{k=0}^{2n-1} f(t_k)\,F_k(t), \tag{13}$$

In formulas (10), (13) $\alpha_k \in \mathbb{R}^2$ are quadrature weights, $\eta_k \in (0,1)$, $k = 1,\dots,N$ – some quadrature points. For 2π-periodic integrals we employ the trapezoidal quadrature rule based on trigonometric interpolation with equidistant points $t_i = i\pi/n,\ i = 0,\dots 2n-1,\ n \in \mathbb{N}$. The weight functions $T_i(t)$ and

$F_k(t)$ are defined as follows

$$T_i(t) = -\frac{1}{n}\sum_{m=1}^{n-1}\sin m(t-t_i) - \frac{1}{2n}\sin n(t-t_i),$$

$$F_k(t) = -\frac{1}{2n}\left(1+2\sum_{m=1}^{n-1}\frac{1}{m}\cos m(t-t_k)+\frac{1}{n}\cos n(t-t_k)\right).$$

The use of these quadratures in (8) and collocation of the approximating equations at quadrature points lead to the linear system

$$\begin{cases} \varphi_{mi} + \sum_{k=1}^{N}\sum_{j=0}^{2n-1}\alpha_k\varphi_{kj}\bar{R}(\eta_m,t_i;\eta_k,t_j)+ \\ \qquad +\frac{1}{2n}\sum_{j=0}^{2n-1}\varphi_{-1j}\widetilde{R}_{-1}(\eta_m,t_i;\xi_{-1},t_j)+ \\ \qquad +\sum_{j=0}^{2n-1}\varphi_{0j}\widetilde{R}_0(\eta_m,t_i;t_j)=0, \\ \\ \frac{1}{2n}\sum_{k=1}^{N}\sum_{j=0}^{2n-1}\alpha_k\varphi_{kj}\check{P}(\xi_{-1},t_i;\eta_k,t_j)+\frac{1}{2n}\sum_{j=0}^{2n-1}\varphi_{0j}\check{P}_0(\xi_{-1},t_i,t_j)+ \\ +\sum_{j=0}^{2n-1}\varphi_{-1j}\left[\check{P}_{-1}^{(1)}(\xi_{-1},t_j)F_j(t_i)+\frac{1}{2n}\check{P}_{-1}^{(2)}(\xi_{-1},t_i;\xi_{-1},t_j)\right]=\widetilde{f}_{1i}, \\ \\ -\frac{1}{2}\widetilde{\varphi}_{0i}+\frac{1}{2n}\sum_{k=1}^{N}\sum_{j=0}^{2n-1}\alpha_k\varphi_{kj}\widehat{P}(t_i;\eta_k,t_j)+ \\ \qquad +\frac{1}{2n}\sum_{j=0}^{2n-1}\varphi_{-1j}\widehat{P}_{-1}(t_i;\xi_{-1},t_j)+ \\ \qquad +\frac{1}{2n}\sum_{j=0}^{2n-1}\varphi_{0j}\widehat{P}_0(t_i,t_j)=\widetilde{f}_{2i}, \end{cases} \tag{14}$$

with

$$\bar{R}(\eta_m,t_i;\eta_k,t_j)=\begin{cases} \frac{1}{2n}\widetilde{R}(\eta_m,t_i;\eta_k,t_j) & \text{for} \quad m\neq k, \\ \\ \frac{1}{2n}\widetilde{R}^{(1)}(\eta_m,t_i;\eta_k,t_j)+T_j(t)\widetilde{R}^{(2)}(\eta_m,t_i;\eta_k,t_j) & \\ \qquad\qquad \text{for} \quad m=k, \end{cases}$$

and the right-hand side $\widetilde{f}_{1i}=\widetilde{f}_1(t_i)$ and $\widetilde{f}_{2i}=\widetilde{f}_2(t_i)$.

Here, we use the following notation $\varphi_{mi}\approx\varphi(\eta_m,t_i)$, $\varphi_{-1i}\approx\varphi_{-1}(t_i)$ and $\varphi_{0i}\approx\varphi_0(t_i)$ for $m=1,\ldots,N$ and $i=0,\ldots,2n-1$. The kernels $\widetilde{R}^{(1)}$ and $\widetilde{R}^{(2)}$ are smooth functions and their representations are provided in [3].

Solving the system (14) we obtain the approximate values of unknown densities. Having these values we can find the approximation of the solution (1)-(3) in the domain D using the following formula

$$\begin{aligned} u(\eta_m, t_i) \approx & \sum_{k=1}^{N} \sum_{j=0}^{2n-1} \alpha_k \varphi_{kj} \overline{P}(\eta_m, t_i; \eta_k, t_j) + \\ & + \frac{1}{2n} \sum_{j=0}^{2n-1} \varphi_{-1j} \widetilde{P}_{-1}(\eta_m, t_i; \xi_{-1}, t_j) + \\ & + \frac{1}{2n} \sum_{j=0}^{2n-1} \varphi_{0j} \widetilde{P}_0(\eta_m, t_i; t_j), \end{aligned} \tag{15}$$

with

$$\overline{P}(\eta_m, t_i; \eta_k, t_j) = \begin{cases} \dfrac{1}{2n} \widetilde{P}(\eta_m, t_i; \eta_k, t_j) & \text{for} \quad m \neq k, \\ \widetilde{P}^{(1)}(\eta_m, t_i; \eta_k, t_j) F_j(t_i) + \dfrac{1}{2n} \widetilde{P}^{(2)}(\eta_m, t_i; \eta_k, t_j) & \\ & \text{for} \quad m = k, \end{cases}$$

where $\widetilde{P}^{(1)}(\eta_m, t_i; \eta_k, t_j)$, $\widetilde{P}^{(2)}(\eta_m, t_i; \eta_k, t_j)$ smooth enough functions.

4. Numerical experiments

In this section, we present some numerical results for two different examples. Together with the approximation of solution in the domain, we will provide numerical results for approximations of the normal derivative on Γ_{-1} (taking into account the jump relations of the single-layer potential normal derivative [6]) and the trace of the solution on Γ_0

$$\begin{aligned} \frac{\partial u}{\partial \nu}(x) = & -\frac{1}{2}\psi_{-1}(x) + \int\limits_D \psi(y) \frac{\partial P(x,y)}{\partial \nu(x)}\, dy + \int\limits_{\Gamma_{-1}} \psi_{-1}(y) \frac{\partial P(x,y)}{\partial \nu(x)}\, ds(y) + \\ & + \int\limits_{\Gamma_0} \psi_0(y) \frac{\partial P(x,y)}{\partial \nu(x)}\, ds(y), \quad x \in \Gamma_{-1}, \end{aligned}$$

$$\begin{aligned} u(x) = & \int\limits_D \psi(y) P(x,y)\, dy + \int\limits_{\Gamma_{-1}} \psi_{-1}(y) P(x,y)\, ds(y) + \\ & + \int\limits_{\Gamma_0} \psi_0(y) P(x,y)\, ds(y), \quad x \in \Gamma_0. \end{aligned}$$

Example 1. Let the domain D (see Fig. 1) is bounded by the two circles:

$$\Gamma_0 = \{x(t) = (1.2\cos(t), 1.2\sin(t)),\ t \in [0, 2\pi)\},$$
$$\Gamma_{-1} = \{x_{-1}(t) = (0.6\cos(t), 0.6\sin(t)),\ t \in [0, 2\pi)\}.$$

Here we have $\xi_{-1} = 0.5$. The function σ is given and equal

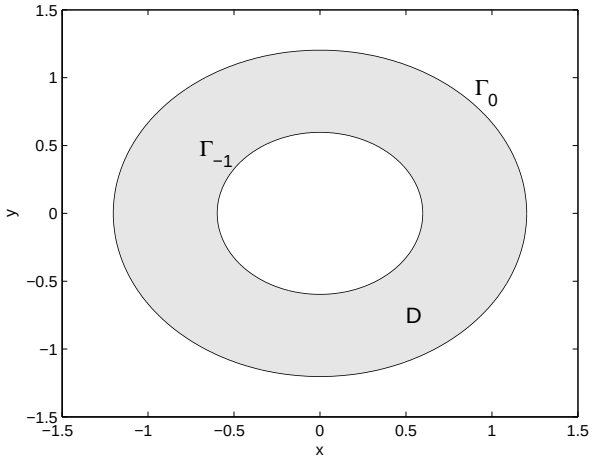


FIG. 1. The solution domain D in Ex. 1

$$\sigma(x) = 4 - x_1^2 + x_2^2, \quad x \in D.$$

Let us choose the boundary functions f_1 and f_2 of the elliptic problem as

$$f_1 = x_1 x_2 \quad \text{on } \Gamma_{-1}, \qquad f_2 = 0.6 x_1 x_2 (4 - x_1^2 + x_2^2) \quad \text{on } \Gamma_0.$$

Easy to verify that $u_{ex} = x_1 x_2$ is the exact solution to (1)-(3).

In (10),(12) we use the midpoint quadrature as a quadrature rule with respect to $\xi \in (\xi_{-1}, 1)$ with weights $\alpha_k = \frac{1-\xi_{-1}}{N}$ and quadrature nodes $\eta_k = 1 - \frac{1-\xi_{-1}}{2N}(2k-1)$, $k = 1, \ldots, N$.

TABL. 1. Absolute error on inner curves $\tilde{\Gamma}_1$-$\tilde{\Gamma}_3$ for Ex. 1

N	n	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_1}$	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_2}$	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_3}$
3	32	2.33E-05	6.64E-05	1.31E-04
	64	8.86E-08	2.52E-07	5.47E-07
6	64	1.16E-05	3.45E-05	7.51E-05
	128	4.97E-08	1.47E-07	3.21E-07
12	128	5.80E-06	1.76E-05	3.85E-05
	256	2.63E-08	7.97E-08	1.74E-07

We will provide the numerical error of the proposed approach on three curves within the domain that are homothetic to the outer boundary and have the following parametric representations

$$\tilde{\Gamma}_k : \tilde{x}_k = (\xi_{-1} + \frac{1-\xi_{-1}}{40}(12k-5))x(t), \quad t \in [0, 2\pi), \quad k = 1, 2, 3. \tag{16}$$

Straightforward calculation gives that homothetic factors related to the curves $\tilde{\Gamma}_1$, $\tilde{\Gamma}_2$, $\tilde{\Gamma}_3$ are 0.5875, 0.7375 and 0.8875 respectively. They correspond to the 4th, 10th, 16th curve counting from the first inner curve after Γ_{-1} in case when

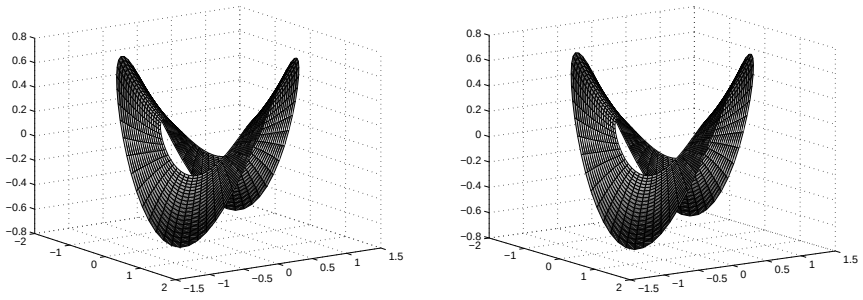
discretisation parameter $N = 20$. The absolute errors for different discretisation parameters N and n are presented in Table 1.

TABL. 2. Absolute error of the normal derivative and the function on boundaries and relative error in D for Ex. 1

N	n	$\|\frac{\partial u_{Nn}}{\partial \nu} - \frac{\partial u_{ex}}{\partial \nu}\|_{\infty,\Gamma_{-1}}$	$\|u_{Nn} - u_{ex}\|_{\infty,\Gamma_0}$	$\frac{\|u_{Nn}-u_{ex}\|_{L_2(D)}}{\|u_{ex}\|_{L_2(D)}} \cdot 100\%$
3	32	3.09E-04	1.03E-04	1.455
	64	1.17E-06	3.38E-07	0.271
6	64	1.89E-04	5.67E-05	0.270
	128	8.08E-07	2.53E-06	0.025
12	128	1.05E-04	3.37E-04	0.277
	256	4.73E-07	7.98E-07	0.276

In Table 2 we present the absolute errors of the normal derivative on the Γ_{-1} and the solution on the Γ_0 together with relative errors with respect to the L_2-norm in the domain D for the same parameters N and n as in Table 1. To calculate the relative error in the domain we use the following approximation with $\tilde{N} = 20$ and $\tilde{n} = 32$

$$\frac{\|u_{Nn} - u_{ex}\|_{L_2(D)}}{\|u_{ex}\|_{L_2(D)}} \approx \left(\frac{\sum_{k=1}^{\tilde{N}} \sum_{j=0}^{2\tilde{n}-1} (u_{Nn} - u_{ex})^2(\tilde{\eta}_k, \tilde{t}_j) J(\tilde{\eta}_k, \tilde{t}_j)}{\sum_{k=1}^{\tilde{N}} \sum_{j=0}^{2\tilde{n}-1} u_{ex}^2(\tilde{\eta}_k, \tilde{t}_j) J(\tilde{\eta}_k, \tilde{t}_j)} \right)^{1/2}. \quad (17)$$



a). exact solution b). approximate solution

FIG. 2. Exact solution and numerical approximation in domain D for Ex. 1

The numerical approximation (for discretisation parameters $N = 6$, $n = 64$) and the exact solution in the domain D are shown in Fig. 2. From the numerical results, we see that parameters N and n are linked between each other – double increase N requires to increase the parameter n at least by two times to decrease

the error. But, in general, presented relative errors in the domain look pretty good as well as absolute errors on inner curves.

Example 2. Let the domain D (see Fig. 3) bounded by the two ellipses:

$$\Gamma_0 = \{x(t) = (a\cos(t), b\sin(t)),\ t \in [0, 2\pi)\},$$
$$\Gamma_{-1} = \{x_{-1}(t) = (0.4a\cos(t), 0.4b\sin(t)),\ t \in [0, 2\pi)\}.$$

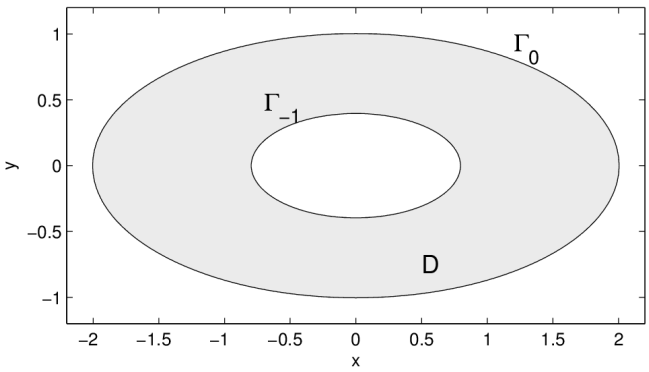


FIG. 3. The solution domain D in Ex. 2

TABL. 3. Absolute error on inner curves $\tilde{\Gamma}_1$-$\tilde{\Gamma}_3$ for Ex. 2

N	n	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_1}$	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_2}$	$\|u_{Nn} - u_{ex}\|_{\infty,\tilde{\Gamma}_3}$
3	32	3.92E-04	9.38E-04	3.07E-03
	64	3.28E-06	1.05E-05	3.76E-05
6	64	2.16E-04	5.30E-04	1.05E-03
	128	1.99E-06	6.24E-06	1.63E-05
12	128	1.18E-04	2.82E-04	5.46E-04
	256	1.14E-06	3.47E-06	8.72E-06

TABL. 4. Absolute error of the normal derivative and the function on boundaries and relative error in D for Ex. 2

N	n	$\|\frac{\partial\tilde{u}}{\partial\nu} - \frac{\partial u_{ex}}{\partial\nu}\|_{\infty,\Gamma_{-1}}$	$\|\tilde{u} - u_{ex}\|_{\infty,\Gamma_0}$	$\frac{\|u_{Nn}-u_{ex}\|_{L_2(D)}}{\|u_{ex}\|_{L_2(D)}} \cdot 100\%$
3	32	5.60E-03	7.67E-02	1.695
	64	3.10E-05	1.59E-04	0.377
6	64	4.07E-03	2.99E-02	0.377
	128	2.65E-05	1.02E-04	0.052
12	128	2.43E-03	5.54E-02	0.094
	256	1.72E-05	8.45E-04	0.077

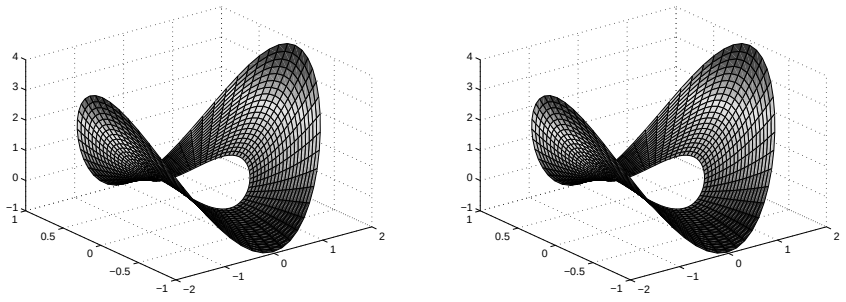
Here we have parameters $a = 2,\ b = 1$ and $\xi_{-1} = 0.4$. The function σ has following representation

$$\sigma(x) = 8 + 2x_1x_2, \quad x \in D.$$

The boundary functions f_1 and f_2 are known

$$f_1 = x_1^2 - x_2^2 \quad \text{on } \Gamma_{-1}, \qquad f_2 = (8+2x_1x_2)(x_1^2 - 4x_2^2)(0.25x_1^2 + 4x_2^2)^{-0.5} \quad \text{on } \Gamma_0.$$

For this example, the exact solution is $u_{ex} = x_1^2 - x_2^2$.



a). exact solution b). approximate solution

FIG. 4. Exact solution and numerical approximation in domain D for Ex. 2

The absolute errors on inner curves (16) are shown in Table 3. Similarly to the Ex. 1., the relative error of the solution in domain D, the absolute errors of its normal derivative on the inner boundary Γ_{-1} and the solution error on the outer boundary Γ_0 are displayed in Table 4. In Fig. 4 the exact solution in the domain D and its approximation for discretisation parameters $N = 6$ and $n = 128$ are shown. Observing the results we can see the same high accuracy of the obtained approximation of the solution as in Ex. 1.

5. CONCLUSION

An indirect integral equation method (based on the solution representation via potentials with densities and using the Levi function) for the numerical solution of a mixed boundary value problem for the generalized Laplace equation in doubly connected domains was applied. The differential problem is reduced to a system of boundary-domain integral equations. As a doubly connected domain, a domain bounded by two homothetic curves is considered. The change of variables in double integrals, quadrature rules application and the collocation of the obtained approximating equations at quadrature nodes lead to a system of the linear equations. Having calculated approximate values of the unknown densities we can find the approximation of the solution in the domain. Applicability of the proposed approach is confirmed by provided numerical results.

Bibliography

1. AL-Jawary M.A. Numerical solution of two-dimensional mixed problems with variable coefficients by the boundary-domain integral and integro-differential equation methods / M.A. AL-Jawary, L.C. Wrobel // Engeneering Analysis with Boundary Elements. – 2011. – Vol. 35 (12). – P. 1279-1287.
2. AL-Jawary M.A. Radial integration boundary integral and integro-differential equation methods for two-dimensional heat conduction problems with variable coefficients / M.A. AL-Jawary, L.C. Wrobel // Engeneering Analysis with Boundary Elements. – 2012. – Vol. 36 (5). – P. 685-695.
3. Beshley A. An integral equation method for the numerical solution of a Dirichlet problem for second-order elliptic equations with variable coefficients / A. Beshley, R. Chapko, B.T. Johansson // Journal of Engineering Mathematics. – 2018. – https://doi.org/10.1007/s10665-018-9965-7.
4. Chapko R.S. On the use of an integral equation approach for the numerical solution of a Cauchy problem for Laplace equation in a doubly connected planar domain / R.S. Chapko, B.T. Johansson, Y. Savka // Inverse Probl. Sci. Eng. – 2014. – Vol. 22. – P. 130-149.
5. Gao X.-W. A boundary element method without internal cells for two-dimensional and three-dimensional elastoplastic problems / X.-W. Gao // J. ApplMech (ASME). – 2002. – Vol. 69. – P. 154-160.
6. Kress R. Linear Integral Equations (3rd ed.) / R. Kress. – New-York: Springer-Verlag, 2014.
7. Kress R. On Trefftz' integral equation for the Bernoulli free boundary value problem / R. Kress // Numerische Mathematik. – 2017. – Vol. 136. – P. 503-522.
8. Mikhailov S.E. Localized boundary-domain integral formulations for problems with variable coefficients / S.E. Mikhailov // Engineering Analysis with Boundary Elements. – 2002. – Vol. 26. – P. 681-690.
9. Morse P.M. Methods of theoretical physics (Vols I,II) / P.M. Morse, H. Feshbach. – New York: McGraw-Hill, 1953.
10. Rangogni R. Numerical solution of the generalized Laplace equation by the boundary element method / R. Rangogni, R. Occhi // Appl. Math. Modelling. – 1987. – Vol. 11. – P. 393-396.

A. Beshley,
Faculty of Applied Mathematics and Informatics,
Ivan Franko National University of Lviv,
1, Universytets'ka Str., Lviv, 79000, Ukraine.

Received 17.08.2018; revised 30.08.2018

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 519.6

A FEW WAYS TO FIND APPROXIMATE SOLUTION TERMS OF THE METHOD OF GENERALIZED SEPARATION OF VARIABLES

V. M. Biletskyy

Резюме. Метод узагальненого розділення змінних будує наближення розв'язку задачі у вигляді суми доданків з розділеними змінними. Доданки знаходять послідовно як розв'язки певних мінімізаційних задач. У цій роботі ми розглянемо деякі способи знаходження наступного доданку наближеного розв'язку та наведемо формальний опис алгоритмів методу.

Abstract. The method of generalized separation of variables approximates a problem solution with a series of terms from a set of elements with separated variables. The terms should be found consecutively as solutions of certain minimization problems. In this paper we consider a few possible ways to find the next series term and give a formal description of the method algorithms.

1. Introduction

The method of generalized separation of variables (MGSV) is an iterative approach to approximate a solution of a linear multidimensional equation. According to the method instead of solving a single multidimensional problem we solve a series of one-dimensional problems and build a solution approximation. The method allows to dramatically decrease a computational complexity of problem solution algorithms. Besides a solution approximation is much more compact than the solution itself, i.e. requires less space.

The method has been originally suggested to solve multidimensional integral and matrix equations [1]. In [4,5] the method description is given for integral Fredholm equations.

The main idea of the method is to represent a solution of a linear d-dimensional equation $Au = f$ as a series of terms with separated variables

$$u\left(x_1, \ldots, x_d\right) = \sum_{k=1}^{\infty} \prod_{j=1}^{d} \phi_j^{(k)}\left(x_j\right),$$

which are found consecutively by minimizing the following functional

$$J_k\left(\phi_1, \ldots, \phi_d\right) = \| f - \sum_{l=1}^{k-1} A\left(\phi_1^{(l)} \otimes \cdots \otimes \phi_d^{(l)}\right)^2 \| \to \min.$$

Here A is a linear continuous operator in the corresponding space.

Key words. Method of generalized separation of variables, linear equation, multidimensional problem, approximate solution.

Paper [6] describes the MGSV modification which builds a solution approximation in a slightly different way

$$u = \sum_{k=1}^{\infty} A^* \left(\phi_1^{(k)} \otimes \cdots \otimes \phi_d^{(k)} \right),$$

where A^* is the adjoint of A and the terms of the series are found from

$$J_k \left(\phi_1, \ldots, \phi_d \right) = \| A^{-1} f - \sum_{l=1}^{k-1} A^* \left(\phi_1^{(l)} \otimes \cdots \otimes \phi_d^{(l)} \right)^2 \| \to \min .$$

In [2,3] a convergence of the solution approximation series to the exact equation solution is proven for both MGSV and its modification, respectively.

2. MGSV

Consider d complex separable Hilbert spaces H_j, $j = 1, \ldots, d$. Let's denote with $(\cdot, \cdot)_j$ an inner product in H_j which defines the corresponding norm $\|[\|j]\cdot$. Let H is a tensor product of the given spaces

$$H = \bigotimes_{j=1}^{d} H_j$$

with a norm $\|\cdot\|$ defined by its inner product $(\cdot, \cdot)$.

Note that H is also a complex separable Hilbert space and for any $h_j^{(1)}, h_j^{(2)} \in H_j$, $j = 1, \ldots, d$

$$\left(h_1^{(1)} \otimes \cdots \otimes h_d^{(1)}, h_1^{(2)} \otimes \cdots \otimes h_d^{(2)} \right) = \prod_{j=1}^{d} \left(h_j^{(1)}, h_j^{(2)} \right)_j .$$

Consider a linear operator equation in H

$$Au = f, \tag{1}$$

where $u, f \in H$ and $A \in \mathfrak{L}(H)$ is a linear continuous operator in H such that there exists its continuous inverse operator $\exists A^{-1} \in \mathfrak{L}(H)$. Note that under such conditions the adjoint operator also exists and is continuous in H as well $\exists A^* \in \mathfrak{L}(H)$. Moreover the equation (1) has a unique solution in H.

The MGSV approximates the solution of (1) with a series where each term has a special form called *separable* with respect to spaces H_j, $j = 1, \ldots, d$. In other words each term is a tensor product of d elements from $H_1, H_2, \ldots, H_d$ respectively. Let's denote with G a set of separable elements of H with respect to H_j, $j = 1, \ldots, d$

$$G = \left\{ \bigotimes_{j=1}^{d} h_j : \; h_j \in H_j, \; j = 1, \ldots, d \right\},$$

Also we define a set G_A as a mapping A applied to the set G

$$G_A = A(G) = \{ Ag : \; g \in G \} . \tag{2}$$

Thus the MGSV solution approximation series consists of elements from G

$$\sum_{k=1}^{\infty} g_k, \quad g_j \in G, \tag{3}$$

where k-th term is found according to the minimum condition

$$\|f - A\left(\sum_{l=1}^{k-1} g_l + g_k\right)\| = \inf_{g \in G} \|f - A\left(\sum_{l=1}^{k-1} g_l + g\right)\|. \tag{4}$$

The terms of (3) are constructed consecutively and produce a sequence of approximate solutions. The sum of the first k terms of (3) is k-th solution approximation of the equation (1)

$$u_k = \sum_{l=1}^{k} g_l, \quad u_0 = 0_H,$$

where 0_H is a zero vector in H.

k-th term of the series (3) is called k-th approximate solution *improvement*

$$g_k = u_k - u_{k-1}.$$

When we have k-th solution approximation u_k by subtracting Au_k from the right-hand side of the equation (1) we get the very same initial equation (1) but with different right-hard side $f - Au_k$ which is called k-th *residual* equation

$$Au = f - Au_k.$$

Let f_k is a right-hand side of k-th residual equation

$$f_k = f - Au_k = f - A\left(\sum_{j=1}^{k} g_j\right), \quad f_0 = f.$$

In [2] it is proven that at least one such element g_k satisfying (4) exists in H. Therefore there always exists a minimization problem solution of the following functional

$$J_k\left(h_1, \ldots, h_d\right) = \|f_{k-1} - A\left(\bigotimes_{j=1}^{d} h_j\right)\|^2, \quad h_j \in H_j,\ j = 1, \ldots, d. \tag{5}$$

By considering the definition of G_A (2) and the condition (4) it's easy to see that element Ag_{k+1} is the best approximation to the right-hand side of k-th residual equation in the set G_A

$$\|f_k - Ag_{k+1}\| = \inf_{g \in G_A} \|f_k - g\|.$$

Algorithm 1 describes a generic approach of MGSV.

The loop break (iteration stop) condition of algorithm 1

$$\frac{\|f_k\|}{\|f\|} = \frac{\|f - Au_k\|}{\|f\|} < \epsilon,$$

Algorithm 1 MGSV

Require: $f \in H$, $A \in \mathfrak{L}(H)$, $\epsilon > 0$
Ensure: $\|f - A\widetilde{u}\| < \epsilon$
  $k \leftarrow 0$
  $u_0 \leftarrow 0_H$
  repeat
    $k \leftarrow k + 1$
    $g_k \leftarrow$ solution of the (4)
    $u_k \leftarrow u_{k-1} + g_k$
    $f_k \leftarrow f - Au_k$
  until $\frac{\|f_k\|}{\|f\|} < \epsilon$
  $\widetilde{u} \leftarrow u_k$
  return $\widetilde{u}$

can be replaced with alternatives, for example

$$\frac{\|g_k\|}{\|u_{k-1}\|} = \frac{\|u_k - u_{k-1}\|}{\|u_{k-1}\|} < \epsilon.$$

The possible ways to find g_k in algorithm 1 are considered below. In [2] the convergence of approximate solution sequence of the MGSV to the exact solution of the equation (1) is proven.

In practice implementations of MGSV consider a discrete case of the equation (1). Assume H_j is a n_j-dimensional space, $j = 1, \ldots, d$. Since H is a tensor product of $H_1, H_2, \ldots, H_d$ it is a n-dimensional space where

$$n = \dim H = \prod_{j=1}^{d} \dim H_j = \prod_{j=1}^{d} n_j.$$

Now the equation (1) is equivalent to a system of n linear equations. In general a space required to store a solution of the equation linearly depends on the number of dimensions n, i.e. the required storage is

$$\mathcal{O}\left(\prod_{j=1}^{d} n_j\right). \tag{6}$$

On the other hand since space H has a special structure a single term of the MGSV approximate solution consumes only

$$\mathcal{O}\left(\sum_{j=1}^{d} n_j\right) \tag{7}$$

of the storage. If we increase number of dimensions in all spaces $H_1, H_2, \ldots, H_d$ simultaneously expression (6) will grow exponentially while expression (7) will grow linearly. Thus as long as the number of terms in the solution approximation is relatively small MGSV produces a compact (in terms of the storage) solution approximation of the equation (1).

In a discrete case MGSV is closely related to approximations of a multidimensional tensor with a product of one-dimensional tensors (vectors) [7]. Indeed, elements u and f can be viewed as d-dimensional tensors of $n_1 \times \cdots \times n_d$ size and operator A as $(2d)$-dimensional tensor of $n_1 \times \cdots \times n_d \times n_1 \times \cdots \times n_d$ size. Then in order to find k-th solution improvement we have to minimize a function with d one-dimensional vector parameters

$$J_k\left(x_1, \ldots, x_d\right) = \left\| f_{k-1} - A\left(x_1 \otimes \cdots \otimes x_d\right)\right\|^2. \tag{8}$$

Here $x_j = \left(x_j^{(1)}, \ldots, x_j^{(n_j)}\right)$ is a one-dimensional vector of coordinates in H_j, $j = 1, \ldots, d$. The norm of d-dimensional tensor t of $n_1 \times \cdots \times n_d$ size can be, for example,

$$\|t\| = \sqrt{\sum_{\substack{1 \le k_j \le n_j \\ 1 \le j \le d}} |t_{k_1, \ldots, k_d}|^2}, \quad t \in H.$$

The function (8) is a polynomial of total degree $2d$ with m variables

$$x_1^{(1)}, \ldots, x_1^{(n_1)}, \ldots, x_d^{(1)}, \ldots, x_d^{(n_d)}.$$

Here

$$m = \sum_{j=1}^{d} n_j.$$

3. Alternating Least Squares

Let's consider the minimization problem of functional (5). In general the problem is nonlinear and can be solved using any nonlinear functional minimization methods. However note that if in (5) we fix all parameter values except of one h_j, $1 \le j \le d$ then we get a functional of a single parameter $h_j \in H_j$ which minimization problem is linear.

Similarly if we fix values of all polynomial (8) variables except of $x_j^{(1)}$, ..., $x_j^{(n_j)}$, $1 \le j \le d$ then we get a quadratic polynomial of n_j variables. The minimization of such polynomial can be done by solving a system of n_j linear equations with n_j variables. This leads us to the method of alternating least squares which can be used to approximate the next term of MGSV series.

The idea of *Alternating Least Squares* method (ALS) is to choose an initial values of (5) variables, fix all of them and then cyclically iterate over the variables, release one of them, solve a linear minimization problem and adjust the current variable value. Algorithm 2 describes a generic approach of ALS method.

Alternatively as a loop break condition in algorithm 2 a relatively small current value of functional (5) can be used instead

$$\frac{J_k\left(h_1^{(l)}, \ldots, h_d^{(l)}\right)}{J_k\left(h_1^{(0)}, \ldots, h_d^{(0)}\right)} < \epsilon.$$

Algorithm 2 ALS

Require: $f_{k-1} \in H$, $A \in \mathfrak{L}(H)$, $h_j^{(0)} \in H_j$, $j = 1, \ldots, d$, $\epsilon > 0$
  $l \leftarrow 0$
  repeat
    $l \leftarrow l + 1$
    for $j = 1$ to d **do**
      fix all variable values of (5) except of h_j and solve

$$J_k\left(h_1^{(l)}, \ldots, h_{j-1}^{(l)}, h_j, h_{j+1}^{(l-1)}, \ldots, h_d^{(l-1)}\right) \longrightarrow \min$$

      $h_j^{(l)} \leftarrow$ the linear minimization problem solution
    end for
  until

$$\frac{\|h_1^{(l)} \otimes \cdots \otimes h_d^{(l)} - h_1^{(l-1)} \otimes \cdots \otimes h_d^{(l-1)}\|}{\|h_1^{(l-1)} \otimes \cdots \otimes h_d^{(l-1)}\|} < \epsilon$$

  for $j = 1$ to d **do**
    $\widetilde{h_j} \leftarrow h_j^{(l)}$
  end for
  return $\widetilde{h_1} \otimes \widetilde{h_2} \otimes \ldots \otimes \widetilde{h_d}$

Consider a numerical sequence

$$\left\{J_k\left(h_1^{(l)}, \ldots, h_d^{(l)}\right)\right\}_{l=0}^{\infty}. \tag{9}$$

According to algorithm 2 the given sequence is monotonically non-increasing

$$\forall l \geq 1 \quad J_k\left(h_1^{(l)}, \ldots, h_d^{(l)}\right) \leq J_k\left(h_1^{(l-1)}, \ldots, h_d^{(l-1)}\right).$$

Since (9) is bounded by zero it converges to some non-negative number L

$$\exists L \geq 0 : \lim_{l \to \infty} J_k\left(h_1^{(l)}, \ldots, h_d^{(l)}\right) = L.$$

However in general sequence (9) does not converge to the infimum of functional (5).

The method of alternating least squares is simple for understanding and implementation, but does not guarantee a convergence to the solution of minimization problem of (5). Besides the method outcome might strongly depend on the initial values $h_1^{(0)}, h_2^{(0)}, \ldots, h_d^{(0)}$.

Note that in some cases a convergence to the minimization problem solution can be proven. For example if the following condition holds

$$\begin{aligned}
&\forall f_{k-1} \in H \quad \forall j, l \quad 1 \leq j < l \leq d \\
&\qquad \forall h_1 \in H_1 \quad \forall h_2 \in H_2 \quad \ldots \quad \forall h_d \in H_d \quad \forall \hat{h_j} \in H_j \quad \forall \hat{h_l} \in H_k \\
&\qquad J_k(h_1, \ldots, h_j, \ldots, h_l, \ldots, h_d) > J_k(h_1, \ldots, \hat{h_j}, \ldots, \hat{h_l}, \ldots, h_d) \quad \Rightarrow \\
&\qquad J_k(h_1, \ldots, h_j, \ldots, h_l, \ldots, h_d) > J_k(h_1, \ldots, h_j, \ldots, \hat{h_l}, \ldots, h_d) \quad \wedge
\end{aligned}$$

$$J_k(h_1,\ldots,h_j,\ldots,h_l,\ldots,h_d) > J_k(h_1,\ldots,\hat{h_j},\ldots,h_l,\ldots,h_d) \quad \wedge$$
$$J_k(h_1,\ldots,\hat{h_j},\ldots,h_l,\ldots,h_d) > J_k(h_1,\ldots,\hat{h_j},\ldots,\hat{h_l},\ldots,h_d) \quad \wedge$$
$$J_k(h_1,\ldots,h_j,\ldots,\hat{h_l},\ldots,h_d) > J_k(h_1,\ldots,\hat{h_j},\ldots,\hat{h_l},\ldots,h_d)$$

then the sequence produced by ALS converges to the (5) minimization problem solution.

Papers [7–9] consider problems of multidimensional tensor decomposition with tensor products of one-dimensional vectors where the numerical ALS method [10, 11] is widely used. Some efficiency improvement techniques are described in [12] while the initial ALS value selection problem is considered in [13].

There are numerous of alternative methods which share the same basic idea with ALS. In [14] some of such methods are compared with ALS:

- DTLD (direct trilinear decomposition);
- ATLD (alternating trilinear decomposition);
- SWATLD (self-weighted alternating trilinear decomposition);
- PALS (pseudo alternating least squares);
- ACOVER (alternating coupled vectors resolution);
- ASD (alternating slice-wise diagonalization);
- ACOMAR (alternating coupled matrices resolution).

According to the paper conclusions none of the methods is superior to ALS in terms of a convergence to the exact solution.

Table 1 contains numerical results of MGSV with ALS for the following equation

$$Au \equiv \int_0^1 \int_0^1 \cos\left(\widehat{x}\widehat{y} + x^2 - y^2\right) u\left(\widehat{x},\widehat{y}\right) \mathrm{d}\widehat{x}\mathrm{d}\widehat{y} - 4u\left(x,y\right) = \sin\left(x^2+y^2\right). \quad (10)$$

For both algorithms $\epsilon = 10^{-5}$. The first column corresponds to MGSV iteration index k, the second column shows the value $\frac{\|f_{k-1}\|}{\|f\|}$ and each of the following columns contains the value

$$\frac{\|f_{k-1} - A\left(h_1^{(l)} \otimes \ldots \otimes h_d^{(l)}\right)\|}{\|f\|}$$

after the l-th iteration of ALS.

Tabl. 1. Numerical results for equation (10)

k	before ALS	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
1	1.000000	0.344960	0.158934	0.153180	0.153133	0.153133	0.153133
2	0.153133	0.147828	0.025094	0.007752	0.007494	0.007493	0.007493
3	0.007493	0.001684	0.000297	0.000293	0.000293		
4	0.000293	0.000288	0.000035	0.000011	0.000010		

4. Nonlinear Least Squares

Better approximation accuracy can be obtained by using *Nonlinear Least Squares* methods (NLS). These gradient methods minimize nonlinear function (8). In particular NLS representatives are Gauss-Newton method [15, 16], damped Gauss-Newton method [17, 18] and PMF methods [19].

Algorithm 3 describes a generic approach of NLS methods for minimization of nonlinear multivariable function (8).

Algorithm 3 NLS

Require: $f_{k-1} \in H$, $A \in \mathfrak{L}(H)$, $\epsilon > 0$
  $l \leftarrow 0$
  $x^{(0)} \leftarrow$ initial value $\{x^{(0)}$ is a variable vector of function (8)$\}$
  repeat
    $l \leftarrow l + 1$
    $x^{(l)} = x^{(l-1)} - \phi\left(x^{(l-1)}\right)$ $\{\phi$ is a mapping which depends on J_k and a particular method$\}$
  until $\frac{\|x^{(l)} - x^{(l-1)}\|}{\|x^{(l-1)}\|} < \epsilon$
  $\widetilde{x} \leftarrow x^{(l)}$
  return $\widetilde{x}$

NLS methods are mostly generalizations and modifications of Newton method. At each iteration based on a gradient we look for an optimal vector and length of the next step.

NLS methods in general produce more accurate approximations than ALS methods, they do not guarantee a convergence to the global minimum of function (8) though. However NLS methods are inferior to ALS in terms of computational complexity. Numerical results provided in [12, 18] show that NLS methods are slower and require more storage than ALS.

5. Stetter-Möller Matrix Method Modification

Papers [20, 21] consider modifications of *Stetter-Möller matrix method* [22, 23] which allows to find a global minimum of a multivariable higher degree polynomial. Suggested approaches lead a polynomial minimization problem to a generalized eigenvalue problem. A set of points where the polynomial global minimum is achieved has several connected components. For each such connected component the method finds at least one point. There are no special application requirements, i.e. the method finds a minimum for an arbitrary polynomial. Thus the method can be used to find a global minimum of function (8).

Let p is a m-variable polynomial of total degree $2d$

$$p(x_1, \ldots, x_m) \in \mathbb{R}[x_1, \ldots, x_m]. \tag{11}$$

Consider a polynomial

$$p_\lambda(x_1, \ldots, x_m) = p(x_1, \ldots, x_m) + \lambda\left(x_1^{2(d+1)} + \ldots + x_m^{2(d+1)}\right), \quad \lambda > 0.$$

According to [20] the global minimum data of (11) can be retreived from p_λ when $\lambda \to 0$.

A polynomial global minimum can be found from the first order conditions by considering its values in critical points. For p_λ if $\lambda > 0$ is fixed this leads to a system of polynomial equations in Gröbner basis [24] which has a finite number of solutions. Thus the Stetter-Möller matrix method can be used.

First, we build matrices $(A_{x_1}, \ldots, A_{x_m})$. Eigenvalues of these matrices which correspond to a common eigenvector form a critical point of polynomial p_λ. Here matrix A_{x_k} $(1 \leq k \leq m)$ represents an operator of multiplication by x_k in quotient space $\mathbb{R}[x_1, \ldots, x_m]/I$ where I is an ideal formed by first order partial derivatives of p_λ.

For an arbitrary polynomial $r(x_1, \ldots, x_m)$ matrix $A_r = r(A_{x_1}, \ldots, A_{x_m})$ contains values of polynomial r in critical points of polynomial p_λ.

Algorithm 4 describes one of the possible approach implementations.

Algorithm 4 Stetter-Möller Matrix Method Modification

Require: $p(x_1, \ldots, x_m) \in \mathbb{R}[x_1, \ldots, x_m]$, $\lambda > 0$, $\epsilon > 0$
  $l \leftarrow 0$
  $\lambda_0 \leftarrow \lambda$
  $\left(x_1^{(0)}, \ldots, x_m^{(0)}\right) \leftarrow (0, \ldots, 0)$
  $v_0 \leftarrow p\left(x_1^{(0)}, \ldots, x_m^{(0)}\right)$
  repeat
    $l \leftarrow l + 1$
    $\lambda_l \leftarrow \frac{\lambda_{l-1}}{2}$
    compute matrices $\left(A_{x_1}^{(l)}, \ldots, A_{x_m}^{(l)}\right)$ for polynomial p_{λ_l}
    compute matrix $A_p^{(l)} = p\left(A_{x_1}^{(l)}, \ldots, A_{x_m}^{(l)}\right)$
    $v_l \leftarrow$ minimum value of $A_p^{(l)}$
    $\left(x_1^{(l)}, \ldots, x_m^{(l)}\right) \leftarrow$ the corresponding vector, i.e. $p\left(x_1^{(l)}, \ldots, x_m^{(l)}\right) = v_l$
  until $\frac{|v_l - v_{l-1}|}{|v_0|} < \epsilon$
  $(\widetilde{x_1}, \ldots, \widetilde{x_m}) \leftarrow \left(x_1^{(l)}, \ldots, x_m^{(l)}\right)$
  return $(\widetilde{x_1}, \ldots, \widetilde{x_m})$

A drawback of the described method is the size of matrix A_r which is equal to $(2d+1)^m$ and grows exponentially with m. However modern ways to solve generalized eigenvalue problems which are based on Jacobi-Davidson or Arnoldi methods [25, 26] do not require a construction of matrix A_r. Thus one of the suggested method modifications [20, 21] can be used instead.

Stetter-Möller matrix method modification unlike ALS and NLS methods always finds a global minimum of a function. However it requires a lot of computational resources. Thus in practice quite often ALS or NLS methods are preferred despite they are not perfectly accurate.

Bibliography

1. Balyash Yu.G. Generalized separation of variables in problems of diffraction and antenna synthesis / Yu.G. Balyash, N.N. Voitovich, S.A. Yaroshko // Proc. of URSI Int. Sympos. on Electromagn. Theory. – 1989. – P. 651-653.
2. Biletskyy V. An iterative method of generalized separation of variables for solving linear operator equations / V. Biletskyy // Journal of Numerical and Applied Mathematics. – 2010. – Vol. 100. – P. 2-9.
3. Biletskyy V. Modification of a method of generalized separation of variables for the solution of multidimensional integral equations / V. Biletskyy Journal of Mathematical Sciences. – 2012. – Vol. 181. – P. 340-349.
4. Biletskyy V. A method of generalized separation of variables for solving many-dimensional linear Fredholm integral equations / V. Biletskyy, S. Yaroshko // Proceedings of XII International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory. – 2007. – P. 94-97.
5. Biletskyy V. A method of generalized separation of variables for solving three-dimensional integral equations / V. Biletskyy, S. Yaroshko // Proceedings of XI International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory. – 2006. – P. 164-168.
6. Biletskyy V. A modification of an iterative method of generalized separation of variables for solving many-dimensional integral equations / V. Biletskyy, S. Yaroshko // Proceedings of international conference "Integral Equations 2010". – 2010. – P. 16-18.
7. Kolda T.G. Tensor Decompositions and Applications / T.G. Kolda, B.W. Bader // SIAM Review. – 2009. – Vol. 51. – P. 455-500.
8. Acar E., Dunlavy D.M., Kolda T.G. A Scalable Optimization Approach for Fitting Canonical Tensor Decompositions / E. Acar, D.M. Dunlavy, T.G. Kolda // Journal of Chemometrics. – 2011. – Vol. 25. – P. 67-86.
9. Zhang T. Rank-one approximation to high order tensors / T. Zhang, G.H. Golub // Siam Journal on Matrix Analysis and Applications. – 2001. – Vol. 23. – P. 534-550.
10. Carroll J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition / J. Carroll, J.-J. Chang // Psychometrika. – 1970. – Vol. 35. – P. 283-319.
11. Harshman R. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis / R. Harshman // UCLA Working Papers in Phonetics. – 1970. – Vol. 16. – P. 1-84.
12. Tomasi G.V Practical and computational aspects in chemometric data analysis / G. Tomasi. – PhD thesis, 2006.
13. Smilde A. Multi-way analysis: Applications in the chemical sciences / A. Smilde, R. Bro, P. Geladi. – Wiley, 2004.
14. Faber N.M. Recent developments in CANDECOMP/PARAFAC algorithms: A critical review / N.M. Faber, R. Bro, P.K. Hopke // Chemometrics and Intelligent Laboratory Systems. – 2003. – Vol. 65. – P. 119-137.
15. Bjõrck A. Numerical Methods for Least Squares Problems / A. Bjõrck. – SIAM, 1996.
16. Nocedal J. Numerical optimization / J. Nocedal, S.J. Wright. – Springer, 1999.
17. Tomasi G. PARAFAC and missing values / G. Tomasi, R. Bro // Chemometrics and Intelligent Laboratory Systems. – 2005. – Vol. 75. – P. 163-180.
18. Tomasi G. A comparison of algorithms for fitting the PARAFAC model / G. Tomasi, R. Bro // Computational Statistics & Data Analysis. – 2006. – Vol. 50. – P. 1700-1734.
19. Paatero P. A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis / P. Paatero // Chemometrics and Intelligent Laboratory Systems. – 1997. – Vol. 38. – P. 223-242.
20. Hanzon B. Global minimization of a multivariate polynomial using matrix methods / B. Hanzon, D. Jibetean // Journal of Global optimization. – 2003. – Vol. 27. – P. 1-23.
21. Bleylevens I. An nD-systems approach to global polynomial optimization with an application to H_2 model order reduction / I. Bleylevens, R. Peeters, B. Hanzon // Decision

and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on. – 2005. – P. 5107-5112.
22. Stetter H.J. Matrix eigenproblems are at the heart of polynomial system solving / H.J. Stetter // ACM SIGSAM Bulletin. – 1996. – Vol. 30. – P. 22-25.
23. Möller H.M. Multivariate polynomial equations with multiple zeros solved by matrix eigenproblems / H.M. Möller, H.J. Stetter // Numerische Mathematik. – 1995. – Vol. 70. – P. 311-329.
24. Cox D.A. Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra / D.A. Cox, J.B. Little, D. O'Shea. – Springer, 1992.
25. Sleijpen G.L.G. A Jacobi–Davidson iteration method for linear eigenvalue problems / G.L.G. Sleijpen, H.A. Van der Vorst // SIAM Review. – 2000. – Vol. 42. – P. 267-293.
26. Fokkema D.R. Jacobi–Davidson Style QR and QZ Algorithms for the Reduction of Matrix Pencils / D.R. Fokkema, G.L.G. Sleijpen, H.A. Van der Vorst // SIAM Journal on Scientific Computing. – 1998. – Vol. 20. – P. 94-125.

V. M. Biletskyy,
Faculty of Applied Mathematics and Informatics,
Ivan Franko National University of Lviv,
1, Universytets'ka Str., Lviv, 79000, Ukraine.

Received 31.05.2018; revised 8.08.2018

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 519.6

ON THE FINITE ELEMENT APPROXIMATION OF A SYSTEM OF ELLIPTIC QUASI-VARIATIONAL INEQUALITIES RELATED TO HAMILTON-JACOBI-BELLMAN EQUATIONS

M. Boulbrachene

Резюме. В роботі розвинуто новий підхід, запропонований в [3], для вивчення скінченно-елементної апроксимації систем еліптичних квазі-варіаційних нерівностей, що пов'язані з рівняннями Гамільтона-Якобі-Бельтрана. Метод поєднує в собі підходи часткових розв'язків, дискретної регулярності для варіаційних нерівностей та геометричну збіжність ітераційної схеми, що наближає розв'язок.

Abstract. In this paper, we exploit a new approach, introduced in [3], to study the finite element approximation of a system of elliptic quasi-variational inequalities (Q.V.I.) related to Hamilton-Jacobi-Bellman (HJB) equations. The method combines the concepts of subsolutions, discrete regularity for variational inequalities, and the geometrical convergence of an iterative scheme approximating the solution.

1. Introduction

We are concerned with the standard finite element approximation of the system of elliptic quasi-variational inequalities (Q.V.I): Find $U = (u_1, ..., u_M) \in \left(H_0^1(\Omega)\right)^M$ such that

$$\begin{cases} a_i(u_i, v - u_i) \geq (f_i, v - u_i) \quad \forall v \in H_0^1(\Omega), \\ u_i \leq k + u_{i+1},\ v \leq k + u_{i+1}, \\ u_{M+1} = u_1, \end{cases} \tag{1}$$

where, Ω is a bounded convex domain of $\mathbb{R}^N$ with sufficiently smooth boundary Γ, $f \geq 0$ is a right hand in $L^\infty(\Omega)$, $k > 0$, $(.,.)$ is the inner product in $L^2(\Omega)$, $a(.,.)$ is the bilinear form defined by: $\forall u, v \in H^1(\Omega)$

$$a_i(u\,, v) = \int_\Omega \left(\sum_{j,k=1}^{N} a_{jk}^i(x) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_k} + \sum_{k=1}^{N} b_k^i(x) \frac{\partial u}{\partial x_k} v + a_0^i(x) uv \right) dx \tag{2}$$

such that

$$a_i(v, v) \geq \delta \left\| v \right\|_{H^1(\Omega)}^2 \ \ \forall v \in H^1(\Omega),$$

where the coefficients $a_{jk}^i(x)$, $b_k^i(x)$, $a_0^i(x)$, $(j, k = 1, ..., N)$, are sufficiently smooth such that

$$a_0^i(x) \geq c_0 > 0, \forall x \in \Omega \tag{3}$$

Key words. Quasi-variational inequalities, Iterative scheme, Finite element, Discrete regularity, Subsolutions, Error estimate.

2000 *Mathematics Subject Classification.* 35 J85, 65N30, 65N15.

and

$$\sum_{1\leq\ j,k\leq\ N} a^i_{jk}(x)\xi_j\xi_k \geq \alpha\,|\xi|^2\,;(x\in\bar{\Omega},\ \xi\in R^N,\ \ \alpha>0). \tag{4}$$

Denoting by $\mathbb{V}_h$, the finite element space consisting of continuous piecewise linear functions vanishing at the boundary, r_h the usual interpolation operator, we define the discrete counterpart of (1) by: find $U_h=(u_{1,h},...,u_{M,h})\in(\mathbb{V}_h)^M$ such that

$$\begin{cases} a_i(u_{i,h},v-u_{i,h})\geq(f,v-u_{i,h}) \quad \forall v\in\mathbb{V}_h,\\ u_{i,h}\leq r_h\left(k+u_{i+1,h}\right),v\leq r_h\left(k+u_{i+1,h}\right),\\ u_{M+1,h}=u_{1,h}. \end{cases} \tag{5}$$

This system appears in stochastic control problems related to Hamilton-Jacobi-Bellman equations (HJB) (see [1], [2]). Its finite element approximation was studied in (cf.,e.g., [4], [5], [6], where different methods were employed.

In this paper, we exploit an idea developed in [3] to derive optimal convergence order for the system of Q.V.I (1).

This method consists, mainly, of combining, in both the continuous and discrete contexts, the concept of subsolutions for variational inequalities and a geometrical convergence of an iterative scheme approximating the solution. For a computational purpose, this method provides an interesting information as it permits to control the error between the continuous iterative scheme and its finite element counterpart.

A brief description of this method is as follows: Let $U^n=(u^n_1,...,u^n_M)$ be the nth iterate of the scheme approximating the solution U, and $U^n_h=(u^n_{1h},...,u^n_{Mh})$ its finite element counterpart, approximating U_h.We construct a sequence of continuous subsolutions $\beta^n=(\beta^n_1,...,\beta^n_M)$ such that

$$\beta^n\leq U^n$$

and

$$\|\beta^n-U^n_h\|_\infty\leq Ch^2\,|\ln h|^2$$

and a sequence of discrete subsolutions $\gamma^n=(\gamma^n_{1,h},...,\gamma^n_{M,h})$ such that:

$$\gamma^n_h\leq U^n_h$$

and

$$\|U^n-\gamma^n_h\|_\infty\leq Ch^2\,|\ln h|^2\,.$$

In this situation, using a concept of discrete regularity, we establish an optimal error estimate for the iterative scheme:

$$\|U^n-U^n_h\|_\infty\leq Ch^2\,|\ln h|^2 \tag{6}$$

and then, combining estimate (6) with the geometrical convergence of the iterative scheme (U^n) and (U^n_h) to the solutions U and U_h of systems (1) and (5), respectively, we also derive error estimate for the system of Q.V.I. (1):

$$\|U-U_h\|_\infty\leq Ch^2\,|\ln h|^2 \tag{7}$$

where

$$\|V\|_\infty=\max\|v_i\|_{L^\infty(\Omega)}\,,\ V=(v_1,...,v_M)$$

and, in all the above error estimates, C is a constant independent of both h and n.

It is worth pointing out that estimate (6) is new for the system (1).

The paper is organized as follows. In sections 2, we recall the construction and convergence of the continuous iterative scheme for system (1). In section 3, we also recall analog discrete results and detail discrete regularity for the discrete iterative scheme. In section 4, we discuss the new approximation approach and derive the main results of this paper. In section 5, we give a numerical example and, finally, in section 6, a short conclusion.

2. The Continuous Problem

2.1. **A Continuous Iterative Scheme.** Let $U^0 = \left(u_1^0, ..., u_M^0\right) \in \left(H^1(\Omega)\right)^M$ be such that u_i^0 solves the equation

$$a(u_i^0, v) = (f_i, v) \quad \forall v \in H_0^1(\Omega); \forall i = 1, ..., M. \tag{8}$$

Then, starting from U^0 solution of (8), we define the continuous sequence (U^n) such that $U^n = (u_1^n, ..., u_M^n)$ and u_i^n solves the variational inequality (V.I)

$$\begin{cases} a(u_i^n, v - u_i^n) \geq (f_i, v - u_i^n) \quad \forall v \in H_0^1(\Omega), \\ u_i^n \leq k + u_{i+1}^{n-1}, v \leq k + u_{i+1}^{n-1}, \\ u_{M+1}^{n-1} = u_1^{n-1}. \end{cases} \tag{9}$$

Theorem 1. *[5] The sequence (U^n) defined in (9) converge decreasingly to the solution U of of system (1). Moreover, there exists $0 < \mu < 1$ such that*

$$\|U^n - U\|_\infty \leq \mu^n \left\|U^0\right\|_\infty. \tag{10}$$

3. The discrete Problem

For the sake of simplicity we suppose that Ω is polyhedral.We then consider a regular and quasi-uniform triangulation τ_h of $\bar{\Omega}$, consisting of n-simplices K. Denote by $h = \max_{K \in \tau_h} h_K$, the meshsize of τ_h with h_K being the diameter of K. For each $K \in \tau_h$, denote by $P_1(K)$ the set of polynomials on K with degree no more than 1. The P_1- conforming finite element space is given by

$$\mathbb{V}_h = \left\{v : v \in H^1(\Omega) \cap C(\bar{\Omega}),\ v_{/K} \in P_1(K), \quad \forall K \in \tau_h\right\}.$$

Let M_i, $1 \leq i \leq N_h$ denote the the vertices of the triangulation τ_h, and let φ_i, $1 \leq i \leq m(h)$, denote the functions of V_h which satisfy

$$\varphi_i(M_j) = \delta_{ij}, 1 \leq i, j \leq N_h$$

so that the functions φ_i form a basis of V_h. For every $v \in H^1(\Omega) \cap C(\bar{\Omega})$, the function

$$r_h v(x) = \sum_{i=1}^{N_h} v(M_i)\varphi_i(x)$$

represents the interpolate of v over τ_h.

Now, in order to establish existence and uniqueness of a solution to V.I (5), the stiffness matrix is required to be an M-Matrix.

Definition 1. A real matrix $d \times d$ matrix $C = (c_{ls})$ with $c_{ls} \leq 0$, $\forall l \neq s$, $1 \leq l, s \leq d$, is called an M-Matrix if C is nonsingular and $C^{-1} \geq 0$ (i.e., all entries of its inverse are nonnegative).

3.1. **Discrete Maximum Principle.** Denote by A^i the matrices with generic coefficient

$$a^i_{ls} = a_i(\varphi_l, \varphi_s), \quad 1 \leq l, s \leq N_h; \quad i = 1, ..., M. \tag{11}$$

Because the bilinear form $a_i(.,.)$ is coercive, we have

$$A^i \text{ is positive definite} \tag{12}$$

and

$$a^i_{ll} > 0 \quad \forall l = 1, ..., m(h). \tag{13}$$

Furthermore, if the matrix (a_{jk}) involved in the bilinear form (2) is symmetric $(a_{jk} = a_{kj})$, then mesh conditions for which the off-diagonal entries of A^i satisfy

$$a^i_{ls} \leq 0, \forall i \neq j, \quad 1 \leq l, s \leq m(h) \tag{14}$$

can be found in [8]. Therefore, combining (12), (13) and (14), we have the following lemma.

Lemma 1. *The matrices A^i, $i = 1, ..., M$ are M-Matrices.*

Proof. See [8], [9]. □

3.2. **A discrete Iterative Scheme.** Let $U^0_h = \left(u^0_{1h}, ..., u^0_{Mh}\right)$ such that $u^0_{i,h} \in \mathbb{V}_h$ solves the equation

$$a_i(u^0_{i,h}, v) = (f_i, v) \quad \forall v \in \mathbb{V}_h; \quad i = 1, ..., M. \tag{15}$$

Now, starting from $U^0_h = 0$, we define the discrete sequence (U^n_h) such that $U^n_h = (u^n_{1h}, ...u^n_{Mh})$ and $u^n_{ih} \in \mathbb{V}_h$ solves the variational inequality (V.I)

$$\begin{cases} a(u^n_{ih}, v - u^n_{ih}) \geq (f_i, v - u^n_{ih}) \quad \forall v \in \mathbb{V}_h, \\ u^n_{ih} \leq k + u^{n-1}_{i+1h}, v \leq k + u^{n-1}_{i+1h}, \\ u^{n-1}_{M+1h} = u^{n-1}_{1h}. \end{cases} \tag{16}$$

Theorem 2. *[5] Under conditions of lemma 1, the sequence (U^n_h) and $(U_{n,h})$ converges decreasingly to the unique solution solution U_h of Q.V.I (5).Moreover, there exists a constant $0 < \mu < 1$ such that*

$$\|U^n_h - U_h\|_\infty \leq \mu^n \left\|U^0_h\right\|_\infty, \tag{17}$$

$$\|U_{n,h} - U_h\|_\infty \leq \mu^n \left\|U^0_h\right\|_\infty. \tag{18}$$

3.3. **Discrete regularity.** Let $\omega \in H_0^1(\Omega)$ be the solution of the V.I

$$\begin{cases} a(\omega, v-\omega) \geq (g, v-\omega) \, \forall v \in H_0^1(\Omega), \\ v \leq r_h\psi, \ \omega \leq r_h\psi \end{cases} \tag{19}$$

and $\omega_h \in \mathbb{V}_h$,its discrete counterpart, the solution of the V.I

$$\begin{cases} a(\omega_h, v-\omega_h) \geq (g, v-\omega_h) \, \forall v \in \mathbb{V}_h, \\ v \leq r_h\psi, \ \omega_h \leq r_h\psi. \end{cases} \tag{20}$$

This concept of "discrete regularity", introduced in [10], can be regarded as the discrete counterpart of the Lewy-Stampaccia estimate $\|\mathcal{A}u\|_\infty \leq C$ ($\mathcal{A}$ being the operator associated with bilinear form $a(.,.)$), extended to the variational form through the $L^1 - L^\infty$ duality. The main role it plays, in the present paper, is in the regularization of the obstacles appearing in the discrete problems (16)

Lemma 2. *[10] We assume that there exists a constant C independent of h such that*

$$|a(\omega_h, \varphi_s)| \leq C \|\varphi_s\|_{L^1(\Omega)} \quad \forall s = 1, 2, ..., N_h. \tag{21}$$

Then, there exists a family of right hand sides $g^{(h)}$ such that

$$\left\|g^{(h)}\right\|_\infty \leq C$$

and

$$a(\omega_h, v) = (g^{n,(h)}, v) \quad \forall v \in \mathbb{V}_h.$$

Theorem 3. *Let conditions of lemma 2 hold. Then, there exists a sequence $\left(g^{n,(h)}\right)_{n\geq 1}$ and a constant $C > 0$ independent of h and n such that*

$$\left\|g^{n,(h)}\right\|_\infty \leq C,$$

$$a(u_{ih}^n, v) = (g^{(h)}, v) \quad \forall v \in \mathbb{V}_h,$$

where u_{ih}^n is defined in (16).

Proof. The proof will be carried out by induction. For $n = 1$, let u_{ih}^1 be the solution of the V.I

$$\begin{cases} a(u_{ih}^1, v - u_h^1) \geq (f_i, v - u_{ih}^1) \ \forall v \in \mathbb{V}_h, \\ v \leq k + u_{ih}^0, \ \ u_{ih}^1 \leq k + u_{ih}^0, \end{cases}$$

where

$$a(u_{ih}^0, v) = (f_i, v) \quad \forall v \in V_h,$$

So, clearly

$$\left|a(u_{ih}^0, \varphi_s)\right| \leq C \|\varphi_s\|_{L^1(\Omega)} \quad \forall s = 1, 2, ..., N_h \tag{22}$$

and, using the discrete Levy-Stampachia inequality [4], we have

$$-(f_i, \varphi_s) \wedge a(k + u_{ih}^0, \varphi_s) \leq a(u_{ih}^1, \varphi_i) \leq (f, \varphi_s).$$

But

$$a(k + u_{ih}^0, \varphi_s) = a(u_{ih}^0, \varphi_s) + (ka_0^i(x), \varphi_s)$$

and, using (22), there exists a constant C such that,

$$-(f_i,\varphi_i)\wedge(-C,\varphi_s)\leq a(u_{ih}^1,\varphi_s)\leq(f,\varphi_s)$$

which implies

$$\left|a(u_{ih}^1,\varphi_s)\right|\leq C\left\|\varphi_s\right\|_{L^1(\Omega)},\forall s=1,2,...,N_h.$$

Hence, making use of lemma 2, there exists a family of right-hands side $\left\{g_i^{1,(h)}\right\}\in L^\infty(\Omega)$ such that

$$\begin{cases} i) & \left\|g_i^{1,(h)}\right\|_\infty\leq C \\ & \text{and} \\ ii) & a(u_{ih}^1,v)=(g_i^{1,(h)},v) \quad \forall v\in\mathbb{V}_h. \end{cases}$$

Now, assume that there exists a constant C independent of n such that

$$a(u_{ih}^{n-1},\varphi_s)\leq C\left\|\varphi_s\right\|_{L^1(\Omega)},\quad \forall s=1,2,...,N_h. \tag{23}$$

So, using the discrete Levy-Stampachia inequality , we get

$$-(f,\varphi_s)\wedge a(k+u_{ih}^{n-1},\varphi_i)\leq a(u_{ih}^n,\varphi_s)\leq(f,\varphi_s)$$

or

$$-(f,\varphi_s)\wedge(a(k+u_{ih}^{n-1},\varphi_s)\leq a(u_{ih}^n,\varphi_s)\leq(f,\varphi_s)$$

and, as

$$a(k+u_h^{n-1},\varphi_s)=a(u_h^{n-1},\varphi_s)+(ka_0^i(x),\varphi_s)$$

using (23) as above, we have

$$-(f_i,\varphi_s)\wedge(-C,\varphi_s)\leq a(u_h^n,\varphi_s)\leq(f,\varphi_s)$$

which implies

$$|a(u_h^n,\varphi_s)|\leq C\left\|\varphi_s\right\|_{L^1(\Omega)}.$$

So, making use of lemma 2, there exists family of right-hands side $\left\{g_i^{n,(h)}\right\}\in L^\infty(\Omega)$ such that

$$\begin{cases} i) & \left\|g_i^{n,(h)}\right\|_\infty\leq C \\ & \text{and} \\ ii) & a(u_{ih}^n,v)=(g_i^{n,(h)},v) \quad \forall v\in\mathbb{V}_h \end{cases}$$

which completes the proof. □

Note that, as

$$a(u_{ih}^n,v)=(g_i^{n,(h)},v)\forall v\in\mathbb{V}_h$$

one can define

$$U^{n,(h)}=\left(u_1^{n,(h)},...,u_M^{n,(h)}\right),$$

the discrete analog of

$$U_h^n=(u_{1h}^n,...,u_{Mh}^n)$$

such that

$$\left\|u_i^{n,(h)}\right\|_{W^{2,p}(\Omega)}\leq C$$

and

$$a(u_i^{n,(h)}, v) = (g_i^{n,(h)}, v) \quad \forall v \in H_0^1(\Omega) \tag{24}$$

and, by standard maximum norm estimates

$$\left\| u_i^{n,(h)} - u_{ih}^n \right\|_\infty \leq Ch^2 \left| \log h \right|. \tag{25}$$

4. L^∞ – Error Analysis

From now on, C will denote an arbitrary constant independent of both h and n.

4.1. **Background.** We begin with recalling some useful properties enjoyed by elliptic variational inequalities. Indeed, let

Definition 2. $w \in H_0^1(\Omega)$ is said to be a subsolution for the VI (19) if

$$\begin{cases} a(w, v) \leq (g, v) \, \forall v \in H_0^1(\Omega), v \geq 0, \\ w \leq \psi. \end{cases} \tag{26}$$

Theorem 4. *[7] The solution ω of V.I (19) is the least upper bound of the set of subsolutions.*

Theorem 5. *[7] Let $\omega = \partial(\tilde{\psi})$ and $\tilde{\omega} = \partial(\tilde{\psi})$. Then, we have*

$$\left\| \omega - \tilde{\omega} \right\|_\infty \leq C \left\| \psi - \tilde{\psi} \right\|_\infty. \tag{27}$$

Remark 1. *Under conditions of lemma 1, the above properties of the solution of V.I (19) remain valid in the discrete case.*

Indeed, let $\omega_h = \partial_h(\psi) \in \mathbb{V}_h$ be the solution of the discrete variational inequality

$$\begin{cases} a(\omega_h, v - \omega_h) \geq (g, v - \omega_h) \, \forall v \in \mathbb{V}_h, \\ \omega_h \leq r_h\psi, v \leq r_h\psi. \end{cases} \tag{28}$$

Next, we shall give the discrete analog of Theorems 3, 4. Their respective proofs will be omitted as they are similar to their continuous counterparts.

Definition 3. $w_h \in \mathbb{V}_h$ is said to be a subsolution for the V.I (28) if

$$\begin{cases} a(w_h, \varphi_s) \leq (g, \varphi_s) \, \forall s = 1, ..., N_h, \\ w_h \leq r_h\psi. \end{cases} \tag{29}$$

Theorem 6. *Under conditions of lemma 1, the solution ω_h of V.I (28) is the least upper bound of the set of discrete subsolutions.*

Theorem 7. *Let $\omega_h = \partial_h(\psi)$ and $\tilde{\omega}_h = \partial_h(\tilde{\psi})$. Then, under conditions of lemma 1, we have*

$$\left\| \omega_h - \tilde{\omega}_h \right\|_\infty \leq C \left\| \psi - \tilde{\psi} \right\|_\infty. \tag{30}$$

Lemma 3. *[11] If $\psi \in W^{2,p}(\Omega)$ and $\omega \in W^{2,p}(\Omega)$, $2 \leq p < \infty$, then the following error estimate holds*

$$\left\| \omega - \omega_h \right\|_\infty \leq Ch^2 \left| \ln h \right|^2. \tag{31}$$

4.2. L^∞- **Error estimate for the Iterative Scheme.** In order to estimate the error between the continuous iterative scheme and its finite element counterpart, we introduce the following sequences of variational inequalities.

An auxiliary sequence of continuous variational inequalities: We introduce the sequence $\bar{U}^n = (\bar{u}_1^n, ..., \bar{u}_M^n)_{n\geq 1}$ such that $\bar{u}_i^n = \partial\left(k + \hat{u}_{i+1,h}^{n-1}\right) \in H_0^1(\Omega)$ solves the continuous V.I**:**

$$\begin{cases} a_i(\bar{u}_i^n, v - \bar{u}^n) \geq (f, v - \bar{u}_i^n) \quad \forall v \in H_0^1(\Omega), \\ \bar{u}_i^n \leq k + u_{i+1}^{n-1,(h)}, \quad v \leq k + u_{i+1}^{n-1,(h)}, \\ u_{M+1,h}^{n-1} = u_{i+1}^{n-1,(h)}, \end{cases} \tag{32}$$

where $u_{i+1}^{n-1,(h)}$ is defined in (24).

An auxiliary sequence of discrete variational inequalities We define the sequence $\bar{U}_h^n = \left(\bar{u}_{1,h}^n, ..., \bar{u}_{i,h}^n\right)_{n\geq 1}$ such that $\bar{u}_{i,h}^n = \partial_h\left(k + u_{i+1}^{n-1}\right) \in \mathbb{V}_h$ solves the discrete V.I

$$\begin{cases} a_i(\bar{u}_{i,h}^n, v - \bar{u}_h^n) \geq (f_i, v - \bar{u}_h^n) \quad \forall v \in \mathbb{V}_h, \\ \bar{u}_h^n \leq r_h\left(k + u_{i+1}^{n-1}\right), \; v \leq r_h\left(k + u_{i+1}^{n-1}\right), \\ u_{M+1}^{n-1} = u_1^{n-1}, \end{cases} \tag{33}$$

where u^0 and u^n are defined in (8) and (9), respectively.

Theorem 8. *We have*

$$\left\|U^n - \bar{U}_h^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2. \tag{34}$$

Proof. As $\bar{u}_{i,h}^n$ is the discrete counterparts of u_i^n and $\|u_i^n\|_{W^{2,p}(\Omega)} \leq C$ (independent of n) (see [5]), making use of (31), we get the desired error estimates. □

Theorem 9.

$$\left\|U^n - U_h^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2. \tag{35}$$

Proof. We proceed by induction. Indeed, consider V.I (32) for $n = 1$:

$$\begin{cases} a_i(\bar{u}_i^1, v - \bar{u}_i^1) \geq (f, v - \bar{u}_i^1) \quad \forall v \in H_0^1(\Omega), \\ \bar{u}_i^1 \leq k + u_{i+1}^{0,(h)}, \quad v \leq k + u_{i+1}^{0,(h)}, \\ u_{M+1}^{0,(h)} = u_1^{0,(h)}. \end{cases}$$

So

$$\left\|\bar{u}_i^1 - u_{i,h}^1\right\|_\infty \leq Ch^2 \left|\ln h\right|^2. \tag{36}$$

Indeed, let $\bar{u}_i^1 = \partial(k + u_{i+1}^{0,(h)})$, $\tilde{u}_{i,h}^1 = \partial_h(k + u_{i+1}^{0,(h)})$ and $u_{i,h}^1 = \partial_h(k + u_{i+1,h}^0)$. Then, as $\tilde{u}_{i,h}^1$ is the discrete analog of $\bar{u}_i^1$, making use of (34), we have

$$\left\|\bar{u}_i^1 - \tilde{u}_{i,h}^1\right\|_\infty \leq Ch^2 \left|\ln h\right|^2. \tag{37}$$

Moreover, using (30) and standard maximum error estimate, we get

$$\begin{aligned}\left\|u_{i,h}^1-\tilde{u}_{i,h}^1\right\|_\infty &\leq \left\|u_{i+1}^{0,(h)}-u_{i+1,h}^0\right\|_\infty \\ &\leq Ch^2\left|\ln h\right|.\end{aligned}$$

Thus

$$\begin{aligned}\left\|\bar{u}_i^1-u_{i,h}^1\right\|_\infty &\leq \left\|\bar{u}_i^1-\tilde{u}_{i,h}^1\right\|_\infty+\left\|\tilde{u}_{i,h}^1-u_{i,h}^1\right\|_\infty \\ &\leq Ch^2\left|\ln h\right|^2.\end{aligned}$$

Now, as $\bar{u}_i^1$ is solution to a V.I, it is also a subsolution, i.e.,

$$\begin{aligned}&a(\bar{u}_i^1,v)\leq(f_i,v)\quad\forall v\in H_0^1(\Omega),v\geq 0,\\ &\bar{u}_i^1\leq k+u_{i+1}^{0,(h)}.\end{aligned}$$

But, as

$$\begin{aligned}\bar{u}_i^1 &\leq k+\left\|u_{i+1}^{0,(h)}-u_{i+1,h}^0\right\|_\infty+u_{i+1}^0\leq\\ &\leq k+Ch^2\left|\ln h\right|^2+u_{i+1}^0,\end{aligned}$$

we have

$$\begin{aligned}&a(\bar{u}_i^1,v)\leq(f,v)\forall v\in H_0^1(\Omega),v\geq 0,\\ &\bar{u}_i^1\leq k+Ch^2\left|\ln h\right|+u_{i+1}^0.\end{aligned}$$

Hence, $\bar{u}_i^1$ is also a subsolution for the V.I with obstacle $k+Ch^2\left|\ln h\right|^2+u_{i+1}^0$. Let $\bar{\omega}_i^1=\partial(k+Ch^2\left|\ln h\right|^2+u_{i+1}^0)$. Then, as $u_i^1=\partial(k+u_{i+1}^0)$, making use of (27) and standard maximum error estimate

$$\left\|u_{i+1}^0-u_{i+1,h}^0\right\|_\infty\leq Ch^2\left|\ln h\right|, \tag{38}$$

we get

$$\begin{aligned}\left\|\bar{\omega}_i^1-u_i^1\right\|_\infty &\leq Ch^2\left|\ln h\right|^2+\left\|u_{i+1}^0-u_{i+1,h}^0\right\|_\infty\leq\\ &\leq Ch^2\left|\ln h\right|^2.\end{aligned}$$

Hence, making use of Theorem 4, we have

$$\bar{u}_i^1\leq\bar{\omega}_i^1\leq u_i^1+Ch^2\left|\ln h\right|^2.$$

Putting

$$\beta_i^1=\bar{u}_i^1-Ch^2\left|\ln h\right|^2,\forall i=1,...,M,$$

we get

$$\beta_i^1\leq u_i^1,\forall i=1,...,M. \tag{39}$$

Further more, using estimate (36), we get

$$\begin{aligned}\left\|\beta_i^1-u_{i,h}^1\right\|_\infty &\leq \left\|\bar{u}_i^1-u_{i,h}^1\right\|_\infty+Ch^2\left|\ln h\right|^2\leq \\ &\leq Ch^2\left|\ln h\right|^2.\end{aligned} \tag{40}$$

Now consider the discrete V.I (33) for $n = 1$:

$$\begin{cases} a_i(\bar{u}^1_{i,h}, v - \bar{u}^1_{i,h}) \geq (f_i, v - \bar{u}^1_{i,h}) \quad \forall v \in \mathbb{V}_h, \\ \bar{u}^1_{i,h} \leq r_h\left(k + u^0_{i+1}\right), \; v \leq r_h\left(k + u^0_{i+1}\right), \end{cases}$$

$\bar{u}^1_{i,h}$ being also a discrete subsolution, we have

$$\begin{aligned} a(\bar{u}^1_{i,h}, \varphi_i) &\leq (f, \varphi_i) \quad \forall \varphi_i, \\ \bar{u}^1_{i,h} &\leq r_h\left(k + u^0_{i+1}\right), \end{aligned}$$

and, from standard maximum error estimate

$$\left\|u^0 - u^0_h\right\|_\infty \leq Ch^2 \left|\ln h\right|.$$

So

$$\begin{aligned} \bar{u}^1_{i,h} &\leq k + \left\|u^0_{i+1} - u^0_{i+1,h}\right\|_\infty + r_h u^0_{i+1,h} \leq \\ &\leq k + Ch^2 \left|\ln h\right|^2 + r_h u^0_{i+1,h}. \end{aligned}$$

then

$$\begin{aligned} a_i(\bar{u}^1_{i,h}, \varphi_i) &\leq (f_i, \varphi_i) \quad \forall \varphi_i, \\ \bar{u}^1_{i,h} &\leq k + Ch^2 \left|\ln h\right|^2 + r_h u^0_{i+1,h} \end{aligned}$$

because r_h is Lipschitz. So, $\bar{u}^1_{i,h}$ is also a discrete subsolution for the V.I with obstacle $k + Ch^2 \left|\ln h\right|^2 + r_h u^0_{i+1,h}$. Let $\bar{\omega}^1_{i,h} = \partial_h(k + Ch^2 \left|\ln h\right|^2 + u^0_{i+1,h})$. As $u^1_{i,h} = \partial_h(k + u^0_{i+1,h})$, making use of (30) and (38), we get

$$\begin{aligned} \left\|\bar{\omega}^1_{i,h} - u^1_{i,h}\right\|_\infty &\leq \left\|u^0_{i+1,h} - u^0_{i+1,h}\right\|_\infty \leq \\ &\leq Ch^2 \left|\ln h\right|^2 \end{aligned}$$

and, applying Theorem 6, we get

$$\bar{u}^1_{i,h} \leq \bar{\omega}^1_{i,h} \leq u^1_{i,h} + Ch^2 \left|\ln h\right|^2.$$

Now, taking

$$\gamma^1_{i,h} = \bar{u}^1_{i,h} - Ch^2 \left|\ln h\right|^2, \quad \forall i = 1, ..., M,$$

we have

$$\gamma^1_{i,h} \leq u^1_{i,h}, \forall i = 1, ..., M. \tag{41}$$

Hence, as $u^1_{i,h}$ is the discrete analog of u^1_i, making use (30) and (34), we get

$$\begin{aligned} \left\|\gamma^1_{i,h} - u^1_i\right\|_\infty &\leq \left\|\bar{u}^1_{i,h} - u^1_i\right\|_\infty + Ch^2 \left|\ln h\right|^2 \leq \\ &\leq Ch^2 \left|\ln h\right|^2. \end{aligned} \tag{42}$$

Thus, combining (39), (40) and (41), (42), we obtain

$$\begin{aligned} u^1_i &\leq \gamma^1_{i,h} + Ch^2 \left|\ln h\right|^2 \\ &\leq u^1_{i,h} + Ch^2 \left|\ln h\right|^2 \\ &\leq \beta^1_i + Ch^2 \left|\ln h\right|^2 \\ &\leq u^1_i + Ch^2 \left|\ln h\right|^2. \end{aligned}$$

That is

$$\left\|u_i^1 - u_{i,h}^1\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 .$$

Let us now assume that

$$\left\|u_i^{n-1} - u_{i,h}^{n-1}\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 . \tag{43}$$

Since $\tilde{u}_{i,h}^n = \partial_h(k + u_{i+1}^{n-1,(h)})$ is the discrete analog of $\bar{u}_i^n = \partial(k + u_{i+1}^{n-1,(h)})$, making use of (34), we get

$$\left\|\bar{u}_i^n - \tilde{u}_{i,h}^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 . \tag{44}$$

Let us now prove that

$$\left\|\bar{u}_i^n - u_{i,h}^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 . \tag{45}$$

Indeed, using (44) , (30), we get

$$\begin{aligned}\left\|\bar{u}_i^n - u_{i,h}^n\right\|_\infty &\leq \left\|\bar{u}_i^n - \tilde{u}_{i,h}^n\right\|_\infty + \left\|\tilde{u}_{i,h}^n - u_{i,h}^n\right\|_\infty \\ &\leq Ch^2 \left|\ln h\right|^2 + \left\|u_{i+1}^{n-1,(h)} - u_{i+1,h}^{n-1}\right\|_\infty \\ &\leq Ch^2 \left|\ln h\right|^2 ,\end{aligned}$$

On the other hand, the solution of V.I (32) is also a subsolution, that is

$$\begin{cases} a_i(\bar{u}_i^n, v) \leq (f_i, v) \quad \forall v \in H^1(\Omega), \quad v \geq 0, \\ \bar{u}_i^n \leq k + u_{i+1}^{n-1,(h)}. \end{cases}$$

So, using (43), we have

$$\begin{aligned}\bar{u}_i^n &\leq k + \left\|u_{i+1}^{n-1} - u_{i+1,h}^{n-1}\right\|_\infty + u_{i+1,h}^{n-1} \\ &\leq k + Ch^2 \left|\ln h\right|^2 + u_{i+1,h}^{n-1}\end{aligned}$$

and thus,

$$\begin{aligned}a_i(\bar{u}_i^n, v) &\leq (f_i, v) \quad \forall v \in H^1(\Omega), \quad v \geq 0, \\ \bar{u}_i^n &\leq k + \left\|u_{i+1}^{n-1} - u_{i+1,h}^{n-1}\right\|_\infty + u_{i+1,h}^{n-1}, \\ &\leq k + Ch^2 \left|\ln h\right|^2 + u_{i+1,h}^{n-1}.\end{aligned}$$

So $\bar{u}_i^n$ is a subsolution for the V.I with obstacle $k + Ch^2 \left|\ln h\right|^2 + u_{i+1,h}^{n-1}$. Let $\bar{\omega}_i^n = \partial(k + Ch^2 \left|\ln h\right|^2 + u_{i+1,h}^{n-1})$. Then, as $u_i^n = \partial(k + u_{i+1}^{n-1})$, making use of (27), and (43), we get

$$\begin{aligned}\left\|\bar{\omega}_i^n - u_i^n\right\|_\infty &\leq Ch^2 \left|\ln h\right|^2 + \left\|u_{i+1,h}^{n-1} - u_{i+1}^{n-1}\right\|_\infty \\ &\leq Ch^2 \left|\ln h\right|^2 .\end{aligned}$$

Hence, applying Theorem 4, we have

$$\bar{u}_i^n \leq \bar{\omega}_i^n \leq u_i^n + Ch^2 \left|\ln h\right|^2 .$$

Now, putting

$$\beta_i^n = \bar{u}_i^n - Ch^2 \left|\ln h\right|^2 , \quad \forall i = 1, ..., M.$$

we obtain

$$\beta_i^n \leq u_i^n, \forall i = 1, ..., M \tag{46}$$

and, using (45) ,

$$\begin{aligned} \left\|\beta_i^n - u_{i,h}^n\right\|_\infty &\leq \left\|\bar{u}_i^n - Ch^2 \left|\ln h\right|^2 - u_{i,h}^n\right\|_\infty \\ &\leq \left\|\bar{u}_i^n - u_{i,h}^n\right\|_\infty + Ch^2 \left|\ln h\right|^2 \\ &\leq Ch^2 \left|\ln h\right|^2 . \end{aligned} \tag{47}$$

Now, consider the discrete V.I (33)

$$\begin{cases} a_i(\bar{u}_{i,h}^n, v - \bar{u}_{i,h}^n) \geq (f_i, v - \bar{u}_{i,h}^n) \quad \forall v \in \mathbb{V}_h, \\ \bar{u}_{i,h}^n \leq r_h\left(k + u_{i+1}^{n-1}\right), \, v \leq r_h\left(k + u_{i+1}^{n-1}\right), \end{cases} \tag{48}$$

$\bar{u}_{i,h}^n$ being also a subsolution, we have

$$\begin{cases} a_i(\bar{u}_{i,h}^n, \varphi_i) \leq (f_i, \varphi_i) \quad \forall i = 1, ..., m(h), \\ \bar{u}_{i,h}^n \leq r_h\left(k + u_{i+1}^{n-1}\right). \end{cases} \tag{49}$$

So, making use of (43), we have

$$\begin{aligned} \bar{u}_{i,h}^n &\leq k + r_h u_{i+1}^{n-1} - r_h u_{i+1,h}^{n-1} + r_h u_{i+1,h}^{n-1} \\ &\leq k + \left\|r_h u_{i+1}^{n-1} - r_h u_{i+1,h}^{n-1}\right\|_\infty + r_h u_{i+1,h}^{n-1} \\ &\leq k + Ch^2 \left|\ln h\right|^2 + r_h u_{i+1,h}^{n-1} \end{aligned}$$

and hence

$$\begin{aligned} a(\bar{u}_{i,h}^n, \varphi_i) &\leq (f_i, \varphi_i) \quad \forall \varphi_i, \\ \bar{u}_{i,h}^n &\leq k + Ch^2 \left|\ln h\right|^2 + r_h \hat{u}_{i+1,h}^{n-1}. \end{aligned}$$

So, $\bar{u}_{i,h}^n$ is a subsolution for the V.I with obstacle $k + Ch^2 \left|\ln h\right|^2 + r_h u_{i+1,h}^{n-1}$. Let $\bar{\omega}_{i,h}^n = \partial_h(k + Ch^2 \left|\ln h\right|^2 + r_h u_{i+1,h}^{n-1})$. Then, as $u_{i,h}^n = \partial_h(k + r_h u_{i+1,h}^{n-1})$, making use of (30) and (43), we get

$$\left\|\bar{\omega}_{i,h}^n - u_{i,h}^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 + \left\|u_{i+1,h}^{n-1} - u_{i+1,h}^{n-1}\right\|_\infty$$

and, due to Theorem 6, we have

$$\bar{u}_{i,h}^n \leq \bar{\omega}_{i,h}^n \leq u_{i,h}^n + Ch^2 \left|\ln h\right|^2 .$$

Now, taking

$$\gamma_{i,h}^n = \bar{u}_{i,h}^n - Ch^2 \left|\ln h\right|^2 , \quad \forall i = 1, ..., M.$$

we obtain

$$\gamma_{i,h}^n \leq u_{i,h}^n. \tag{50}$$

Moreover, $\bar{u}_h^n$ being the discrete counterpart of u^n, using (34), we have

$$\left\|\bar{u}_{i,h}^n - u_i^n\right\|_\infty \leq Ch^2 \left|\ln h\right|^2 , \quad \forall i = 1, ..., M$$

and therefore

$$\begin{aligned}\left\|\gamma_{i,h}^n - u_i^n\right\|_\infty &\le \left\|\bar{u}_{i,h}^n - u_i^n\right\|_\infty + Ch^2 \left|\ln h\right|^2 \\ &\le Ch^2 \left|\ln h\right|^2 .\end{aligned} \tag{51}$$

Finally, combining (46), (47) and (50), (51), we obtain

$$\begin{aligned}u_i^n &\le \gamma_{i,h}^n + Ch^2 \left|\ln h\right|^2 \\ &\le u_{i,h}^n + Ch^2 \left|\ln h\right|^2 \\ &\le \beta_i^n + Ch^2 \left|\ln h\right|^2 \\ &\le u_i^n + Ch^2 \left|\ln h\right|^2 .\end{aligned}$$

That is

$$\left\|u_i^n - u_{i,h}^n\right\|_\infty \le Ch^2 \left|\ln h\right|^2 \quad \forall i = 1, ..., M.$$

□

4.3. **L^∞-Error estimate for the system of QVIs.** Now combining estimates (10) , (17), and (35), we have:

Theorem 10.

$$\left\|U - U_h\right\|_\infty \le Ch^2 \left|\ln h\right|^2 . \tag{52}$$

Proof. Indeed,

$$\begin{aligned}\left\|U - U_h\right\|_\infty &\le \left\|U - U^n\right\|_\infty + \left\|U^n - U_h^n\right\|_\infty + \left\|U_h^n - U_h\right\|_\infty \\ &\le \mu^n \left\|U^0\right\|_\infty + Ch^2 \left|\ln h\right|^2 + \mu^n \left\|U_h^0\right\|_\infty .\end{aligned} \tag{53}$$

So, passing to the limit, as $n \to \infty$, the desired result follows. □

Remark 2. *For practical purposes, it is interesting to estimate the error between the exact solution and the actually computed approximations U_h^n, that is,*

$$\left\|U - U_h^n\right\|_\infty \le \mu^n \left\|U^0\right\|_\infty + Ch^2 \left|\ln h\right|^2 . \tag{54}$$

Proof. Indeed,

$$\begin{aligned}\left\|U - U_h^n\right\|_\infty &\le \left\|U - U^n\right\|_\infty + \left\|U^n - U_h^n\right\|_\infty \\ &\le \mu^n \left\|U^0\right\|_\infty + Ch^2 \left|\ln h\right|^2 .\end{aligned}$$

□

5. Numerical example

Let $\Omega = (0,1) \times (0,1)$, $M = 3$, $\mathcal{A}^i = -\triangle$, $f_1 = \sin^2 x$, $f_2 = \cos^2 x$, $f_3 = e^x$. We divide Ω into squares with edge $h = \frac{1}{10}$, then by diagonals with same direction divide every square into two triangles. Then the finite dimensional quasi-variational inequalities system is

$$\begin{cases} U_i \in K_i, \\ \left(A^i U_i - F_i, V - U_i\right) \ge 0, \quad \forall V \in K_i, \quad i = 1, ..., M, \end{cases} \tag{55}$$

where A^i are the stiffness matrices defined in (11), and the right-hand side $F_i = (f_i, \varphi_l), l = 1, ..., N_h,, K_i = \left\{V \in R^{N_h} \text{ such that } V \leq K + U_{i+1}\right\}, U_{M+1} = U_1$, $K = (k, ..., k))^T$ .The iterative scheme is,

$$\begin{cases} U_i^{n+1} \in K^{i,n+1}, \\ \left(A^i U^{i,n+1} - F^i, V - U^{i,n+1}\right) \geq 0, \quad \forall V \in K^{i,n+1}, \quad i = 1, ..., M, \end{cases} \tag{56}$$

where $K^{i,n+1} = \left\{V \in R^{N_h} \text{ such that } V \leq K + U^{i,n}\right\}$, $U^{M+1,n} = U^{1,n}$.

We take $k = 0.01$ and solve (56) (Jacobi type) with projected Gauss-Seidel as inner iteration. The stopping criteria for the inner iteration and outer iteration both are $\epsilon = 10^{-6}$, the initial value is $U^0 = (U_1^0, ..., U_M^0)$, such that $A^i U_i^0 = F^i, \quad i = 1, ..., M$.

The computation of the solution for h, $h/2$ and $h/4$ leads to a convergence order $p = 2.062$, which is in good agreement with the theory.

6. Conclusion

This paper addresses the finite element of the Dirichlet problem for an elliptic quasi-variational inequalities system. The optimal error estimate is derived, combining geometric convergence of an iterative scheme and its finite element error estimate, obtained by means of the concept of subsolutions and discrete regularity for variational inequalities. A numerical example is also given to support the theory.

In light of the findings of this work, we wonder whether these can be exploited to:

1. Extend the study to the noncoercive problem.
2. Derive a posteriori error estimate for this system of Q.V.I.

This will be the focus of our attention in future works.

Acknowledgments The author would like to thank the referee for her-his careful reading of the paper and for valuable comments and suggestions.

Bibliography

1. Evans L.C. Optimal stochastic switching and the Dirichlet Problem for the Bellman equations / L.C. Evans, A. Friedman // Transactions of the American Mathematical Society. – 1979. – 253. – P. 365-389.
2. Lions P.L. Optimal control of stochastic integrals and Hamilton Jacobi Bellman equations (part I) / P.L. Lions, J.L. Menaldi // SIAM control and optimization. – 1979. – 20.
3. Boulbrachene M. On the finite element approximation of variational inequalities with noncoercive operators / M. Boulbrachene // Numerical Functional Analysis and Optimization. – 2015. – 36. – P. 1107-1121.
4. Cortey Dumont P. Sur l' analyse numerique des equations de Hamilton-Jacobi-Bellman / P. Cortey Dumont, // Math. Meth in Appl. Sci. – 1987. – 9. – P. 198-209.
5. Boulbrachene M. The Finite element approximation of Hamilton-Jacobi-Bellman equations / M. Boulbrachene, M. Haiour // Computers & Mathematics with Applications. – 2001. – Vol. 41. – 993-1007.
6. Boulbrachene M. Optimal L^∞ -error estimates of a finite element method for Hamilton-Jacobi-Bellman Equations / M. Boulbrachene, P. Cortey Dumont // Numerical Functional Analysis and Optimization. – 2009. – No. 30, (5-6). – P. 421-435.
7. Bensoussan A. Applications of variational inequalities in stochastic control problems. / A. Bensoussan, J.L. Lions. – North Holland, 2000.

8. Lu C. Maximum principle in linear finite element approximations of anisotropic diffusion–convection–reaction problems / C. Lu, W. Huang, J. Qiu // Numer. Math. – 2014. – 127. – P. 515-537.
9. Vejchodsky T. The discrete maximum principle for Galerkin solutions of elliptic problems / T. Vejchodsky // Centr. Eur. J. Math. – 2012. – 10 (1). – P. 25-43.
10. Cortey Dumont P. Contribution a l' approximation des inequations variationnelles en norme L^∞ / P. Cortey-Dumont // C.R.Acad. Sci. Paris Ser. I Math. – 1983. – 296, 17. – P. 753-756.
11. Cortey-Dumont P. On the finite element approximation in the L^∞ norm of variational inequalities with nonlinear operators, / P. Cortey-Dumont // Numer.Num. – 1985. – 47. – P. 45-57.

M. Boulbrachene,
Department of Mathematics,
Sultan Qaboos University,
P.O. Box 36, Muscat 123, Sulltanate of Oman.

Received 06.03.2018; revised 09.07.2018

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 517.9

COUPLING OF LAGUERRE TRANSFORM AND FAST BEM FOR SOLVING DIRICHLET INITIAL-BOUNDARY VALUE PROBLEMS FOR THE WAVE EQUATION

A. R. Hlova, S. V. Litynskyy, Yu. A. Muzychuk, A. O. Muzychuk

Резюме. Подано поглиблений аналіз двох підходів до розв'язування початково-крайової задачі Діріхле для однорідного хвильового рівняння, який базується на поєднанні перетворення Лагера за часовою змінною і методу граничних елементів (МГЕ) у необмеженій просторовій області. В результаті обидва підходи приводять до тієї ж самої нескінченної трикутної системи граничних інтегральних рівнянь. Аналіз проведено у вагових просторах Соболєва, елементами яких є функції часової змінної, які набувають значень у відповідних просторах Соболєва.

Для зменшення потреби в обчислювальних ресурсах реалізовано швидкий МГЕ, використовуючи адаптивну перехресну апроксимацію отриманих матриць. Крім того, метод поширено на розв'язування задачі Діріхле в області з включенням. Також подано чисельні результати для модельних задач, які ілюструють точність і очікуваний порядок збіжності запропонованого методу.

Abstract. We present an improved analysis of two approaches to solving of the Dirichlet initial-boundary value problem for a homogeneous wave equation, which are based on the combination of the Laguerre transform for the time variable with the Galerkin-BEM in an unbounded spatial domain. Both approaches lead to the same infinite triangular system of boundary integral equations as a result. The analysis is done in weighted Sobolev spaces of functions of the time variable taking values in suitable Sobolev spaces.

For reducing both storage and computational costs we implement the fast BEM using adaptive cross approximation of obtained matrices. Furthermore, we extend this method for solving the Dirichlet problem in the domain with an inclusion. We also present numerical results for some model problems which illustrate the accuracy and estimated convergence order of the proposed method.

1. Introduction

In recent years, many studies have been dedicated to the development of effective methods for the numerical solution of time domain boundary integral equations (TDBIEs), which arise from initial-boundary value problems (IBVPs) for the wave equation. Comprehensive lists of related works are presented in [11, 35]. A common feature of these studies is the usage of deep analytical concepts to take into account the dependence of the solutions on the time variable. However, as noted in [10], the computational complexity of proposed

Key words. Dirichlet initial-boundary value problem, wave equation, Laguerre transform, Fast Galerkin-BEM, time domain boundary integral equations, boundary integral equation, retarded single layer potential, half-space with inclusion, adaptive cross approximation.

approaches is still high for problems in 3D domains and the development of effective numerical methods remains actual.

In this paper we present new results of solving both IBVPs and TDBIEs by approach, which is based on the Laguerre transform (LT) [18, 25] in the time variable. The advantage of this transform is that an inverse LT is easy to calculate. Moreover, for solving both boundary value problems (BVPs) and boundary integral equations (BIEs) in the Laguerre domain, efficient recursive algorithms can be constructed using techniques well developed for elliptic problems and their BIEs.

We distinguish two approaches with respect to the order in which the LT is applied in solving IBVPs. In the first case, the transform is applied directly to the IBVP, and as a result, a BVP for infinite triangular system of elliptic equations is obtained. Such approach was used (without much theoretical justification) for solving different evolutional IBVPs in papers [4,5,13,28,29,33,37], in which for the problems in the Laguerre domain a suitable representation of the solution was also constructed and corresponding BIEs were derived. Variational formulations for such problems and associated BIEs were proposed and justified for the first time in [30].

Theoretical aspects of another approach, when the LT is directly applied to retarded potentials, were investigated in [24, 25]. The results for Dirichlet and Neumann IBVPs obtained therein have enabled to substantiate the equivalence between each of these problems and infinite triangular systems of corresponding BIEs in the Laguerre domain and also to define the scope of the problems that can be solved with help of the LT.

Both aforementioned approaches lead to the same infinite triangular system of BIEs. This fact creates a basis for the justification of the first approach, as well as for the effective implementation of the BEM for numerical solution of the system of BIEs. These two aspects determine the main research goal of this article.

We begin in Section 1 with a brief description of the second approach, where the LT is applied to the TDBIE, which arose from the Dirichlet IBVP by using a retarded single layer potential. We introduce the needed functional spaces, give a definition of the LT and obtain an infinite sequence of BIEs.

In Section 2 we transform the IBVP to the BVP for an infinite system of elliptic equations and explain how this approach leads to a sequence of BIEs. After that we derive the representation of the solution of the IBVP in the form of the Fourier-Laguerre series, which coefficients represent the solution of the BVP in the Laguerre domain. Then in Section 3 we consider the IBVP in the half-space with some inclusion and obtain the representation of its solution using a Green's function for such domain. At the end in Section 4 we demonstrate the implementation of the Galerkin-BEM and its fast modification, and present the results of the numerical experiments.

2. Reduction of the IBVP to the infinite system of BIEs

Let Ω^- be a bounded domain in $\mathbb{R}^3$ with Lipschitz boundary Γ, $\Omega := \mathbb{R}^3 \setminus \overline{\Omega^-}$, $\mathbb{R}_+ := (0, \infty)$, $Q := \Omega \times \mathbb{R}_+$ and $\Sigma := \Gamma \times \mathbb{R}_+$. We consider the initial-boundary

value problem for the homogeneous wave equation

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u(x,t) = 0, \quad (x,t) \in Q, \tag{1}$$

where $\Delta := \sum_{i=1}^{3} \partial^2/\partial x_i^2$ is the Laplace operator. We find a function $u(x,t)$, $(x,t) \in \overline{Q}$, which satisfies (in some sense) the equation(1), homogeneous initial conditions

$$u(x,0) = 0, \quad \frac{\partial u(x,0)}{\partial t} = 0, \quad x \in \Omega, \tag{2}$$

and the Dirichlet boundary condition

$$u(x,t) = g(x,t), \quad (x,t) \in \Sigma, \tag{3}$$

where function g is given on Σ. We also call (1)-(3) a Dirichlet problem.

To solve the IBVP (1)-(3) we use a retarded single layer potential

$$(\mathcal{S}\mu)(x,t) := \frac{1}{4\pi} \int\limits_{\Gamma} \frac{\mu(y, t-|x-y|)}{|x-y|} d\Gamma_y, \quad (x,t) \in \overline{Q}, \tag{4}$$

where $\mu : \Gamma \times \mathbb{R} \to \mathbb{R}$ is an unknown density. It is known (see, e.g., [34]) that if an arbitrary function $\mu(y,\tau)$ is smooth enough and $\mu(y,\tau) = 0$ for $y \in \Gamma$ and $\tau \leq 0$, then function

$$u(x,t) = (\mathcal{S}\mu)(x,t), \quad (x,t) \in \overline{Q}, \tag{5}$$

satisfies (in the classical sense) the wave equation and initial conditions. The function u satisfies also the boundary condition (3), if μ is a solution of such TDBIE

$$(\mathcal{V}\mu)(x,t) := \frac{1}{4\pi} \int\limits_{\Gamma} \frac{\mu(y, t-|x-y|)}{|x-y|} d\Gamma_y = g(x,t), \quad (x,t) \in \Sigma. \tag{6}$$

Let X be a Hilbert space with an inner product $(\cdot,\cdot)_X$ and an induced norm $||\cdot||_X$. In order to construct a generalized solution of the IBVP (1)-(3) we consider spaces of functions of the time variable which have values in some Hilbert space X. For such functions the weighted Lebesgue space $L^2_\sigma(\mathbb{R}_+; X)$ [9] with weight $\rho_\sigma(t) = e^{-\sigma t}$ ($t \in \mathbb{R}_+$ and parameter $\sigma > 0$) is the simplest Hilbert space. Elements $v \in L^2_\sigma(\mathbb{R}_+; X)$ are measurable functions $v : \mathbb{R}_+ \to X$ such that $\int\limits_{\mathbb{R}_+} ||v(t)||^2_X \, e^{-\sigma t} dt < \infty$. This space is equipped with the inner product

$$(v,w)_{L^2_\sigma(\mathbb{R}_+;X)} := \int\limits_{\mathbb{R}_+} \big(v(t), w(t)\big)_X \, e^{-\sigma t} dt, \quad v, w \in L^2_\sigma(\mathbb{R}_+; X), \tag{7}$$

and the norm

$$\|v\|_{L^2_\sigma(\mathbb{R}_+;X)} := \sqrt{(v,v)_{L^2_\sigma(\mathbb{R}_+;X)}}, \quad v \in L^2_\sigma(\mathbb{R}_+; X). \tag{8}$$

We also consider the weighted Sobolev spaces

$$H^m_\sigma(\mathbb{R}_+;X) := \{\ v \in L^2_\sigma(\mathbb{R}_+;X)\,|\, v^{(k)} \in L^2_\sigma(\mathbb{R}_+;X), \quad v^{(k)}(0)=0,\ k=\overline{0,m}\} \tag{9}$$

where $m \in \mathbb{N}$ ($\mathbb{N}$ is the set of natural numbers), with norm

$$\|v\|_{H^m_\sigma(\mathbb{R}_+;X)} := \left(\sum_{k=0}^{m} \left\|v^{(k)}\right\|^2_{L^2_\sigma(\mathbb{R}_+;X)}\right)^{1/2}. \tag{10}$$

Here derivatives $v^{(k)},\ k \in \mathbb{N}$, are understood in terms of the space $\mathcal{D}'(\mathbb{R}_+;X)$, elements of which are distributions with values in the space X. We assume that elements of the space $H^m_\sigma(\mathbb{R}_+;X)$ are extended with zero for non-positive arguments.

It is well known [18], that Laguerre polynomials $\{L_k(\sigma\cdot)\}_{k\in\mathbb{N}_0:=\mathbb{N}\cup\{0\}}$ form an orthogonal basis in the space $L^2_\sigma(\mathbb{R}_+) := L^2_\sigma(\mathbb{R}_+;\mathbb{R})$, that is, for every function $f \in L^2_\sigma(\mathbb{R}_+)$ there exists its expansion in the Fourier-Laguerre series

$$f(t) = \sum_{k=0}^{\infty} f_k\, L_k(\sigma t),\ t \in \mathbb{R}_+, \tag{11}$$

where Fourier-Laguerre coefficients $f_0, f_1, ..., f_k, ...$ have the representation formula

$$f_k := \sigma \int\limits_{\mathbb{R}_+} f(t)\, L_k(\sigma t)\, e^{-\sigma t} dt, \quad k \in \mathbb{N}_0. \tag{12}$$

We write a sequence of any elements of the set X as a vector-column $\mathbf{v} := (v_0, v_1, ...)^\top$ and denote by X^∞ a set of all possible sequences of elements of the set X. In particular, we consider a space of numerical sequences $l^2 := \left\{\mathbf{v} \in \mathbb{R}^\infty \,|\, \sum\limits_{j=0}^{\infty} |v_j|^2 < +\infty\right\}$ with the inner product $(\mathbf{v},\mathbf{w}) = \sum\limits_{j=0}^{\infty} v_j w_j$ and the norm $||\mathbf{v}||_{l^2} := \left(\sum\limits_{j=0}^{\infty} |v_j|^2\right)^{1/2}$ for $\mathbf{v},\mathbf{w} \in l^2$.

We recall [18] that the Laguerre transform (LT) is a mapping $\mathcal{L} : L^2_\sigma(\mathbb{R}_+) \to l^2$, which maps an arbitrary function f to a sequence $\mathbf{f} = (f_0, f_1, ..., f_k, ...)^\top$ according to the rule (12). We will also use the notation $\mathcal{L}_k f \equiv (\mathcal{L}f)(k) := f_k\ \ \forall k \in \mathbb{N}_0$. Note that the Parseval equality holds

$$||f||^2_{L^2_\sigma(\mathbb{R}_+)} = \frac{1}{\sigma}\sum_{k=0}^{\infty} |f_k|^2. \tag{13}$$

The LT $\mathcal{L}$ is a bijective mapping and its inverse $\mathcal{L}^{-1} : l^2 \to L^2_\sigma(\mathbb{R}_+)$ maps an arbitrary sequence $\mathbf{h} = (h_0, h_1, ..., h_k, ...)^\top$ to a function

$$(\mathcal{L}^{-1}\mathbf{h})(t) := \sum_{k=0}^{\infty} h_k\, L_k(\sigma t),\ t \in \mathbb{R}_+. \tag{14}$$

For the arbitrary function $f \in L^2_\sigma(\mathbb{R}_+)$ we have an equality

$$\mathcal{L}^{-1}\mathcal{L}f = f. \tag{15}$$

In [24] the LT was extended on functions of time variable with values in the Hilbert space X. LT was considered as a mapping $\mathcal{L} : L^2_\sigma(\mathbb{R}_+; X) \to X^\infty$ which operates according to the rule (12).

Let

$$l^2(X) := \left\{\mathbf{v} \in X^\infty \mid \quad \sum_{j=0}^{\infty} \|v_j\|_X^2 < +\infty\right\}$$

be a Hilbert space with the inner product $(\mathbf{v}, \mathbf{w}) = \sum\limits_{j=0}^{\infty} (v_j, w_j)_X$ and the norm $||\mathbf{v}||_{l^2(X)} := \left(\sum\limits_{j=0}^{\infty} \|v_j\|_X^2\right)^{1/2}$, $\mathbf{v}, \mathbf{w} \in l^2(X)$.

Proposition 1 ([24], Theorem 2). *The mapping $\mathcal{L} : L^2_\sigma(\mathbb{R}_+; X) \to X^\infty$ that maps an arbitrary function f to a sequence $\boldsymbol{f} := (f_0, f_1, ..., f_k, ...)^\top$ according to the formula (12), is injective and its image is the space $l^2(X)$, and*

$$\|f\|^2_{L^2_\sigma(\mathbb{R}_+;X)} = \frac{1}{\sigma} \sum_{k=0}^{\infty} ||f_k||_X^2. \tag{16}$$

In addition, for the arbitrary function $f \in L^2_\sigma(\mathbb{R}_+; X)$ we have an equality

$$\mathcal{L}^{-1}\mathcal{L}f = f, \tag{17}$$

where the mapping $\mathcal{L}^{-1} : l^2(X) \to L^2_\sigma(\mathbb{R}_+; X)$ is the inverse to $\mathcal{L}$ and maps the arbitrary sequence $\boldsymbol{h} := (h_0, h_1, ..., h_k, ...)^\top$ to the function h according to the formula (14).

Definition 4 ([24]). Let $\sigma > 0$ and X be a Hilbert space. Mappings

$$\mathcal{L} : L^2_\sigma(\mathbb{R}_+; X) \to l^2(X) \quad \text{and} \quad \mathcal{L}^{-1} : l^2(X) \to L^2_\sigma(\mathbb{R}_+; X),$$

mentioned in theorem 1, are called, respectively, *direct and inverse Laguerre transforms*, and the formula (16) is an analogue of the Parseval equality.

Definition 5 ([23]). Let X, Y, Z be arbitrary sets and $q : X \times Y \to Z$ be some mapping. By a *q-convolution* of sequences $\mathbf{u} \in X^\infty$ and $\mathbf{v} \in Y^\infty$ we understand the sequence $\mathbf{w} := (w_0, w_1, ..., w_j, ...)^\top \in Z^\infty$, whose elements are obtained by the rule

$$w_j := \sum_{i=0}^{j} q\left(u_{j-i}, v_i\right) \equiv \sum_{i=0}^{j} q\left(u_i, v_{j-i}\right), \quad j \in \mathbb{N}_0; \tag{18}$$

the q-convolution of $\mathbf{u}$ and $\mathbf{v}$ is shortly written in form $\mathbf{w} = \mathbf{u} \underset{q}{\circ} \mathbf{v}$.

If $X = \mathcal{L}(Y, Z)$ is a space of linear operators acting from the space Y into the space Z and $q(A, v) = Av$, $A \in \mathcal{L}(Y, Z)$, $v \in Y$, then components of

the q-convolution of arbitrary sequences $\mathbf{A} \in \big(\mathcal{L}(Y,Z)\big)^{\infty}$ and $\mathbf{v} \in Y^{\infty}$ are represented by the formula

$$w_j = \sum_{i=0}^{j} A_{j-i} v_i, \quad j \in \mathbb{N}_0. \tag{19}$$

In this case we write $\mathbf{w} = \mathbf{A} \underset{Z}{\circ} \mathbf{v}$.

Note that for any function $f \in L^2_\sigma(\, R_+; X)$ the Fourier-Laguerre series of the function $f(t-a)$, $a > 0$, can be expressed in terms of the sequence $\mathbf{f} := \mathcal{L} f$ [24, Lemma 1]:

$$f\big(\,\cdot - a\big) = e^{-\sigma a} \sum_{j=0}^{\infty} \Bigg(\sum_{i=0}^{j} \zeta_{j-i}(\sigma a) f_i \Bigg) L_j(\sigma \cdot) \ \text{ in } L^2_\sigma(\mathbb{R}_+; X), \tag{20}$$

where

$$\zeta_0(s) := 1, \ \ \zeta_k(s) := L_k(s) - L_{k-1}(s), \ \ s \in \overline{\mathbb{R}_+} = [0, \infty), \qquad k \in \mathbb{N}. \tag{21}$$

Let $H^1(\Omega)$ and $H^{1/2}(\Gamma)$ denote the usually defined (see, e.g., [17]) Sobolev spaces and $H^{-1/2}(\Gamma) := (H^{1/2}(\Gamma))'$. Consider now the retarded single layer potential (4) and TDBIE (6). Assuming the density $\mu \in L^2_\sigma(\mathbb{R}_+; H^{-1/2}(\Gamma))$ is sufficiently smooth, we can write the expansion [24]:

$$(\mathcal{S}\mu)(x,t) = \sum_{j=0}^{\infty} u_j(x)\, L_j(\sigma t), \ \ (x,t) \in Q, \tag{22}$$

where coefficients $u_j := \mathcal{L}_j\, \mathcal{S}\mu$, $j \in \mathbb{N}_0$, are components of the q-convolution

$$\mathbf{u}(x) := \big(\mathbf{S} \underset{H^1(\Omega)}{\circ} \boldsymbol{\mu}\big)(x), \ \ x \in \Omega. \tag{23}$$

Here $\boldsymbol{\mu} := \mathcal{L}\mu$ and the sequence $\mathbf{S}$ consists of operators $S_k : H^{-1/2}(\Gamma) \to H^1(\Omega)$, $k \in \mathbb{N}_0$, acting on any function $\xi \in L^2(\Gamma)$ according to the rule

$$(S_k \xi)(x) := \int_{\Gamma} \xi(y) e_k(x-y)\, d\Gamma_y, \ \ x \in \Omega, \tag{24}$$

where

$$e_0(z) := \frac{e^{-\sigma|z|}}{4\pi|z|}, \ e_k(z) := \frac{e^{-\sigma|z|}}{4\pi|z|}\big(L_k(\sigma|z|) - L_{k-1}(\sigma|z|)\big), \ \ z \in \mathbb{R}^3 \setminus \{0\}, \ \ k \in \mathbb{N}. \tag{25}$$

One can extend the expression (24) to the $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$ duality product $(S_k\xi)(x) = \big\langle \xi(\cdot), e_k(x-\cdot)\big\rangle_\Gamma$, $x \in \Omega$, for elements $\xi \in H^{-1/2}(\Gamma)$ [24].

Similarly, applying the LT to the equation (6), we obtain an infinite triangular system of BIEs

$$\mathbf{V} \underset{H^{1/2}(\Gamma)}{\circ} \boldsymbol{\mu} = \mathbf{g} \quad \text{on } \Gamma, \tag{26}$$

where $\mathbf{g} := \mathcal{L} g$ and $\mathbf{V}$ is a sequence of boundary operators $V_k : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$, $k \in \mathbb{N}_0$, which may be expressed as a composition $V_k := \gamma_0 \circ S_k$ of

operator S_k with trace operator γ_0. In case of $\xi \in L^2(\Gamma)$ we have

$$(V_k\xi)(x) = \int\limits_{\Gamma} \xi(y) e_k(x-y)\, d\Gamma_y, \quad x \in \Gamma. \tag{27}$$

Proposition 2 ([24], Theorem 1). *Let $g \in H^{m+4}_{\sigma_0}(\mathbb{R}_+; H^{1/2}(\Gamma))$ for some $\sigma_0 > 0$ and $m \in \mathbb{N}_0$. Then there exists a unique generalized solution of the problem (1)-(3), it belongs to the space $H^{m+1}_{\sigma_0}(\mathbb{R}_+; H^1(\Omega))$ and for any $\sigma \geq \sigma_0$ such an inequality holds*

$$||u||_{H^{m+1}_{\sigma}(\mathbb{R}_+; H^1(\Omega))} \leq C||g||_{H^{m+4}_{\sigma}(\mathbb{R}_+; H^{1/2}(\Gamma))}, \tag{28}$$

where $C > 0$ is a constant that is not dependent on g.

In addition, the generalized solution of the problem (1)-(3) can be represented as a sum of series (22), that is convergent in the space $L^2_{\sigma_0}(\mathbb{R}_+; H^1(\Omega))$, which coefficients $\mathbf{u}$ are defined by formula (23), where the sequence $\boldsymbol{\mu} \in l^2(H^{-1/2}(\Gamma))$ is a solution of the system of the BIEs (26) with $\mathbf{g} := \mathcal{L}g$.

Note that the assumption about the function g in the proposition 2 guarantees the applicability of the LT at all stages of constructing of the numerical solution to the problem (1)-(3) without any additional assumption about relation between parameters m and σ_0. On theoretical aspects of generalized solutions to such problems in other functional spaces, see, for example, in [21].

3. System of the convolutional type and its solution

We can also obtain both the representation (22) of the generalized solution of the problem (1)-(3) and the system of the BIEs (26) in another way. For this we use such property of the LT for the derivatives of the function $f \in H^2_{\sigma}(\mathbb{R}_+; X)$:

$$\mathcal{L}_k\left(\frac{\partial^2 f(t)}{\partial t^2}\right) = \sigma^2 \sum_{l=0}^{k} (k-l+1)\mathcal{L}_l\big(f(t)\big), \quad k \in \mathbb{N}_0. \tag{29}$$

By applying the LT to the wave equation (1) directly and using (29), in Ω we obtain the following infinite triangular system of elliptic equations

$$\begin{cases} Pu_0 = 0, \\ c_1u_0 + Pu_1 = 0, \\ c_2u_0 + c_1u_1 + Pu_2 = 0, \\ \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ c_ku_0 + c_{k-1}u_1 + ... + Pu_k = 0, \\ \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \end{cases} \tag{30}$$

where $u_k := \mathcal{L}_k u$, $k \in \mathbb{N}_0$, are the unknown functions and $P := c_0 I - \Delta$, $c_k := (k+1)\sigma^2$, I is the identity operator. Henceforth we denote $\mathbf{u} := (u_0, u_1, ...)^\top$ and $\mathbf{G}$ the infinite triangular matrix in the left hand side of (30). This allows us to rewrite the system in form

$$\mathbf{Gu}{=}\mathbf{0} \text{ in } \Omega. \tag{31}$$

By the LT we obtain from the condition (3) a sequence of boundary conditions regarding the unknown functions

$$\gamma_0 \mathbf{u} = \mathbf{g} := \mathcal{L} g \text{ on } \Gamma. \tag{32}$$

Theorem 1. *Let the given function g satisfies the condition of the proposition 2, that is, $g \in H^{m+4}_{\sigma_0}(\mathbb{R}_+; H^{1/2}(\Gamma))$ for some $\sigma_0 > 0$ and $m \in \mathbb{N}_0$. Then the unique generalized solution $u \in H^{m+1}_{\sigma_0}(\mathbb{R}_+; H^1(\Omega))$ of the problem (1)-(3) can be represented by the solution $\mathbf{u} := (u_0, u_1, ...)^\top$ of the boundary value problem (31), (32) as the sum of a series*

$$u(x,t) = \sum_{j=0}^{\infty} u_j(x)\, L_j(\sigma t), \quad (x,t) \in Q. \tag{33}$$

Proof. Let us consider a top part $\mathbf{G}^k \mathbf{u}^k = \mathbf{0}$ of the system (31) for any fixed $k \in \mathbb{N}_0$, which consists from the $k+1$ equations. According to the [30, Lemma 2] its any solution $\mathbf{u}^k := (u_0, u_1, ..., u_k)^\top$ can be represented in Ω by the formula

$$u_j(x) = \sum_{i=0}^{j} \big\langle \mu_i(\cdot), e_{j-i}(x - \cdot) \big\rangle_\Gamma, \ x \in \Omega, \ j \in \mathbb{N}_0, \tag{34}$$

where $\mu_j, \ j \in \mathbb{N}_0$, are some elements of the space $H^{-1/2}(\Gamma)$ and functions $e_j, \ j \in \mathbb{N}_0$, may be expressed through a fundamental solution $\mathbf{E} := (E_0, E_1, ...)^\top$ of the operator $\mathbf{G}$ in form

$$e_0 := E_0, \ \ e_j := E_j - E_{j-1}, \ j \in \mathbb{N}. \tag{35}$$

In addition, if the sequence $\boldsymbol{\mu}^k := \big(\mu_0, \mu_1, ..., \mu_k\big)^\top$ is obtained as a solution of the system of BIEs

$$\sum_{i=0}^{j} \big\langle \mu_i(\cdot), e_{j-i}(x - \cdot) \big\rangle_\Gamma = g_j, \ x \in \Gamma, \ j \in \overline{0,k}, \tag{36}$$

then the sequence $\mathbf{u}^k$ will be the solution of suitable Dirichlet problem for the system $\mathbf{G}^k \mathbf{u}^k = \mathbf{0}$.

Notice that (35) may be reduced to form (25) [31, Theorem 1]. Therefore, the formula (34) coincides with the representation of the Fourier-Laguerre coefficients of the retarded potential (4) and BIEs in the system (36) are the same as in the infinite system (26). So sequence $\boldsymbol{\mu} := \big(\mu_0, \mu_1, ...\big)^\top$ coincides with LT of the solution μ of the TDBIE (6) and, as a consequence, the solution $\mathbf{u}$ of the problem (31), (32) coincides with LT of the solution u of the problem (1)-(3). As a conclusion from the Proposition 2 we have that $\boldsymbol{\mu} \in l^2(H^{-1/2}(\Gamma))$ and $\mathbf{u} \in l^2(H^1(\Omega))$.

Using the notation (21), in the case $\mu_i \in L^2_\sigma(\Gamma)$ we can rewrite the formula (34)

$$\begin{aligned} u_j(x) &= \sum_{i=0}^{j} \int\limits_{\Gamma} \mu_i(y) e_{j-i}(x-y) d\Gamma_y = \\ &= \int\limits_{\Gamma} \frac{e^{-\sigma|x-y|}}{4\pi|x-y|} \sum_{i=0}^{j} \mu_i(y) \zeta_{j-i}(\sigma|x-y|) d\Gamma_y. \end{aligned} \tag{37}$$

By substituting the expression (37) into the partial sum

$$\widetilde{u}^k(x,t) := \sum_{j=0}^{k} u_j(x) L_j(\sigma t), \quad (x,t) \in Q, \tag{38}$$

and taking the external sum into the integral over Γ we obtain

$$\widetilde{u}^k(x,t) = \int\limits_{\Gamma} \frac{e^{-\sigma|x-y|}}{4\pi|x-y|} \sum_{j=0}^{k} \sum_{i=0}^{j} \mu_i(y) \zeta_{j-i}(\sigma|x-y|) L_j(\sigma t) d\Gamma_y, \quad (x,t) \in Q. \tag{39}$$

Taking into account, that $\boldsymbol{\mu} \in l^2(H^{-1/2}(\Gamma))$ and formula (20) holds for this sequence, putting $k \to \infty$ we finally get

$$u(x,t) = \int\limits_{\Gamma} \frac{1}{4\pi|x-y|} \mu(y, t-|x-y|) d\Gamma_y, \quad (x,t) \in Q, \tag{40}$$

where $\mu = \mathcal{L}^{-1}\boldsymbol{\mu}$. Since μ is the solution of the TDBIE (6), the retarded potential (40) coincides with potential (4). Therefore, (40) is the solution of the problem (1)-(3). □

Taking into account that the system (26) is triangular we rewrite it as a sequence of BIEs

$$\begin{cases} (V_0\mu_0)(x) = g_0(x), \\ (V_0\mu_1)(x) = \tilde{g}_1(x), \\ \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \\ (V_0\mu_k)(x) = \tilde{g}_k(x), \quad k \in \mathbb{N}, \quad x \in \Gamma, \\ \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \end{cases} \tag{41}$$

with recurrent expressions in right-hand sides

$$\tilde{g}_k(x) := g_k(x) - \sum_{i=0}^{k-1} (V_{k-i}\mu_i)(x), \quad k \in \mathbb{N}. \tag{42}$$

Since the boundary operator V_0 is $H^{-1/2}(\Gamma)$-elliptical [6,17], for arbitrary fixed $k \in \mathbb{N}_0$ the $k-$th equation in (41) with $g_k \in H^{1/2}(\Gamma)$ has a unique solution $\mu_k \in H^{-1/2}(\Gamma)$. We can choose (by some criteria) the value of parameter N and find from (41) the first components for the sequence $\boldsymbol{\mu}^N := \big(\mu_0, \mu_1, ..., \mu_N, 0, 0, ...\big)^\top$.

Using it for calculation a sequence $\mathbf{u}^N := \big(u_0, u_1, ..., u_N, 0, 0, ...\big)^\top$ by the formula

$$\mathbf{u}^N(x) = \big(\mathbf{S} \underset{H^{1/2}(\Gamma)}{\circ} \boldsymbol{\mu}^N\big)(x), \quad x \in \Omega, \tag{43}$$

we obtain an approximate solution $\widetilde{u}^N(x,t)$ of the problem (1)-(3) as a partial sum (38) of the expansion (22) of the exact solution $u(x,t)$.

4. Problems in the domain with an inclusion

Reducing the IBVP (1)-(3) to the BVP (31), (32) allows us to solve it by numerical approaches, which have been successfully used for solution of the elliptic problems. In particular, it concerns the use of surface potentials, which are based on Green's function [8] for specific domain Ω_0 instead of the fundamental solution (25) for operator $\mathbf{G}$ in $\mathbb{R}^3$. Suppose Γ_0 is a Lipschitz boundary of Ω_0.

Definition 6 ([31]). Let $\mathbf{N}(x,y) := (N_0(x,y), N_1(x,y), ...)^\top$, $(x,y) \in \overline{\Omega}_0 \times \Omega_0$ be a solution of the equation

$$\mathbf{G}\mathbf{u} = \bar{\boldsymbol{\delta}}_y \text{ in } \big(\mathcal{D}'(\Omega_0)\big)^\infty, \tag{44}$$

where $\bar{\boldsymbol{\delta}}_y := (\,\delta(\cdot - y),\ 0,\ 0,\ ...\,)^\top$. We say that $\mathbf{N}$ is *Green's function for the Dirichlet problem for the system* (31) *in the domain* Ω_0 if all its components vanish for $(x,y) \in \Gamma_0 \times \Omega_0$.

Building the Green's function for the domain with arbitrary geometry isn't a simple task in general. But for domains with a certain type of symmetry it can be built analytically by the reflection method [31]. Without loss of generality we present here the Green's function for the Dirichlet problem in case of the half-space $\Omega_0 = \mathbb{R}^2 \times \mathbb{R}_+$:

$$N_k(x,y) = e_k(x-y) - e_k(x-y^*),\ k \in \mathbb{N}_0, \tag{45}$$

where y^* is a point symmetric to the point y in regards to the plane Γ_0 and functions e_k are defined by (25).

Let us denote the unit exterior normal vector to the surface Γ_0 as $\boldsymbol{\nu}$. Consider a sequence $\mathbf{D}$ which consists of operators $D_k : H^{1/2}(\Gamma_0) \to H^1(\Omega))$, $k \in \mathbb{N}_0$, that act on an arbitrary function $\xi \in H^{1/2}(\Gamma_0)$ according to the rule

$$(D_k\xi)(x) := \int\limits_{\Gamma_0} \xi(y)\, \partial_\nu N_k(x,y)\, d\Gamma_y, \quad x \in \Omega_0, \tag{46}$$

where ∂_ν is the notation of the normal derivative. If $\boldsymbol{\lambda} \in l^2(H^{1/2}(\Gamma_0))$ is an arbitrary sequence then a sequence

$$\mathbf{u}(x) := -\big(\mathbf{D} \underset{H^1(\Omega)}{\circ} \boldsymbol{\lambda}\big)(x), \quad x \in \Omega_0, \tag{47}$$

satisfies the system (31) [31].

Let bounded domain Ω^- with a Lipschitz boundary Γ is an inclusion in the domain Ω_0 ($\Gamma_0 \cap \Gamma = \oslash$) and $\Omega := \Omega_0 \setminus \overline{\Omega^-}$. For an arbitrary function

$\mu \in L^2_\sigma(\mathbb{R}_+; H^{-1/2}(\Gamma))$ let us consider q-convolution

$$\mathbf{u}(x) := \big(\widetilde{\mathbf{S}} \underset{H^1(\Omega)}{\circ} \boldsymbol{\mu}\big)(x), \quad x \in \Omega, \tag{48}$$

of sequences $\boldsymbol{\mu} := \mathcal{L}\mu$ and $\widetilde{\mathbf{S}} := (\widetilde{S}_0, \widetilde{S}_1, ...)^\top$, where operators $\widetilde{S}_k : H^{-1/2}(\Gamma) \to H^1(\Omega))$, $k \in \mathbb{N}_0$, act on an arbitrary $\xi \in L^2(\Gamma)$ according to the rule

$$(\widetilde{S}_k \xi)(x) := \int\limits_\Gamma \xi(y)\, N_k(x,y)\, d\Gamma_y, \quad x \in \Omega. \tag{49}$$

For $\xi \in H^{-1/2}(\Gamma)$ one can extend the expression (49) to the $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$ duality product $(\widetilde{S}_k \xi)(x) = \big\langle \xi(\cdot), N_k(x - \cdot)\big\rangle_\Gamma$ with $x \in \Omega$. It is easy to see that for arbitrary functions $\mu \in L^2_\sigma(\mathbb{R}_+; H^{-1/2}(\Gamma))$ and $\lambda \in L^2_\sigma(\mathbb{R}_+; H^{1/2}(\Gamma_0))$ a combination of the sequences

$$\mathbf{u}(x) := \big(\widetilde{\mathbf{S}} \underset{H^1(\Omega)}{\circ} \boldsymbol{\mu}\big)(x) - \big(\mathbf{D} \underset{H^1(\Omega)}{\circ} \boldsymbol{\lambda}\big)(x), \; x \in \Omega, \tag{50}$$

satisfies the system (31) in Ω and the boundary condition $\gamma_0 \mathbf{u} = \boldsymbol{\lambda}$ on Γ_0.

Suppose u satisfies the wave equation (1) and initial conditions (2) in Ω and traces $\gamma_{0,0} u = \lambda$ and $\gamma_{0,1} u = g$ are given on the cylinders $\Sigma_0 := \Gamma_0 \times \mathbb{R}_+$ and $\Sigma = \Gamma \times \mathbb{R}_+$ respectively. Then unknown sequence $\boldsymbol{\mu}$ for the representation (50) can be obtained from the system of BIEs

$$\widetilde{\mathbf{V}} \underset{H^{1/2}(\Gamma)}{\circ} \boldsymbol{\mu} = \mathbf{g} + \gamma_{0,1}\big(\mathbf{D} \underset{H^1(\Omega)}{\circ} \boldsymbol{\lambda}\big) \quad \text{on } \Gamma, \tag{51}$$

where $\mathbf{g} := \mathcal{L} g$ and the components of the sequence $\widetilde{\mathbf{V}}$ are boundary operators $\widetilde{V}_k := \gamma_{0,1} \circ \widetilde{S}_k$, $\widetilde{V}_k : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$, $k \in \mathbb{N}_0$. Note that the resulting system can be reduced to the sequence of BIEs similar to (41) and has only one solution.

5. Fast BEM and results of numerical experiments

Both (26) and (51) systems are triangular so one can solve their equations sequentially. For this we use Galerkin-BEM and it fast modification [16,36].

Let $\Gamma^M = \bigcup_{l=1}^M \overline{\tau}_l$ be some approximation of the boundary Γ by triangular boundary elements $\{\tau_l\}_{l=1}^M$ and $\{\varphi_l^0\}_{l=1}^M$ be a set of linearly-independent on Γ^M piece-wise constant functions

$$\varphi_l^0(x) = \begin{cases} 1, & x \in \tau_l, \\ 0, & x \notin \tau_l. \end{cases} \tag{52}$$

Treating a value $h := \max\limits_{l=\overline{1,M}} \big(\int_{\tau_l} \mathrm{d}s\big)^{1/2}$ as a parameter of the spatial approximation, we will consider a finite-dimensional space $S_h^0(\Gamma) := \operatorname{span}\{\varphi_l^0\}_{l=1}^M$ and represent a numerical solution of the system (41) by a sequence $\boldsymbol{\mu}^{N,h} :=$

$\left(\mu_0^h, \mu_1^h, ..., \mu_N^h, 0, 0, ...\right)^\top$ which components are linear combinations of piecewise constant functions

$$\mu_k^h = \sum_{l=1}^{M} \mu_{k,l}^h \varphi_l^0 \ \in S_h^0(\Gamma), \ \ k \in \mathbb{N}_0. \tag{53}$$

Here $\left\{\mu_{k,l}^h\right\}_{l=1}^{M} =: \boldsymbol{\mu}_k^h \in \mathbb{R}^M$ is a vector of unknown coefficients which can be found from the following system of linear algebraic equations

$$\mathbf{V}_0^h \boldsymbol{\mu}_k^h = \mathbf{g}_k^h - \sum_{j=0}^{k-1} \mathbf{V}_{k-j}^h \boldsymbol{\mu}_j^h, \ \ k \in \mathbb{N}_0, \tag{54}$$

where $g_k^h[i] = \int_{\tau_i} g_k(x) ds_x, \quad i = \overline{1, M}$, and elements of the matrix $\mathbf{V}_p^h$ have following form

$$V_p^h[i,l] = \int_{\tau_i} \int_{\tau_l} e_p(x-y) ds_y ds_x, \ \ i,l = \overline{1,M}, \ p \in \mathbb{N}_0. \tag{55}$$

Notice, that for any $k \geq 1$ the components $\mu_0, \mu_1, ..., \mu_{k-1}$, obtained from BIE (41) on previous steps, are included into the expression in the right-hand side of the current equation. The evaluation of the surfaces integrals (55) has been discussed in [32].

We interpret sequences

$$\boldsymbol{\mu}^{N,h} := \left(\mu_0^h, \mu_1^h, ..., \mu_N^h, 0, 0, ...\right)^\top$$

and

$$\mathbf{u}^{N,h} := \left(u_0^h, u_1^h, ..., u_N^h, 0, 0, ...\right)^\top$$

with some fixed value of the parameter N as numerical solutions of the systems of BIEs (26) and the BVP (31)-(32), respectively. As well, a partial sum $\widetilde{u}^{N,h}(x,t) := \sum\limits_{j=0}^{N} u_j^h(x) L_j(\sigma t)$ we use as a numerical solution of the problem (1)-(3).

Let us assess the accuracy of the proposed method. Taking into account an obvious inequality $\|u - \widetilde{u}^{N,h}\|_{H_\sigma^1(\mathbb{R}_+;H^1(\Omega))} \leq \|u - \widetilde{u}^N\|_{H_\sigma^1(\mathbb{R}_+;H^1(\Omega))} + \|\widetilde{u}^N - \widetilde{u}^{N,h}\|_{H_\sigma^1(\mathbb{R}_+;H^1(\Omega))}$, in this paper we restrict ourselves to examining the posteriori error of the numerical solution, which corresponds to the second term in the right hand part of this inequality. An asymptotic error of the numerical solution in this case has been investigated in [22].

In the following we demonstrate numerical solutions of some model problems for the wave equation in the domain $\Omega = \mathbb{R}^3 \backslash \overline{\Omega^-}$, where $\Omega^- = (-1,1) \times (-1,1) \times (-1,1)$. For generating boundary values we use a spherical impulse represented by the formula

$$w(x,t) := |x|^{-1} w^*(t - |x| + 1) \vartheta(t - |x| + 1), \ (x,t) \in \mathbb{R}^3 \setminus \{0\} \times \mathbb{R}_0, \tag{56}$$

with a cubic B-spline w^* and the Heaviside step function $\vartheta(t)$. Notice that the function w satisfies (1) and (2).

Example 1. We consider the problem (1)-(3) in $\Omega \times \mathbb{R}_+$ with the given trace data $g = w$ on Σ and analyze accuracy and convergence of numerical solutions u_k^h and $\widetilde{u}^{N,h}$ on the sequence of discretization Γ^M with increasing M and with $N = 20$.

TABL. 1. Convergence analysis of u_0^h, u_{10}^h and $\widetilde{u}^{N,h}$ for Example 1 with $\sigma = 4$, $N = 20$ and increasing M

M	$u_0^h(x)$			u_{10}^h			$\widetilde{u}^{N,h}$		
	δ_0^h	eoc_0	ϵ_0^h	δ_{10}^h	eoc_{10}	ϵ_{10}^h	$\widetilde{\delta}^{N,h}$	$\widetilde{eoc}^{N,h}$	$\widetilde{\epsilon}^{N,h}$
108	$1.92{\cdot}10^{-4}$		3.24	$2.92{\cdot}10^{-3}$		22.21	$2.40{\cdot}10^{-2}$		4.66
300	$7.01{\cdot}10^{-5}$	2.03	1.18	$8.46{\cdot}10^{-4}$	2.43	6.43	$8.11{\cdot}10^{-3}$	2.13	1.57
768	$3.22{\cdot}10^{-5}$	2.42	0.54	$2.97{\cdot}10^{-4}$	2.23	2.26	$3.09{\cdot}10^{-3}$	2.05	0.60
1452	$1.83{\cdot}10^{-5}$	2.24	0.31	$1.49{\cdot}10^{-4}$	2.16	1.14	$1.62{\cdot}10^{-3}$	2.03	0.31
1728	$1.55{\cdot}10^{-5}$	2.16	0.26	$1.24{\cdot}10^{-4}$	2.14	0.94	$1.36{\cdot}10^{-3}$	2.02	0.26
2700	$1.02{\cdot}10^{-5}$	2.14	0.17	$7.72{\cdot}10^{-5}$	2.12	0.59	$8.63{\cdot}10^{-4}$	2.03	0.17
4800	$5.93{\cdot}10^{-6}$	2.11	0.10	$4.22{\cdot}10^{-5}$	2.10	0.32	$4.83{\cdot}10^{-4}$	2.02	0.09

At first we consider the impact of the parameter h on the approximation error of numerical solutions u_k^h, $k \in \overline{0,N}$, and $\widetilde{u}^{N,h}$ with some fixed value of the parameter N. For this we compute values $\delta_k^h := ||u_k^h - u_k||_{L^2(\Omega_{a,b})}$ and $\epsilon_k^h := \delta_k^h / ||u_k||_{L^2(\Omega_{a,b})} * 100\ \%$, and also values $\widetilde{\delta}^{N,h} := ||\widetilde{u}^{N,h} - \widetilde{u}^N||_{L^2_\sigma(\mathbb{R}_+;L^2(\Omega_{a,b}))}$ and $\widetilde{\epsilon}^{N,h} := \widetilde{\delta}^{N,h} / ||\widetilde{u}^N||_{L^2_\sigma(\mathbb{R}_+;L^2(\Omega_{a,b}))} * 100\ \%$, where $(a,b) =: \Omega_{(a,b)}$ is a spatial interval from which observation points are taken. Notice that we provide estimates in the norm of such Lebesgue space with aim to simplify calculations in the unbounded exterior domain Ω. Using a sequence of finite-dimensional spaces $S_h^0(\Gamma)$ with decreasing h for both kinds of numerical solutions we evaluate estimated orders of convergence [36] $eoc_k := \ln(\delta_k^{h_{j-1}} / \delta_k^{h_j}) / \ln(h_{j-1}/h_j)$, $k \in \overline{0,N}$, and $\widetilde{eoc}^{N,h} := \ln(\widetilde{\delta}^{N,h_{j-1}} / \widetilde{\delta}^{N,h_j}) / \ln(h_{j-1}/h_j)$, where h_{j-1} and h_j are consequent values of the parameter h.

Computed in $\Omega_{(a,b)}$ with $a = (1.2, 0, 0)$ and $b = (10, 0, 0)$, some results of the series of numerical experiments are given in Table 1. They highlight that $eoc \approx 2$ for both numerical solutions u_k^h and $\widetilde{u}^{N,h}$.

Now we assume that the cube Ω^- is included in the half space $\Omega_0 = \mathbb{R}^2 \times (-2, \infty)$ and $\Omega = \Omega_0 \setminus \overline{\Omega^-}$. For generating boundary functions in this case we use a function $\widehat{w}(x,t) := w(x,t) - w(x^*,t)$, where x^* is a point symmetric to the point x with respect to the plane $\Gamma_0 = \{(x_1, x_2, x_3) \mid x_3 = -2\}$. It is obvious that function $\widehat{w}$ satisfies (1) and (2) and $\widehat{w}(x,t) \equiv 0$ on Γ_0.

Example 2. We consider the problem (1)-(3) in $\Omega \times \mathbb{R}_+$ with traces $\gamma_{0,0} u = \lambda \equiv 0$ and $\gamma_{0,1} u = g = \widehat{w}$ given on the cylinders $\Sigma_0 := \Gamma_0 \times \mathbb{R}_+$ and $\Sigma = \Gamma \times \mathbb{R}_+$ respectively, and analyze accuracy and convergence of numerical solutions u_k^h and $\widetilde{u}^{N,h}$ on the sequence of discretization Γ^M with increasing M and with $N = 20$.

We solve this problem by modified BEM using the representation (50) based on Green's functions for the Dirichlet problem for the system (31) in the domain Ω_0. In this approach after discretization of BIEs we obtain matrices $\widetilde{\mathbf{V}}_k^h$ similar to the $\mathbf{V}_k^h$, $k \in \mathbb{N}_0$. Results of the numerical experiment are plotted in Figure 1.

As we can see from the Table 2 numerical solutions, obtained in this approach, have the same accuracy and the convergence order as in the previous example. Notice that some complication of the method due to the use of Green's functions does not lead to significant increase of computational resources for solving the problem in the domain with inclusion. The fact that we have avoided solving BIEs on the unbounded surface Γ_0 is an advantage of the modified BEM in solving such problems.

TABL. 2. Convergence analysis of u_0^h, u_{10}^h and $\widetilde{u}^{N,h}$ for Example 2 with $\sigma = 4$, $N = 20$ and increasing M

M	$u_0^h(x)$			u_{10}^h			$\widetilde{u}^{N,h}$		
	δ_0^h	eoc_0	ϵ_0^h	δ_{10}^h	eoc_{10}	ϵ_{10}^h	$\widetilde{\delta}^{N,h}$	$\widetilde{eoc}^{N,h}$	$\widetilde{\epsilon}^{N,h}$
108	$8.58{\cdot}10^{-5}$		3.24	$1.36{\cdot}10^{-3}$		7.59	$1.78{\cdot}10^{-2}$		3.35
300	$3.14{\cdot}10^{-5}$	2.03	1.19	$3.33{\cdot}10^{-4}$	2.76	1.85	$4.96{\cdot}10^{-3}$	2.50	0.94
768	$1.44{\cdot}10^{-5}$	2.42	0.55	$9.97{\cdot}10^{-5}$	2.56	0.56	$1.77{\cdot}10^{-3}$	2.20	0.33
1452	$8.14{\cdot}10^{-6}$	2.23	0.31	$4.64{\cdot}10^{-5}$	2.40	0.26	$9.06{\cdot}10^{-4}$	2.10	0.17
1728	$6.93{\cdot}10^{-6}$	2.16	0.26	$3.79{\cdot}10^{-5}$	2.31	0.21	$7.57{\cdot}10^{-4}$	2.05	0.14
2700	$4.56{\cdot}10^{-6}$	2.13	0.17	$2.27{\cdot}10^{-5}$	2.29	0.13	$4.79{\cdot}10^{-4}$	2.06	0.09

We now wish to notice that matrices $\mathbf{V}_k^h$ and $\widetilde{\mathbf{V}}_k^h$, $k \in \overline{0,N}$, which arise after discretization of boundary operators in equations (26) and (51), are fully populated and can reach large sizes. So for their calculation we apply the Fast BEM which based on adaptive cross approximation (ACA) of these matrices [3, 12]. Because this approach is universal in relation to the function in the kernel of boundary operators, an efficient algorithm can be constructed for calculating all the above matrices.

It can be checked that functions in the sequence $\mathbf{e}(x-y) = \big(e_0(x-y), e_1(x-y), ...e_k(x-y), ...\big)^\top$ are asymptotically smooth [3, Definition 3.2.]. This ensures that for each of the matrices $\mathbf{V}_k^h$ ACA algorithm admits admissible partitions into blocks that can be approximated by the product of matrices of smaller rank. For example, if some block $\mathbf{A} \in \mathbb{R}^{m\times n}$ in $\mathbf{V}_k^h$ is admissible it can be approximated with arbitrary small error ε in Frobenius norm by the matrix $\mathbf{S}_r := \mathbf{Q}\mathbf{T}^\top$, where $\mathbf{Q} \in \mathbb{R}^{m\times r}$ and $\mathbf{T} \in \mathbb{R}^{n\times r}$ are matrices of rank $r \leq \min(m,n)$. To do this we have to calculate and store in RAM only a subset of elements of the block $\mathbf{A}$ [3, Chapter 3].

In order to demonstrate efficiency of the ACA we apply Fast BEM to the problem which we have considered in the Example 1. As we can see from the Figure 2, memory consumption for storing data of the approximated matrix $\mathbf{V}_0^h$ depends on the parameter M almost linearly. By contrast, we need to store

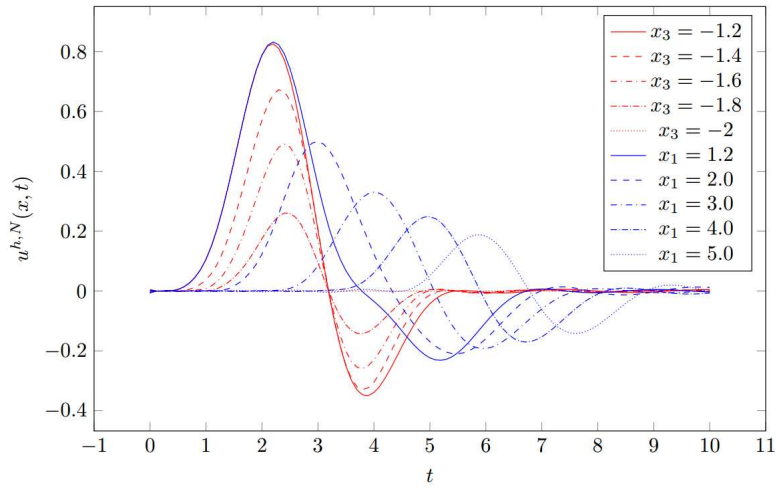


FIG. 1. Numerical solution of the problem in Example 2 in two sets of the observation points $\{(x_1, 0, 0)\}$ and $\{(0, 0, x_3)\}$

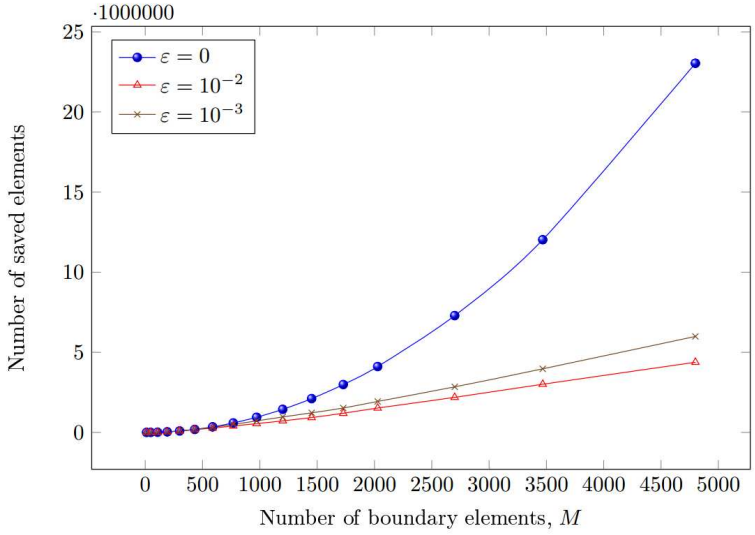


FIG. 2. Memory consumption for storing data of the matrix $\mathbf{V}_0^h$ for the Fast BEM ($\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$) and for the ordinary BEM ($\varepsilon = 0$)

M^2 elements of $\mathbf{V}_0^h$ using ordinary BEM. The same dependency concerns the time needed for calculating data of $\mathbf{V}_0^h$ by the fast and the ordinary BEM.

Note that according to the ACA algorithm admissible blocks are allocated outside of the main diagonal of the matrix. So their approximation doesn't require high accuracy. On Figure 3 we demonstrate the error of the numerical

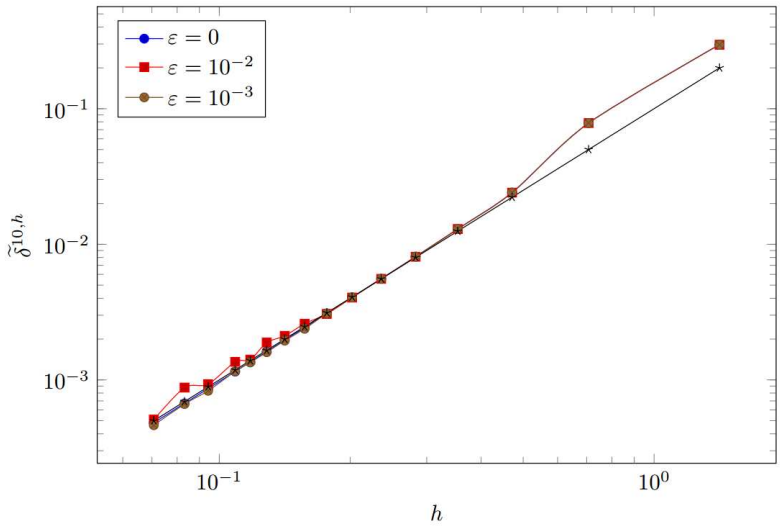


FIG. 3. Error $\widetilde{\delta}^{10,h}$ of numerical solutions for Example 1, which was obtained by the Fast BEM ($\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-3}$) and by the ordinary BEM ($\varepsilon = 0$)

solutions for Example 1, which were obtained by the Fast BEM with approximation of admissible blocks in matrices $\mathbf{V}_k^h$ with some fixed values of the error ε. As we can see, the numerical solution in case of $\varepsilon = 10^{-3}$ has almost the same error $\widetilde{\delta}^{N,h}$ as in case of the application the ordinary BEM, when all elements of matrices $\mathbf{V}_k^h$ were calculated (on the figure we denote this solution by $\varepsilon = 0$).

6. CONCLUSIONS

We have described two approaches based on the Laguerre transform in the time domain, that require the solution of a sequence of boundary integral equations to obtain an approximate solution of the Dirichlet problem for the wave equation. After an additional justification for such transform, we have shown the application of the boundary elements method for solving integral equations in the Laguerre domain and derived a representation of the approximate solution of the wave equation.

In solving evolutional problems the coupling of the LT and the BEM makes it possible to use other techniques, that have been developed for elliptical problems. In particular, we have modified this method for solving Dirichlet problem in the domain with an inclusion, using Green's functions for the representation of the solution. Also we have implemented the Fast BEM using adaptive cross approximation for reducing both the storage and computational costs.

Finally, we can point out that in this article we have confined ourselves to considering a problem with a Dirichlet boundary condition in order to simplify the presentation. For other boundary conditions the approaches considered above will lead to other boundary integral equations that will need to be solved

by another implementation of the BEM. We also remark that the Laguerre transform can be combined with other suitable methods. For example, for solving more general second-order hyperbolic equations, which coefficients are variable in the space domain, the Laguerre transform can be similarly combined with the finite elements method.

Bibliography

1. Bamberger A. Formulation variationnelle pour le calcul de la diffraction d'une onde acoustique par une surface rigide / A. Bamberger, T. Ha Duong // Math. Methods Appl. Sci. – 1986. – № 8. – P. 598-608.
2. Banjai L. Fully discrete Kirchhoff formulas with CQ-BEM / L. Banjai, A.R. Laliena, F.-J. Sayas // IMA J. Numer. Anal. – 2015. – № 35. – P. 859-884.
3. Bebendorf M. Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems / M. Bebendorf. – Springer Science & Business Media, 2008.
4. Chapko R. Numerical solution of the Dirichlet initial boundary value problem for the heat equation in exterior 3-dimensional domains using integral equations / R. Chapko, B. Johansson // Journal of Engineering Mathematics, Springer. – 2016. – DOI 10.1007/s10665-016-9858-6. – 17 P.
5. Chapko R. On the numerical solution of initial boundary value problems by the Laguerre transformation and boundary integral equations / R. Chapko, R. Kress // Integral and Integrodifferential Equations: Theory, Methods and Applications. Series in Mathematical Analysis and Applications. – 2000. – № 2. – P. 55-69.
6. Costabel M. Boundary integral operators on Lipschitz domains: elementary results / M. Costabel // SIAM Journal on Mathematical Analysis. – 1988. – № 19. – P. 613-626.
7. Costabel M. Time-dependent problems with the boundary integral equation method / M. Costabel // Encyclopedia of Computational Mechanics. – Chichester, 2004. – John Wiley and Sons, Ltd. – P. 703-721.
8. Dautray R. Mathematical analysis and numerical methods for science and technology. Volume 4 Integral Equations and Numerical Methods / R. Dautray, J.L. Lions. – Berlin: Springer-Verlag, 1992. – 493 p.
9. Dautray R. Mathematical analysis and numerical methods for science and technology. Volume 5 Evolution problems I / R. Dautray, J.L. Lions. – Berlin: Springer-Verlag, 1992. – 742 p.
10. Davies P.J. Numerical approximation of first kind Volterra convolution integral equations with discontinuous kernels / Penny J. Davies, Dugald B. Duncan // J. Integral Equations Applications. – 2017. – № 29 (1). – P. 41-73.
11. Dominguez V. Recent progress in time domain boundary integral equations / V. Dominguez, N. Salles, F.-G. Sayas // Journal of Integral Equations and Applications. – 2017. – № 29 (1). – P. 1-4.
12. Hackbusch W. Hierarchical Matrices: Algorithms and Analysis / W. Hackbusch // Springer Series in Computational Mathematics. – 2015. – Vol. 49. – 511 p.
13. Halazyuk V.A. Metod integral'nykh rivnyan' u nestatsionarnykh zadachakh dyfraktsiyi / V.A. Halazyuk, Y.V. Lyudkevych, A.O. Muzychuk. – L'viv. un-t. – 1984. – Dep. v UkrNIINTI, # 601 Uk-85 Dep. (in Ukrainian).
14. Hassell M. A new and improved analysis of the time domain boundary integral operators for acoustics / M. Hassell, T. Qiu, T. Sanchez-Vizuet, F.-J. Sayas // Department of Mathematical Sciences. – 2015. – P. 1-22.
15. Ha-Duong T. On retarded potential boundary integral equations and their discretisation / T. Ha-Duong. – In P. Davies, D. Duncan, P. Martin, B. Rynne, eds. // Topics in computational wave propagation. Direct and inverse problems. – Berlin: Springer-Verlag, 2003. – P. 301-336.

16. Hsiao G.C. Boundary element methods: foundation and error analysis / G.C. Hsiao, W.L. Wendland. – In E. Stein, R. de Borst, T. Hughes, eds. // Encyclopaedia of Computational Mechanics. – John Wiley and Sons, Ltd., 2004. – P. 339-373.
17. Hsiao G.C. Boundary Integral Equations / G.C. Hsiao, W.L. Wendland // Applied Mathematical Sciences. – Springer-Verlag Berlin Heidelberg, 2008. – 620 p.
18. Keilson J. The bilateral Laguerre transform / J. Keilson, W. Nunn, U. Sumita // Applied Mathematics and Computation. – 1981. – Vol. 8, Issue 2. – P. 137-174.
19. Laliena A.R. A distributional version of Kirchhoff's formula / A.R. Laliena, F.J. Sayas // Journal of Mathematical Analysis and Applications. – 2009. – № 1. – P. 197-208.
20. Laliena A.R. Theoretical aspects of the application of convolution quadrature to scattering of acoustic waves / A.R. Laliena, F.J. Sayas // Numer. Math. – 2009. – № 112 (4). – P. 637-678.
21. Lions J.L. Non-Homogeneous Boundary Value Problems and Applications. Volume 2 / J.L. Lions, E. Magenes. – Berlin: Springer-Verlag, 1972. – 242 p.
22. Litynskyy S.V. Combination of the Laguerre transform and the boundary elements method for the solution of retarded potential integral equations / S.V. Litynskyy, Yu.A. Muzychuk, A.O. Muzychuk // Journal of Mathematical Sciences. – 2017. – Vol. 224, No. 3. – C. 89-101.
23. Litynskyy S. On week solutions of boundary value problems for an infinite system of elliptic equations / S. Litynskyy, Yu. Muzychuk, A. Muzychuk // Visn. L'viv. un-tu. Ser. prykl. matem. ta inform. – 2009. – Issue 15. – P. 52-70. (in Ukrainian).
24. Litynskyy S. Solving of the initial-boundary value problems for the wave equation by the use of retarded potential and the Laguerre transform / S. Litynskyy, A. Muzychuk // Matematychni Studii. – 2015. – № 2 (44). – P. 185-203.
25. Litynskyy S. On the generalized solution of the initial-boundary value problems with Neumann condition for the wave equation by the use of retarded double layer potential and the Laguerre transform / S. Litynskyy, A. Muzychuk // J. of Computational and Appl. Math. – 2016. – No 2 (122). – P. 21-39.
26. Litynskyy S. On the numerical solution of the initial boundary value problem with Neumann condition for the wave equation by the use of the laguerre transform and boundary elements method / S. Litynskyy, Y. Muzychuk, A. Muzychuk // Acta Mechanica et Automatica, The J. of Bialystok Techn. Univ. – 2016. – Vol. 10, No 4. – P. 285-290.
27. Lubich Ch. On the multistep time discretization of linear/newline initial-boundary value problems and their boundary integral equations / C. Lubich // Numerische Mathematik. – 1994. – № 3. – P. 365-389.
28. Lyudkevych Y.V. Numerical solution of boundary problems for wave equation / Y.V. Lyudkevych, A.E. Muzychuk. – L'viv: LDU, 1990. (in Russian).
29. Lyudkevych Y.V. Chysel'ne rozvyazuvannya krayovoi zadachi Dirikhle dlya rivnyannya teploprovidnosti metodamy integral'nykh peretvoren' ta integral'nykh rivnyan' u vypadku nezamknutykh osesymetrychnykh poverkhon' / Y.V. Lyudkevych, R.B. Skaskiv // Visn. L'viv. un-t. Ser. mekh.-mat. – 1989. – № 31. – P. 2-8.
30. Muzychuk Y.A. On variational formulations of inner boundary value problems for infinite systems of elliptic equations of special kind / Y.A. Muzychuk, R.S. Chapko // Matematychni Studii. – 2012. – № 1 (38). – P. 12-34.
31. Muzychuk Yu. On the boundary integral equation method for boundary-value problems for a system of elliptic equations of the special type in partially semi-infinite domains / Yu. Muzychuk, R. Chapko // Reports of the National Academy of Sciences of Ukraine. – 2012. – № 11. – P. 20-27. (in Ukrainian).
32. Muzychuk Yu. On the boundary integral equation method for exterior boundary value problems for infite systems of elliptic equations of special kind / Y. Muzychuk // J. of Computational and Appl. Math. – 2014. – № 2 (116). – P. 96-116.
33. Pasichnyk R.M. Chyslennoe reshenye hranychno-vremennoho intehral'noho uravnenyya tipa volnovoho potentsyala: Intehral'nye uravneniya v prikladnom modelyrovanii / R.M. Pasichnyk // Tezysy dokl. 2-y resp. nauch.-tekhn. konf. – Kyiv, 1986. – P. 175-176.

34. Polozhyy H. Equations of mathematical physics / H. Polozhyy. – M: Nauka. 1964. (in Russian).
35. Sayas F.-J. Retarded potentials and time domain boundary integral equations: a road-map / F.-J. Sayas. – Springer International Publishing, 2016. – 241 p.
36. Steinbach O. Numerical Approximation Methods for Elliptic Boundary Value Problems. Finite and Boundary Elements / O. Steinbach. – New-York: Springer Science, 2008. – 383 p.
37. Vavrychuk V.H. Numerical solution of mixed non-stationary problem of thermal conductivity in partially unbounded domain / V.H. Vavrychuk // Visnyk of the Lviv university. Series of Applied mathematics and informatics. – 2011. – Issue 17. – P. 62-72. (in Ukrainian).

A. R. Hlova, S. V. Litynskyy, Yu. A. Muzychuk, A. O. Muzychuk,
Faculty of Applied Mathematics and Informatics,
Ivan Franko National University of Lviv,
1, Universytets'ka Str., Lviv, 79000, Ukraine.

Received 16.03.2018; revised 02.05.2018

Журнал обчислювальної та прикладної математики | **2018** **№ 2 (128)** | Journal of Numerical & Applied Mathematics

UDC 517.988

LAGRANGE INTERPOLATION FORMULA IN LINEAR SPACES

O. F. Kashpur, V. V. Khlobystov

Резюме. В лінійному нескінченновимірному просторі зі скалярним добутком і в скінченновимірному евклідовому просторі досліджена точність формули Лагранжа на поліномах відповідного степеня.

Abstract. In a linear infinite-dimensional space with scalar product and in a finite-dimensional Euclidean space the accuracy of the Lagrange formula on polynomials of the corresponding degree is investigated.

The problem of polynomial approximation of nonlinear operators is an actual in both the theoretical and in the applied senses. One of the methods of its solution is interpolation. A partial case of this problem is the polynomial interpolation of many-variable functions. It was shown in [1] that for the construction of the unique interpolation polynomial in the Euclidean space E_k it is necessary that the relation (between the n-th degree of the polynomial and the number of nodes m) $m = (n+k)!/n!k!$ be executed. Moreover constructing an n-th degree interpolant in E_k induces some difficulties. In practice, there are cases where the number of interpolation nodes is given less than what is needed to construct of the unique interpolant of the corresponding degree. In [2], it is shown that the number of nodes can be chosen less than dimension of the space of polynomials used for seeking the solution, with the problem will be invariantly solvable and will be have the unique solution with minimum norm generated by a scalar product by the Gaussian measure [3, 7]. We call an interpolation task invariantly solvable if it has a solution at arbitrary values of the function in the nodes.

In [4] interpolation operator polynomials in Hilbert spaces are given. In the article one of these interpolants is considered. It is shown that it is an interpolation Lagrange formula with fundamental functional polynomials in a linear space with a scalar product. This interpolation Lagrange formula (the number of nodes m and the degree of polynomial n are not interconnected) is studied both for the case of an infinite-dimensional linear space and for the case of the finite-dimensional Euclidean space E_k, the conditions for the accuracy of the Lagrange formula on polynomials of the corresponding degree are determined.

Key words. Hilbert space, Euclidean space, operator, interpolation polynom, invariance of solution.

It was shown in [4] that the interpolation operator polynomial of n-th degree for the operator f has the form

$$P_n(x) = \left\langle \overline{f}, \Gamma_m^+ \sum_{p=0}^{n} (x_i, x)^p |_{i=1}^m \right\rangle, \tag{1}$$

where x_i is an interpolation node, $P_n(x_i) = f(x_i) = f_i$, $i = \overline{1,m}$, $\overline{f} = (f_1, f_2, ..., f_m)$, x_i, $x \in H, H$ is the Hilbert space, $f : H \to Y$, Y is a linear space, $f_i \in Y$, Γ_m^+ is the Moore-Penrose pseudo-inverse matrix to the matrix

$$\Gamma_m = \| \sum_{p=0}^{n} (x_i, x_j)^p \|, \; < \cdot, \cdot >= \sum_{i=1}^{m} f_i \alpha_i, \; \alpha_i \in R_1.$$

In [4], in the event of fulfillment of the necessary and sufficient condition for solvability of operator interpolation task, such as

$$A_0 \overline{f} = \overline{0}, \; A_0 = E - \Gamma_m^+ \Gamma = E - \Gamma \Gamma_m^+, \tag{2}$$

A_0 is an idempotent symmetric matrix. Based on (2), we get: if the matrix Γ_m is nonsingular ($\Gamma_m^+ = \Gamma_m^{-1}$), then the problem will be invariantly solvable, that is, the solution will exist for any values of the operator in the nodes.

We denote $\Gamma_m^k = \|(x_i, x_j)^k\|$. In [4] it is shown that in the case of fulfillmenting of the condition

$$rg(\Gamma_m^0 + \Gamma_m^1) + n - 1 \geq m \tag{3}$$

the operator interpolation problem is invariantly solvable.

Consequently, let us consider the case when the problem is invariantly solvable: $\Gamma_m^+ = \Gamma_m^{-1}$, and the formula (1) turn in to the form:

$$P_n(x) = \left\langle \overline{f}, \Gamma_m^{-1} \sum_{p=0}^{n} (x_i, x)^p |_{i=1}^m \right\rangle. \tag{4}$$

In the following, the formula (4) will be rewritten in a different form and we reduce it to the Lagrange formula in a linear space with a scalar product. Let X, Y be linear spaces, X with a scalar product $(\cdot, \cdot)$, $f : X \to Y$, $P_n(x)$ be an interpolation operator polynomial of n-th degree for f with nodes $x_1, x_2, ..., x_m$, $P_n(x_i) = f(x_i) = f_i$, $x, x_i \in X$, $i = \overline{1,m}$, and the nodes x_i are chosen in such a way that the matrix $\|P_{ni}(x_j)\|$ will be nonsingular, where

$$P_{ni}(x) = \sum_{k=0}^{n} L_{ki} x^k, L_{ki} x^k = (x_i, x)^k, \; L_{0i} = 1, \; P_{ni} : X \to R_1, \; i = \overline{1,m}.$$

The invertibility of the matrix for a finite-dimensional Euclidean space is considered in [2] by the choice of independent vectors related with nodes. In the following, we denote: $\overline{P}_n(x) = (P_{n1}(x), P_{n2}(x), ..., P_{nm}(x))$, and by $P_{ni}^{-1}(x_j)$ the elements of the matrix $\|P_{ni}(x_j)\|^{-1}$. According to [4] we get

$$P_n(x) = \langle \overline{f}, \|P_{ni}(x_j)\|^{-1} \overline{P}_n(x) \rangle =$$

$$= \langle \overline{f}, \|P_{ni}^{-1}(x_j)\| \overline{P}_n(x) \rangle =$$

$$= \sum_{i=1}^{m} f_i \sum_{j=1}^{m} P_{ni}^{-1}(x_j) P_{nj}(x) = \sum_{i=1}^{m} f_i l_i(x), \tag{5}$$

where

$$l_i(x) = \sum_{j=1}^{m} P_{ni}^{-1}(x_j) P_{nj}(x),$$

$$l_i(x_k) = \sum_{j=1}^{m} P_{ni}^{-1}(x_j) P_{nj}(x_k) = \delta_{ik}, \tag{6}$$

δ_{ik} is the Kronecker symbol. Since (5), (6), we obtain

$$P_n(x_k) = \sum_{i=1}^{m} f_i l_i(x_k) = f_k = f(x_k), k = \overline{1,m}.$$

Thus, the formula (5) is the Lagrange formula for an interpolation polynomial in a linear space with a scalar product

$$P_n(x) = \sum_{i=1}^{m} f_i l_i(x),\ l_i(x_k) = \delta_{ik},\ i,k = \overline{1,m}, \tag{7}$$

where $l_i(x)$ are fundamental functional Lagrange polynomials of n-th degree, $l_i : X \to R_1$.

Note that the interpolant (7) with the nodes $x_i, i = \overline{1,m}$ is not a unique polynomial in X. Indeed, if $p_n : X \to Y$ is an arbitrary operator polynomial of n-th degree [5], then formula

$$P_n(x) = p_n(x) + \sum_{i=1}^{m} (f_i - p_n(x_i)) l_i(x) \tag{8}$$

defines the set of interpolation operator polynomials of n-th degree for the operator f,

$$P_n(x_k) = p_n(x_k) + \sum_{i=1}^{m} (f_i - p_n(x_i)) l_i(x_k) =$$

$$= p_n(x_k) + \sum_{i=1}^{m} (f_i - p_n(x_i)) \delta_{ik} = f_k = f(x_k), k = \overline{1,m}.$$

In [4] it is proved that the interpolant (7) belonging to the set (8) has a minimal norm generated by a scalar product by the Gaussian measure [3, 7].

It is known that in infinite-dimensional spaces, the finite set of nodes does not guarantee the uniqueness of the interpolant and its invariance with respect to polynomials of the corresponding degree. It was shown in [6-8] that the continuum information used to construct an interpolation polynomial does not provide the uniqueness of the interpolation formula. The so-called "Kergin insterpolation" for many-variable functions and in the Banach space was considerated in the paper [8]. We note, firstly, that the interpolation formulas (see [8])are convergence with the formulas from [6, 7] obtained in the 1960s up to equivalent integral transformations, and secondly, the classical Newton

interpolation formulas for many-variable functions can not be derived from this formulas [9].

It has been known that the expression

$$p_n(x) - \sum_{i=1}^{m} p_n(x_i) l_i(x) \tag{9}$$

does not turn into a zero element of the infinite-dimensional linear space Y [5], that is, the Lagrange formula is not exact on the operator polynomial of the corresponding degree, and when constructing polynomial (5) the numbers m and n are not related.

Example 1. Let's put in (9) $n = 1$, where $p_1 : C[0,1] \to C[0,1], p_1(x) = \int_0^1 K(t,s)x(s)ds, K(t,s)$ is a continuous function on $[0,1] \times [0,1]$. Taking into account the form $l_i(x)$, we obtain that $p_1(x) - \sum_{i=1}^{m} p_1(x_i) l_i(x) \neq 0$. Consequently, in an infinite-dimensional linear space, the Lagrange formula is not exact on polynomials of the corresponding degree.

Let us consider the partial case where X is a finite-dimensional Euclidean space on an example of the space $E_2, f : E_2 \to R_1, u \in E_2, u = (x,y)$, $u_i = (x_i, y_i), i = \overline{1,m}$, where u_i is selected so that the matrix $\| \sum_{p=0}^{n} (x_i x_j + y_i y_j)^p \|$ has to be nonsingular (see [2]). From (5) we get

$$\begin{aligned} P_n(x,y) &= \left(\overline{f}, \left\| \sum_{p=0}^{n} (x_i x_j + y_i y_j)^p \right\|^{-1} \sum_{p=0}^{n} (x x_i + y y_i)^p |_{i=1}^{m} \right) = \\ &= \sum_{i=1}^{m} f_i l_i(x,y). \end{aligned} \tag{10}$$

Then

$$l_i(x,y)|_{i=1}^{m} = \left\| \sum_{p=0}^{n} (x_i x_j + y_i y_j)^p \right\|^{-1} \sum_{p=0}^{n} (x x_i + y y_i)^p |_{i=1}^{m},$$

$$l_i(x_k, y_k) = \delta_{ik},\ i,k = \overline{1,m}.$$

Taking into account (10), we obtain

$$P_n(x_k, y_k) = \sum_{i=1}^{m} f_i l_i(x_k, y_k) = f_k = f(x_k, y_k),\ k = \overline{1,m}$$

and the formula (6) is the interpolation Lagrange formula for $f : E_2 \to R_1$, where $l_i(x,y)$ are the fundamental Lagrange n-th degree polynomials of two variables. Also on the basis of [4] $P_n(x,y)$ is the minimum norm interpolant [3, 7] on the set of n-th degree interpolants of two variables.

In the following, we assume that the number m is given (fixed), and the n-th degree of the interpolation polynomial is chosen from the inequality $m \leq \min p = \overline{p}$, where p is the dimension of the space of n-th degree polynomials in $E_2, p = (n+1)(n+2)/2$ [10].

Example 2. Let $m = 2$, $u_i = (x_i, y_i)$, $i = 1, 2$, $u_1 = (0, 1)$, $u_2 = (1, 0)$. Then

$$m = 2 \leq \min(n+1)(n+2)/2 = \overline{p} = 3,\ n = 1.$$

Let us verify the condition (3) of the invariant solvability of the problem:

$$rg(\Gamma_m^0 + \Gamma_m^1) + n - 1 = 2 + 1 - 1 = 2 \geq m,\ m = 2.$$

Thus, with such a choice of nodes, the problem is invariantly solvable, that is, the matrix Γ_2 has an invertible.

Let us construct the interpolation polynomial. We get

$$\left\| \sum_{p=0}^{1} (u_i, u_j)^p \right\|^{-1} = \frac{1}{3} \begin{Vmatrix} 2 & -1 \\ -1 & 2 \end{Vmatrix},$$

$$\left\| \sum_{p=0}^{1} (u_i, u_j)^p \right\|^{-1} \sum_{p=0}^{1} (u_i, u_j)^p |_{i=1}^{2} = \frac{1}{3} \begin{Vmatrix} 1 - x + 2y \\ 1 + 2x - y \end{Vmatrix},$$

$$l_1(x, y) = \frac{1}{3}(1 - x + 2y),$$

$$l_2(x, y) = \frac{1}{3}(1 + 2x - y),$$

$$l_i(u_j) = \delta_{ij},\ i, j = 1, 2,$$

$$P_1(x, y) = \sum_{i=1}^{2} f_i l_i(x, y).$$

Let $f(x, y) = 1 + 2x + 3y$. Then $f_1 = f(0, 1) = 4$, $f_2 = f(1, 0) = 3$,

$$P_1(x, y) = 4 \cdot \frac{1}{3}(1 - x + 2y) + 3 \cdot \frac{1}{3}(1 + 2x - y) =$$

$$= \frac{1}{3}(7 + 2x + 5y) \neq 1 + 2x + 3y,$$

that is, in the case of $m = 2, \overline{p} = 3, n = 1$, the interpolant $P_1(x, y)$ is not exact on the polynomial of the 1-st degree.

Example 3. Let $m = 3, u_i = (x_i, y_i), i = 1, 2, 3, u_1 = (0, 1), u_2 = (1, 0), u_3 = (0, -1)$. Then

$$m = 3 \leq \min(n+1)(n+2)/2 = \overline{p} = 3,\ n = 1.$$

Check the condition (3):

$$rg(\Gamma_m^0 + \Gamma_m^1) + n - 1 = 3 + 1 - 1 = 3 \geq m,\ m = 3.$$

The condition is fulfilled, hence there exists Γ_3^{-1}. Let us construct the interpolation polynomial. We obtain

$$\left\| \sum_{p=0}^{1} (u_i, u_j)^p \right\|^{-1} = \frac{1}{4} \begin{Vmatrix} 3 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 3 \end{Vmatrix},$$

$$\left\|\sum_{p=0}^{1}(u_i,u_j)^p\right\|^{-1}\sum_{p=0}^{1}(u_i,u)^p|_{i=1}^{3}=\frac{1}{2}\left\|\begin{matrix}1-x+y\\2x\\1-x-y\end{matrix}\right\|,$$

$$P_3(u)=\sum_{i=1}^{3}f_il_i(u),$$

$$l_1(x,y)=1/2(1-x+y),\ l_2(x,y)=x,\ l_3(x,y)=1/2(1-x-y),$$

$$l_i(u_j)=\delta_{ij},i,j=1,2,3.$$

Let $f(u)=1+2x+3y$, then

$$f_1=f(0,1)=4,\ f_2=f(1,0)=3,\ f_3=f(0,-1)=-2.$$

We get

$$P_1(x,y)=4\cdot\frac{1}{2}(1-x+y)+3x-2\cdot\frac{1}{2}(1-x-y)=1+2x+3y,$$

that is, in the case of $m=3,\overline{p}=3,n=1$, the Lagrange interpolant (10) is exact on the first degree polynomial of two variables.

Thus, for the finite-dimensional Euclidean space E_2, the conclusion is as follows: in the case of $m<\overline{p}$ we have the unique Lagrange interpolant with minimum norm, herewith it is not exact on polynomials of the corresponding degree (Example 2). In the paper [2] this interpolant is called underdetermined. If $m=\overline{p}$, then the Lagrange interpolation polynomial is unique and is exact on the polynomial of the corresponding degree [1] (example 3).

Similar considerations and transformations can be made for the Euclidean space E_k, $u\in E_k$, $u=(x_1,x_2,...,x_k)$, where the number of nodes m is given (fixed), and the n-th degree of the interpolant is determined from the condition

$$m\le\min p=\overline{p},p=(n+k)!/n!k!,\ k\ge 2, \tag{11}$$

where p is the dimension of the space of n-th degree polynomials in E_k [1]. We select the nodes $u_1,u_2,...,u_m$ in such a way that there exists the inverse matrix in (5), and the degree of the interpolation polynomial is determined from inequality (11).

Let us formulate the following conclusion for the space E_k. We get

Theorem 1. Let $f:E_k\to R_1$, $k\ge 2$, m be given. Then, if $m=\overline{p}$, then the Lagrange interpolant $P_n(u)$, $u\in E_k$ will be exact on all polynomials of degree not higher than n, and if $m<\overline{p}$, then the minimum norm interpolant $P_n(u)$, does not have such a property.

We fix the degree of the interpolation polynomial and the number of nodes, for example, $n=2,m=4$. For this case, we construct interpolants in spaces R_1, E_k, $k=2,3,\ldots$. Let p_k be the dimension of polynomials of the second degree in E_k.

In the space E_2, when $n=2,m=4$, we obtain that $p_2=6$. So, for unambiguous definition of $P_2(x,y)$, there are not enough two interpolation nodes. If we consider the construction of the interpolation polynomial of the second degree in E_3, in the case of $m=4$, we obtain that $p_3=10$ and for the unambiguous construction of the interpolant there are not enough 6 nodes. If we

continue this process, then it is clear that as the dimension of the space E_k grows, the dimension of the polynomial space of the two variables p_k increases, and therefore, when constructing the interpolation polynomial of the 2-nd degree for 4 nodes, we are in a situation of "underdeterminacy". As you can see, the larger the dimension of the space E_k, the more indeterminacy (uncertainty) and less accurate of the constructed interpolation polynomial. We arrive at the following conclusion: in the case of decreasing of the Euclidean space dimension, the "underdeterminacy" of the Lagrange interpolant is decreases, and in the case $f : R_1 \to R_1$ we have $m = \overline{p} = n + 1$, that is, we obtain the classical n-th degree Lagrange polynomial with $n + 1$ nodes for the function of one variable. In the space R_1 for $m = 4$ we get that $\overline{p} = 3$, that is, we can construct the interpolation polynomial of the third degree, herewith the resulting interpolant is unique.

As regards the linear space X with a scalar product, the following statement holds. If the interpolation nodes are chosen so that the corresponding matrix is nonsingular, then there is always the unique Lagrange interpolation polynomial with minimum norm [3, 7], but this interpolant is not exact on the operator polynomials of the corresponding degree (Example 1). We note that, the numbers m (number of nodes) and n (interpolation degree) are not related to each other when the interpolation operator Lagrange polynomial is constructed[4].

Remark. We consider the polynomial (8) in the following form

$$P_n(x) = p_n(x, f) + \sum_{i=1}^{m} (f_i - p_n(x_i, f)) l_i(x),\ x \in X, \tag{12}$$

where $p_n(x, f)$ is a c-polynomial, that is $p_n(x, f) = f$, if $f = p_n(x)$ is an arbitrary polynomial operator of degree not higher than n [4]. Then the formula (12) defines an exact interpolant on polynomials of the corresponding degree. Several examples of constructing a c-polynomial are considered in [4].

Bibliography

1. Babenko K.I. Foundations of numerical analysis / K.I. Babenko. – Moscow: Izhevsk: RC "Regular and chaotic dynamics", 2002. (in Russian).
2. Kashpur O.F. To some questions of a polynomial interpolation in euclidean spaces / O.F. Kashpur, V.V. Khlobystov // Dopov. Nac. Akad. Nauk. Ukr. – 2016. – № 10. – P. 10-14. (in Ukrainian).
3. Yegorov A.D. Approximate methods for computation of continual integral / A.D. Yegorov, P.I. Sobolevskiy, L.A. Yanovich. – Minsk: Nauka i Tehnika, 1985. (in Russian).
4. Makarov V.L. Interpolation of operators / V.L. Makarov, V.V. Khlobystov, L.A. Yanovich. – Kiev: Nauk. Dumka, 2000. (in Russian).
5. Trenogin V.A. Functional analysis / V.A. Trenogin. – Moscow: Nauka, 1980. (in Russian).
6. Ulm S. On the construction of generalized separated differences / S. Ulm, V. Polly // Izv. AN ESSR Sir phys.-mat. sciences. – 1969. – 18, № 1. – P. 100-102. (in Russian).
7. Sobolevskiy P.I. Interpolation of functionals and some approximate formulas for integrals with the Gaussian measure / P.I. Sobolevskiy // Izv. Academy of Sciences of the BSSR phys.-mate sciences. – 1975. – № 2. – P. 5-12. (in Russian).
8. Filipsson L. Kergin interpolation in Banach space / L. Filipsson // J. Approx. Theory. – 2004. – № 127. – P. 108-123.

9. Makarov V.L. Interpolation of the many-variable functionals / V.L. Makarov, V.V. Khlobystov, I.I. Demkiv // Dopov. Nac. Akad. Nauk. Ukr.– 2009.– № 5.– P. 29-35. (in Ukrainian).
10. Berezin I.S. Methods of computations / I.S. Berezin, N.P. Zhidkov. – Moscow: Fizmatgiz, 1962. – Vol. 1. (in Russian).

O.F. Kashpur,
Taras Shevchenko National University of Kyiv,
4D, Glushkova Str., Kyiv, 03187, Ukraine;
V. V. Khlobystov,
Institute of Mathematics,
National Academy of Sciences of Ukraine,
3, Tereshchenkovskaya Str., 01601, Kyiv, Ukraine.

Received 02.09.2018; revised 24.09.2018

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 519.65 + 517.548.5 + 519.622

ALGEBRAIC AND TRIGONOMETRIC GENERALIZED INTERPOLATION OF HERMITE-BIRKHOFF TYPE FOR OPERATORS DEFINED ON FUNCTIONAL SPACES AND FUNCTIONS OF MATRIX VARIABLE, AND THEIR APPLICATIONS

A. P. Khudyakov, Ye. V. Panteleyeva, A. A. Trofimuk

Резюме. У роботі побудовано алгебраїчну формулу типу Ерміта для операторів, визначених у функціональних просторах. Інтерполяційна формула подібного виду, яка містить значення диференціалів Гато довільного порядку, побудована на множині матриць. Отримано матрицю, аналогічну до формули Лейбніца. Сконструйовано формулу апроксимації диференціалів Гато довільного порядку з матричними аргументами. На основі матричної інтерполяційної формули типу Ерміта побудовано чисельний метод для розв'язування задачі Коші для матрично-диференціального рівняння. Продемонстровано приклад чисельного розв'язування задачі Коші для матрично-диференціального рівняння першого порядку. Побудовано і досліджено параметричне сімейство тригонометричних матричних інтерполяційних поліномів типу Ерміта-Біркгофа.

Abstract. For operators defined in function spaces, the algebraic interpolation formula of Hermite type is constructed. The interpolation formula of similar type, containing the value of the Gateaux differential of an arbitrary order, is constructed for operators on the set of matrices. Matrix analogues of the Leibniz formula are obtained. The formula for approximate calculation of the Gateaux differential of an arbitrary order of the matrix argument function is constructed. Based on the matrix interpolation formula of the Hermite type, the approximate method for solving the Cauchy problem for the matrix-differential equation is obtained. The illustrative example of approximate solving the Cauchy problem for a first-order matrix-differential equation is constructed. A parametric family of trigonometric matrix interpolation polynomials of Hermite-Birkhoff type is constructed and investigated.

1. Introduction

The fundamentals of the theory of operator interpolation are given in [1,2]. Here, in particular, the problem of operator interpolation of Hermite-Birkhoff type is investigated. The complexity of this problem lies in the fact that even with different interpolation nodes it can either have a non-unique solution, or do not have a solution at all. Some basics of matrix interpolation are also contained in [1,2]. The theory of matrix interpolation is quite fully given in [3]. The papers [4–6] are devoted to the construction and research of Hermite-Birkhoff generalized matrix interpolation formulas for concrete Chebyshev systems.

Key words. Generalized interpolation of Hermite-Birkhoff type, Gateaux differential, Leibniz formula, matrix argument function, Cauchy problem for the matrix-differential equation.

2010 *Mathematics Subject Classification.* 65D05, 39B42, 65F60, 65Q10, 65L05.

In the given work the interpolation formulas for functions of a scalar argument, constructed and investigated in [7, 8], are summarized to the case of operators defined in functional spaces and on the set of matrices. When proving the theorems on the fulfillment of interpolation conditions for the respective polynomials, matrix analogues of the Leibniz formula are used, which are also obtained in this work. The parametric family of trigonometric matrix Hermite-Birkhoff polynomials is constructed.

2. Algebraic interpolation

Let X be a certain given set of functions $x = x(s)$, defined on the segment $[a, b]$, $Y = \{y(s,t), t \in T \subset \mathbb{R}^N\}$ – some function space where T is a given numerical set of N-dimensional space $\mathbb{R}^N$, and let $F(x) \equiv F(t; x(s))$ be an operator mapping X into Y. Let's assume that in the various elements $x_k = x_k(s)$ $(k = 0, 1, \ldots, n)$ of the set X, such that $x_k(s) \neq x_\nu(s)$ on $[a, b]$, the values $F(x_k)$ of the operator $F(x)$, $x \in X$ are known. We choose in the set X functions $h_1(s)$, $h_2(s)$, $\ldots$, $h_{n+1}(s)$ such that $h_1(s)h_2(s)\cdots h_{n+1}(s) \neq 0$ on $[a, b]$. Let the value $D_{n+1}(F;\, x_{n+1})$ of the operator of the form

$$D_{n+1}F(x) = \delta^{n+1}F[x; h_1h_2\cdots h_{n+1}],$$

where $\delta^{n+1}F[x; h_1h_2\cdots h_{n+1}]$ is the Gateaux differential of the order $n+1$ of the operator $F(x)$ at the point x in the directions h_1, h_2, $\ldots$, h_{n+1}, be known in the node $x_{n+1} = x_{n+1}(s) \in X$.

We now consider further the operator polynomials $P_{n+1} : X \to Y$ of the form

$$P_{n+1}(x) = \sum_{\nu=0}^{n+1} a_\nu(t, s)x^\nu(s), \tag{1}$$

where $a_\nu(t, s)$ are some functions of the variables t and s.

We introduce the polynomials $l_{n,k}(x) = (x - x_0)(x - x_1)\cdots(x - x_{k-1})\times$ $\times(x - x_{k+1})\cdots(x - x_n)$, $\omega_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$.

Theorem 1. *The interpolation polynomial*

$$\tilde{L}_{n+1}(x) = L_n(x) + \frac{\omega_n(x)D_{n+1}F(x_{n+1})}{(n+1)!h_1h_2\cdots h_{n+1}},$$

where

$$L_n(x) = \sum_{k=0}^{n} \frac{l_{n,k}(x)F(x_k)}{l_{n,k}(x_k)}, \tag{2}$$

satisfies the interpolation conditions

$$\tilde{L}_{n+1}(x_k) = F(x_k)\ (k = 0, 1, \ldots, n);$$

$$D_{n+1}\left(\tilde{L}_{n+1}; x_{n+1}\right) = D_{n+1}(F; x_{n+1}). \tag{3}$$

The formula (2) *is exact for the operator polynomials of the type* (1) *of the degree not higher than* $n + 1$.

Proof. Since $l_{n,k}(x_i) = \delta_{ki} l_{n,k}(x_k)$, where δ_{ki} is the Kronecker symbol, and $\omega_n(x_k) = 0$, $k, i = 0, 1, \ldots, n$, then the fulfillment of the first group of interpolation conditions in (3) is obvious.

Since $\delta^{n+1} P_n[x; h_1 h_2 \cdots h_{n+1}] \equiv 0$, where $P_n(x)$ is an arbitrary operator algebraic polynomial of a degree not higher than n, then $\delta^{n+1} L_n[x; h_1 h_2 \cdots h_{n+1}] \equiv$ $\equiv 0$. It is also obvious that $\delta^{n+1} \omega_n[x; h_1 h_2 \cdots h_{n+1}] = (n+1)! h_1 h_2 \cdots h_{n+1}$. Taking into account the structure of the polynomial (2), we will obtain that the last condition in (3) also holds.

We now prove the invariance of the formula (2) with respect to the polynomials of the form (1) of the degree not higher than $n+1$. If $F(x) = P_n(x)$, where $P_n(x)$ is a polynomial of the form (1) of the degree not higher than n, then as is known in [2, p. 361], $L_n(P_n;\ x) \equiv P_n(x)$. And since in this case $D_{n+1} P_n(x) \equiv 0$, then $\tilde{L}_{n+1}(P_n; x) \equiv P_n(x)$. Let further suppose $F(x) = \tilde{P}_{n+1}(x) = = x^{n+1}(s)$, then $D_{n+1} \tilde{P}_{n+1}(x) = (n+1)! h_1 h_2 \cdots h_{n+1}$, and

$$\tilde{L}_{n+1}(\tilde{P}_{n+1}; x) = L_n(\tilde{P}_{n+1}; x) + \omega_n(x).$$

By analogy with to the scalar case [7, p. 6], $\tilde{L}_{n+1}\left(\tilde{P}_{n+1};\, x\right) \equiv \tilde{P}_{n+1}(x)$. Thus, the formula (2) is exact for operator polynomials of the form (1) of the degree not higher than $n+1$. □

We now consider the problem of interpolating operators on the set of matrices. Let X be the set of functional or stationary square matrices $A = A(t)$, $t \in T \subset \subset \mathbb{R}$. Let's introduce differential operator of type

$$D^n F(A) = \left.\frac{d^n F(z)}{dz^n}\right|_{z=A}, \quad D = \frac{d}{dz},\ z \in \mathbb{C},\ A \in X, \tag{4}$$

where $F(z)$ is the entire function.

The value of the operator (4) for the matrix function of the type $B_1 F(A) B_2$, where B_1 and B_2 are some fixed matrices from X, is calculated by the formula $D^n\left(B_1 F(A) B_2\right) = B_1 D^n F(A) B_2$. The operator D, which is included in (4), for the function of the type $F(cA+B)$, where $c \in \mathbb{C}$, and B is a certain fixed matrix of X, defined by the equality $DF(cA+B) = cF'(z)|_{z=cA+B}$, and for the product $U(A)V(A)$ by the formula $D\left(U(A)V(A)\right) = DU(A)V(A) + U(A)DV(A)$. In the last expression, it is important in what order the multipliers in matrix products are taken. For example, $D\left(V(A)U(A)\right) = DV(A)U(A) + V(A)DU(A)$, and in the general case, $D\left(U(A)V(A)\right) \neq D\left(V(A)U(A)\right)$. Similarly, the values of higher-order operators are calculated, as well as operators from the products of functions with a number of multipliers more than two.

In mathematical analysis, the Leibniz formula for the derivative of n-th order ($n \in \mathbb{N}$) of the product of two scalar functions is known [9]

$$(u(z) \cdot v(z))^{(n)} = \sum_{k=0}^{n} C_n^k u^{(n-k)}(z) v^{(k)}(z), \text{ where } C_n^k = \frac{n!}{k!(n-k)!}, \tag{5}$$

which holds if the functions $u(z)$ and $v(z)$ are n times differentiable at the point $z \in \mathbb{C}$. We generalize this formula to the case of functions of the matrix argument and operator of the type (4).

Theorem 2. *If the functions $U(z)$ and $V(z)$, $z \in \mathbb{C}$, are differentiable n times, then the formula*

$$D^n\left(U(A)V(A)\right) = \sum_{k=0}^{n} C_n^k D^k U(A) D^{n-k} V(A),\ A \in X, \tag{6}$$

is valid.

Proof. We apply the method of mathematical induction. When $n = 1$ we will have

$$D^1\left(U(A)V(A)\right) = DU(A)V(A) + U(A)DV(A) =$$
$$= C_1^0 D^1 U(A)V(A) + C_1^1 U(A) D^1 V(A).$$

Let's assume that the formula (6) is exact for $n = k$. We prove that it also holds for $n = k + 1$.

$$D^{k+1}\left(U(A)V(A)\right) = D\left[\sum_{k=0}^{n} C_n^k D^k U(A) D^{n-k} V(A)\right] =$$

$$= \sum_{k=0}^{n} C_n^k \left[D^{k+1} U(A) D^{n-k} V(A) + D^k U(A) D^{n-k+1} V(A)\right] =$$

$$= C_n^0 D^0 U(A) D^{n+1} V(A) + \sum_{k=1}^{n} \left(C_n^{k-1} + C_n^k\right) D^k U(A) D^{n-k+1} V(A) +$$
$$+ C_n^n D^{n+1} U(A) D^0 V(A).$$

Since $C_n^{k-1} + C_n^k = C_{n+1}^k$, $C_n^0 = C_{n+1}^0 = 1$, $C_n^n = C_{n+1}^{n+1} = 1$, then

$$D^{k+1}\left(U(A)V(A)\right) = \sum_{k=0}^{n+1} C_{n+1}^k D^k U(A) D^{n+1-k} V(A).$$

□

We now introduce the differential operator of the form

$$\tilde{D}_{n+1} F(A) \equiv \tilde{D}_{n+1} F(A;\, H_{n+1} H_n \cdots H_1) = \delta^{n+1} F[A;\, H_{n+1} H_n \cdots H_1], \tag{7}$$

where $\delta^{n+1} F[A;\, H_{n+1} H_n \cdots H_1]$ is Gateaux differential of order $n + 1$ at the point $A \in X$ in the directions $H_1, H_2, \ldots, H_{n+1}$ from X. We assume that $\tilde{D}_0 F(A) \equiv F(A)$.

Theorem 3. *If the functions $U(A)$ and $V(A)$ are Gateaux differentiable n times at the point $A \in X$, then the formula*

$$\tilde{D}_n\left(U(A)V(A);\, H_n H_{n-1} \cdots H_1\right) = \tag{8}$$

$$= \sum_{k=0}^{n} \sum_{\substack{i_1,\ldots,i_k \\ j_1,\ldots,j_{n-k}}} \tilde{D}_k U(A;\, H_{i_k} H_{i_{k-1}} \cdots H_{i_1}) \tilde{D}_{n-k} V(A;\, H_{j_{n-k}} H_{j_{n-k-1}} \cdots H_{j_1})$$

holds true.

Here, for each value of k $(0 \le k \le n)$ the summation is over for all disjoint sets $(i_1, i_2, \ldots, i_k)$ and $(j_1, j_2, \ldots, j_{n-k})$ such that $1 \le i_1 < i_2 < \ldots < i_k \le n$; $1 \le j_1 < j_2 < \ldots < j_{n-k} \le n$.

Proof. We use, as in the proof of theorem 2, the method of mathematical induction. If $n = 1$ by the definition of the Gateaux differential we will have

$$\tilde{D}_1 (U(A)V(A);\, H_1) = \delta [U(A)V(A);\, H_1] = \lim_{\lambda \to 0} \left(\frac{U(A+\lambda H_1)V(A+\lambda H_1)}{\lambda} - \right.$$

$$\left. - \frac{U(A)V(A)}{\lambda} \right) = \lim_{\lambda \to 0} \left(\frac{U(A+\lambda H_1)V(A+\lambda H_1) - U(A)V(A+\lambda H_1)}{\lambda} + \right.$$

$$\left. + \frac{U(A)V(A+\lambda H_1) - U(A)V(A)}{\lambda} \right) = \delta U[A;\, H_1]V(A) + U(A)\delta V[A;\, H_1] =$$

$$= \tilde{D}_1 U(A;\, H_1)V(A) + U(A)\tilde{D}_1 V(A;\, H_1). \qquad (9)$$

Hereinafter the expression of the form $\delta [U(A)V(A);\, H_1]$ should be understood as the Gateaux differential $\delta W[A;\, H_1]$, respectively, of the function $W(A) = U(A)V(A)$ at the point A in the direction H_1.

Let's suppose that formula (8) is true when $n = m$. We show that it holds for $n = m + 1$. From (7) – (9) we have

$$\tilde{D}_{m+1} (U(A)V(A);\, H_{m+1} \cdots H_1) = \delta \left[\tilde{D}_m (U(A)V(A);\, H_m \cdots H_1);\, H_{m+1} \right] =$$

$$= \sum_{k=0}^{n} \sum_{\substack{i_1,\ldots,i_k \\ j_1,\ldots,j_{n-k}}} \left(\tilde{D}_{k+1} U (A;\, H_{n+1} H_{i_k} \cdots H_{i_1})\, \tilde{D}_{n-k} V \left(A;\, H_{j_{n-k}} \cdots H_{j_1}\right) + \right.$$

$$\left. + \tilde{D}_k U (A;\, H_{i_k} \cdots H_{i_1})\, \tilde{D}_{n+1-k} V \left(A;\, H_{n+1} H_{j_{n-k}} \cdots H_{j_1}\right) \right) =$$

$$= \sum_{k=0}^{n+1} \sum_{\substack{i_1,\ldots,i_k \\ j_1,\ldots,j_{n+1-k}}} \tilde{D}_k U (A;\, H_{i_k} \cdots H_{i_1})\, \tilde{D}_{n+1-k} V \big(A;\, H_{j_{n+1-k}} \cdots H_{j_1}\big).$$

Here the summation is carried out in the same way as in the formulation of the theorem, while $1 \le i_1 < i_2 < \ldots < i_k \le n+1$; $1 \le j_1 < j_2 < \ldots < j_{n+1-k} \le n+1$. □

In the special case, for example, for $n = 3$ the formula (8) has the form

$$\tilde{D}_3 (U(A)V(A); H_3 H_2 H_1) = \tilde{D}_3 U (A; H_3 H_2 H_1)\, V(A) + \tilde{D}_2 U (A; H_3 H_2) \times$$

$$\times \tilde{D}_1 V (A; H_1) + \tilde{D}_2 U (A; H_3 H_1)\, \tilde{D}_1 V (A; H_2) + \tilde{D}_2 U (A; H_2 H_1) \times$$

$$\times \tilde{D}_1 V (A; H_3) + \tilde{D}_1 U (A; H_1)\, \tilde{D}_2 V (A; H_3 H_2) + \tilde{D}_1 U (A; H_2) \times$$

$$\times \tilde{D}_2 V (A; H_3 H_1) + \tilde{D}_1 U (A; H_3)\, \tilde{D}_2 V (A; H_2 H_1) + U(A)\tilde{D}_3 V (A; H_3 H_2 H_1).$$

We suppose that in the elements $A_k(t)$ of the set X such that $A_k(t) - A_\nu(t)$ are invertible matrices, $t \in T$, $k, \nu = 0, 1, \ldots, n$, $k \ne \nu$, the values of the operator $F(A)$ are known, as well as at the node $A_{n+1}(t)$ the value $\tilde{D}_m F(A_{n+1}) \equiv \tilde{D}_m F(A_{n+1}; H_m H_{m-1} \cdots H_1)$ of the operator (7) from $F(A)$, where $1 \le m \le n$, $H_k \in X$ $(k = 1, 2, \ldots, m)$ is known. Let's introduce the notations $\omega(A) = (A - A_0)(A - A_1) \cdots (A - A_n)$, $l_k(A) = (A - A_0) \cdots (A - A_{k-1})(A - A_{k+1}) \cdots (A - A_n)$, $B_k = \tilde{D}_m l_k(A_{n+1})$, $\tilde{A}_k = B_k A_{n+1} + B_k^{-1} \sum_{i=1}^{m} \tilde{D}_{m-1} l_k(A_{n+1};$

$H_m \cdots H_{i+1} H_{i-1} \cdots H_1) B_k H_i$ $(k = 0, 1, \ldots, n)$. We will assume that the matrices B_k, $l_k(A_k)$, $B_k A_k - \tilde{A}_k$ $(k = 0, 1, \ldots, n)$ and $\tilde{D}_m \omega(A_{n+1})$ are invertible.

Theorem 4. *The matrix polynomial of the degree not higher than* $n + 1$

$$\tilde{L}_{n+1}(F; A) = \sum_{k=0}^{n} l_k(A)(B_k A - \tilde{A}_k) \left[l_k(A_k)(B_k A_k - \tilde{A}_k) \right]^{-1} F(A_k) + \\ + \omega(A) \left[\tilde{D}_m \omega(A_{n+1}) \right]^{-1} \tilde{D}_m F(A_{n+1}) \tag{10}$$

satisfies the interpolation conditions

$$\tilde{L}_{n+1}(A_k) = F(A_k) \ (k = 0, 1, \ldots, n); \quad \tilde{D}_m \tilde{L}_{n+1}(A_{n+1}) = \tilde{D}_m F(A_{n+1}). \tag{11}$$

Proof. Since $l_k(A_i) = \delta_{ki} l_k(A_k)$ $(k, i = 0, 1, \ldots, n)$, where δ_{ki} is the Kronecker symbol, and $\omega(A_k) = 0$ for the same values of k, then the first group of the conditions in (11) is satisfied. By the formula (8)

$$\tilde{D}_m \Big(l_k(A)(B_k A - \tilde{A}_k); H_m \cdots H_1 \Big) = \tilde{D}_m l_k(A; H_m \cdots H_1)(B_k A - \tilde{A}_k) + \\ + \sum_{i=1}^{m} \tilde{D}_{m-1} l_k(A; H_m \cdots H_{i+1} H_{i-1} \cdots H_1) \tilde{D}_1(B_k A - \tilde{A}_k; H_i).$$

Due to the fact that $\tilde{D}_1(B_k A - \tilde{A}_k; H_i) = B_k H_i$, then for $A = A_{n+1}$

$$\tilde{D}_m \Big(l_k(A)(B_k A - \tilde{A}_k); H_m \cdots H_1 \Big) \Big|_{A = A_{n+1}} = B_k(B_k A_{n+1} - \tilde{A}_k) + \\ + \sum_{i=1}^{m} \tilde{D}_{m-1} l_k(A; H_m \cdots H_{i+1} H_{i-1} \cdots H_1) B_k H_i = 0.$$

Taking into account the structure of the formula (10), we will obtain that the last condition in equation (11) also holds. □

Using the interpolation polynomial (10), we can construct a formula for approximate calculation of the Gateaux differential of the m-th $(1 \le m \le n)$ order from the function of the matrix argument $F(A)$ by its values at the nodes $A_0, A_1, \ldots, A_n$. Indeed, the relation

$$F(A) = \sum_{k=0}^{n} l_k(A)(B_k A - \tilde{A}_k) \left[l_k(A_k)(B_k A_k - \tilde{A}_k) \right]^{-1} F(A_k) + \\ + \omega(A) \left[\tilde{D}_m \omega(A_{n+1}) \right]^{-1} \tilde{D}_m F(A_{n+1}) + R_n(F; A),$$

where $R_n(F; A)$ is the remainder term of the formula (10), holds true. Then, expressing from the last equality $\tilde{D}_m F(A_{n+1})$, we will have

$$\tilde{D}_m F(A_{n+1}) = \tilde{D}_m \omega(A_{n+1}) \omega^{-1}(A) \left(F(A) - \sum_{k=0}^{n} l_k(A)(B_k A - \tilde{A}_k) \times \right. \\ \left. \times \Big[l_k(A_k)(B_k A_k - \tilde{A}_k) \Big]^{-1} F(A_k) - R_n(F; A) \right). \tag{12}$$

Discarding in (12) the remainder term $R_n(F;A)$ of the formula (10), we will obtain the required approximate formula for calculating the Gateaux differential

$$\delta^m F[A;H_mH_{m-1}\cdots H_1] \cong \tilde{D}_m\omega(A_{n+1})\omega^{-1}(A)\times$$

$$\times\left(F(A)-\sum_{k=0}^{n} l_k(A)(B_kA-\tilde{A}_k)\left[l_k(A_k)(B_kA_k-\tilde{A}_k)\right]^{-1}F(A_k)\right). \qquad (13)$$

Here, the matrix A must be such that the matrices entering into the formula are invertible.

3. The solving matrix-differential equations

Let X be the set of square stationary matrices of fixed size. We consider the matrix equation containing the first-order Gateaux differential of the matrix function

$$\delta U[A;H] = F(U,A),\ U(A_0)=U_0,\ A,H\in X, \qquad (14)$$

where $U(A)$ is a function of the matrix argument, F is some generally non-linear function of two arguments, $\delta U[A;H]$ is the Gateaux differential at the point A in the direction H satisfying the specified in (14) initial condition.

For the approximate solving the Cauchy problem (14), we use the formula (13) for approximating the Gateaux differential of the matrix argument function. In our case it takes the form

$$\delta U[A;H] = \delta\omega[A;H]\omega^{-1}(A_{n+1})\times \qquad (15)$$

$$\times\left(U(A_{n+1})-\sum_{k=0}^{n} l_k(A_{n+1})(B_kA_{n+1}-\tilde{A}_k)\left[l_k(A_k)(B_kA_k-\tilde{A}_k)\right]^{-1}U(A_k)\right),$$

where $B_k = B_k(A) = \delta l_k[A;H]$, $\tilde{A}_k = \tilde{A}_k(A) = B_k(A)A + B_k^{-1}(A)l_k(A)\times$ $\times B_k(A)H$. Here $A_0, A_1, \ldots, A_n$ are the matrices from X such that the inverse matrices in (15) exist.

Substituting (15) into (14), we obtain

$$\delta\omega[A;H]\omega^{-1}(A_{n+1})\left(Y_{n+1}-\sum_{k=0}^{n} l_k(A_{n+1})(B_kA_{n+1}-\tilde{A}_k)\times\right.$$

$$\left.\times\left[l_k(A_k)(B_kA_k-\tilde{A}_k)\right]^{-1}Y_k\right) = F(Y,A),\ Y_0=U_0, \qquad (16)$$

where $Y_0, Y_1, \ldots, Y_{n+1}$ is approximate solution of the problem (14) in the matrix nodes $A_0, A_1, \ldots, A_{n+1}$. If now we substitute the matrix nodes A_k ($k = 1,2,\ldots,n+1$) instead of A in (16), then we obtain the system (in the general case, non-linear) matrix equations. Solving this system by some direct or iterative method, we obtain the required approximate solution of the problem (14).

Example. Let X be the set of square matrices of size 2. We consider the Cauchy problem for the function of the matrix variable $U(A), A\in X$

$$\delta U[A;H] = 3U(A)+2A,\ U(A_0)=U_0, \qquad (17)$$

where $A_0 = \begin{pmatrix} 0.312 & 0.467 \\ 0.457 & 0.02 \end{pmatrix}$, $U_0 = \begin{pmatrix} 0.316 & 0.338 \\ 0.23 & 0.002 \end{pmatrix}$, $H = \begin{pmatrix} 0.021 & 0.43 \\ 0.405 & 0.223 \end{pmatrix}$. Let's introduce the matrix nodes $A_1 = \begin{pmatrix} 0.11 & 0.032 \\ 0.223 & 0.155 \end{pmatrix}$, $A_2 = \begin{pmatrix} 0.004 & 0.085 \\ 0.5 & 0.305 \end{pmatrix}$, $A_3 = \begin{pmatrix} 0.234 & 0.028 \\ 0.2 & 0.004 \end{pmatrix}$, $A_4 = \begin{pmatrix} 0.051 & 0.291 \\ 0.176 & 0.498 \end{pmatrix}$.

For the approximate solving of the problem (14) we use the formula (16) for $n = 3$. We construct a system of matrix equations. In this case, it is linear. We have

$$Y_0 = U_0 = \begin{pmatrix} 0.316 & 0.338 \\ 0.23 & 0.002 \end{pmatrix}, \ \delta\omega[A_i; H]\omega^{-1}(A_4) \left(Y_4 - \sum_{k=0}^{3} l_k(A_4) \times \right.$$

$$\left. \times \Big(B_k(A_i)A_4 - \tilde{A}_k(A_i) \Big) \Big[l_k(A_k) \Big(B_k(A_i)A_k - \tilde{A}_k(A_i) \Big) \Big]^{-1} Y_k \right) =$$

$$= 3Y_i + 2A_i, \ i = 1, 2, 3, 4. \tag{18}$$

Let's present numerically the system of the matrix equations (18) to within 3 significant digits to determine the unknowns Y_0, Y_1, Y_2, Y_3, Y_4

$$\begin{gathered} Y_0 = U_0, \ - \begin{pmatrix} 0.992 & 0.186 \\ 0.180 & 0.0380 \end{pmatrix} Y_0 - \begin{pmatrix} 292 & 302 \\ 47.5 & 51.9 \end{pmatrix} Y_1 + \begin{pmatrix} 0.142 & 4.05 \\ 0.268 & 6.00 \end{pmatrix} Y_2 + \\ + \begin{pmatrix} 2.49 & -15.5 \\ 2.00 & -12.3 \end{pmatrix} Y_3 + \begin{pmatrix} 3.33 & 4.20 \\ 0.815 & 0.606 \end{pmatrix} Y_4 = \begin{pmatrix} 0.22 & 0.064 \\ 0.446 & 0.31 \end{pmatrix}, \\ \begin{pmatrix} 2.48 & 14.1 \\ -2.12 & -12.1 \end{pmatrix} Y_0 - \begin{pmatrix} 1368 & 2630 \\ -1190 & -2289 \end{pmatrix} Y_1 - \begin{pmatrix} 246 & 297 \\ -235 & -285 \end{pmatrix} Y_2 + \\ + \begin{pmatrix} -50.8 & 6.08 \\ 52.1 & -6.20 \end{pmatrix} Y_3 + \begin{pmatrix} -8.96 & -14.4 \\ 7.56 & 12.5 \end{pmatrix} Y_4 = \begin{pmatrix} 0.008 & 0.17 \\ 1.0 & 0.61 \end{pmatrix}, \\ \begin{pmatrix} 8.20 & -2.04 \\ 1.83 & -0.441 \end{pmatrix} Y_0 - \begin{pmatrix} 211 & 135 \\ 49.2 & 32.5 \end{pmatrix} Y_1 + \begin{pmatrix} 13.7 & 21.9 \\ 2.06 & 3.15 \end{pmatrix} Y_2 + \\ + \begin{pmatrix} -10.2 & -34.7 \\ 1.20 & 8.53 \end{pmatrix} Y_3 - \begin{pmatrix} 7.12 & 12.0 \\ 1.92 & 2.75 \end{pmatrix} Y_4 = \begin{pmatrix} 0.468 & 0.056 \\ 0.4 & 0.008 \end{pmatrix}, \\ \begin{pmatrix} 0.149 & 0.662 \\ -0.286 & -0.975 \end{pmatrix} Y_0 + \begin{pmatrix} 230 & 340 \\ -363 & -539 \end{pmatrix} Y_1 + \begin{pmatrix} 2.60 & 3.26 \\ -1.86 & -2.36 \end{pmatrix} Y_2 + \\ + \begin{pmatrix} -0.991 & 0.424 \\ 0.727 & -0.138 \end{pmatrix} Y_3 + \begin{pmatrix} -14.4 & -15.6 \\ 15.9 & 21.2 \end{pmatrix} Y_4 = \begin{pmatrix} 0.102 & 0.582 \\ 0.352 & 0.996 \end{pmatrix}. \end{gathered} \tag{19}$$

The system of the matrix equations (19) can be written element-by-element, having obtained a system of 20 linear algebraic equations with respect to 20 unknowns (elements of matrices Y_0, Y_1, Y_2, Y_3, Y_4). Immediately excluding Y_0 from the remaining matrix equations in (19), we will obtain the system of 16 linear algebraic equations that can be solved, for example, by the Gauss method. According to this method, the solution of the system (19) has the form

$$Y_0 = U_0, \ Y_1 = \begin{pmatrix} 0.00221 & 0.00618 \\ -0.00177 & -0.00416 \end{pmatrix}, \ Y_2 = \begin{pmatrix} -0.0393 & 0.00504 \\ 0.0264 & -0.0223 \end{pmatrix},$$

$$Y_3 = \begin{pmatrix} 0.133 & 0.132 \\ -0.0130 & -0.0395 \end{pmatrix},\ Y_4 = \begin{pmatrix} -0.171 & -0.546 \\ 0.148 & 0.455 \end{pmatrix}.$$

The solution of the problem (17) obtained in the matrix nodes can be restored using the matrix interpolation formula [2, p. 459] of the form $L_{n0}(A) = \sum\limits_{k=0}^{n} l_k(A) l_k^{-1}(A_k) F(A_k)$, where, as before, $l_k(A) = (A - A_0) \cdots (A - A_{k-1}) \times \times (A - A_{k+1}) \cdots (A - A_n)$ $(k = 0, 1, \ldots, n)$, satisfying the interpolation conditions $L_{n0}(A_k) = F(A_k)$ for $k = 0, 1, \ldots, n$. In our case, $n = 4$, $F(A_k) = Y_k$ $(k = 0, 1, 2, 3, 4)$ and $U(A) \approx Y(A) = L_{4,0}(A)$.

We introduce the matrices of the form $\bar{A}_i = (A_{i-1} + A_i)/2$ $(i = 1, 2, 3, 4)$ and define the norms of the residual matrices between the left and right sides of the matrix-differential equation of the problem (14). We calculate the Gateaux differential $\delta Y[A; H] = \delta L_{4,0}[A; H]$ by the known [10] formula $\delta Y\left[\bar{A}_i; H\right] = \lim\limits_{\lambda \to 0} \left\{\lambda^{-1} \left[Y(\bar{A}_i + \lambda H) - Y(\bar{A}_i)\right]\right\}$.

We denote by $R_i = \|\delta Y\left[\bar{A}_i; H\right] - 3Y\left(\bar{A}_i\right) - 2\bar{A}_i\|_2$, $i = 1, 2, 3, 4$, where $\|\cdot\|_2$ is the spectral norm of the corresponding matrix [11]. In our case, these norms are equal to $R_1 = 0.699$, $R_2 = 0.528$, $R_3 = 0.959$, $R_4 = 0.250$. The numerical experiment shows that the discrepancy between the left and right sides of the equation (14) is small, however, the accuracy of the approximation is not high. To obtain a higher accuracy of the solution it is necessary to involve more nodes or to use other methods of approximating the matrix-differential operator.

Analogous methods for solving matrix-differential equations can be obtained using the formulas of trigonometric, exponential, and other types of matrix generalized Hermite-Birkhoff interpolation.

4. Trigonometric interpolation

In [7] for 2π-periodic scalar functions the parametric family of trigonometric interpolation polynomials of degree not higher than $n + 1$ of the form

$$T_{n+1}^{\alpha,\beta}(x) = H_n(x) + \frac{\Omega_{n+1}^{\alpha,\beta}(x) D_{2n+1}(f; x_j)}{D_{2n+1}(\Omega_{n+1}^{\alpha,\beta}; x_j)}, \tag{20}$$

where $\Omega_{n+1}^{\alpha,\beta}(x) = \left(\alpha \sin\frac{x}{2} + \beta \cos\frac{x}{2}\right) \prod\limits_{k=0}^{2n} \sin\frac{x - x_k}{2}$, $\alpha^2 + \beta^2 \neq 0$, $H_n(x)$ is a trigonometric interpolation polynomial of degree not higher than n of Lagrange type, and the differential operator $D_{2n+1} f(x)$ is defined by the formula

$$D_{2n+1} f(x) = (D^2 + n^2) \cdots (D^2 + 1^2) D f(x),\ D = \frac{d}{dx},$$

is constructed. The polynomial (20) satisfies the interpolation conditions

$$T_{n+1}^{\alpha,\beta}(x_i) = f(x_i)\ (i = 0, 1, \ldots, 2n);\ \ D_{2n+1}(T_{n+1}^{\alpha,\beta}; x_j) = D_{2n+1}(f; x_j).$$

We generalize the formula (20) in the case of functions of the matrix argument. Let X be the set of square matrices, $F(z)$ be an entire 2π-periodic function, $z \in \mathbb{C}$. In different matrix nodes A_k such that the matrices $A_k - A_\nu$

$(k, \nu = 0, 1, \ldots, 2n)$ are invertible, the values $F(A_k)$ of the function $F(A)$, $A \in X$, are known. The value $D_{2n+1}(F; A_j)$ of the matrix-differential operator

$$D_{2n+1}F(A) = (D^2 + n^2) \cdots (D^2 + 1^2) DF(z)\big|_{z=A}, \; D = \frac{d}{dz}, \tag{21}$$

is also known in one of the nodes A_j.

Let's consider the differential operator of even order

$$D_{2n}F(A) = \left(D^2 + (n-1)^2\right) \cdots \left(D^2 + 1^2\right) D^2 F(z)\big|_{z=A}. \tag{22}$$

The values of the operator for functions of the forms $B_1F(A)B_2$, $F(cA + B)$ and $U(A)V(A)$ are calculated similarly, as are the values of the operator (4) for functions of this type. We assume that $D_0F(A) \equiv F(A)$.

Let's generalize the Leibniz formula (5) to the case of functions of the matrix argument, and when the differential operators (21) and (22) are taken instead of the derivatives. Is valid

Theorem 5. *If the functions $U(z)$ and $V(z)$, $z \in \mathbb{C}$, are differentiable m times, then the formula*

$$D_m\left(U(A)V(A)\right) = D_{2p+1}\left(U(A)V(A)\right) = \sum_{k=0}^{m} C_m^k D_{m-k}U(A)D_kV(A), \tag{23}$$

$$D_m\left(U(A)V(A)\right) = D_{2p+2}\left(U(A)V(A)\right) = \sum_{k=0}^{m} C_m^k D_{m-k}U(A)D_kV(A) -$$

$$-\frac{m(m-1)}{4} \sum_{k=1,3,\ldots}^{m-3} C_{m-2}^k D_{m-k-2}U(A)D_kV(A), \; A \in X, \; p = 0, 1, \ldots,$$

is valid.

The proof of the theorem 5 repeats the proof of the analogous theorem for the scalar case [8, p. 18-21]. In this case, the order of the multipliers in the matrix products must be strictly preserved: the values of the operators (21), (22) from the function $U(A)$ should be located to the left of the values of these operators from the function $V(A)$.

Lemma 1. *For trigonometric polynomials of the form*

$$P_n(A) = \sin\frac{A - B_1}{2} \sin\frac{A - B_2}{2} \cdots \sin\frac{A - B_{2n}}{2},$$

where $B_1, B_2, \ldots, B_{2n}$ are some matrices from X, the following identities are valid

$$D_jP_n(A) \equiv 0, \; j = 2n + 1, 2n + 2, \ldots \tag{24}$$

Proof. Let's apply the method of mathematical induction. When $n = 1$

$$P_1(A) = \sin\frac{A - B_1}{2} \sin\frac{A - B_2}{2},$$

and by the formula (23) for $m = 3$ we have

$$D_3P_1(A) = D_3 \sin\frac{A - B_1}{2} \cdot \sin\frac{A - B_2}{2} + 3D_2 \sin\frac{A - B_1}{2} \cdot D_1 \sin\frac{A - B_2}{2} +$$

$$+3D_1 \sin\frac{A-B_1}{2}\cdot D_2 \sin\frac{A-B_2}{2} + \sin\frac{A-B_1}{2}\cdot D_3 \sin\frac{A-B_2}{2}.$$

Since

$$D_1 \sin\frac{A-B_k}{2} = D\sin\frac{A-B_k}{2} = \frac{1}{2}\cos\frac{A-B_k}{2},$$

$$D_2 \sin\frac{A-B_k}{2} = D^2\sin\frac{A-B_k}{2} = -\frac{1}{4}\sin\frac{A-B_k}{2},$$

$$D_3 \sin\frac{A-B_k}{2} = \left(D^3+D\right)\sin\frac{A-B_k}{2} = \frac{3}{8}\cos\frac{A-B_k}{2} \quad (k=1,2),$$

then $D_3P_1(A)\equiv 0$.

For the operator (21), (22) the properties $D_{2n+2}F(A) = DD_{2n+1}F(A)$, $D_{2n+3}F(A) = \left(D^2+(n+1)^2\right)D_{2n+1}F(A)$, $n\in\mathbb{N}$, where $F(A)$ is some matrix function for which the values of the operators (21) and (22) at the point $A\in X$ exist, are hold. Then it is obvious that $D_jP_1(A)\equiv 0$ when $j=4,5,\ldots$

Let's suppose that the relations (24) hold when $n=k$. We will show that they are true when $n=k+1$. By the formula (23) for $m=2k+3$ we have

$$D_{2k+3}P_{k+1}(A) = D_{2k+3}\left(P_k(A)\tilde{P}_1(A)\right) = \sum_{i=0}^{2k+3} C_{2k+3}^i D_{2k+3-i}P_k(A)\cdot D_i\tilde{P}_1(A),$$

where

$$\tilde{P}_1(A) = \sin\frac{A-B_{2k+1}}{2}\sin\frac{A-B_{2k+2}}{2}.$$

For $i\le 2$, by assumption, the identities $D_{2k+3-i}P_k(A)\equiv 0$ hold, and when $i>2$ the identities $D_i\tilde{P}_1(A)\equiv 0$ are valid. Therefore $D_{2k+3}P_{k+1}(A)\equiv 0$. □

Let α and β be some fixed matrices from X that are not simultaneously zero.

Theorem 6. *The trigonometric polynomial*

$$T_{n+1}(A)\equiv T_{n+1}(A;\alpha,\beta) =$$

$$= H_n(A) + \Omega_{n+1}(A)\left[D_{2n+1}(\Omega_{n+1};A_{n+1})\right]^{-1} D_{2n+1}(F;A_{n+1}), \tag{25}$$

where

$$H_n(A) = \sum_{k=0}^{2n}\Psi_k(A)\Psi_k^{-1}(A_k)F(A_k), \tag{26}$$

$$\Psi_k(A) = \sin\frac{A-A_0}{2}\cdots\sin\frac{A-A_{k-1}}{2}\sin\frac{A-A_{k+1}}{2}\cdots\sin\frac{A-A_{2n}}{2},$$

$$\Omega_{n+1}(A)\equiv\Omega_{n+1}(A;\alpha,\beta) = \left(\alpha\sin\frac{A}{2}+\beta\cos\frac{A}{2}\right)\prod_{k=0}^{2n}\sin\frac{A-A_k}{2},$$

satisfies the interpolation conditions

$$T_{n+1}(A_k) = F(A_k)\ (k=0,1,\ldots,2n);$$

$$D_{2n+1}(T_{n+1};A_{2n+1}) = D_{2n+1}(F;A_{2n+1}). \tag{27}$$

Proof. Since $\Psi_k(A_i) = \delta_{ki}\Psi_k(A_k)$, where δ_{ki} is the Kronecker symbol $(k, i = 0, 1, \ldots, 2n)$, then the polynomial (26) coincides with the operator $F(A)$ at the interpolation nodes $A_0, A_1, \ldots, A_{2n}$. It's obvious that $\Omega_{n+1}(A_k) = 0$ when $k = \overline{0, 2n}$. Therefore, the polynomial (25) coincides with $F(A)$ at the above-mentioned interpolation nodes.

We show that the last condition in (27) also holds. By the lemma $D_{2n+1}\Psi_k(A) = 0$ for $k = 0, 1, \ldots, 2n$, so $D_{2n+1}H_n(A) = 0$. Taking into account the structure of the formula (25), we obtain that the condition stated above for the polynomial $T_{n+1}(A)$ is satisfied. □

5. Conclusion

In this work we obtained the following new results: interpolation formulas for functions of a scalar argument are generalized to the case of operators defined in functional spaces and on the set of matrices. The algebraic operator and matrix interpolation Hermite–Birkhoff polynomials are constructed, as well as the parametric family of trigonometric matrix interpolation polynomials of Hermite type. Theorems on the fulfillment of the interpolation conditions are proved. For the operator interpolation formula, a class of polynomials for which it is exact is found. Matrix analogues of the Leibniz formula for linear matrix-differential operators of a special form are constructed. Based on the matrix algebraic interpolation polynomial, the formula for the approximation of the Gateaux differential of an arbitrary order of the matrix argument function is obtained. This formula is used in the construction of the approximate method for solving the Cauchy problem with a matrix-differential equation of the first order. In the computer algebra system, the illustrative example of a numerical solving the Cauchy problem of the indicated type is realized.

Acknowledgements. The work has been carried out with the financial support of the Belarusian Republican Foundation for Fundamental Research (project No. F16M-055).

Bibliography

1. Makarov V.L. Interpolation of operators / V.L. Makarov, V.V. Khlobystov, L.A. Yanovich. – K.: Naukova Dumka, 2000. – 407 p. (in Russian).
2. Makarov V.L. Methods of Operator Interpolation / Volodymyr L. Makarov, Volodymyr V. Khlobystov, Leonid A. Yanovich. – K.: Works of the Insitute of Mathematics of the National Academy of Sciences of Ukraine, 2010. – Vol. 83. – 517 p.
3. Yanovich L.A. Fundamentals of the interpolation theory of functions of matrix variables / L.A. Yanovich, M.V. Ignatenko. – Minsk: Belaruskaya Navuka, 2016. – 281 p. (in Russian).
4. Khudyakov A.P. Some problems in the theory of interpolation / A.P. Khudyakov. – Saarbrucken, Deutschland: LAP LAMBERT Academic Publishing, 2014. – 132 p. (in Russian).
5. Yanovich L.A. On one class of interpolating formulas for functions of matrix variables / L.A. Yanovich, A.P. Hudyakov // J. Numer. Appl. Math. – 2011. – No. 2 (105). – P. 136-147.
6. Khudyakov A.P. Generalized Hermite interpolation polynomials for functions of the matrix variable / A.P. Khudyakov, L.A. Yanovich // Trudy Instituta matematiki NAN Belarusi. – 2011. – Vol. 19, No. 2. – P. 103-114. (in Russian).

7. Khudyakov A.P. Generalized Hermite-Birkhoff interpolation formulas for the case of Chebyshev systems of functions / A.P. Khudyakov, L.A. Yanovich // Vesci NAN Belarusi. Ser. fiz-mat. navuk. – 2015. – No. 2. – P. 5-14. (in Russian).
8. Khudyakov A.P. The Hermite-Birkhoff interpolation formulas for algebraic and trigonometric systems of functions with one special node / A.P. Khudyakov, A.A. Trofimuk // Vesci NAN Belarusi. Ser. fiz-mat. navuk. – 2017. – No. 1. – P. 14-28. (in Russian).
9. Zorich V.A. Mathematical analysis. Part 1 / V.A. Zorich. – 4th ed. – M.: MCCME, 2002. – 664 p. (in Russian).
10. Trenogin V.A. Functional Analysis / V.A. Trenogin. – M.: Nauka. Chief. ed. of phys. and mat. lit., 1980. – 496 p. (in Russian).
11. Gantmakher F.R. Theory of matrices / F.R. Gantmakher. – 3rd ed. – M.: Nauka, 1967. – 575 p.

A. P. Khudyakov, Ye. V. Panteleyeva, A. A. Trofimuk,
Faculty of Physics and Mathematics,
Brest State University named after A.S. Pushkin,
Kosmonavtov Boulevard, 21, Brest, 224016, Republic of Belarus;
Faculty of Mathematics and Programming Technologies,
Francisk Skorina Gomel State University,
Sovetskaya str., 104, Gomel, 246019, Republic of Belarus.

Received 19.03.2018; revised 30.05.2018

Журнал обчислювальної та прикладної математики | **2018** **№ 2 (128)** | Journal of Numerical & Applied Mathematics

UDC 519.6

CONVERGENCE OF A TWO-STEP METHOD FOR THE NONLINEAR LEAST SQUARES PROBLEM WITH DECOMPOSITION OF OPERATOR

S. M. Shakhno, R. P. Iakymchuk, H. P. Yarmola

Резюме. У роботі запропоновано двокроковий метод для розв'язування нелінійної задачі найменших квадратів з декомпозицією оператора та досліджено його збіжність за класичних умов Ліпшиця для похідних першого і другого порядків диференційовної частини та поділених різниць першого порядку недиференційовної частини декомпозиції. Встановлено порядок і радіус збіжності методу, а також область єдиності розв'язку нелінійної задачі про найменші квадрати. Проведено чисельні експерименти на ряді тестових задачах.

Abstract. In this article, we propose a two-step method for the nonlinear least squares problem with the decomposition of the operator. We investigate the convergence of this method under the classical Lipschitz condition for the first- and second-order derivatives of the differentiable part and for the first-order divided differences of the non-differentiable part of the decomposition. The convergence order as well as the convergence radius of the method are studied and the uniqueness ball of the solution of the nonlinear least squares problem is examined. Finally, we carry out numerical experiments on a set of test problems.

1. Introduction

Let us consider the nonlinear least squares problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}F(x)^T F(x), \tag{1}$$

where F is a Fréchet differentiable operator defined on $\mathbb{R}^n$ with its values on $\mathbb{R}^m$, $m \geq n$. The best known method for finding an approximate solution of the problem (1) is the Gauss-Newton method, which is defined as

$$x_{k+1} = x_k - [F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k),\ k = 0, 1, 2, ... \tag{2}$$

The convergence analysis of the method (2) under various conditions was conducted in [6–8]. In paper [18], three free-derivative iterative methods were investigated under the classical Lipschitz conditions. The radius of the convergence ball and the convergence order of these methods were determined. The study of these methods was conducted in the case of both zero and nonzero residuals.

Key words. Nonlinear least squares problem, two-step method, Gauss-Newton method, decomposition of operator, Lipschitz conditions, radius of convergence, uniqueness ball.

In particular, Shakhno [18] proposed the Secant-type method, which was later also studied by Ren and Argyros in [12], as follows

$$x_{k+1} = x_k - [F(x_k, x_{k-1})^T F(x_k, x_{k-1})]^{-1} F(x_k, x_{k-1})^T F(x_k), \quad k = 0, 1, 2, ... \tag{3}$$

This study [18] also determines the convergence order of the method (3) in case of zero residual, which equals to $\frac{1+\sqrt{5}}{2} = 1,618....$

In [2, 4, 10, 11], there was considered a two-step modification of the Gauss-Newton method for solving the problem (1)

$$\begin{cases} x_{k+1} = x_k - [F'(z_k)^T F'(z_k)]^{-1} F'(z_k)^T F(x_k), \\ y_{k+1} = x_{k+1} - [F'(z_k)^T F'(z_k)]^{-1} F'(z_k)^T F(x_{k+1}), \ k = 0, 1, 2, ..., \end{cases} \tag{4}$$

where $z_k = (x_k + y_k)/2$; x_0 and y_0 are given. In case when $m = n$, this method is equivalent to the methods proposed by Bartish [3] and Werner [23]. On each iteration, the method (4) computes the inversion of the matrix $[F'(z_k)^T F'(z_k)]^{-1}$ only once.

In [17], we proposed the difference variant of the method (4) that uses divided differences instead of derivatives as follows

$$\begin{cases} x_{k+1} = x_k - [F(x_k, y_k)^T F(x_k, y_k)]^{-1} F(x_k, y_k)^T F(x_k), \\ y_{k+1} = x_{k+1} - [F(x_k, y_k)^T F(x_k, y_k)]^{-1} F(x_k, y_k)^T F(x_{k+1}), \ k = 0, 1, 2, ... \end{cases} \tag{5}$$

This method is built on top of the Secant-type method [12,18] (3) for solving the nonlinear least squares problem. This method can also be applied to problems with non-differentiable operators.

However, for some problems the nonlinear function in (1) is composed of the differentiable and non-differentiable parts. In this case, the problem (1) can be written as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} (F(x) + G(x))^T (F(x) + G(x)), \tag{6}$$

where the residual function $F + G$ is defined on $\mathbb{R}^n$ with its values on $\mathbb{R}^m$ and it is nonlinear by x; F is a continuously differentiable function; G is a continuous function, differentiability of which, in general, is not required. To solve the problem (6), we proposed in [14,19] a method that takes into account the specific features of both F and G as

$$x_{k+1} = x_k - [A_k^T A_k]^{-1} A_k^T (F(x_k) + G(x_k)), \quad k = 0, 1, ..., \tag{7}$$

where $A_k = F'(x_k) + G(x_k, x_{k-1})$; $F'(x_k)$ is a Fréchet derivative of $F(x)$; $G(x_k, x_{k-1})$ is the divided difference of the first-order of the function $G(x)$ at points x_k, x_{k-1}; x_0, x_{-1} are given starting points. This method has the convergence order of $\frac{1+\sqrt{5}}{2}$ for solving the problem (6) with zero residual. In case when $m = n$, the method (7) reassembles the well-know Newton-Secant method for nonlinear equations [1, 5, 15].

In this article, we propose a two-step iterative method, for solving the problem (6), which considers the decomposition of the nonlinear operator, as follows

$$\begin{cases} x_{k+1} = x_k - [A_k^T A_k]^{-1} A_k^T (F(x_k) + G(x_k)), \\ y_{k+1} = x_{k+1} - [A_k^T A_k]^{-1} A_k^T (F(x_{k+1}) + G(x_{k+1})), \quad k = 0, 1, ..., \end{cases} \tag{8}$$

where $A_k = F'(\frac{x_k + y_k}{2}) + G(x_k, y_k)$. The main goal of this paper is to analyze the local convergence of the method (8) for the problem (6) with zero as well as non-zero residuals. Additionally, we study both the order and the radius of the convergence of the method (8) as well as the uniqueness ball of the solution of the problem (6). To note, this method as well as the method (5) have the same convergence order of $1 + \sqrt{2}$ in case of zero residual.

In case of $m = n$, the problem (6) reduces to solving a system of n nonlinear equations with n unknown and the method (8) reduces to the method [16,20,21].

2. Preliminaries

Let us denote $B(x_*, r) = \{x \in D \subseteq \mathbb{R}^n : \|x - x_*\| < r\}$ as an open ball with the radius r ($r > 0$) at x_*, D is an open convex subset of $\mathbb{R}^n$.

Let $\mathbb{R}^{m \times n}$, $m \geq n$, denote a set of all $m \times n$ matrices. Then, for a full rank matrix $A \in \mathbb{R}^{m \times n}$, its Moore-Penrose pseudo-inverse [8] is defined as $A^\dagger = (A^T A)^{-1} A^T$.

Lemma 1 ([13,22]). *Let $A, E \in \mathbb{R}^{m \times n}$. Assume that $C = A + E$, $\|A^\dagger\|\|E\| < 1$, and $rank(A) = rank(C)$. Then,*

$$\|C^\dagger\| \leq \frac{\|A^\dagger\|}{1 - \|A^\dagger\|\|E\|}.$$

If $rank(A) = rank(C) = \min(m, n)$, we can obtain

$$\|C^\dagger - A^\dagger\| \leq \frac{\sqrt{2}\|A^\dagger\|^2\|E\|}{1 - \|A^\dagger\|\|E\|}.$$

Lemma 2 ([6]). *Let $A, E \in \mathbb{R}^{m \times n}$. Assume that $C = A + E$, $\|EA^\dagger\| < 1$, and $rank(A) = n$, then $rank(C) = n$.*

3. Local Convergence Analysis of the Method (8)

In this section, we investigate the convergence of the method (8) and determine its convergence radius.

Theorem 1. *Let $F + G : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, be continuous operator, where F is a twice Fréchet differentiable operator and G is a continuous operator on a subset $D \subseteq \mathbb{R}^n$. Assume that the problem (6) has a solution $x_* \in D$ and an operator $A_* = F'(x_*) + G(x_*, x_*)$ has full rank. Suppose that Fréchet derivatives $F'(x)$ and $F''(x)$ satisfy the Lipschitz conditions on D*

$$\|F'(x) - F'(y)\| \leq L\|x - y\|, \tag{9}$$

$$\|F''(x) - F''(y)\| \leq N\|x - y\|, \tag{10}$$

and the function G has the first order divided difference $G(x,y)$ and

$$\|G(x,y)-G(u,v)\| \le M(\|x-u\|+\|y-v\|) \tag{11}$$

for all $x,y,u,v \in D$; L, N, and M are non-negative numbers.

Also, the radius $r>0$ is a root of the equation

$$\beta N p^2 + 120\beta T p + 48\sqrt{2}\alpha\beta^2 T - 24 = 0, \tag{12}$$

where

$$2\sqrt{2}\alpha\beta^2 T < 1. \tag{13}$$

Then, for all $x_0, y_0 \in B(x_*, r) \subseteq D$ the sequences $\{x_k\}$ and $\{y_k\}$, which are generated by the method (8), are well defined, remain in $B(x_*, r)$ for all $k \ge 0$, and converge to x_* such that

$$\rho(x_{k+1}) \le \frac{\beta}{1-\beta T\tau_k}\big((N/24)\rho(x_k)^3 + T\rho(x_k)\rho(y_k) + \sqrt{2}\alpha\beta T\tau_k\big), \tag{14}$$

$$\begin{aligned}\rho(y_{k+1}) \le{}& \frac{\beta}{1-\beta T\tau_k}\big((N/24)\rho(x_{k+1})^3 + \\ &+T(\rho(x_{k+1})+\rho(x_k)+\rho(y_k))\rho(x_{k+1}) + \sqrt{2}\alpha\beta T\tau_k\big),\end{aligned} \tag{15}$$

$$r_{k+1} = \max\{\rho(x_{k+1}), \rho(y_{k+1})\} \le q r_k \le \cdots \le q^{k+1} r_0, \tag{16}$$

where

$$0 < q = \frac{\beta\big((N/24)\rho(x_0)^2 + T(2\rho(x_0)+\rho(y_0)) + 2\sqrt{2}\alpha\beta T\big)}{1-\beta T\tau_0} < 1, \tag{17}$$

$\rho(x) = \|x-x_*\|$, $\tau_k = \tau(x_k, y_k) = \|x_k - x_*\| + \|y_k - x_*\|$, $r_0 = \max\{\rho(x_0), \rho(y_0)\}$, $\alpha = \|F(x_*) + G(x_*)\|$, $\beta = \|(A_*^T A_*)^{-1} A_*^T\|$, $T = \dfrac{L+2M}{2}$, $\beta T\tau_0 < 1$.

Proof. From (13) it follows that (12) has the unique positive root, which we annotate as r.

Let choose arbitrary $x_0, y_0 \in B(x_*, r)$ and denote $A_k = F'\Big(\dfrac{x_k+y_k}{2}\Big) + G(x_k, y_k)$. For $k=0$, we have the following estimate

$$\begin{aligned}\|A_0 - A_*\| &= \left\|F'\Big(\frac{x_0+y_0}{2}\Big) + G(x_0,y_0) - (F'(x_*) + G(x_*,x_*))\right\| = \\ &= \left\|F'\Big(\frac{x_0+y_0}{2}\Big) - F'(x_*) + G(x_0,y_0) - G(x_*,x_*)\right\| \le \\ &\le \left\|F'\Big(\frac{x_0+y_0}{2}\Big) - F'(x_*)\right\| + \|G(x_0,y_0) - G(x_*,x_*)\| \le \\ &\le \frac{L}{2}(\|x_0-x_*\| + \|y_0-x_*\|) + M(\|x_0-x_*\| + \|y_0-x_*\|) \le \\ &\le \frac{L+2M}{2}(\|x_0-x_*\| + \|y_0-x_*\|) = T(\|x_0-x_*\| + \|y_0-x_*\|)\end{aligned}$$

and

$$\|(A_*^T A_*)^{-1} A_*^T [A_0 - A_*]\| \le \beta T(\|x_0-x_*\| + \|y_0-x_*\|) = \beta T\tau_0 < 1.$$

According to Lemma 1

$$\|(A_0^TA_0)^{-1}A_0^T\| \le \frac{\beta}{1-\beta T(\|x_0-x_*\|+\|y_0-x_*\|)} = \frac{\beta}{1-\beta T\tau_0},$$

and to Lemma 2

$$\|(A_0^TA_0)^{-1}A_0^T-(A_*^TA_*)^{-1}A_*^T\| \le \frac{\sqrt{2}\beta^2T(\|x_0-x_*\|+\|y_0-x_*\|)}{1-\beta T(\|x_0-x_*\|+\|y_0-x_*\|)} = \frac{\sqrt{2}\beta^2T\tau_0}{1-\beta T\tau_0}.$$

For x_1, y_1 that are generated by (8), we have

$$\begin{aligned}
x_1 \;-\; & x_* = x_0 - x_* - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_0)+G(x_0)) = \\
= \; & \left[A_0^TA_0\right]^{-1}A_0^T\left[A_0(x_0-x_*)-(F(x_0)+G(x_0))+(F(x_*)+G(x_*))\right]+ \\
& +\left[A_*^TA_*\right]^{-1}A_*^T(F(x_*)+G(x_*)) - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_*)+G(x_*)) = \\
= \; & \left[A_0^TA_0\right]^{-1}A_0^T\left[F'\left(\frac{x_0+x_*}{2}\right)(x_0-x_*)-F(x_0)+F(x_*)+\right. \\
& +G(x_0,x_*)(x_0-x_*)-G(x_0)+G(x_*)+ \\
& \left.+\left(A_0-F'\left(\frac{x_0+x_*}{2}\right)-G(x_0,x_*)\right)(x_0-x_*)\right]+ \\
& +\left[A_*^TA_*\right]^{-1}A_*^T(F(x_*)+G(x_*)) - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_*)+G(x_*));
\end{aligned}$$

$$\begin{aligned}
y_1 \;-\; & x_* = x_1 - x_* - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_1)+G(x_1)) = \\
= \; & \left[A_0^TA_0\right]^{-1}A_0^T\left[A_0(x_1-x_*)-(F(x_1)+G(x_1))+(F(x_*)+G(x_*))\right]+ \\
& +\left[A_*^TA_*\right]^{-1}A_*^T(F(x_*)+G(x_*)) - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_*)+G(x_*)) = \\
= \; & \left[A_0^TA_0\right]^{-1}A_0^T\left[F'\left(\frac{x_1+x_*}{2}\right)(x_1-x_*)-F(x_1)+F(x_*)+\right. \\
& +G(x_1,x_*)(x_1-x_*)-G(x_1)+G(x_*)+ \\
& \left.+\left(A_0-F'\left(\frac{x_1+x_*}{2}\right)-G(x_1,x_*)\right)(x_1-x_*)\right]+ \\
& +\left[A_*^TA_*\right]^{-1}A_*^T(F(x_*)+G(x_*)) - \left[A_0^TA_0\right]^{-1}A_0^T(F(x_*)+G(x_*)).
\end{aligned}$$

According to Lemma 1 from [23] with the value $\omega = 1/2$ we can write

$$F(x)-F(y)-F'\left(\frac{x+y}{2}\right)(x-y) =$$

$$= \frac{1}{4}\int_0^1 (1-t)\left[F''\left(\frac{x+y}{2}+\frac{t}{2}(x-y)\right)-\right.$$

$$\left.-F''\left(\frac{x+y}{2}+\frac{t}{2}(y-x)\right)\right](x-y)^2dt.$$

By setting $x = x_*$ and $y = x_0$ in the equation above, we receive

$$\left\|F(x_*) - F(x_0) - F'\left(\frac{x_0+x_*}{2}\right)(x_* - x_0)\right\| =$$
$$= \frac{1}{4}\left\|\int_0^1 (1-t)\left[F''\left(\frac{x_0+x_*}{2} + \frac{t}{2}(x_* - x_0)\right) -\right.\right.$$
$$\left.\left. - F''\left(\frac{x_0+x_*}{2} + \frac{t}{2}(x_0 - x_*)\right)\right](x_* - x_0)^2 dt\right\| \leq$$
$$\leq \frac{1}{4}\int_0^1 t(1-t)N\|x_0 - x_*\|^3 dt = \frac{1}{24}N\rho(x_0)^3.$$

Using to the Lipschitz conditions (9) and (11), we get the following estimates

$$\begin{aligned}\left\|A_0 - F'\Big(\frac{x_0+x_*}{2}\Big) - G(x_0, x_*)\right\| &= \left\|F'\Big(\frac{x_0+y_0}{2}\Big) - F'\Big(\frac{x_0+x_*}{2}\Big) + \right. \\ &\left. + G(x_0, y_0) - G(x_0, x_*)\right\| \leq T\|y_0 - x_*\|,\end{aligned}$$

$$\begin{aligned}\left\|A_0 - F'\Big(\frac{x_1+x_*}{2}\Big) - G(x_1, x_*)\right\| &= \left\|F'\Big(\frac{x_0+y_0}{2}\Big) - F'\Big(\frac{x_1+x_*}{2}\Big) + \right. \\ &\left. + G(x_0, y_0) - G(x_1, x_*)\right\| \leq \\ &\leq T(\|x_0 - x_1\| + \|y_0 - x_*\|) \leq \\ &\leq T(\|x_0 - x_*\| + \|x_1 - x_*\| + \|y_0 - x_*\|).\end{aligned}$$

Hence, from (12) it follows that

$$\begin{aligned}0 < q &= \frac{\beta\big((N/24)\rho(x_0)^2 + T(2\rho(x_0) + \rho(y_0)) + 2\sqrt{2}\alpha\beta T\big)}{1 - \beta T\tau_0} < \\ &< \frac{\beta\big((N/24)r^2 + 3Tr + 2\sqrt{2}\alpha\beta T\big)}{1 - 2\beta Tr} = 1.\end{aligned}$$

Thus, by Lemmas 1, 2, conditions (9), (10) and (11), we obtain

$$\|x_1 - x_*\| \leq \frac{\beta\big((N/24)\rho(x_0)^3 + T\rho(x_0)\rho(y_0) + \sqrt{2}\alpha\beta T\tau_0\big)}{1 - \beta T\tau_0} \leq qr_0 < r.$$

Similarly,

$$\begin{aligned}\|y_1 - x_*\| &\leq \frac{\beta\big((N/24)\rho(x_1)^3 + T(\rho(x_0) + \rho(x_1) + \rho(y_0))\rho(x_1)\big)}{1 - \beta T\tau_0} + \\ &+ \frac{\sqrt{2}\alpha\beta^2 T\tau_0}{1 - \beta T\tau_0} \leq qr_0 < r.\end{aligned}$$

Therefore, $x_1, y_1 \in B(x_*, r)$ and both (14) and (15) follow. Also, (16) is satisfied

$$r_1 = \max\{\|x_1 - x_*\|, \|y_1 - x_*\|\} \leq qr_0.$$

Using mathematical induction, assume that $x_k, y_k \in B(x_*, r)$ and (16) holds for $k > 0$. Then, for $k+1$ from (8) we obtain that

$$\begin{aligned}\|x_{k+1} - x_*\| &\leq \frac{\beta\big((N/24)\rho(x_k)^3 + T\rho(x_k)\rho(y_k) + \sqrt{2}\alpha\beta T\tau_k\big)}{1 - \beta T\tau_k} \leq \\ &\leq \frac{\beta\big((N/24)\rho(x_0)^2 + T\rho(x_0) + 2\sqrt{2}\alpha\beta T\big)r_k}{1 - \beta T\tau_0} \leq qr_k < r\end{aligned}$$

and

$$\begin{aligned}\|y_{k+1} - x_*\| &\leq \frac{\beta\big((N/24)\rho(x_{k+1})^3 + T(\rho(x_k) + \rho(x_{k+1}) + \rho(y_k))\rho(x_{k+1})\big)}{1 - \beta T\tau_k} + \\ &+ \frac{\sqrt{2}\alpha\beta^2 T\tau_k}{1 - \beta T\tau_k} \leq \frac{\beta\big((N/24)\rho(x_0)^2 + T(2\rho(x_0) + \rho(y_0))\big)r_k}{1 - \beta T\tau_0} + \\ &+ \frac{2\sqrt{2}\alpha\beta^2 T r_k}{1 - \beta T\tau_0} = qr_k < r.\end{aligned}$$

According to (17) and both inequalities (14) and (15), we receive

$$r_{k+1} = \max\{\|x_{k+1} - x_*\|, \|y_{k+1} - x_*\|\} \leq qr_k \leq q^2 r_{k-1} \leq \cdots \leq q^{k+1} r_0.$$

Thus, $x_{k+1}, y_{k+1} \in B(x_*, r)$ as well as (14), (15) and (16) hold. □

From (12) it follows that the convergence radius of the method (8) is

$$r = \frac{2(1 - 2\sqrt{2}\alpha\beta^2 T)}{5\beta T + \sqrt{(5\beta T)^2 + \frac{1}{6}\beta N(1 - 2\sqrt{2}\alpha\beta^2 T)}}.$$

Remark 3. *Note that the condition (11) can be replaced with the weaker one*

$$\|G(x, y) - G(u, v)\| \leq M_1\|x - u\| + M_2\|y - v\| \tag{18}$$

for all $x, y, u, v \in D$, M_1 and M_2 are positive numbers. This enlarges applicability of the method (8).

For zero residual ($F(x_*) + G(x_*) = 0$), the Theorem 1 can be formulated as

Theorem 2. *Let $F + G : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, be continuous operator, where F is a twice Fréchet differentiable operator and G is a continuous operator on a subset $D \subseteq \mathbb{R}^n$. Assume that the problem (6) has a solution $x_* \in D$, and the operator $A_* = F'(x_*) + G(x_*, x_*)$ has full rank. Suppose that Fréchet derivatives $F'(x)$ and $F''(x)$ on D satisfy the classic Lipschitz conditions as in (9) and (10), respectively; the function G has the first order divided difference $G(x, y)$ that satisfies the Lipschitz conditions as in (11). Moreover, the radius $r > 0$ is a unique positive root of the following equation*

$$\beta N p^2 + 120\beta T p - 24 = 0.$$

Then, the combined method (8) converges to x_* for all x_0, $y_0 \in B(x_*, r) \subseteq D$ such that

$$\rho(x_{k+1}) \leq \frac{\beta}{1-\beta T\tau_k}\big((N/24)\rho(x_k)^3 + T\rho(x_k)\rho(y_k)\big), \tag{19}$$

$$\rho(y_{k+1}) \leq \frac{\beta\big((N/24)\rho(x_{k+1})^3 + T(\rho(x_{k+1}) + \rho(x_k) + \rho(y_k))\rho(x_{k+1})\big)}{1-\beta T\tau_k} \tag{20}$$

$$r_{k+1} = \max\{\rho(x_{k+1}), \rho(y_{k+1})\} \leq qr_k \leq \cdots \leq q^{k+1}r_0,$$

where $\rho(x) = \|x - x_*\|$, $\tau_k = \tau(x_k, y_k) = \|x_k - x_*\| + \|y_k - x_*\|$, $r_0 = \max\{\rho(x_0), \rho(y_0)\}$, $\beta = \|(A_*^T A_*)^{-1} A_*^T\|$, $\beta T\tau_0 < 1$,

$$0 < q = \frac{\beta\big((N/24)\rho(x_0)^2 + T(2\rho(x_0) + \rho(y_0))\big)}{1-\beta T\tau_0} < 1.$$

From Theorem 2, the convergence radius is

$$r = \frac{2}{5\beta T + \sqrt{(5\beta T)^2 + \frac{1}{6}\beta N}} < \frac{1}{5\beta T}.$$

This radius is two times smaller than the convergence radius of the differential method (4) from [11] (a two-step modification of the Gauss-Newton method) and equals to the convergence radius of the difference method (5) from [17].

Corollary 1. *Convergence order of the iterative method (8) in case of zero residual is equal to* $1 + \sqrt{2}$.

Proof. Let us denote $\gamma = \dfrac{\beta N/24}{1-\beta T\tau_0}$, $\eta = \dfrac{\beta T}{1-\beta T\tau_0}$, $a_k = \rho(x_k)$, $b_k = \rho(y_k)$, $k = 0, 1, 2, ...$ Since the residual is zero, i.e. $\alpha = \|F(x_*) + G(x_*)\| = 0$, from the inequalities (19) and (20) we have

$$a_{k+1} \leq a_k(\gamma a_k^2 + \eta b_k), \tag{21}$$

$$\begin{aligned} b_{k+1} &\leq a_{k+1}\left[\gamma a_{k+1}^2 + \eta/3(a_k + a_{k+1} + b_k)\right] \leq \\ &\leq a_{k+1}\left[(\gamma a_k + 2\eta/3)a_k + \eta b_k/3\right] \leq \\ &\leq a_{k+1}a_k\left[\gamma r + \eta\right] = a_{k+1}a_k\phi_1. \end{aligned} \tag{22}$$

From (21) and (22) for large enough k, it follows

$$a_{k+1} \leq a_k(\gamma a_k^2 + \eta b_k) \leq a_k(\gamma a_k^2 + \eta\phi_1 a_k a_{k-1}) \leq a_k^2 a_{k-1}(\gamma + \eta\phi_1) = a_k^2 a_{k-1}\phi_2.$$

From this inequality, we obtain an equation

$$\rho^2 - 2\rho - 1 = 0.$$

The positive root of the latter, which is $\rho_* = 1 + \sqrt{2}$, is the order of convergence of the iterative method (8). □

Under the classic Lipschitz condition a theorem for the uniqueness of the solution can be written as follow

Theorem 3. *Suppose x_* satisfies (6) and $F(x)$ has a continuous derivative $F'(x)$ and $G(x)$ has a divided difference $G(x,y)$ in D. Moreover, operator $F'(x_*)+G(x_*,x_*)$ has full rank; $F'(x)$ satisfies the Lipschitz condition as in (9); the divided difference $G(x,y)$ satisfies the Lipschitz condition as in (11). Let $r>0$ satisfies*

$$\beta(Lr/2+M)+\alpha\beta_0(L+2M)\le 1,$$

where $\beta_0=\|(F'(x_*)+G(x_*,x_*))^T(F'(x_*)+G(x_*,x_*))\|$. Then, x_* is a unique solution of the problem (6) in $\overline{B(x_*,r)}$.

The proof of this theorem is analogous to the one in [6].

To note, in case when $G(x)=0$, we obtain the same results as in Theorem 2 in [11].

4. Numerical experiments

In this section, we give two examples to show the application of our results. We consider method (8) and its partial cases, namely the two-step Gauss-Newton method ($G\equiv 0$) and the two-step Secant method ($F\equiv 0$). We use the norm $\|x\|=\sqrt{\sum\limits_{i=1}^{p}x_i^2}$ for $x\in\mathbb{R}^p$.

Example 1. Consider function $F+G: D=\mathbb{R}\to\mathbb{R}^2$ given by [12]:

$$F(x)+G(x)=\begin{pmatrix} x+\mu \\ \lambda x^2+x-\mu \end{pmatrix},$$

where $\lambda,\mu\in\mathbb{R}$ are two parameters.

It is known, that $x_*=0$ is the unique solution of the considered problem. Therefore, we can define constants α and β as follows:

$$\alpha=\sqrt{2}|\mu|,\ \beta=\frac{1}{\sqrt{2}}.$$

Let $G(x)=(0,0)^T$. Then

$$F'(x)=\begin{pmatrix} 1 \\ 2\lambda x+1 \end{pmatrix},\quad F''(x)=\begin{pmatrix} 0 \\ 2\lambda \end{pmatrix}$$

and

$$\|F'(x)-F'(y)\|=\left\|\begin{pmatrix} 0 \\ 2\lambda(x-y) \end{pmatrix}\right\|=2|\lambda||x-y|,$$

$$\|F''(x)-F''(y)\|=\left\|\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right\|=0|x-y|.$$

That is, we can set constants $L=2|\lambda|$, $N=0$, $M=0$, $T=\dfrac{L}{2}=\dfrac{2|\lambda|}{2}=|\lambda|$.

Let $F(x)=(0,0)^T$. Then

$$G(x,y)=\begin{pmatrix} \dfrac{x+\mu-y-\mu}{x-y} \\ \dfrac{\lambda x^2+x-\mu-\lambda y^2-y+\mu}{x-y} \end{pmatrix}=\begin{pmatrix} 1 \\ \lambda(x+y)+1 \end{pmatrix}$$

and

$$\|G(x,y)-G(u,v)\| = \left\| \begin{pmatrix} 0 \\ \lambda(x-u+y-v) \end{pmatrix} \right\| \leq |\lambda|(|x-y|+|u-v|).$$

That is, we can set constants $L=0$, $N=0$, $M=|\lambda|$, $T=M=|\lambda|$.

Then equation (12) for both methods has form

$$5\sqrt{2}|\lambda|r+4|\lambda\mu|-2=0.$$

It has unique positive solution

$$r=\frac{\sqrt{2}-2\sqrt{2}|\lambda\mu|}{5|\lambda|}$$

if parameters λ and μ satisfy

$$\lambda\neq 0,\ |\lambda\mu|<\frac{1}{2}.$$

Let $x_0=0.2$, $y_0=0.2001$. For this problem $A_k=\begin{pmatrix} 1 \\ \lambda(x_k+y_k)+1 \end{pmatrix}$ in both cases. Therefore, we get the same result by the two-step Gauss-Newton method and the two-step Secant method.

TABL. 1. The results for $\lambda=1$, $\mu=0$

k	$\rho(x_{k+1})$	The right side of (14)	$\rho(y_{k+1})$	The right side of (15)
0	1.893e-002	3.946e-002	3.412e-003	7.821e-003
1	3.229e-005	4.640e-005	3.600e-007	5.190e-007
2	5.812e-012	8.220e-012	9.487e-017	1.342e-016
3	0	3.899e-028	0	0

TABL. 2. The results for $\lambda=0.5$, $\mu=0.2$

k	$\rho(x_{k+1})$	The right side of (14)	$\rho(y_{k+1})$	The right side of (15)
0	2.624e-002	6.308e-002	1.881e-002	5.121e-002
1	2.326e-003	4.755e-003	2.230e-003	4.617e-003
2	2.284e-004	4.578e-004	2.274e-004	4.564e-004
3	2.280e-005	4.560e-005	2.279e-005	4.559e-005
4	2.279e-006	4.558e-006	2.279e-006	4.558e-006
5	2.279e-007	4.558e-007	2.279e-007	4.558e-007
6	2.279e-008	4.558e-008	2.279e-008	4.558e-008
7	2.279e-009	4.558e-009	2.279e-009	4.558e-009
8	2.279e-010	4.558e-010	2.279e-010	4.558e-010

If $\lambda=1$ and $\mu=0$ we obtain $2\sqrt{2}\alpha\beta^2T=0<1$, $\beta T\tau_0\approx 0.2829134232<1$, $q\approx 0.5917483231<1$, $r\approx 0.2828427125$ and $B(x_*,r)\subset D$. If $\lambda=0.5$ and $\mu=0.2$ we obtain $2\sqrt{2}\alpha\beta^2T=0.2<1$, $\beta T\tau_0\approx 0.1414567116<1$, $q\approx 0.4800775864<1$, $r\approx 0.4525483400$ and $B(x_*,r)\subset D$. From Tables 1,

2, we can see that sequences $\{x_k\}$ and $\{y_k\}$ converges to the solution x_* and error estimates (14) and (15) are true for all $k \geq 0$.

Example 2. Consider function $F + G : D \subseteq \mathbb{R} \to \mathbb{R}^3$ given by:

$$F(x) + G(x) = \begin{pmatrix} x + \mu \\ \lambda x^3 + x - \mu \\ \lambda|x^2 - 1| - \lambda \end{pmatrix},$$

$$F(x) = \begin{pmatrix} x + \mu \\ \lambda x^3 + x - \mu \\ 0 \end{pmatrix}, G(x) = \begin{pmatrix} 0 \\ 0 \\ \lambda|x^2 - 1| - \lambda \end{pmatrix},$$

where $\lambda, \mu \in \mathbb{R}$ are two parameters.

The unique solution of this problem is $x_* = 0$. Therefore, we can set constants α and β as follows:

$$\alpha = \sqrt{2}|\mu|, \beta = \frac{1}{\sqrt{2}}.$$

Let $D = \{x : |x| < 0.5\}$. Then

$$F'(x) = \begin{pmatrix} 1 \\ 3\lambda x^2 + 1 \\ 0 \end{pmatrix}, \quad F''(x) = \begin{pmatrix} 0 \\ 6\lambda x \\ 0 \end{pmatrix}$$

and

$$\|F'(x) - F'(y)\| = \left\| \begin{pmatrix} 0 \\ 3\lambda(x^2 - y^2) \\ 0 \end{pmatrix} \right\| =$$

$$= 3|\lambda||x + y||x - y| \leq 3|\lambda||x - y|,$$

$$\|F''(x) - F''(y)\| = \left\| \begin{pmatrix} 0 \\ 6\lambda(x - y) \\ 0 \end{pmatrix} \right\| = 6|\lambda||x - y|;$$

$$G(x, y) = \begin{pmatrix} 0 \\ 0 \\ \dfrac{\lambda|x^2 - 1| - \lambda - \lambda|y^2 - 1| + \lambda}{x - y} \end{pmatrix} =$$

$$= \begin{pmatrix} 0 \\ 0 \\ \dfrac{\lambda(1 - x^2 - 1) - \lambda(1 - y^2)}{x - y} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\lambda(x + y) \end{pmatrix}$$

and

$$\|G(x, y) - G(u, v)\| = \left\| \begin{pmatrix} 0 \\ 0 \\ -\lambda(x - u + y - v) \end{pmatrix} \right\| \leq$$

$$\leq |\lambda|(|x - u| + |y - v|).$$

That is, we can set constants $L = 3|\lambda|$, $N = 6|\lambda|$, $M = |\lambda|$, $T = \dfrac{5|\lambda|}{2}$.

Then equation has form

$$\sqrt{2}|\lambda|r^2 + 50\sqrt{2}|\lambda|r + 40|\lambda\mu| - 8 = 0.$$

It has unique positive solution

$$r = \frac{\sqrt{5000|\lambda|^2 - 4\sqrt{2}|\lambda|(40|\lambda\mu| - 8)} - 50\sqrt{2}|\lambda|}{2\sqrt{2}|\lambda|}$$

if parameters λ and μ satisfy

$$\lambda \neq 0,\ |\lambda\mu| < \frac{1}{5}.$$

Let $x_0 = 0.1$, $y_0 = 0.1001$. If $\lambda = 1$ and $\mu = 0$ we obtain $2\sqrt{2}\alpha\beta^2 T = 0 < 1$, $\beta T\tau_0 \approx 0.3537301673 < 1$, $q \approx 0.8236105147 < 1$, $r \approx 0.1128822370$ and $B(x_*, r) \subset D$. If $\lambda = 0.5$ and $\mu = 0.2$ we obtain $2\sqrt{2}\alpha\beta^2 T = 0.5 < 1$, $\beta T\tau_0 \approx 0.1768650836 < 1$, $q \approx 0.9307554564 < 1$, $r \approx 0.1128822370$ and $B(x_*, r) \subset D$.

TABL. 3. The results for $\lambda = 1$, $\mu = 0$

k	$\rho(x_{k+1})$	The right side of (14)	$\rho(y_{k+1})$	The right side of (15)
0	1.002e-003	2.765e-002	1.503e-005	5.509e-004
1	1.216e-010	2.684e-008	1.063e-016	2.189e-013
3	0	2.285e-026	0	0

TABL. 4. The results for $\lambda = 0.5$, $\mu = 0.2$

k	$\rho(x_{k+1})$	The right side of (14)	$\rho(y_{k+1})$	The right side of (15)
0	1.980e-003	7.163e-002	1.494e-003	6.120e-002
1	4.549e-007	8.738e-004	4.526e-007	8.712e-004
2	3.090e-014	2.269e-007	3.090e-014	2.269e-007
3	1.185e-017	1.545e-014	1.185e-017	1.545e-014

Therefore, all conditions in Theorem 1 are satisfied for the two-step method (8). Hence, Theorem 1 applies.

5. CONCLUSIONS

We studied the local convergence of the method (8) for the nonlinear least squares problem with the decomposition of the operator under the classic Lipschitz conditions for the first- and second-order derivatives and for the divided differences of the first order. We determined the convergence order and the radius of the method (8) as well as proved the uniqueness ball of the solution of the nonlinear least squares problem (6). We gave examples that confirm the theoretical results. Furthermore, the method (8) has promising characteristics for parallelization, which we plan to utilize for constructing and developing new parallel methods for solving the problem (6).

Bibliography

1. Argyros I.K. Convergence and Applications of Newton-type Iterations / I. K. Argyros. – New York: Springer-Verlag, 2008.
2. Bartish M.Ia. About applications of a modification of the Gauss-Newton method / M. Ia. Bartish, A. I. Chypurko // Visnyk of Lviv Univ. Ser. Appl. Math. and Infor. – 1999. – Vol. 1. – P. 3-7. (in Ukrainian).
3. Bartish M.Ia. About one iterative method for solving functional equations / M. Ia. Bartish // Dopov. AN URSR. Ser. A. – 1968. – Vol. 30, № 5. – P. 387-391. (in Ukrainian).
4. Bartish M.Ia. About one modification of the Gauss-Newton method / M. Ia. Bartish, A. I. Chypurko, S. M. Shakhno // Visnyk of Lviv Univ. Ser. Mech. Math. – 1995. – Vol. 42. – P. 35-38. (in Ukrainian).
5. Catinas E. On some iterative methods for solving nonlinear equations / E. Catinas // Revue d'Analyse Numerique et de Theorie de l'Approximation. – 1994. – Vol. 23, № 1. – P. 47-53.
6. Chen J. Convergence of Gauss-Newton method's and uniqueness of the solution / J. Chen, W. Li // Applied Mathematics and Computation. – 2005. – Vol. 170. – P. 686-705.
7. Chong C. Convergence behavior of Gauss-Newton's method and extensions of the Smale point estimate theory / C. Chong, N. Hu, J. Wang // Journal of Complexity. – 2010. – Vol. 26. – P. 268-295.
8. Dennis J.M. Numerical Methods for Unconstrained Optimization and Nonlinear Equations / J. M. Dennis, R. B. Schnabel. – New York: Prentice-Hall, 1983.
9. Hernandez M.A. The Secant method for nondifferentiable operators / M. A. Hernandez, M. J. Rubio // Appl. Math. Lett. – 2002. – Vol. 15. – P. 395-399.
10. Iakymchuk R. On the Convergence Analysis of a Two-Step Modification of the Gauss-Newton Method / R. P. Iakymchuk, S. M. Shakhno // PAMM. – 2014. – Vol. 14. – P. 813-814.
11. Iakymchuk R. Convergence analysis of a two-step modification of the Gauss-Newton method and its Applications / R. P. Iakymchuk, S. M. Shakhno, H. P Yarmola // Journal of Numerical and Applied Mathematics. – 2017. – Vol. 126, № 3. – P. 61-74.
12. Ren H. Local convergence of a secant type method for solving least squares problems / H. Ren, I. K. Argyros // AMC (Appl. Math. Comp.). – 2010. – Vol. 217. – P. 3816-3824.
13. Steward G.W. On the continuity of the generalized inverse / G. W. Steward // SIAM J. Appl. Math. – 1960. – Vol. 17, № 1. – P. 33-45.
14. Shakhno S.M. An iterative method for solving nonlinear least squares problems with nondifferentiable operator / S. M. Shakhno, R. P. Iakymchuk, H. P Yarmola // Mat. Stud. – 2017. – Vol. 48, № 1. – P. 97-107.
15. Shakhno S.M. Convergence analysis of combined method for solving nonlinear equations / S. M. Shakhno, I. V. Mel'nyk, H. P Yarmola // J. Math. Sci. – 2016. – Vol. 212. – P. 16-26.
16. Shakhno S.M. Convergence of the two-step combined method and uniqueness of the solution of nonlinear operator equations / S. M. Shakhno // Journal of Computational and Applied Mathematics. – 2014. – **261**. – P. 378-386.
17. Shakhno S.M. On a difference method with superquadratic convergence for solving nonlinear least squares problems / S. M. Shakhno, O. P. Gnatyshyn, R. P. Iakymchuk // Visnyk Lviv. Univ. Ser. Appl. Math. Inform. – 2007. – Vol. 13. – P. 51-58. (in Ukrainian).
18. Shakhno S.M. On an iterative algorithm of order 1.839... for solving the nonlinear least squares problems / S. M. Shakhno, O. P. Gnatyshyn // Applied Mathematics and Computation. – 2005. – Vol. 161. – P. 253-264.
19. Shakhno S. One combined method for solving nonlinear least squares problems / S. Shakhno, Yu. Shunkin // Visnyk Lviv. Univ. Ser. Appl. Math. Inform. – 2017. – Vol. 13. – P. 51-58. (in Ukrainian).
20. Shakhno S.M. On the two-step method for solving nonlinear equations with nondifferentiable operator / S. M. Shakhno, H. P Yarmola // Proc. Appl. Math. Mech. – 2012. – **1**. – P. 617-618.

21. Shakhno S.M. Two-point method for solving nonlinear equations with nondifferentiable operator / S. M. Shakhno, H. P Yarmola // Mat. Stud. – 2011. – Vol. 36, № 2. – P. 213-220. (in Ukrainian).
22. Wedin P.-Å. Perturbation theory for pseudo-inverses / P.-Å. Wedin // BIT Numerical Mathematics. – 1973. – Vol. 13, № 2. – P. 217-232.
23. Werner W. Über ein Verfahren der Ordnung $1+\sqrt{2}$ zur Nullstellenbestimmung / W. Werner // Numer. Math. – 1979. – Vol. 32. – P. 333-342.

S. M. Shakhno, H. P. Yarmola,
Faculty of Applied Mathematics and Informatics,
Ivan Franko National University of Lviv,
1, Universytets'ka Str., Lviv, 79000, Ukraine.

R. P. Iakymchuk,
Department of Computational Science and Technology,
School of Electrical Engineering and Computer Science,
KTH Royal Institute of Technology
5 Lindstedtsvägen 5, 100 44 Stockholm, Sweden.

Received 03.08.2018; revised 16.10.2018

Журнал обчислювальної та прикладної математики | 2018 № 2 (128) | Journal of Numerical & Applied Mathematics

UDC 517.988:519.632

METHOD OF TWO-SIDED APPROXIMATIONS FOR FINDING POSITIVE SOLUTIONS OF BOUNDARY VALUE PROBLEMS FOR SEMILINEAR ELLIPTIC SYSTEMS: THE USE OF THE GREEN-RVACHEV'S QUASI-FUNCTION

M. V. Sidorov

Резюме. Розглядається однорідна задача Діріхле для системи напівлінійних еліптичних рівнянь. Для побудови двобічних наближень до додатного розв'язку цієї системи використовується перехід до еквівалентної системи нелінійних інтегральних рівнянь (за допомогою квазіфункції Гріна-Рвачова) з подальшим її аналізом методами теорії напівупорядкованих просторів. Робота і ефективність розробленого метода продемонстрована обчислювальним експериментом для тестової системи з експоненціальною нелінійністю.

Abstract. A homogeneous Dirichlet problem for a system of semilinear elliptic equations is investigated. To construct two-sided approximations to a positive solution of this system, the transition to an equivalent system of nonlinear integral equations (with the help of the Green-Rvachev's quasi-function) with its subsequent analysis by methods of the theory of semiordered spaces is used. The work and efficiency of the developed method are demonstrated by a computational experiment for a test system with exponential nonlinearity.

1. Introduction

Let us consider the problem of finding a positive solution of a system of n semilinear elliptic equations with a homogeneous Dirichlet condition:

$$-\Delta u_i = f_i(\mathbf{x}, u_1, ..., u_n), \quad \mathbf{x} \in \Omega, \tag{1}$$

$$u_i(\mathbf{x}) > 0, \quad \mathbf{x} \in \Omega, \tag{2}$$

$$u_i|_{\partial\Omega} = 0, \quad i = 1, ..., n, \tag{3}$$

or in a vector form

$$-\Delta \mathbf{u} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \quad \mathbf{x} \in \Omega,$$

$$\mathbf{u} > \boldsymbol{\theta}, \quad \mathbf{x} \in \Omega,$$

$$\mathbf{u}|_{\partial\Omega} = \boldsymbol{\theta},$$

where Ω is a bounded Jordan-measurable domain from $\mathbb{R}^m$ with piecewise smooth boundary $\partial\Omega$ ($\bar{\Omega} = \Omega \cup \partial\Omega$), $\mathbf{x} = (x_1, ..., x_m)$, $\mathbf{u} = (u_1, ..., u_n)$, $-\Delta\mathbf{u} = (-\Delta u_1, ..., -\Delta u_n)$, $\mathbf{f} = (f_1, ..., f_n)$, $\boldsymbol{\theta} = (0, ..., 0)$, Δ is the Laplace operator, $\Delta = \frac{\partial^2}{\partial x_1^2} + ... + \frac{\partial^2}{\partial x_m^2}$.

Let us assume that the functions $f_i(\mathbf{x}, u_1, ..., u_n)$ are continuous and positive for $\mathbf{x} \in \bar{\Omega}$, $u_1, ..., u_n > 0$, for all $i = 1, 2, ..., n$.

Key words. Positive solution; semilinear elliptic systems; heterotone operator; two-sided approach; Green-Rvachev's quasi-function.

The problem (1) – (3) is a mathematical model of many stationary processes, which are considered in chemical kinetics, biology, combustion theory, etc. [12], and the condition of positivity (2) naturally arises from the physical meaning of the functions $u_1, \ldots, u_n$ as the substance concentration, population size, temperature, etc. Many studies are devoted to the investigation of problem (1) – (3) [1,2,6,9,10,12,19, etc.], but the focus in these papers was mainly on elucidating the conditions of the existence and uniqueness of a positive solution of the problem or on the conditions of the presence a solution with radial symmetry for the case when Ω is the unit ball. In the paper [17] for numerical analysis of the problem (1) – (3) a method of two-sided approximations, which consists in the transition to an equivalent system of Hammerstein integral equations with its subsequent investigations by methods of the theory of nonlinear operators in semiordered spaces, in particular, using the theory of heterotone operators developed by V. I. Opoĭcev, was proposed. The method showed effectiveness in solving the test problem, but it has some limitations in practical application. They are related to the fact that an analytic expression for the Green's function must be known. This significantly limits the range of regions Ω, in which a numerical solution can be found, to the cases presented in the reference literature [15].

The purpose of the paper is to develop iterative methods for solving the boundary value problem (1) – (3), which have a two-sided nature of convergence to the desired solution and would not be tied to the presence of a known Green's function. Two-sided approximate methods for solving nonlinear operator equations based on the theory of nonlinear operators in semiordered spaces were developed in [4,5,7,8,13,14]. This paper continues the research begun in [17,18], and extends them to areas of arbitrary geometry.

2. Some information from the theory of nonlinear operators in spaces with cones

Let us give from the theory of nonlinear operators in semiordered spaces some concepts and facts, which will be used further [7,13,14].

Let $\mathcal{E}$ be a real Banach space, θ is a zero element of space $\mathcal{E}$. A closed convex set $\mathcal{K} \subset \mathcal{E}$ is called a cone, if from the fact that $u \in \mathcal{E}$, $u \neq \theta$, follows $\alpha u \in \mathcal{K}$ with $\alpha \geq 0$ and $-u \notin \mathcal{K}$.

Any cone $\mathcal{K} \subset \mathcal{E}$ allows to enter in space $\mathcal{E}$ a semiordering by the rule: $v \leqslant w$, if $w - v \in \mathcal{K}$. The elements $u \geqslant \theta$ (i.e. $u \in \mathcal{K}$) are called positive. The set of elements $< v, w >$ of a semiordered space, which consists of those $u \in \mathcal{E}$ for which $v \leqslant u \leqslant w$, is called a cone segment.

An important class of cones for the applications of the theory of semiordered spaces in computational mathematics is normal cones. A cone $\mathcal{K}$ is called normal if there exists a number $N(\mathcal{K}) > 0$, that from $\theta \leqslant x \leqslant y$ follows $\|x\| \leq N(\mathcal{K})\,\|y\|$. In this case, it is said that the norm is semimonotonic. If $N(\mathcal{K}) = 1$, then the cone is called acute and it is said that the norm is monotonous.

The operator $T : \mathcal{E} \to \mathcal{E}$ is called positive if it leaves invariant the cone $\mathcal{K}$, i.e. $T(u) \in \mathcal{K}$ for any $u \in \mathcal{K}$.

The operator $T : \mathcal{E} \to \mathcal{E}$ is called heterotone (or mixed monotone [3,11, etc.]), if it allows a diagonal representation $T(u) \equiv \hat{T}(u,u)$, where the companion operator $\hat{T} : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$ monotonically increases with respect to the first argument and decreases with respect to the second one, i.e.

a) if $v_1 \leqslant v_2$, then $\hat{T}(v_1, w) \leqslant \hat{T}(v_2, w)$ for all $w \in \mathcal{E}$;
b) if $w_1 \leqslant w_2$, then $\hat{T}(v, w_1) \geqslant \hat{T}(v, w_2)$ for all $v \in \mathcal{E}$.

A cone segment $< v^0, w^0 >$ is called strongly invariant for a heterotone operator T, if

$$\hat{T}(v^0, w^0) \geq v^0, \quad \hat{T}(w^0, v^0) \leq w^0. \tag{4}$$

For the equation $u = T(u)$ with the heterotone operator T, let us form two iterative processes

$$v^{(k+1)} = \hat{T}(v^{(k)}, w^{(k)}), \quad w^{(k+1)} = \hat{T}(w^{(k)}, v^{(k)}), \quad k = 0, 1, 2, ..., \tag{5}$$

starting from the point (v^0, w^0) formed by the ends of the strongly invariant cone segment $< v^0, w^0 >$.

From the heterotony of the operator T for which the operator $\hat{T}$ is a companion one, it follows that the sequence $\{v^{(k)}\}$ does not increase, and the sequence $\{w^{(k)}\}$ does not decrease with respect to the cone $\mathcal{K}$. If the cone $\mathcal{K}$ is normal and the operator $\hat{T}$ is completely continuous, then the limits v^* and w^* of these sequences exist. Thus, the chain of inequalities holds:

$$\begin{aligned} v^0 = v^{(0)} \leqslant v^{(1)} \leqslant ... \leqslant v^{(k)} \leqslant ... \leqslant v^* \leqslant w^* \leqslant ... \leqslant \\ \leqslant w^{(k)} \leqslant ... \leqslant w^{(1)} \leqslant w^{(0)} = w^0. \end{aligned}$$

In this case, two cases are possible: $v^* < w^*$ and $v^* = w^*$. In the second case, $u^* := v^* = w^*$ is the unique on $< v^0, w^0 >$ fixed point of the operator T, that is, it is the unique on $< v^0, w^0 >$ solution of the equation $u = T(u)$.

The elements v^* and w^* are a solution of the system

$$v^{(k+1)} = \hat{T}(v^{(k)}, w^{(k)}), \quad w^{(k+1)} = \hat{T}(w^{(k)}, v^{(k)}), \quad k = 0, 1, 2, ... \tag{6}$$

The equality $v^* = w^*$ will hold if the system (6) does not have on $< v^0, w^0 >$ such solutions (v, w) that $v \neq w$.

Then the results of [7] imply the following fact.

Theorem 1. *Let the cone segment $< v^0, w^0 >$ be strongly invariant for the heterotone operator T for which the operator $\hat{T}$ is a companion one, the cone $\mathcal{K}$ be normal, and the operator $\hat{T}$ be completely continuous. Then the successive approximations, which are formed according to scheme (5), where $v^{(0)} = v^0$, $w^{(0)} = w^0$, converge to the unique on $< v^0, w^0 >$ fixed point u^* of the operator T and the following inequalities*

$$\begin{aligned} v^0 = v^{(0)} \leqslant v^{(1)} \leqslant ... \leqslant v^{(k)} \leqslant ... \leqslant u^* \leqslant ... \leqslant \\ \leqslant w^{(k)} \leqslant ... \leqslant w^{(1)} \leqslant w^{(0)} = w^0 \end{aligned} \tag{7}$$

are satisfied.

The chain of inequalities (7) characterizes the iterative process (5) as a method of two-sided approximations.

The condition that the system (6) does not have on $< v^0, w^0 >$ such solutions (v, w) that $v \neq w$, can be complicated for practical employment. A sufficient condition of the fulfilment of the equality $v^* = w^*$ is the existence of such $\gamma \in (0; 1)$ that

$$\left\| \hat{T}(v, w) - \hat{T}(w, v) \right\| \leq \gamma \left\| v - w \right\| \tag{8}$$

for all $v, w \in < v^0, w^0 >$ [3].

If the condition (8) is satisfied, it is obtained the estimate

$$\left\| w^{(k)} - v^{(k)} \right\| = \left\| \hat{T}(w^{(k-1)}, v^{(k-1)}) - \hat{T}(v^{(k-1)}, w^{(k-1)}) \right\| \leq$$
$$\leq \gamma \left\| w^{(k-1)} - v^{(k-1)} \right\| \leq ... \leq \gamma^k \left\| w^0 - v^0 \right\|.$$

Then, if

$$u^{(k)} = \frac{1}{2}(w^{(k)} + v^{(k)}) \tag{9}$$

is taken as the approximate solution of the operator equation $u = T(u)$ on the k-th iteration, then the following error estimate holds:

$$\left\| u^* - u^{(k)} \right\| \leq \frac{\gamma^k}{2} \left\| w^0 - v^0 \right\|. \tag{10}$$

Thus, the following theorem holds.

Theorem 2. *Let the cone segment $< v^0, w^0 >$ be strongly invariant for the heterotone operator T for which the operator $\hat{T}$ is a companion one, the cone $\mathcal{K}$ be normal, and the operator $\hat{T}$ be completely continuous. Then, if condition (8) is satisfied, the successive approximations that are formed according to the scheme (5), where $v^{(0)} = v^0$, $w^{(0)} = w^0$, two-sided in the sense of (7) converge to the unique on $< v^0, w^0 >$ fixed point u^* of the operator T and for the approximate solution of the form (9) on the k-th iteration the estimate (10) holds.*

From estimation (10) it follows that for a faster convergence of iterations (5) it is necessary to choose a strongly invariant cone segment $< v^0, w^0 >$ of as short as possible length $\left\| w^0 - v^0 \right\|$.

If the accuracy $\varepsilon > 0$ with which it is necessary to find an approximate solution of the equation $u = T(u)$, is given, then, using the estimate (10), from the inequality $\left\| u^* - u^{(k)} \right\| < \varepsilon$, it is obtained that to achieve the specified accuracy it is necessary to do

$$k_0(\varepsilon) = \left[\frac{\ln \frac{\left\| w^0 - v^0 \right\|}{2\varepsilon}}{\ln \frac{1}{\gamma}} \right] + 1 \tag{11}$$

iterations, where the square brackets denote the integer part of the number.

3. Construction of two-sided approximations

To analyze the problem (1) – (3) and construct two-sided approximations to its positive solution, let us use the methods of the theory of nonlinear operators in semiordered spaces [7,13,14] and the Green-Rvachev's quasi-function [16,18].

Let the boundary $\partial\Omega$ of the domain consists of a finite number of pieces of lines $\sigma_i(\mathbf{x}) = 0$, $i = 1, 2, ..., r$, where each $\sigma_i(\mathbf{x})$ is an elementary function. Then with the help of the R-functions method [15] one can construct in the form of a single analytic expression an elementary function $\omega(\mathbf{x})$, which describes the geometry of the region Ω, that is:

a) $\omega(\mathbf{x}) > 0$ in Ω;
b) $\omega(\mathbf{x}) = 0$ on $\partial\Omega$;
c) $|\nabla\omega(\mathbf{x})| \neq 0$ on $\partial\Omega$.

Also, the function $\omega(\mathbf{x})$ can have certain properties of differentiation due to the use of various sufficiently complete systems of R-functions [16].

Definition 7. Let $g_m(r)$ be a fundamental solution of the equation $\Delta u = 0$ in $\mathbb{R}^m$. The Green-Rvachev's quasi-function of the first boundary value problem for the Laplace operator in $\mathbb{R}^m$ is the function

$$Q_m(\mathbf{x}, \boldsymbol{\xi}) = g_m(r) - \tilde{g}_m(\mathbf{x}, \boldsymbol{\xi}), \tag{12}$$

where $\mathbf{x} = (x_1, ..., x_m)$, $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)$,

$$r = |\mathbf{x} - \boldsymbol{\xi}| = \sqrt{\sum_{i=1}^{m} (x_i - \xi_i)^2}, \quad \tilde{g}_m(\mathbf{x}, \boldsymbol{\xi}) = g_m\left(\sqrt{r^2 + 4\omega(\mathbf{x})\omega(\boldsymbol{\xi})}\right),$$

$\omega(\mathbf{x})$ is the function that describes the geometry of the domain Ω.

Let us note [16] that for the case when Ω is a ball of radius R in $\mathbb{R}^m$, and $\omega(\mathbf{x}) = \frac{1}{2R}(R^2 - x_1^2 - ... - x_m^2)$, the Green-Rvachev's quasi-function (12) turns into the exact Green's function of the first boundary value problem for the Laplace operator considered in a ball Ω.

The fundamental solutions of the Laplace equation have the form

$$g_2(r) = \frac{1}{2\pi} \ln \frac{1}{r},$$

$$g_3(r) = \frac{1}{4\pi} \cdot \frac{1}{r},$$

$$g_m(r) = \frac{1}{|S_1|\,(m-2)} \cdot \frac{1}{r^{m-2}}, \quad m > 3,$$

where $|S_1|$ is the area of a single sphere in $\mathbb{R}^m$, consequently, the Green-Rvachev's quasi-function acquires the form

$$Q_2(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{2\pi} \ln \sqrt{1 + \frac{4\omega(\mathbf{x})\omega(\boldsymbol{\xi})}{r^2}} \text{ in } \mathbb{R}^2, \tag{13}$$

$$Q_3(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{4\pi} \cdot \frac{\sqrt{r^2 + 4\omega(\mathbf{x})\omega(\boldsymbol{\xi})} - r}{r\sqrt{r^2 + 4\omega(\mathbf{x})\omega(\boldsymbol{\xi})}} \text{ in } \mathbb{R}^3, \tag{14}$$

$$Q_m(\mathbf{x},\boldsymbol{\xi}) = \frac{1}{|S_1|\,(m-2)} \cdot \frac{\left(r^2 + 4\omega(\mathbf{x})\omega(\boldsymbol{\xi})\right)^{\frac{m}{2}-1} - r^{m-2}}{r^{m-2}\left(r^2 + 4\omega(\mathbf{x})\omega(\boldsymbol{\xi})\right)^{\frac{m}{2}-1}} \text{ in } \mathbb{R}^m, \quad m > 3. \quad (15)$$

From (13) – (15) and Definition 7 the following lemma on the properties of the Green-Rvachev's quasi-function follows.

Lemma 1. *The Green-Rvachev's quasi-function (12) has the following properties:*

a) $Q(\mathbf{x},\boldsymbol{\xi}) = 0$ *on* $\partial\Omega$;
b) *is a symmetric function:* $Q(\mathbf{x},\boldsymbol{\xi}) = Q(\boldsymbol{\xi},\mathbf{x})$;
c) *has the same feature for* $\mathbf{x} = \boldsymbol{\xi}$ *as the usual Green's function;*
d) *is positive in the area* Ω: $Q(\mathbf{x},\boldsymbol{\xi}) > 0$, $\mathbf{x},\boldsymbol{\xi} \in \Omega$, $\mathbf{x} \neq \boldsymbol{\xi}$.

According to [16, 18], from each of the equations (1) let us proceed to an integral equation of the form

$$\begin{aligned} u_i(\mathbf{x}) = &\int_\Omega K_m(\mathbf{x},\boldsymbol{\xi}) u_i(\boldsymbol{\xi}) d\boldsymbol{\xi} + \\ &+ \int_\Omega Q_m(\mathbf{x},\boldsymbol{\xi}) f_i(\boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi})) d\boldsymbol{\xi}, \quad i = 1, ..., n, \end{aligned} \quad (16)$$

where $K_m(\mathbf{x},\boldsymbol{\xi}) = -\frac{\partial^2}{\partial\xi_1^2}\tilde{g}_m(\mathbf{x},\boldsymbol{\xi}) - ... - \frac{\partial^2}{\partial\xi_m^2}\tilde{g}_m(\mathbf{x},\boldsymbol{\xi})$.

The system of equations (16) can be written in the form of a vector equation of Urysohn

$$\mathbf{u}(\mathbf{x}) = \int_\Omega \mathbf{P}(\mathbf{x},\boldsymbol{\xi},\mathbf{u}(\boldsymbol{\xi}))d\boldsymbol{\xi},$$

where

$$\mathbf{P}(\mathbf{x},\boldsymbol{\xi},\mathbf{u}(\boldsymbol{\xi})) = (P_1(\mathbf{x},\boldsymbol{\xi},u_1(\boldsymbol{\xi}),...,u_n(\boldsymbol{\xi})),...,P_n(\mathbf{x},\boldsymbol{\xi},u_1(\boldsymbol{\xi}),...,u_n(\boldsymbol{\xi}))),$$

$$P_i(\mathbf{x},\boldsymbol{\xi},u_1(\boldsymbol{\xi}),...,u_n(\boldsymbol{\xi})) = K_m(\mathbf{x},\boldsymbol{\xi})u_i(\boldsymbol{\xi}) + Q(\mathbf{x},\boldsymbol{\xi})f_i(\boldsymbol{\xi},u_1(\boldsymbol{\xi}),...,u_n(\boldsymbol{\xi}))),$$
$$i = 1, ..., n.$$

If the boundary value problem (1) – (3) has a classical solution, then it also satisfies the system of equations (16). If the classical solution of the problem does not exist, then the system of equations (16) can be used to introduce the concept of a generalized solution of the boundary value problem (1) – (3).

The system of equations (16) will be considered in a Banach space $\mathbf{C}_n(\bar{\Omega}) = \{\mathbf{u} = (u_1, ..., u_n) : u_i \in C(\bar{\Omega}), i = 1, ..., n\}$ of vector functions continuous in $\bar{\Omega}$ with the norm $\|\mathbf{u}\|_n = \max\{\|u_1\|, ..., \|u_n\|\}$, where $\|u_i\| = \max\limits_{\mathbf{x}\in\bar{\Omega}} |u_i(\mathbf{x})|$, $i = 1, ..., n$. Let us select in $\mathbf{C}^n(\bar{\Omega})$ the cone $\boldsymbol{\mathcal{K}}_+ = \{\mathbf{u} = (u_1, ..., u_n) \in \mathbf{C}^n(\bar{\Omega}) : u_i(\mathbf{x}) \geq 0, \mathbf{x} \in \bar{\Omega}, i = 1, ..., n\}$ of vector functions with non-negative coordinates. Note that the cone $\boldsymbol{\mathcal{K}}_+$ in $\mathbf{C}^n(\bar{\Omega})$ is normal (and even acute).

With the help of the cone $\boldsymbol{\mathcal{K}}_+$ in the space $\mathbf{C}^n(\bar{\Omega})$ let us introduce a semiordering by the rule:

$$\text{for } \mathbf{u}, \mathbf{v} \in \mathbf{C}^n(\bar{\Omega}) \quad \mathbf{u} \leqslant \mathbf{v}, \text{ if } \mathbf{v} - \mathbf{u} \in \boldsymbol{\mathcal{K}}_+,$$

that is,

$$\mathbf{u} \leqslant \mathbf{v}, \text{ if } u_i(\mathbf{x}) \leq v_i(\mathbf{x}) \text{ for all } \mathbf{x} \in \bar{\Omega} \text{ and for all } i = 1, ..., n.$$

Definition 8. By a solution (generalized) of the problem (1) – (3) will be meant a vector-valued function $\mathbf{u}^* \in \boldsymbol{\mathcal{K}}_+$, which is a solution of the system of integral equations (16).

Let us construct a process of two-sided approximations for finding the solution of the integral equations system (16) (and consequently, the solution of the boundary value problem (1) – (3)).

Let us introduce a nonlinear integral operator $\mathbf{T}$ acting in $\mathbf{C}_n(\bar{\Omega})$ by the rule, which is determined by the right-hand side of the equations system (16)

$$\mathbf{T}(\mathbf{u})(\mathbf{x}) = \int\limits_{\Omega} \mathbf{P}(\mathbf{x}, \boldsymbol{\xi}, \mathbf{u}(\boldsymbol{\xi}))d\boldsymbol{\xi} = (T_1(\mathbf{u})(\boldsymbol{\xi}), ..., T_n(\mathbf{u})(\boldsymbol{\xi})), \tag{17}$$

where

$$T_i(\mathbf{u})(\mathbf{x}) = \int\limits_{\Omega} P_i(\mathbf{x}, \boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi}))d\boldsymbol{\xi} =$$

$$= \int\limits_{\Omega} K_m(\mathbf{x}, \boldsymbol{\xi})u_i(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_{\Omega} Q_m(\mathbf{x}, \boldsymbol{\xi})f_i(\boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi}))d\boldsymbol{\xi}. \tag{18}$$

The operator $\mathbf{T}$ of the form (17) can be represented as the sum of a linear integral operator $\mathbf{T}_1$ acting in $\mathbf{C}_n(\bar{\Omega})$ by the rule

$$\mathbf{T}_1(\mathbf{u})(\mathbf{x}) = \left(\int\limits_{\Omega} K_1(\mathbf{x}, \boldsymbol{\xi})u_1(\boldsymbol{\xi})d\boldsymbol{\xi}, ..., \int\limits_{\Omega} K_n(\mathbf{x}, \boldsymbol{\xi})u_n(\boldsymbol{\xi})d\boldsymbol{\xi} \right),$$

and a nonlinear Hammerstein operator $\mathbf{T}_2$ acting in $\mathbf{C}_n(\bar{\Omega})$ by the rule

$$\mathbf{T}_2(\mathbf{u})(\mathbf{x}) = \left(\int\limits_{\Omega} Q_m(\mathbf{x}, \boldsymbol{\xi})f_1(\boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi}))d\boldsymbol{\xi}, ..., \right.$$

$$\left. \int\limits_{\Omega} Q_m(\mathbf{x}, \boldsymbol{\xi})f_n(\boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi}))d\boldsymbol{\xi} \right).$$

From the item d) of Lemma 1 it follows that the operator $\mathbf{T}_2$ is a positive operator, because it leaves the cone $\boldsymbol{\mathcal{K}}_+$ invariant, but because there is no assurance in the sign of the function $K_m(\mathbf{x}, \boldsymbol{\xi})$ for $\mathbf{x}, \boldsymbol{\xi} \in \Omega$ $(\mathbf{x} \neq \boldsymbol{\xi})$, the question of the positivity of the operator $\mathbf{T}_1$ is an open one. Therefore, we can not say that the operator $\mathbf{T}$ is positive. However, the operator $\mathbf{T}$ of the form (17) can be represented as a difference of positive operators.

Let us denote

$$K_m^+(\mathbf{x}, \boldsymbol{\xi}) = \max\{0, K_m(\mathbf{x}, \boldsymbol{\xi})\}, \quad K_m^-(\mathbf{x}, \boldsymbol{\xi}) = \max\{0, -K_m(\mathbf{x}, \boldsymbol{\xi})\}.$$

It is clear that $K_m^+(\mathbf{x}, \boldsymbol{\xi}) \geq 0$ and $K_m^-(\mathbf{x}, \boldsymbol{\xi}) \geq 0$ for $\mathbf{x}, \boldsymbol{\xi} \in \Omega$ $(\mathbf{x} \neq \boldsymbol{\xi})$.

Then

$$K_m(\mathbf{x},\boldsymbol{\xi}) = K_m^+(\mathbf{x},\boldsymbol{\xi}) - K_m^-(\mathbf{x},\boldsymbol{\xi}), \quad |K_m(\mathbf{x},\boldsymbol{\xi})| = K_m^+(\mathbf{x},\boldsymbol{\xi}) + K_m^-(\mathbf{x},\boldsymbol{\xi})$$

and the operators T_i, $i = 1, ..., n$, of the form (17) will be written in the form

$$T_i(\mathbf{u})(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi})u_i(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi})u_i(\boldsymbol{\xi})d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi})f_i(\boldsymbol{\xi}, u_1(\boldsymbol{\xi}), ..., u_n(\boldsymbol{\xi}))d\boldsymbol{\xi}, \quad i = 1, ..., n. \tag{19}$$

Suppose that the vector-valued function $\mathbf{f}(\mathbf{x},\mathbf{u})$ allows a diagonal representation $\mathbf{f}(\mathbf{x},\mathbf{u}) = \hat{\mathbf{f}}(\mathbf{x},\mathbf{u},\mathbf{u}) = (\hat{f}_1(\mathbf{x},\mathbf{u},\mathbf{u}), ..., \hat{f}_n(\mathbf{x},\mathbf{u},\mathbf{u}))$, besides, continuous on the sets of variables $\mathbf{x}$, $\mathbf{v}$, $\mathbf{w}$ functions $\hat{f}_i(\mathbf{x},\mathbf{v},\mathbf{w}) = \hat{f}_i(\mathbf{x}, v_1, ..., v_n, w_1, ..., w_n)$ monotonically increase with respect to all v_i and monotonically decrease with respect to all w_i, $i = 1, ..., n$, for all $\mathbf{x} \in \Omega$. Then the operator $\mathbf{T}$ of the form (17) will be heterotone with the companion operator

$$\hat{\mathbf{T}}(\mathbf{v},\mathbf{w})(\mathbf{x}) = (\hat{T}_1(\mathbf{v},\mathbf{w})(\mathbf{x}), ..., \hat{T}_n(\mathbf{v},\mathbf{w})(\mathbf{x})), \tag{20}$$

where

$$\hat{T}_i(\mathbf{v},\mathbf{w})(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi})v_i(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi})w_i(\boldsymbol{\xi})d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi})\hat{f}_i(\boldsymbol{\xi}, v_1(\boldsymbol{\xi}), ..., v_n(\boldsymbol{\xi}), w_1(\boldsymbol{\xi}), ..., w_n(\boldsymbol{\xi}))d\boldsymbol{\xi}, \quad i = 1, ..., n. \tag{21}$$

It is clear that the operators $\mathbf{T}$ and $\hat{\mathbf{T}}$ are completely continuous, and the operator T_i of the form (18) will be heterotone with the companion operator $\hat{T}_i$ of the form (21).

In the cone $\boldsymbol{\mathcal{K}}_+$ let us select a strongly invariant cone segment $< \mathbf{v}^0, \mathbf{w}^0 >$, $\mathbf{v}^0 = (v_1^0, ..., v_n^0)$, $\mathbf{w}^0 = (w_1^0, ..., w_n^0)$, by conditions (4), which for the operator $\hat{\mathbf{T}}$ that is defined by (20), will have the form: for all $\mathbf{x} \in \bar{\Omega}$

$$\int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi})v_i^0(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi})w_i^0(\boldsymbol{\xi})d\boldsymbol{\xi} + \tag{22}$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi})\hat{f}_i(\boldsymbol{\xi}, v_1^0(\boldsymbol{\xi}), ..., v_n^0(\boldsymbol{\xi}), w_1^0(\boldsymbol{\xi}), ..., w_n^0(\boldsymbol{\xi}))d\boldsymbol{\xi} \geq v_i^0(\mathbf{x}), \quad i = 1, ..., n,$$

$$\int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi})w_i^0(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi})v_i^0(\boldsymbol{\xi})d\boldsymbol{\xi} + \tag{23}$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi})\hat{f}_i(\boldsymbol{\xi}, w_1^0(\boldsymbol{\xi}), ..., w_n^0(\boldsymbol{\xi}), v_1^0(\boldsymbol{\xi}), ..., v_n^0(\boldsymbol{\xi}))d\boldsymbol{\xi} \leq w_i^0(\mathbf{x}), \quad i = 1, ..., n.$$

Let us form an iterative process by the scheme (5):

$$v_i^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi}) v_i^{(k)}(\boldsymbol{\xi}) d\xi - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi}) w_i^{(k)}(\boldsymbol{\xi}) d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi}) \hat{f}_i(\boldsymbol{\xi}, v_1^{(k)}(\boldsymbol{\xi}), ..., v_n^{(k)}(\boldsymbol{\xi}), w_1^{(k)}(\boldsymbol{\xi}), ..., w_n^{(k)}(\boldsymbol{\xi})) d\boldsymbol{\xi}, \qquad (24)$$

$$w_i^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi}) w_i^{(k)}(\boldsymbol{\xi}) d\xi - \int\limits_\Omega K_m^-(\mathbf{x},\xi) v_i^{(k)}(\boldsymbol{\xi}) d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi}) \hat{f}_i(\boldsymbol{\xi}, w_1^{(k)}(\boldsymbol{\xi}), ..., w_n^{(k)}(\boldsymbol{\xi}), v_1^{(k)}(\boldsymbol{\xi}), ..., v_n^{(k)}(\boldsymbol{\xi})) d\boldsymbol{\xi}, \qquad (25)$$

$$i = 1, ..., n, \quad k = 0, 1, 2, ...; \qquad (26)$$

$$v_i^{(0)}(\mathbf{x}) = v_i^0(\mathbf{x}), \quad w_i^{(0)}(\mathbf{x}) = w_i^0(\mathbf{x}), \quad i = 1, ..., n. \qquad (27)$$

Taking into account Theorem 1, such conditions for the existence of a unique solution of the problem (1) – (3) and the convergence of successive approximations (24) – (27) to it can be given.

Theorem 3. *Let* $< \mathbf{v}^0, \mathbf{w}^0 >$ *be a strongly invariant cone segment for the heterotone operator* $\mathbf{T}$ *of the form (17) with the companion operator* $\hat{\mathbf{T}}$ *of the form (20) and the system of* $2n$ *integral equations*

$$v_i(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi}) v_i(\boldsymbol{\xi}) d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi}) w_i(\boldsymbol{\xi}) d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi}) \hat{f}_i(\boldsymbol{\xi}, v_1(\boldsymbol{\xi}), ..., v_n(\boldsymbol{\xi}), w_1(\boldsymbol{\xi}), ..., w_n(\boldsymbol{\xi})) d\boldsymbol{\xi}, \quad i = 1, ..., n,$$

$$w_i(\mathbf{x}) = \int\limits_\Omega K_m^+(\mathbf{x},\boldsymbol{\xi}) w_i(\boldsymbol{\xi}) d\boldsymbol{\xi} - \int\limits_\Omega K_m^-(\mathbf{x},\boldsymbol{\xi}) v_i(\boldsymbol{\xi}) d\boldsymbol{\xi} +$$

$$+ \int\limits_\Omega Q_m(\mathbf{x},\boldsymbol{\xi}) \hat{f}_i(\boldsymbol{\xi}, w_1(\boldsymbol{\xi}), ..., w_n(\boldsymbol{\xi}), v_1(\boldsymbol{\xi}), ..., v_n(\boldsymbol{\xi})) d\boldsymbol{\xi}, \quad i = 1, ..., n,$$

does not have on $< \mathbf{v}^0, \mathbf{w}^0 >$ *solutions such that* $\mathbf{v} \neq \mathbf{w}$. *Then the iterative process (24) – (27) converges in the norm of the space* $\mathbf{C}_n(\bar{\Omega})$ *to the unique on* $< \mathbf{v}^0, \mathbf{w}^0 >$ *continuous positive solution* $\mathbf{u}^*$ *of the boundary value problem (1) – (3), and a chain of inequalities hold:*

$$\mathbf{v}^0 = \mathbf{v}^{(0)} \leqslant \mathbf{v}^{(1)} \leqslant ... \leqslant \mathbf{v}^{(k)} \leqslant ... \leqslant \mathbf{u}^* \leqslant ... \leqslant \mathbf{w}^{(k)} \leqslant ... \leqslant \mathbf{w}^{(1)} \leqslant \mathbf{w}^{(0)} = \mathbf{w}^0.$$

Let us now use Theorem 2. Let for each i, $i = 1, ..., n$, there exist such number $L_i > 0$ that the function $\hat{f}_i(\mathbf{x}, v_1, ..., v_n, w_1, ..., w_n)$ for all numbers v_1,

$\ldots$, v_n, w_1, $\ldots$, w_n such that $0 < v_i, w_i < M_0^i$, where $M_0^i = \max\limits_{\mathbf{x}\in\bar{\Omega}} w_i^0(\mathbf{x})$, $i = 1, ..., n$, and for all $\mathbf{x} \in \Omega$ satisfies the inequality

$$\left|\hat{f}_i(\mathbf{x}, v_1, ..., v_n, w_1, ..., w_n) - \hat{f}_i(\mathbf{x}, w_1, ..., w_n, v_1, ..., v_n)\right| \leq$$
$$\leq L_i \max\{|v_1 - w_1|, ..., |v_n - w_n|\}. \qquad (28)$$

Let us consider for an arbitrary i, $i = 1, ..., n$, the difference $\hat{T}_i(\mathbf{w}, \mathbf{v})(\mathbf{x}) - \hat{T}_i(\mathbf{v}, \mathbf{w})(\mathbf{x})$:

$$\hat{T}_i(\mathbf{w}, \mathbf{v})(\mathbf{x}) - \hat{T}_i(\mathbf{v}, \mathbf{w})(\mathbf{x}) = \int\limits_{\Omega} [K_m^+(\mathbf{x}, \boldsymbol{\xi}) + K_m^-(\mathbf{x}, \boldsymbol{\xi})][w_i(\boldsymbol{\xi}) - v_i(\boldsymbol{\xi})]d\boldsymbol{\xi} +$$
$$+ \int\limits_{\Omega} Q_m(\mathbf{x}, \mathbf{s})[\hat{f}_i(\boldsymbol{\xi}, w_1(\boldsymbol{\xi}), ..., w_n(\boldsymbol{\xi}), v_1(\boldsymbol{\xi}), ..., v_n(\boldsymbol{\xi})) -$$
$$- \hat{f}_i(\boldsymbol{\xi}, v_1(\boldsymbol{\xi}), ..., v_n(\boldsymbol{\xi}), w_1(\boldsymbol{\xi}), ..., w_n(\boldsymbol{\xi}))]d\boldsymbol{\xi}.$$

Then, taking into account the inequality (28), we get an estimate

$$\left\|\hat{\mathbf{T}}(\mathbf{w}, \mathbf{v}) - \hat{\mathbf{T}}(\mathbf{v}, \mathbf{w})\right\|_n = \max_{i=1,...,n} \max_{\mathbf{x}\in\bar{\Omega}} \left|\hat{T}_i(\mathbf{w}, \mathbf{v})(\mathbf{x}) - \hat{T}_i(\mathbf{v}, \mathbf{w})(\mathbf{x})\right| \leq$$
$$\leq \max_{i=1,...,n} \{M_1 + L_i M\} \cdot \max_{i=1,...,n} \max_{\mathbf{x}\in\bar{\Omega}} |w_i(\mathbf{x}) - v_i(\mathbf{x})| = (M_1 + LM)\|\mathbf{w} - \mathbf{v}\|_n,$$

where

$$M = \max_{\mathbf{x}\in\bar{\Omega}} \int\limits_{\Omega} Q_m(\mathbf{x}, \boldsymbol{\xi})d\boldsymbol{\xi}, \qquad (29)$$

$$M_1 = \max_{\mathbf{x}\in\bar{\Omega}} \int\limits_{\Omega} [K_m^+(\mathbf{x}, \boldsymbol{\xi}) + K_m^-(\mathbf{x}, \boldsymbol{\xi})]d\boldsymbol{\xi}, \qquad (30)$$

$$L = \max_{i=1,...,n} L_i. \qquad (31)$$

Therefore,

$$\left\|\hat{\mathbf{T}}(\mathbf{w}, \mathbf{v}) - \hat{\mathbf{T}}(\mathbf{v}, \mathbf{w})\right\|_n \leq \gamma\|\mathbf{w} - \mathbf{v}\|_n,$$

where $\gamma = M_1 + LM$.

Thus, the following theorem holds.

Theorem 4. *Let $< \mathbf{v}^0, \mathbf{w}^0 >$ be a strongly invariant cone segment for the heterotone operator $\mathbf{T}$ of the form (17) with the companion operator $\hat{\mathbf{T}}$ of the form (20) and the condition (28) holds, besides, $\gamma = M_1 + LM < 1$, where the constants M, M_1 and L are defined by the equalities (29), (30) and 31) respectively. Then, the iterative process (24) – (27) two-sided converges in the norm of the space $\mathbf{C}_n(\bar{\Omega})$ to the unique on $< \mathbf{v}^0, \mathbf{w}^0 >$ continuous positive solution $\mathbf{u}^*$ of the boundary value problem (1) – (3).*

On the k-th iteration, in accordance with (9), as an approximate solution of the boundary value problem (1) – (3) the vector function

$$\mathbf{u}^{(k)}(\mathbf{x}) = \frac{1}{2}(\mathbf{w}^{(k)}(\mathbf{x}) + \mathbf{v}^{(k)}(\mathbf{x}))$$

is accepted.

Then there will be a posteriori estimate of the error of the approximation $\mathbf{u}^{(k)}(\mathbf{x})$:

$$\left\|\mathbf{u}^* - \mathbf{u}^{(k)}\right\|_n \le \frac{1}{2} \max_{i=1,\dots,n} \max_{\mathbf{x}\in\bar{\Omega}} (w_i^{(k)}(\mathbf{x}) - v_i^{(k)}(\mathbf{x})).$$

If the accuracy $\varepsilon > 0$ is given, then the iterative process should be carried out until the inequality

$$\max_{i=1,\dots,n} \max_{\mathbf{x}\in\bar{\Omega}} (w_i^{(k)}(\mathbf{x}) - v_i^{(k)}(\mathbf{x})) < 2\varepsilon \tag{32}$$

will be satisfied and then with an accuracy ε it can be expected that $\mathbf{u}^*(\mathbf{x}) \approx \mathbf{u}^{(k)}(\mathbf{x})$.

If the conditions of Theorem 4 are satisfied, then an a priori estimate of the error will be:

$$\left\|\mathbf{u}^* - \mathbf{u}^{(k)}\right\|_n \le \frac{\gamma^k}{2} \max_{i=1,\dots,n} \max_{\mathbf{x}\in\bar{\Omega}} (w_i^0(\mathbf{x}) - v_i^0(\mathbf{x})),$$

from which it is obtained that to achieve the accuracy ε it is necessary to do

$$k_0(\varepsilon) = \left[\frac{\ln \frac{\max\limits_{i=1,\dots,n} \max\limits_{\mathbf{x}\in\bar{\Omega}} (w_i^0(\mathbf{x}) - v_i^0(\mathbf{x}))}{2\varepsilon}}{\ln \frac{1}{\gamma}} \right] + 1 \tag{33}$$

iterations, where the square brackets denote the integer part of the number.

4. Numerical experiment

The construction of two-sided approximations to the solution of the boundary value problem (1) – (3) will be demonstrated on the system of two equations with exponential nonlinearities:

$$-\Delta u_1 = e^{u_2}, \quad -\Delta u_2 = e^{-u_1}, \quad \mathbf{x} \in \Omega, \tag{34}$$

$$u_1(\mathbf{x}) > 0, \quad u_2(\mathbf{x}) > 0, \quad \mathbf{x} \in \Omega, \tag{35}$$

$$u_1|_{\partial\Omega} = u_2|_{\partial\Omega} = 0, \tag{36}$$

where $\Omega = \{\mathbf{x} = (x_1, x_2) : 0 < x_1, x_2 < 1\}$.

The functions $f_1(\mathbf{x}, u_1, u_2) = e^{u_2}$, $f_2(\mathbf{x}, u_1, u_2) = e^{-u_1}$ are positive and continuous with respect to the set of variables, if $u_1, u_2 > 0$ and allow a diagonal representation with the help of functions

$$\hat{f}_1(\mathbf{x}, v_1, v_2, w_1, w_2) = e^{v_2}, \quad \hat{f}_2(\mathbf{x}, v_1, v_2, w_1, w_2) = e^{-w_1}. \tag{37}$$

The problem (34) – (36) is replaced by an equivalent system of integral equations

$$u_1(\mathbf{x}) = \int_\Omega K_2(\mathbf{x}, \boldsymbol{\xi}) u_1(\boldsymbol{\xi}) d\boldsymbol{\xi} + \int_\Omega Q_2(\mathbf{x}, \boldsymbol{\xi}) e^{u_2(\boldsymbol{\xi})} d\boldsymbol{\xi}, \tag{38}$$

$$u_2(\mathbf{x}) = \int_\Omega K_2(\mathbf{x}, \boldsymbol{\xi}) u_2(\boldsymbol{\xi}) d\boldsymbol{\xi} + \int_\Omega Q_2(\mathbf{x}, \boldsymbol{\xi}) e^{-u_1(\boldsymbol{\xi})} d\boldsymbol{\xi}, \tag{39}$$

where $Q_2(\mathbf{x},\boldsymbol{\xi})$ is determined by the formula (13),

$$K_2(\mathbf{x},\boldsymbol{\xi}) = -\frac{\partial^2}{\partial\xi_1^2}\tilde{g}_2(\mathbf{x},\boldsymbol{\xi}) - \frac{\partial^2}{\partial\xi_2^2}\tilde{g}_2(\mathbf{x},\boldsymbol{\xi}),$$

$$\tilde{g}_2(\mathbf{x},\boldsymbol{\xi}) = \frac{1}{2\pi}\ln\frac{1}{\sqrt{r^2+4\omega(\mathbf{x})\omega(\boldsymbol{\xi})}},$$

$$\omega(\mathbf{x}) = [x_1(1-x_1)]\wedge_0[x_2(1-x_2)] \equiv$$
$$\equiv x_1(1-x_1)+x_2(1-x_2)-\sqrt{x_1^2(1-x_1)^2+x_2^2(1-x_2)^2}.$$

With the system (38) – (39) let us associate a heterotone operator

$$\mathbf{T}(u_1,u_2) = \left(\int\limits_\Omega K_2(\mathbf{x},\boldsymbol{\xi})u_1(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{u_2(\boldsymbol{\xi})}d\boldsymbol{\xi},\right.$$
$$\left.\int\limits_\Omega K_2(\mathbf{x},\boldsymbol{\xi})u_2(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-u_1(\boldsymbol{\xi})}d\boldsymbol{\xi}\right), \tag{40}$$

for which the companion operator has the form

$$\hat{\mathbf{T}}(v_1,v_2,w_1,w_2) = \left(\int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})v_1(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})w_1(\boldsymbol{\xi})d\boldsymbol{\xi} + \right.$$
$$+\int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{v_2(\boldsymbol{\xi})}d\boldsymbol{\xi},\ \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})v_2(\boldsymbol{\xi})d\boldsymbol{\xi} -$$
$$\left.-\int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})w_2(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-w_1(\boldsymbol{\xi})}d\boldsymbol{\xi}\right),$$

where

$$K_2^+(\mathbf{x},\boldsymbol{\xi}) = \max\{0,K_2(\mathbf{x},\boldsymbol{\xi})\},\quad K_2^-(\mathbf{x},\boldsymbol{\xi}) = \max\{0,-K_2(\mathbf{x},\boldsymbol{\xi})\}.$$

For the operator $\mathbf{T}$ of the form (40) a strongly invariant cone segment will be sought in the form $< \mathbf{v}^0,\mathbf{w}^0 >$, where $\mathbf{v}^0(\mathbf{x}) = (v_1^0(\mathbf{x}),v_2^0(\mathbf{x})) = (\alpha_1\omega(\mathbf{x}),\alpha_2\omega(\mathbf{x}))$, $\mathbf{w}^0(\mathbf{x}) = (w_1^0(\mathbf{x}),w_2^0(\mathbf{x})) = (\beta_1\omega(\mathbf{x}),\beta_2\omega(\mathbf{x}))$, $0<\alpha_1<\beta_1$, $0<\alpha_2<\beta_2$.

For the chosen vector-valued functions $\mathbf{v}^0$, $\mathbf{w}^0$ the system of inequalities (22), (23) for determining the constants α_1, α_2, β_1, β_2 has the form: for all $\mathbf{x}\in\bar{\Omega}$

$$\alpha_1\int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} - \beta_1\int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} +$$
$$+\int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{\alpha_2\omega(\boldsymbol{\xi})}d\boldsymbol{\xi} \geq \alpha_1\omega(\mathbf{x}),$$

$$\alpha_2 \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} - \beta_2 \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-\beta_1\omega(\boldsymbol{\xi})}d\boldsymbol{\xi} \geq \alpha_2\omega(\mathbf{x}),$$

$$\beta_1 \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} - \alpha_1 \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{\beta_2\omega(\boldsymbol{\xi})}d\boldsymbol{\xi} \leq \beta_1\omega(\mathbf{x}),$$

$$\beta_2 \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} - \alpha_2 \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})\omega(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-\alpha_1\omega(\boldsymbol{\xi})}d\boldsymbol{\xi} \leq \beta_2\omega(\mathbf{x}).$$

These inequalities are satisfied, for example, by the numbers $\alpha_1 = 0,01$, $\alpha_2 = 0,01$, $\beta_1 = 0,59$, $\beta_2 = 0,55$.

Because for $0 < v_1, w_1 < \frac{\sqrt{2}-1}{2\sqrt{2}}\beta_1$, $0 < v_2, w_2 < \frac{\sqrt{2}-1}{2\sqrt{2}}\beta_2$ $(\max\limits_{\mathbf{x}\in\bar{\Omega}}\omega(\mathbf{x}) = \frac{\sqrt{2}-1}{2\sqrt{2}})$

$$\left|\hat{f}_1(\mathbf{x},v_1,v_2,w_1,w_2) - \hat{f}_1(\mathbf{x},w_1,w_2,v_1,v_2)\right| = |e^{v_2} - e^{w_2}| \leq$$

$$\leq e^{\frac{\sqrt{2}-1}{2\sqrt{2}}\beta_2}|v_2 - w_2| \leq e^{\frac{\sqrt{2}-1}{2\sqrt{2}}\beta_2}\max\{|v_1 - w_1|, |v_2 - w_2|\},$$

$$\left|\hat{f}_2(\mathbf{x},v_1,v_2,w_1,w_2) - \hat{f}_2(\mathbf{x},w_1,w_2,v_1,v_2)\right| = \left|e^{-w_1} - e^{-v_1}\right| \leq$$

$$\leq |v_2 - w_2| \leq \max\{|v_1 - w_1|, |v_2 - w_2|\},$$

then

$$L = \max\left\{e^{\frac{\sqrt{2}-1}{2\sqrt{2}}\beta_2}, 1\right\} = \max\{1,08388; 1\} = 1,08388.$$

Further we find

$$M = \max_{\mathbf{x}\in\bar{\Omega}} \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})d\boldsymbol{\xi} = 0,04093,$$

$$M_1 = \max_{\mathbf{x}\in\bar{\Omega}} \int\limits_\Omega [K_2^+(\mathbf{x},\boldsymbol{\xi}) + K_2^-(\mathbf{x},\boldsymbol{\xi})]d\boldsymbol{\xi} = 0,70819,$$

$$\gamma = M_1 + LM = 0,753.$$

Thus, $\gamma < 1$ and by Theorem 4, the successive approximations that are formed by the scheme

$$v_1^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})v_1^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})w_1^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{v_2^{(k)}(\boldsymbol{\xi})}d\boldsymbol{\xi},$$

$$v_2^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})v_2^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})w_2^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} + \\ + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-w_2^{(k)}(\boldsymbol{\xi})}d\boldsymbol{\xi},$$

$$w_1^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})w_1^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})v_1^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} + \\ + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{w_2^{(k)}(\boldsymbol{\xi})}d\boldsymbol{\xi},$$

$$w_2^{(k+1)}(\mathbf{x}) = \int\limits_\Omega K_2^+(\mathbf{x},\boldsymbol{\xi})w_2^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} - \int\limits_\Omega K_2^-(\mathbf{x},\boldsymbol{\xi})v_2^{(k)}(\boldsymbol{\xi})d\boldsymbol{\xi} + \\ + \int\limits_\Omega Q_2(\mathbf{x},\boldsymbol{\xi})e^{-v_2^{(k)}(\boldsymbol{\xi})}d\boldsymbol{\xi}, \quad k = 0,1,2,...,$$

$$v_1^{(0)}(\mathbf{x}) = \alpha_1\omega(\mathbf{x}), \quad v_2^{(0)}(\mathbf{x}) = \alpha_2\omega(\mathbf{x}),$$
$$w_1^{(0)}(\mathbf{x}) = \beta_1\omega(\mathbf{x}), \quad w_2^{(0)}(\mathbf{x}) = \beta_2\omega(\mathbf{x}),$$

two-sided converge to the solution of problem (34) – (36).

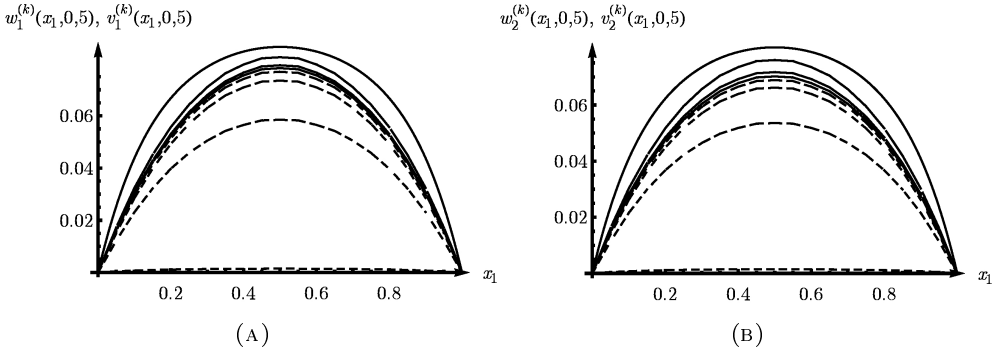
TABL. 1. The values of the estimate of the approximate solution error

Iteration number k	$\varepsilon_1^{(k)}$	$\varepsilon_2^{(k)}$
0	$0,42 \cdot 10^{-1}$	$0,40 \cdot 10^{-1}$
1	$0,23 \cdot 10^{-1}$	$0,22 \cdot 10^{-1}$
2	$0,12 \cdot 10^{-1}$	$0,11 \cdot 10^{-1}$
3	$0,60 \cdot 10^{-2}$	$0,56 \cdot 10^{-2}$
4	$0,29 \cdot 10^{-2}$	$0,28 \cdot 10^{-2}$
5	$0,14 \cdot 10^{-2}$	$0,13 \cdot 10^{-2}$
6	$0,70 \cdot 10^{-3}$	$0,66 \cdot 10^{-3}$
7	$0,34 \cdot 10^{-3}$	$0,32 \cdot 10^{-3}$
8	$0,17 \cdot 10^{-3}$	$0,16 \cdot 10^{-3}$
9	$0,80 \cdot 10^{-4}$	$0,76 \cdot 10^{-4}$

Let us choose $\varepsilon = 10^{-4}$. Then, in accordance with (33), to achieve this accuracy, it is necessary to make $k_0(\varepsilon) = \left[\frac{\ln \frac{\max\{\beta_1,\beta_2\}}{2\varepsilon}}{\ln \frac{1}{\gamma}}\right] + 1 = 28$ iterations. In fact, the accuracy $\varepsilon = 10^{-4}$ was achieved at the ninth iteration. As one can see, the theoretical error estimate turned out to be greatly overestimated. As an approximate solution of problem (34) – (36), the functions $u_1^{(9)}(\mathbf{x}) = \frac{v_1^{(9)}(\mathbf{x})+w_1^{(9)}(\mathbf{x})}{2}$, $u_2^{(9)}(\mathbf{x}) = \frac{v_2^{(9)}(\mathbf{x})+w_2^{(9)}(\mathbf{x})}{2}$ will be accepted.

TABL. 2. The values of the approximate solution in points $\mathbf{x}_i = (0,1i;0,5)$, $i = 0,1,...,10$

$\mathbf{x}_i = (0,1i;0,5)$	$u_1^{(9)}(\mathbf{x}_i)$	$u_2^{(9)}(\mathbf{x}_i)$
$(0;0,5)$	0	0
$(0,1;0,5)$	$0,0301$	$0,0274$
$(0,2;0,5)$	$0,0520$	$0,0471$
$(0,3;0,5)$	$0,0666$	$0,0599$
$(0,4;0,5)$	$0,0751$	$0,0672$
$(0,5;0,5)$	$0,0778$	$0,0696$
$(0,6;0,5)$	$0,0751$	$0,0672$
$(0,7;0,5)$	$0,0666$	$0,0599$
$(0,8;0,5)$	$0,0520$	$0,0471$
$(0,9;0,5)$	$0,0301$	$0,0274$
$(1;0,5)$	0	0

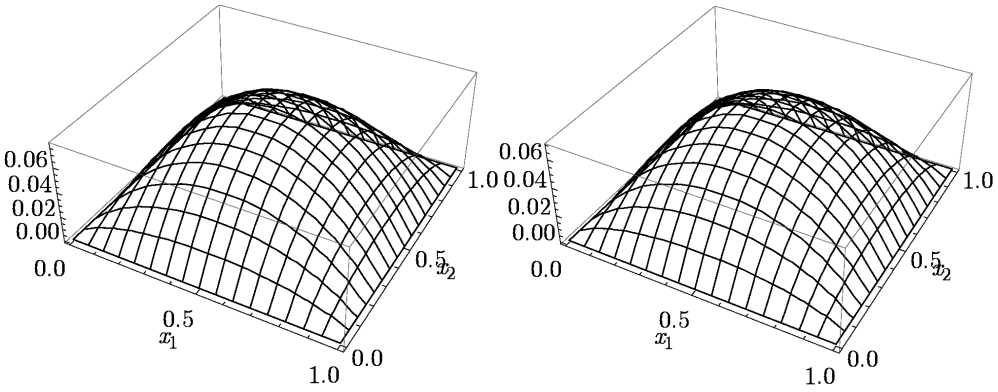


(A) (B)

FIG. 1. Graphs of the cross-sections of upper and lower approximations $w_1^{(k)}(x_1,0,5)$, $v_1^{(k)}(x_1,0,5)$ (a) and $w_2^{(k)}(x_1,0,5)$, $v_2^{(k)}(x_1,0,5)$ (b), $k = 0,2,6,8$

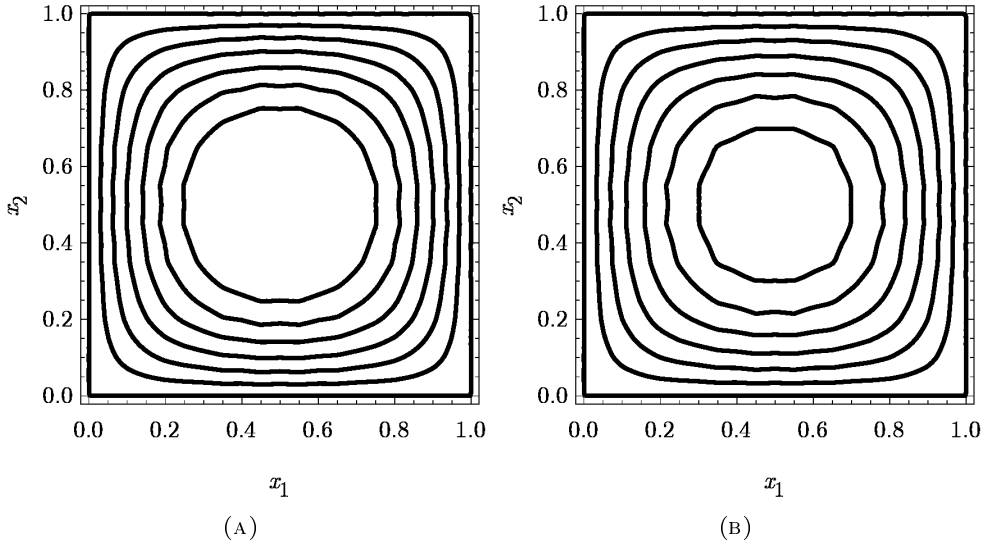
Table 1 gives the data how the estimate $\varepsilon_i^{(k)} = \max\limits_{\mathbf{x}\in\bar{\Omega}} \frac{1}{2}(w_i^{(k)}(\mathbf{x}) - v_i^{(k)}(\mathbf{x}))$ of the norm of the error $\left\|u_i^* - u_i^{(k)}\right\|$ of the approximate solution $u_i^{(k)}(\mathbf{x})$, $i = 1,2$, varies depending on the iteration number k, $k = 0,1,...,9$. Table 2 shows the values, found with accuracy $\varepsilon = 10^{-4}$ of the approximate solution $u_1^{(9)}(\mathbf{x})$, $u_2^{(9)}(\mathbf{x})$ at the points located on the straight line $x_2 = 0,5$ with the step $0,1$, and also it was found that $\left\|u_1^{(9)}\right\| = 0,0778$, $\left\|u_2^{(9)}\right\| = 0,0696$.

Fig. 1 shows the graphs of the cross-sections of the upper $w_1^{(k)}(\mathbf{x})$, $w_2^{(k)}(\mathbf{x})$ and the lower $v_1^{(k)}(\mathbf{x})$, $v_2^{(k)}(\mathbf{x})$ approximations at $x_2 = 0,5$ for $k = 0,2,6,8$. Fig. 2, 3 show the surfaces of the approximate solutions $u_1^{(9)}(\mathbf{x})$, $u_2^{(9)}(\mathbf{x})$ and



(A) (B)

FIG. 2. Graphs of the approximate solutions $u_1^{(9)}(\mathbf{x})$ (a) and $u_2^{(9)}(\mathbf{x})$ (b)



(A) (B)

FIG. 3. Contour lines of the approximate solutions $u_1^{(9)}(\mathbf{x})$ (a) and $u_2^{(9)}(\mathbf{x})$ (b)

their contour lines (with the step $0,01$) respectively. Considering the relationship $\frac{\varepsilon_i^{(k+1)}}{\varepsilon_i^{(k)}}$, $k = 0, 1, ..., 10$, $i = 1, 2$, according to the table 1, it was received that $\frac{\varepsilon_1^{(k+1)}}{\varepsilon_1^{(k)}} \approx \frac{\varepsilon_2^{(k+1)}}{\varepsilon_2^{(k)}} \approx 0,486$, that indicates the geometric rate of convergence of the iterative sequence with the corresponding index. Let us note that the

convergence exponent turned out to be less than the exponent γ estimated in accordance with Theorem 4.

5. Conclusions

In the paper a method of two-sided approximations of the solution of the homogeneous Dirichlet problem for a system of semilinear elliptic equations is proposed on the basis of the Green-Rvachev's quasi-function method. A computational experiment carried out for a system with exponential nonlinearity demonstrated the possibilities and effectiveness of the method. The proposed approach to the numerical solution of semilinear systems can be used in solving various applied problems, the mathematical model of which is the problem (1) – (3). The proposed method is more universal than the existing methods, and it allows to solve the problem in question in areas of arbitrary geometry, provided that this region can be described by the R-function method.

Bibliography

1. Cui R. Uniqueness of the positive solution for a class of semilinear elliptic systems / R. Cui, Y. Wang, J. Shi // Nonlinear Analysis: Theory, Methods & Applications. – 2007. – Vol. 67, № 6. – P. 1710-1714.
2. Dalmasso R. Existence and uniqueness of positive radial solutions for the Lane-Emden system / R. Dalmasso // Nonlinear Analysis: Theory, Methods & Applications. – 2004. – Vol. 57, № 3. – P. 341-348.
3. Guo D. Coupled fixed points of nonlinear operators with applications / D. Guo, V. Lakshmikantham // Nonlinear Analysis: Theory, Methods & Applications. – 1987. – Vol. 11, № 5. – P. 623-632.
4. Kolosov A.I. A class of boundary value problems reducible to an equation with a heterotonic operator / A.I. Kolosov // Differentsial'nye Uravneniya. – 1985. – Vol. 21, № 11. – P. 1884-1891.
5. Kolosova S.V. On the construction of two-sided approximations to the positive solution of the Lane-Emden equation / S.V. Kolosova, V.S. Lukhanin, M.V. Sidorov // Visnyk of Zaporizhzhya National University. Physical and mathematical Sciences. – 2015. – No. 3. – P. 107-120. (in Russian).
6. Korman P. On Lane-Emden type systems / P. Korman, J. Shi // Discrete Contin. Dyn. Syst. – 2005. – P. 510-517.
7. Krasnosel'skij M.A. Positive Solutions of Operator Equations / M.A. Krasnosel'skij. – Moscow: Fizmatgiz, 1962. (in Russian).
8. Kurpel' N.S. Two-sided Operator Inequalities and their Application / N.S. Kurpel', B.A. Shuvar. – Kiev: Naukova Dumka, 1980. (in Russian).
9. Li C. A degree theory framework for semilinear elliptic systems / C. Li, J. Villavert // Proceedings of the American Mathematical Society. – 2016. – Vol. 144, № 9. – P. 3731-3740.
10. Maniwa M. Uniqueness and existence of positive solutions for some semilinear elliptic systems / M. Maniwa // Nonlinear Analysis: Theory, Methods & Applications. – 2004. – Vol. 59, № 6. – P. 993-999.
11. Moore J. Existence of multiple quasifixed points of mixed monotone operators by iterative techniques / J. Moore // Applied Mathematics and Computation. – 1981. – Vol. 9, № 2. – P. 135-141.
12. Pao C.V. Nonlinear Parabolic And Elliptic Equations / C.V. Pao. – New-York: Plenum Press, 1992.
13. Opoĭcev V.I. Generalization of the theory of monotone and concave operators / V.I. Opoĭcev // Trudy Moskov. Mat. Obshch. – 1978. – Vol. 36. – P. 237-273. (in Russian).

14. Opoĭcev V.I. Nonlinear Operators in Spaces with a Cone / V.I. Opoĭcev, T.A. Khurodze. – Tbilisi: Izdatel'stvo Tbilisskogo Universiteta, 1984. (in Russian).
15. Polyanin A.D. Handbook of Linear Partial Differential Equations for Engineers and Scientists / A.D. Polyanin, V.E. Nazaikinskii. – 2nd. ed. Boca Raton-London-New-York: Chapman and Hall/CRC Press, 2016.
16. Rvacev V.L. Theory of R-functions and its Some Applications / V.L. Rvacev. – Kiev: Naukova Dumka, 1982. (in Russian).
17. Sidorov M.V. Construction of two-sided approximations to positive solutions of boundary value problems for semilinear elliptic systems / M.V. Sidorov // Journal of Numerical & Applied Mathematics. – 2017. – No. 3 (126). – P. 110-123.
18. Sidorov M.V. Construction two-sided iterative processes for solving nonlinear boundary value problems using methods of Green's functions and the quasi-functions of Green-Rvachev / M.V. Sidorov // Visnyk of Zaporizhzhya National University. Physical and mathematical Sciences. – 2017. – No. 2. – P. 250-259. (in Ukrainian).
19. Troy W.C. Symmetry properties in systems of semilinear elliptic equations / W.C. Troy // Journal of Differential Equations. – 1981. – Vol. 42, № 3. – P. 400-413.

M. V. Sidorov,
Kharkiv National University of Radio Electronics,
14, Nauki Ave., Kharkiv, 61166, Ukraine.

Received 03.09.2018; revised 26.09.2018

Taras Shevchenko National University of Kyiv

Journal of Numerical and Applied

2018 **M A T H E M A T I C S** 2 (128)

Founded in 1965

C O N T E N T S