Journal of Optimization, Differential Equations and Their Applications

ISSN (print)
2617-0108

ISSN (online)
2663-6824

Volume 30

Issue 1

2022

DNU

Online Edition at http://model-dnu.dp.ua

Journal of Optimization, Differential Equations and Their Applications

Editor-in-Chief
Peter I. Kogut
Department of Differential Equations
Oles Honchar Dnipro National University
72, Gagarin av., Dnipro 49010, Ukraine
(+380) 67631-6755

p.kogut@i.ua

Journal of Optimization, Differential Equations and Their Applications welcomes application-oriented articles with strong mathematical content in scientific areas such as deterministic and stochastic ordinary and partial differential equations, mathematical control theory, modern optimization theory, variational, topological and viscosity methods, qualitative analysis of solutions, approximation and numerical aspects. JODEA also provides a forum for research contributions on nonlinear differential equations and optimal control theory motivated by applications to applied sciences.
JODEA is published twice a year (in June and December).

ISSN (print) 2617–0108, ISSN (on-line) 2663–6824, DOI 10.15421/14223001.

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **30**, Issue **1**, June **2022**, pp. 1–41, DOI 10.15421/142201

ISSN (print) 2617–0108
ISSN (on-line) 2663–6824

MODELING OF CHAOTIC PROCESSES BY MEANS OF ANTISYMMETRIC NEURAL ODES

Vasiliy Ye. Belozyorov[*], Danylo V. Dantsev[†]

Abstract. The main goal of this work is to construct an algorithm for modeling chaotic processes using special neural ODEs with antisymmetric matrices (antisymmetric neural ODEs) and power activation functions (PAFs). The central part of this algorithm is to design a neural ODEs architecture that would guarantee the generation of a stable limit cycle for a known time series. Then, one neuron is added to each equation of the created system until the approximating properties of this system satisfy the well-known Kolmogorov theorem on the approximation of a continuous function of many variables. In addition, as a result of such an addition of neurons, the cascade of bifurcations that allows generating a chaotic attractor from stable limit cycles is launched. We also consider the possibility of generating a homoclinic orbit whose bifurcations lead to the appearance of a chaotic attractor of another type. In conclusion, the conditions under which the found attractor adequately simulates the chaotic process are discussed. Examples are given.

Key words: system of ordinary autonomous differential equations, neural network, antisymmetric matrix, power activation function, Lyapunov stability, limit cycle, homoclinic orbit, strange non-chaotic attractor, search algorithm.

2010 Mathematics Subject Classification: 13A50, 14L24, 15A72, 93C05.

Communicated by Prof. P. I. Kogut

1. Introduction

Recurrent neural networks (RNN) are one of the main tools for modeling various dynamic processes. It should be noted that the quality of modeling with the help of RNN depends on activation functions used in the network [1–7].

As for activation functions, the good results of modeling various processes were obtained precisely with the help of those neural networks in which the well-known rectified linear units (ReLU) were used [4, 6, 8]. Naturally, any generalizations of ReLU deserve attention. Therefore, we will consider some of them.

We will not focus on the advantages or disadvantages of one or another activation function, but will focus only on those properties that are essentially used in this work.

In recent years, an interesting idea has appeared to interpret a system of ordinary differential equations in the form of a suitable neural network (residual network) [9–12]. Precisely this interpretation is implemented in the present work:

[*] Department of Applied Mathematics, Oles Honchar Dnipro National University, 72, Gagarin's Avenue, 49010, Dnipro, Ukraine, belozvye2017@gmail.com

[†] Chaotic Dynamics Incorporated, 201-45793, Luckakuck Way, Chilliwack BC V2R 5S3, Canada, danylo@chaodyna.com

a system of differential equations (a system of so-called neural ODEs) is considered as a continuous analogue of some RNN [13–17]. It should be noted that in [17] the neural network was considered as a linear control system closed by nonlinear feedback. In this case, the ReLU activation functions played the role of the functions constituting the feedback. The task of modeling was not to bring the trajectories of the model and the real process closer together, but to find the algebraic invariants that determine the behavior of the model built for a known time series. If the corresponding invariants for different lengths of this time series turn out to be equal, then we can talk about the adequacy of the model and the real process.

Below we will focus on two areas of research, which can be formulated in the following questions.

1) If a neural network models a certain dynamic process, then how to guarantee the stability or boundedness of solutions of the system of differential equations describing a continuous analog of the aforementioned neural network?

2) In the theory of bifurcations, the following result is well known: in any determinate system, chaotic processes arise as a result of bifurcations of limit cycles or homoclinic orbits [18–21]. Therefore, how to design the architecture of neural ODEs system so that the resulting architecture would generate a limit cycle? (It is now known that most types of chaos in systems of differential equations begin with bifurcations of limit cycles [18–20].)

The answer to the first question will be successful if the activation functions are chosen so that Lyapunov analysis can be done for the resulting system of neural ODEs [22–24]. Piecewise continuous functions, each part of which is a power function, can be proposed as such functions. The use of power activation functions (PAFs) in neural networks is a generalization of the rectified linear units. In the present time ReLU are standard functions to increase the depth of learning. Therefore, power activation functions are on obvious generalization of ReLU.

Note that for systems of neural ODEs with PAF, the answer to the first question has already been partially obtained in article [25]. (In the present paper, the results proved in [25] will be generalized.) As for the second question, the main part of the article will be devoted to the answer to this question.

It should be said that in a large part of the article neural ODEs with chaotic modes are discussed. Chaos constitutes the basic form of collective neural activity for all processes and functions of perception. It acts as a controlled noise source to ensure uninterrupted access to previous memorized images and memorizing new ones. Chaos allows the system to be always active, ridding it of the need to wake up or enter a stable state every time the input changes [26]. Many researchers agree that the best from the point of view of storing and processing information is the regime of ordered chaos [27]. On the one hand, this mode has all the advantages of chaos, on the other hand, this mode can be controlled. The set of states through which the trajectory of a chaotic system passes is called a chaotic attractor. Therefore, the conditions for the existence of chaotic attractors in systems of neural ODEs are the subject of research in this paper.

The final sections of the article are devoted to the reconstruction of neural ODE systems. For this purpose, several algorithms have been developed for determining the parameters of ODE systems for known time series. The essence of these algorithms lies in the fact that they use the special structure of neural ODEs (antisymmetric neural ODEs) with which it is possible to generate a limit cycle. After that, by choosing the weight coefficients, we obtain such bifurcations of the indicated cycle, which lead to the simulation of a real chaotic process.

This article is organized as follows. Section 2 presents well-known results on the theory of approximation of continuous functions of several variables, which are necessary for further research. Section 3 is devoted to the search for conditions under which periodic solutions and homoclinic orbits can appear in neural ODEs. Conditions for the emergence of chaos in neural ODEs with PAFs are studied in Section 4. In Section 5, some generalizations of power activation functions are studied. Section 6 provides algorithms for tuning the weight coefficients of neural ODEs with PAFs.

The whole Section 7 is devoted to applying the algorithms of Section 6 to the problem of restoring ordinary differential equations from known time series describing the dynamics of certain processes. Finally, some conclusions based on the results presented in the article are given in Section 8.

2. Mathematical preliminaries

We now recall several well-known results from the theory of approximation of real functions of n variables [28–30].

Let $\mathbb{X}$ be an arbitrary set in the linear space $\mathbb{R}^n$. By $\mathbf{C}(\mathbb{X})$ denote a set of continuous real functions of n variables with domain of definition $\mathbb{X}$.

Definition 2.1. A set of real functions $\mathbb{F} \subset \mathbf{C}(\mathbb{X})$ is called separating points of the set $\mathbb{X} \subset \mathbb{R}^n$ if for any different $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}$, there exists a function $f \in \mathbb{F}$ such that $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$.

Definition 2.2. A collection of functions $\mathbb{F} \subset \mathbf{C}(\mathbb{X})$ is called closed with respect to a function of one variable $\phi : \mathbb{R} \to \mathbb{R}$ if $\phi(f) \in \mathbb{F}$ for any $f \in \mathbb{F}$.

Theorem 2.1. *([28]). Let $\mathbb{X} \subset \mathbb{R}^n$ be a compact space and $\mathbf{C}(\mathbb{X})$ be the algebra of continuous real functions on $\mathbb{X}$. Let also the set $\mathbb{F} \subset \mathbf{C}(\mathbb{X})$ containing the constant 1 be the linear subspace closed with respect to the nonlinear continuous function $\phi : \mathbb{R} \to \mathbb{R}$ and separating points of the set $\mathbb{X}$. Then $\mathbb{F}$ is dense in $\mathbf{C}(\mathbb{X})$.*

Theorem 2.1 can be interpreted as a statement about the universal approximation possibilities of arbitrary nonlinearity: using linear operations and a single nonlinear element ϕ, one can construct an algorithm that builds an analytical model of any continuous function with any desired accuracy.

From an applied point of view Theorem 2.1 can be presented as follows.

Let $(\mathbf{u}, \mathbf{v}) \equiv ((u_1, \ldots, u_n), (v_1, \ldots, v_n))$ be a scalar product of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

Let $F(x_1,\dots,x_n)$ be a real continuous function defined on a closed bounded domain $\mathbb{D} \subset \mathbb{R}^n$. Let also $\epsilon > 0$ be any arbitrarily small number, which means the accuracy of the approximation.

Theorem 2.2. *([28–30]).Let ψ be a continuous nonlinear real function of one variable such that $\forall f \in \mathbb{F}$ we have $\psi(f) \in \mathbb{F}$. Then there exist an integer $m > 0$, a set of vectors $\mathbf{a}_j \in \mathbb{R}^n$, and sets of real numbers ξ_j and b_j; $j = 1,\dots,m$, such that the function*

$$H(\mathbf{x}) \equiv H(x_1,\dots,x_n) = \sum_{j=1}^{m} \xi_j \psi((\mathbf{a}_j,\mathbf{x}) + b_j) \tag{2.1}$$

approximates the given function $F(x_1,\dots,x_n)$ with the error ϵ in the domain $\mathbb{D}$.

Thus, $\forall (x_1,\dots,x_n)^T \in \mathbb{D}$, we have $|F(x_1,\dots,x_n) - H(x_1,\dots,x_n)| < \epsilon$.

In terms of neural networks, this theorem can be formulated as follows. Any continuous function of several variables can be realized with any accuracy using a two-layer neural network with a sufficient number of neurons and one nonlinear activation function in the hidden layer [1, 2, 28].

3. Periodic solutions of neural ODEs

We assume that we know the dimension n of the real phase space in which the considered dynamic process $\mathbf{x}(t) = (x_1(t),\dots,x_n(t))^T \in \mathbb{R}^n$ takes place [17, 25, 26]. We will also assume that the functions $x_1(t),\dots,x_n(t)$ are continuous and differentiable with respect to time t on the interval $[0,\infty)$. Our goal will be to model this process with a suitable system of neural ODEs. This system will be based on new concepts, which are demonstrated below.

Introduce the following power functions [25]:

$$g(u,\alpha \vee \beta) = \begin{cases} -(-u)^\beta & \text{if}(u < 0 \text{ and } \beta > 0);\ 0 \text{ if}(u < 0 \text{ and } \beta = 0) \\ u^\alpha & \text{if}(u \geq 0 \text{ and } \alpha > 0);\ 0 \text{ if}(u \geq 0 \text{ and } \alpha = 0) \end{cases} \tag{3.1}$$

and

$$g(u,\alpha \vee \beta) = \begin{cases} (-u)^\beta & \text{if}(u < 0 \text{ and } \beta > 0);\ 0 \text{ if}(u < 0 \text{ and } \beta = 0) \\ u^\alpha & \text{if}(u \geq 0 \text{ and } \alpha > 0);\ 0 \text{ if}(u \geq 0 \text{ and } \alpha = 0). \end{cases} \tag{3.2}$$

Definition 3.1. [25]. Representation (3.1) ((3.2)) is called an odd (even) activation function.

It is obvious that the linear combination

$$\mathcal{H}(u,\mathbf{s}) = s_1 g(u,\alpha_1 \vee \beta_1) + \dots + s_k g(u,\alpha_k \vee \beta_k), \tag{3.3}$$

where $\mathbf{s} = (s_1,\dots,s_k,\alpha_1,\dots,\alpha_k,\beta_1,\dots,\beta_k)$ and $s_1 \geq 0,\dots,s_k \geq 0$, of the odd (even) functions of form (3.1) (form (3.2)) again is the odd (even) function (see [25]).

Using these concepts, we can refine the form of function $\psi(u)$ in Theorem 2.2.

1) If the activation function $\psi(u) := \psi(u, \alpha \vee \beta)$ is odd, then in representation (2.1) we can consider $\xi_j = 1$; $j = 1, \ldots, m$.

Indeed, let $\xi_j < 0$. Then

$$\xi_i \psi(u, \alpha \vee \beta) = \left\{ \begin{array}{l} ((-\xi_j)^{1/\beta}(-u))^\beta \text{ if}(u < 0) \\ -((-\xi_j)^{1/\alpha}u)^\alpha \text{ if}(u \geq 0) \end{array} \right\} = \psi(-(-\xi_j)^{(1/\alpha)\vee(1/\beta)}u, \alpha \vee \beta).$$

Similarly, let $\xi_j > 0$. Then

$$\xi_i \psi(u, \alpha \vee \beta) = \left\{ \begin{array}{l} -(-\xi_j^{1/\beta}u)^\beta \text{ if}(u < 0) \\ (\xi_j^{1/\alpha}u)^\alpha \text{ if}(u \geq 0) \end{array} \right\} = \psi(\xi_j^{(1/\alpha)\vee(1/\beta)}u, \alpha \vee \beta).$$

Thus, we have

$$\xi_j \psi((\mathbf{a}_j, \mathbf{x}) + b_j) = \left\{ \begin{array}{l} \psi[\xi_j^{(1/\alpha)\vee(1/\beta)}(\mathbf{a}_j, \mathbf{x}) + \xi_j^{(1/\alpha)\vee(1/\beta)}b_j] \text{ if}(\xi_j > 0) \\ \psi[-(-\xi_j)^{(1/\alpha)\vee(1/\beta)}(\mathbf{a}_j, \mathbf{x}) - (-\xi_j)^{(1/\alpha)\vee(1/\beta)}b_j] \text{ if}(\xi_j < 0) \end{array} \right\}.$$

If we now introduce redesignations

$$\psi[\xi_j^{(1/\alpha)\vee(1/\beta)}(\mathbf{a}_j, \mathbf{x}) + \xi_j^{(1/\alpha)\vee(1/\beta)}b_j] \equiv \psi[(\overline{\mathbf{a}_j}, \mathbf{x}) + \overline{b_j}],$$

then we get formula (2.1) in which $\xi_j = 1$; $j = 1, \ldots, m$.

2) Let us turn to Definition 2.1 in which we will consider $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{F} := \{\mathcal{H}(a_{1i}x_1 + \cdots + a_{ni}x_n, \mathbf{s}_i)\}$ is the union of all odd functions of the form (3.3); $i = 1, 2, \ldots$. (Note that any element of the set $\mathbb{F}$ is a function of one variable: $a_{1i}x_1 + \cdots + a_{ni}x_n = u_i$.) Since for any $f \in \mathbb{F}$ and any $a \in \mathbb{R}$ equation $f(u) = a$ has a single root, then it is clear that the set $\mathbb{F}$ separates the points of the set $\mathbb{X}$.

3) Let $f(u, \gamma \vee \delta), g(u, \alpha \vee \beta) \in \mathbb{F}$, where $\mathbb{F}$ is the set of all odd functions. Then, we have

$$f(g(u, \alpha \vee \beta), \gamma \vee \delta) = \left\{ \begin{array}{l} -((-u)^\beta)^\delta = -(-u)^{\beta\delta} \text{ if}(u < 0) \\ (u^\alpha)^\gamma = u^{\alpha\gamma} \text{ if}(u \geq 0) \end{array} \right\}.$$

Thus, we have $\forall f, g \in \mathbb{F}$ $f(g) \in \mathbb{F}$ and the requirement of Definition 2.2 is satisfied.

As follows from items 1)–3), all conditions of Theorems 2.1 and 2.2 are true for odd activation functions. In this regard, we can reduce the parameters $\xi_j, j = 1, \ldots, m$, in formula (2.1) of Theorem 2.2.

Corollary of Theorem 2.2. *Let ψ be the odd activation nonlinear function of one variable such that $\forall f \in \mathbb{F}$ we have $\psi(f) \in \mathbb{F}$. Then there exist an integer $m > 0$, a set of vectors $\mathbf{a}_j \in \mathbb{R}^n$, and a set of real numbers b_j; $j = 1, \ldots, m$, such that the function*

$$H(\mathbf{x}) \equiv H(x_1, \ldots, x_n) = \sum_{j=1}^{m} \psi((\mathbf{a}_j, \mathbf{x}) + b_j) \tag{3.4}$$

approximates the given function $F(x_1, \dots, x_n)$ with the error ϵ in the domain $\mathbb{D}$.

Now, we can apply this Corollary to approximate the derivatives $\dot{x}_i(t)$ of the functions $x_i(t)$; $i = 1, \dots, n$.

As a result, we can get the following system of ordinary differential equations:

$$\begin{cases} \dot{x}_1(t) = h_1([A - rI]_1 \cdot \mathbf{x} + b_1) + f_{11}(p_{11}^{(1)} x_1 + \cdots + p_{1n}^{(1)} x_n + d_1^{(1)}) + \cdots \\ \qquad + f_{1k_1}(p_{k_1 1}^{(1)} x_1 + \cdots + p_{k_1 n}^{(1)} x_n + d_{k_1}^{(1)}), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots , \\ \dot{x}_n(t) = h_n([A - rI]_n \cdot \mathbf{x} + b_n) + f_{n1}(p_{11}^{(n)} x_1 + \cdots + p_{1n}^{(n)} x_n + d_1^{(n)}) + \cdots \\ \qquad + f_{nk_n}(p_{k_n 1}^{(n)} x_1 + \cdots + p_{k_n n}^{(n)} x_n + d_{k_n}^{(n)}) \end{cases} \tag{3.5}$$

with the known vector of initial values $(x_{10}, \dots, x_{n0})^T$.

Here $h_i(u_i, \alpha_i \vee \beta_i)$ and $f_{ij}(u_i, \gamma_{ij} \vee \delta_{ij})$ are real power odd functions of one variable u_i; $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ is the vector of states, $A \in \mathbb{R}^{n \times n}$, I is the identity $n \times n$ matrix; $[A - rI]_i$ is the i-th row of the matrix $A - rI$; r, $b_i, p_{11}^{(i)}, \dots, p_{1n}^{(i)}, \dots, p_{k_i 1}^{(i)}, \dots, p_{k_i n}^{(i)}, d_1^{(i)}, \dots, d_{k_i}^{(i)}$ are real parameters; k_i is a nonnegative integer (if $k_i = 0$, then $f_{ij}(u) \equiv 0$); $j = 1, \dots, k_i; i = 1, \dots, n$. (The meaning of the first terms on the right-hand sides of the equations of system (3.5) will be explained below.)

In what follows, we will assume that the conditions of Theorem 2.2 (on local existence and uniqueness of a solution [31]) are fulfilled for system (3.5) with initial data vector $(x_{10}, \dots, x_{n0})^T$.

System (3.5) was created for solving approximation problems. However, in various issues of modeling, it can be interesting in itself. Therefore, in the next theorem we can weaken the conditions under which system (3.5) was constructed.

Theorem 3.1. *(***Main Theorem***). Suppose that $r \geq 0$ is a sufficiently large number such that the symmetric matrix $A + A^T - 2rI$ is negative definite. Let $\alpha_i > 0$ and $\beta_i > 0$, and the power function $h_i(u_i, \alpha_i \vee \beta_i)$ be odd; $i = 1, \dots, n$. Let also $q = \min(\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n)$ and let $f_{ij}(u_i, \gamma_{ij} \vee \delta_{ij})$ be power (even or odd) activation functions such that $q > \gamma_{ij} \geq 0$ and $q > \delta_{ij} \geq 0$; $j = 1, \dots, k_i; i = 1, \dots, n$. Then any solution of system (3.5) is bounded.*

Proof. The proof of this theorem basically repeats the proof of Theorem 4.1 [25].

Let us introduce a new variable into system (3.5) according to the formula $\mathbf{x} \to \mathbf{y} = (A - rI)\mathbf{x} + \mathbf{b}$, where $\mathbf{b} = (b_1, \dots, b_n)^T$. Further, the number $r \geq 0$ can be taken large enough so that the matrix $(A - rI)$ will be invertible. Therefore, the specified replacement will be correct.

In this case, system (3.5) can be rewritten as

$$\dot{\mathbf{x}}(t) = (A - rI)\mathbf{h}(\mathbf{x}) + \mathbf{f}(\mathbf{x}), \tag{3.6}$$

where $\mathbf{h}(\mathbf{x}) = (h_1(x_1, \alpha_1 \vee \beta_1), \dots, h_n(x_n, \alpha_n \vee \beta_n))^T$, $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))^T$. (For simplicity, in the newly obtained system, we will leave the previous notation of phase variables.)

We define the applicant for the role of the Lyapunov function for system (3.6) the real function $V(x_1,\dots,x_n)$ by the following rule:

$$\begin{aligned} V(x_1,\dots,x_n) &= \frac{h_1(x_1,\gamma_1+1)}{\gamma_1+1}+\dots+\frac{h_n(x_n,\gamma_n+1)}{\gamma_n+1}\\ &= 0.5\Big[\Big(\frac{h_1(x_1,\gamma_1)}{\gamma_1+1},\dots,\frac{h_n(x_n,\gamma_n)}{\gamma_n+1}\Big),(x_1,\dots,x_n)\Big]\\ &\quad+0.5\Big[(x_1,\dots,x_n),\Big(\frac{h_1(x_1,\gamma_1)}{\gamma_1+1},\dots,\frac{h_n(x_n,\gamma_n)}{\gamma_n+1}\Big)\Big], \end{aligned} \tag{3.7}$$

where the scalar products of the corresponding vectors are placed in square brackets and $\forall i \in \{1,\dots,n\}$

$$\gamma_i = \begin{cases} \beta_i, \text{if } x_i < 0,\\ \alpha_i, \text{if } x_i \ge 0. \end{cases}$$

(a1) The case of strictly odd function $h_i(x_i,\alpha_i \vee \beta_i)$; $i=1,\dots,n$ [25].

Since fractions β_i+1 and α_i+1 have an even numerator and odd denominator, then the function $V(x_1,\dots,x_n)$ will be positive definite. Further, from the definition of function $V(x_1,\dots,x_n)$ and system (3.6) it follows that

$$\begin{aligned} &\dot V_t(x_1(t),\dots,x_n(t)) = \Big(\frac{h_1(x_1,\gamma_1+1)}{\gamma_1+1}+\dots+\frac{h_n(x_n,\gamma_n+1)}{\gamma_n+1}\Big)'_t\\ &= h_1(x_1(t),\alpha_1\vee\beta_1)\cdot\dot x_1(t)+\dots+h_n(x_n(t),\alpha_n\vee\beta_n)\cdot\dot x_n(t)\\ &= \Big(h_1(x_1(t),\alpha_1\vee\beta_1),\dots,h_n(x_n(t),\alpha_n\vee\beta_n)\Big)S\begin{pmatrix} h_1(x_1(t),\alpha_1\vee\beta_1)\\ h_2(x_2(t),\alpha_2\vee\beta_2)\\ \vdots\\ h_n(x_n(t),\alpha_n\vee\beta_n)\end{pmatrix}\\ &\quad+(\mathbf{h}(\mathbf{x}),\mathbf{f}(\mathbf{x})), \end{aligned} \tag{3.8}$$

where $S := 0.5(A+A^T-2rI)$.

Introduce the norm of matrix $Q=\{q_{ij}\}\in\mathbb{R}^n$ by the following formula:

$$\|Q\| = \sum_{1\le i,j\le n}|q_{ij}|.$$

Similarly, we define the norm of vector $\mathbf{u}=(u_1,\dots,u_n)^T$: $\|\mathbf{u}\| = \sum\limits_{1\le i\le n}|u_i|$.

Now we estimate the derivative $\dot V_t(x_1(t),\dots,x_n(t))$ of function $V(x_1(t),\dots,x_n(t))$, taking into account the fact that matrix S is negative definite:

$$\begin{aligned} \dot V_t(\mathbf{x}(t)) \le \lambda_{\max}(S)\cdot(&h_1^2(x_1(t),\alpha_1\vee\beta_1)+h_2^2(x_2(t),\alpha_2\vee\beta_2)\\ &+\dots+h_n^2(x_n(t),\alpha_n\vee\beta_n))+W(x_1(t),\dots,x_n(t)), \end{aligned}$$

where $\lambda_{\max}(S)$ denotes the maximal eigenvalue of symmetric matrix $S \in \mathbb{R}^{n\times n}$ and $W(x_1(t),\dots,x_n(t)) := \|\mathbf{h}(\mathbf{x})\| \cdot \|\mathbf{f}(\mathbf{x})\|$ is a positive definite function.

The last inequality can be rewritten as follows:

$$\begin{aligned}&\frac{d}{dt}V(x_1(t),\dots,x_n(t))\\ &\le W(x_1(t),\dots,x_n(t)) - b\cdot(h_1(x_1(t),2\gamma_1)+\dots+h_n(x_n(t),2\gamma_n)),\end{aligned} \tag{3.9}$$

where $b = -\lambda_{\max}(S) > 0$.

The solution of inequality (3.9) can be found by the formula

$$\begin{aligned}&V(x_1(t),\dots,x_n(t)) \le V_0\\ &\times \exp\int\limits_0^t \left[\frac{W(x_1(\tau),\dots,x_n(\tau)) - b\cdot(h_1(x_1(\tau),2\gamma_1)+\dots+h_n(x_n(\tau),2\gamma_n)}{V(x_1(\tau,\dots,x_n(\tau))}\right]d\tau,\end{aligned} \tag{3.10}$$

where the constant $V_0 = V(x_1(0),\dots,x_n(0)) > 0$.

Note that the functions $W(x_1,...,x_n), V(x_1,\dots,x_n)$, and $H(x_1,\dots, x_n) = h_1(x_1, 2\gamma_1)+\dots+h_n(x_n,2\gamma_n)$ are positive definite power functions. In addition, $\deg W(x_1, \dots, x_n) = q\cdot\max(\gamma_{ij},\delta_{ij}) < 2q$, and $\deg H(x_1,\dots,x_n) = 2q$; $j = 1,\dots,k_i$; $i = 1,\dots,n$.

In this case, on the one hand, there exists a moment $T_0 > 0$ such that if $t > T_0$, then $W(x_1,\dots,x_n) - bH(x_1,\dots,x_n) < 0$, and

$$\lim_{t\to\infty}\frac{W(x_1(t),\dots,x_n(t)) - bH(x_1(t),\dots,x_n(t))}{V(x_1(t),\dots,x_n(t))} < 0. \tag{3.11}$$

Thus, we have $V(x_1(t),\dots,x_n(t)) \to 0$ at $t\to\infty$. But on the other hand this fact means that if the function $V(x_1(t),\dots,x_n(t))$ is small enough, then there exists the moment $T_1 > T_0 > 0$ such that if $t > T_1$, then $W(x_1(t),\dots,x_n(t)) - bH(x_1(t),\dots,x_n(t)) > 0$ and the positive function $V(x_1(t),\dots,x_n(t))$ increases, and so on.

Let R be some positive constant. Denote by $\mathbb{V} = \{x_1,\dots,x_n\} \subset \mathbb{R}^n$ the set of all points in $\mathbb{R}^n$ satisfying the condition $V(x_1,\dots,x_n) - R^2 \le 0$. Let us assume that R is a minimal constant for which the set $\mathbb{H} = \{(x_1,\dots,x_n) \in \mathbb{V} \mid W(x_1,\dots,x_n) - bH(x_1,\dots,x_n) \ge 0\}$ is not empty. Since $H(x_1,\dots,x_n)$ is positive definite, it is clear that the set $\mathbb{H}$ is compact.

Now we assume that there exists a moment T_u such that for $t = T_u$ $\mathbf{x}(T_u) \notin \mathbb{H}$, but $\mathbf{x}(T_u) \in \mathbb{V}$. Then inequality (3.11) is not satisfied.

Further, since $\mathbb{H} \subset \mathbb{V}$, then $\mathbb{V} - \mathbb{H}$ is the compact positively invariant set with respect to (3.6). Therefore, if solution $\mathbf{x}(t)$ of system (3.6) belongs to $\mathbb{V} - \mathbb{H}$, then it is bounded. It means that the solution $V(x_1,\dots,x_n)$ of equation (3.8) also should be bounded.

Denote by T_s the moment of time such that $\mathbf{x}(T_s) \in \mathbb{H}$. Then, by virtue of (3.11) and according to LaSalle's Theorem [25], we get that solution $\mathbf{x}(t)$ of system

(3.6) starting at $\mathbb{H}$ belongs to $\mathbb{H}$. In addition, $\mathbf{x}(t)$ is attracted to the boundary of $\mathbb{H}$ as $t \to +\infty$. Thus, it is bounded.

Now we use Comparison Principle [25]. Then it remain to compare the solution $V(x_1, \dots, x_n)$ of equation (3.8) and a similar solution of inequality (3.9). From here it follows the boundedness of solution $\mathbf{x}(t)$ of system (3.6) for any initial condition $\mathbf{x}_0 \in \mathbb{R}^n$. This completes the proof of case (a1) for strictly odd functions.

(a2) The case of odd function $h_i(x_i, \alpha_i \vee \beta_i)$; $i = 1, \dots, n$ [25].

Now we can apply Theorem 4.1 [25] to equation (3.6). Then all the ideas that were used in the proof of Theorem 4.1 [25] can be directly carried over to the proof of Theorem 3.1. Since

$$\min(\deg h_1(x_1, \alpha_1 \vee \beta_1), ..., \deg h_n(x_n, \alpha_n \vee \beta_n))$$
$$> \max(\deg f_1(x_1, \dots, x_n), ..., \deg f_n(x_1, \dots, x_n)),$$

it remains to verify only one condition of Theorem 4.1 [25]: the symmetric matrix $A + A^T - 2rI$ must be negative definite. The last condition can always be achieved by choosing the sufficiently large parameter $r \geq 0$. It is clear that the same statement will also hold for system (3.5). This remark completes the proof. □

Comment 3.1. In the general case the function $V(x_1, \dots, x_n)$ is not the Lyapunov function for system (3.6). It is guaranteed to be the Lyapunov function if $\mathbf{f}(\mathbf{x}) \equiv 0$.

Let us compose from functions (3.3) the following vector- function:

$$\mathcal{H}(\mathbf{x}, \mathbf{s}_1, ..., \mathbf{s}_n) = (\mathcal{H}_1(x_1, \mathbf{s}_1), \dots, \mathcal{H}_n(x_n, \mathbf{s}_n))^T, \tag{3.12}$$

where $\mathbf{s}_j = (s_{j1}, \dots, s_{jk_j}), j = 1, \dots, n$.

Corollary of Theorem 3.1. *Let the vector* $\mathbf{h}(\mathbf{x})$ *in system (3.6) be replaced by the vector* $\mathcal{H}(\mathbf{x}, \mathbf{s}_1, \dots, \mathbf{s}_n)$*. Then, under the conditions of Theorem 3.1, any solution of system* $\dot{\mathbf{x}}(t) = (A - rI)\mathcal{H}(\mathbf{x}, \mathbf{s}_1, \dots, \mathbf{s}_n) + \mathbf{f}(\mathbf{x})$ *is bounded.*

Proof repeats the proof of Theorem 3.1 if in this theorem we replace the function $V(x_1, \dots, x_n)$ (3.7) by the function

$$V(x_1, \dots, x_n) = \int \mathcal{H}_1(x_1, \mathbf{s}_1) dx_1 + \dots + \int \mathcal{H}_n(x_n, \mathbf{s}_n) dx_n,$$

where all the indicated integrals are integrals of power functions. □

Now, we will assume that in system (3.6) $r = 0$ and $\mathbf{f}(\mathbf{x}) = 0$. Then we get the following system

$$\dot{\mathbf{x}}(t) = A\mathbf{h}(\mathbf{x}). \tag{3.13}$$

Theorem 3.2. *Let* $\alpha_i > 0$ *and* $\beta_i > 0$*, and the power function* $h_i(u_i, \alpha_i \vee \beta_i)$ *be odd;* $i = 1, \dots, n$*. Let also* $A \in \mathbb{R}^{n \times n}$ *be an antisymmetric matrix such that* $\operatorname{rank} A = m \leq n$*. Let* $\mathbb{V} \subset \mathbb{R}^n$ *be a* $(n - m)$*-dimensional subspace of* $\mathbb{R}^n$ *such that* $A\mathbb{V} = \mathbf{0}$ *(if* $m = n$*, then* $\mathbb{V} = \mathbf{0}$*; for odd* $n > 1$*, we always have* $m < n$ *and* $\mathbb{V} \neq \mathbf{0}$*). Then any solution* $\mathbf{x}(\mathbf{x}_0, t)$ *of system (3.13) starting from point* $\mathbf{x}_0 = (x_{10}, \dots, x_{n0})^T \notin \mathbb{V}$ *is periodic.*

Proof. We will again use the Lyapunov function $V(x_1, \dots, x_n)$ of form (3.7), which was introduced in the proof of Theorem 3.1.

Let $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^n | V(\mathbf{x}) \le V(\mathbf{x}_0)\}$. Then the set $\mathbb{S} \subset \mathbb{R}^n$ is compact positively invariant with respect to (3.13). In addition, from here it follows that $V_t^{'}(\mathbf{x}) = 0.5\mathbf{h}^T(\mathbf{x})(A+A^T)\mathbf{h}(\mathbf{x}) = 0$. This means that the solution $\mathbf{x}(\mathbf{x}_0, t)$ of system (3.13) starting from a point $\mathbf{x}_0 \in \mathbb{R}^n$ forms a closed curve on the boundary of the set $\mathbb{S} \subset \mathbb{R}^n$ and therefore the solution $\mathbf{x}(\mathbf{x}_0, t)$ is periodic [31]. □

Now consider the following system of ordinary differential equations:

$$\dot{\mathbf{x}}(t) = \mathbf{h}(A\mathbf{x} + \mathbf{b}) \Longleftrightarrow \left\{ \begin{array}{l} \dot{x}_1(t) = h_1(A_1 \cdot \mathbf{x} + b_1) \\ \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ , \\ \dot{x}_n(t) = h_n(A_n \cdot \mathbf{x} + b_n) \end{array} \right. \tag{3.14}$$

with the known vector of initial values $\mathbf{x}_0$. Here A_i is the i-th row of the matrix A; $i = 1, ..., n$.

Theorem 3.3. *Let, under the conditions of Theorem 3.2, $m = n - 1$ and n be odd. If $\alpha_i = \beta_i$ and $b_i = 0$, $i = 1, \dots, n$, then any solution of system (3.14) is periodic; if at least for one $k \in \{1, \dots, n\}$, we have $\alpha_k \ne \beta_k$ or $b_k \ne 0$, then any solution of system (3.14) is a winding of an infinite cylinder with generatrix $\mathbb{V}$. In addition, there exist numbers $T > 0$ and $\lambda = \lambda(T) \in \mathbb{R}$ that do not depend on $\mathbf{x}_0$, such that for any solution $\mathbf{x}(t)$ of system (3.14), we have*

$$\mathbf{x}(t + T) - \mathbf{x}(t) = \lambda \mathbb{V},$$

where $\lambda = 0$ if and only if $\alpha_i = \beta_i$ and $b_i = 0$, $i = 1, \dots, n$.

Proof. (b1) With the help of suitable changes of variables $\mathbf{y} = A\mathbf{x}+\mathbf{b}$, we transform system (3.14) into the following system (generally speaking, not equivalent to system (3.14)):

$$\frac{d}{dt}(A\mathbf{x}(t) + \mathbf{b}) = A \cdot \left(\begin{array}{c} h_1(A_1\mathbf{x}(t) + b_1) \\ . \ . \ . \ . \ . \ . \ . \ . \\ h_n(A_n\mathbf{x}(t) + b_n) \end{array} \right), \tag{3.15}$$

where $\mathbf{b} = (b_1, \dots, b_n)^T$. Then in the new variables $\mathbf{y}$ we get the system $\dot{\mathbf{y}}(t) = A\mathbf{h}(\mathbf{y})$. (This is system (3.13).)

It is clear that in this case all the conditions of Theorem 3.2 are fulfilled and we obtain a periodic solution $\mathbf{p}(t)$ of the system $\dot{\mathbf{y}}(t) = A\mathbf{h}(\mathbf{y})$. In addition, taking into account the new variables $\mathbf{y}$, system (3.14) can be written in the form $\dot{\mathbf{x}}(t) = \mathbf{h}(\mathbf{y})$. From here it follows that

$$\mathbf{x}(t) = \int \mathbf{h}(\mathbf{y}(t))dt. \tag{3.16}$$

It is important to note that the solutions $\mathbf{q}(t)$ of system (3.14) are not (generally speaking) periodic. (It is found from the solution of the equation $\mathbf{p} = A\mathbf{q} + \mathbf{b}$, where for odd n matrix A is singular.)

It is known that the indefinite integral of a continuous periodic function of period T is the sum of a periodic function of the same period T and some linear function. Thus, from (3.16), (3.14) it follows that $\mathbf{q}(t) = \mathbf{q}_T(t) + t\lambda\mathbb{V} \subset \mathbb{R}^n$ and

$$\frac{d}{dt}(\mathbf{q}_T(t) + t\lambda) = \mathbf{h}(A(\mathbf{q}_T(t) + t\lambda\mathbb{V}) + \mathbf{b}).$$

Here $\mathbf{q}_T(t) = (q_{T1}(t), ..., q_{Tn}(t))^T \in \mathbb{R}^n$ is a periodic vector function.

Let $\mathbb{W} = A(\mathbb{R}^n)$ be a linear subspace in $\mathbb{R}^n$. Denote by $A|_{\mathbb{W}}$ the restriction of A to $\mathbb{W}$. Then, we have $\mathbf{q}_T(t) + t\lambda\mathbb{V} = (A|_{\mathbb{W}})^{-1}(\mathbf{p}(t) - \mathbf{b})$, where $t\lambda\mathbb{V}$ is a straight line in $\mathbb{R}^n$ passing through the origin. Thus, the set $\{\mathbf{q}_T(t) + t\lambda\mathbb{V}\} \subset \mathbb{R}^n$ is a curve wound on the cylinder with generatrix $\mathbb{V}$. (The projection of the periodic curve $\mathbf{q}_T(t) + t\lambda\mathbb{V}, t_0 \leq t \leq t_0 + T$ onto any $(n-1)$-dimensional hyperplane $\mathbb{P} \subset \mathbb{R}^n$ such that $\mathbb{P} \perp \mathbb{V}$ is a directrix of cylinder.)

(b2) Let $\mathbb{V} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$. Consider the situation $\alpha_1 = \beta_1$ and $b_1 = 0$. Then at $A_1 \cdot \mathbf{x} \geq 0$ the first equation of system (3.14) will not change, and at $A_1 \cdot \mathbf{x} < 0$ this equation will be represented as $\dot{x}_1(t) = -h_1(A_1 \cdot \mathbf{x})$.

Thus, if $A_1 \cdot \mathbf{x} \geq 0$, then $x_1(t) = q_{T1}(t) + t\lambda \cdot v_1$; if $A_1 \cdot \mathbf{x} < 0$, then $x_1(t) = q_{T1}(t) - t\lambda \cdot v_1$. Since the initial conditions are the same for both equations, then, in accordance with the well-known Cauchy theorem on the existence and uniqueness of solution, there must be $\lambda = 0$. Repeating the same reasoning for each of the equations of system (3.14) for $\alpha_i = \beta_i$ and $b_i = 0$, we finally obtain $\lambda = 0; i = 2, \ldots, n$. The last statement completes the proof of Theorem 3.3. □

Definition 3.2. System (3.14) in which matrix A is antisymmetric, power activation functions $h_i(u_i, \alpha_i \vee \beta_i)$, $i = 1, \ldots, n$, are odd is called the unperturbed system of antisymmetric neural ODEs.

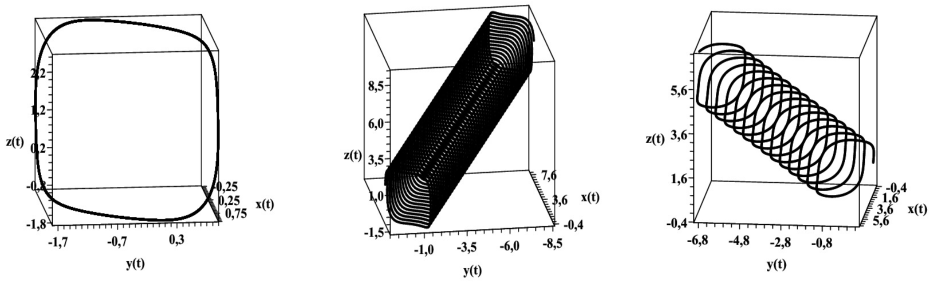
In addition, the following system

$$\dot{\mathbf{y}}(t) = T^{-1} \cdot \begin{bmatrix} h_1(A_1 \cdot T \cdot \mathbf{y} + b_1) \\ \ldots\ldots\ldots, \\ h_n(A_n \cdot T \cdot \mathbf{y} + b_n) \end{bmatrix} \tag{3.17}$$

obtained from system (3.14) of antisymmetric neural ODEs with the help of the linear invertible transformation $T \in \mathbb{R}^{n \times n}$ $(\mathbf{x} = T \cdot \mathbf{y})$ will also be called the unperturbed system of antisymmetric neural ODEs.

Definition 3.3. Let the matrix A of system (3.5) be antisymmetric and $r \geq 0$. Then, under the conditions of Theorems 3.1, system (3.5) is called a perturbed system of antisymmetric neural ODEs.

The next figure demonstrates the statements of Theorem 3.3 (see Fig. 3.1):



(a1) (a2) (a3)

Fig. 3.1. Trajectories of the system
$\dot{x} = (-y - z + b_1)^{r_1}, \dot{y} = (x - z + b_2)^{r_2}, \dot{z} = (x + y + b_3)^{r_3}$:
(a1) $r_1 = (1.5 \vee 1.5), r_2 = (2.5 \vee 2.5), r_3 = (2 \vee 2)$, and $b_1 = b_2 = b_3 = 0$;
(a2) $r_1 = (1.5 \vee 1.5), r_2 = (2.5 \vee 2.5), r_3 = (2 \vee 2)$, and $b_1 = 1, b_2 = b_3 = 0$;
(a3) $r_1 = (1.5 \vee 2.5), r_2 = (2.5 \vee 2.5), r_3 = (2 \vee 2)$, and $b_1 = b_2 = b_3 = 0$.

4. On conditions for the appearance of chaos in system (3.6)

It is known that the following question often arises when modeling chaotic processes: can the created model generate chaotic behavior? The same question can be extended to model (3.6) (or (3.5)).

Let us denote by the symbol

$$\deg \mathbf{h}(\mathbf{x}) = (\deg h_1(x_1, \alpha_1 \vee \beta_1), \ldots, \deg h_n(x_n, \alpha_n \vee \beta_n)) \in \mathbb{N}^n,$$

where $\mathbb{N}$ is the set of natural numbers. Then the inequality $\deg \mathbf{h}(\mathbf{x}) > \deg \mathbf{v}(\mathbf{x}) = (\deg v_1(x_1, \gamma_1 \vee \delta_1), \ldots, \deg v_n(x_n, \gamma_n \vee \delta_n))$ means that $\alpha_i > \gamma_i > 0$ and $\beta_i > \delta_i > 0$; $i = 1, \ldots, n$.

Consider the following simplified version of system (3.6):

$$\dot{\mathbf{x}}(t) = B\mathbf{x} + \mathbf{v}(\mathbf{x}) + (A - rI)\mathbf{h}(\mathbf{x}). \tag{4.1}$$

Here matrices $A, B \in \mathbb{R}^{n \times n}$, and the matrix A is antisymmetric; the power vector functions $\mathbf{v}(\mathbf{x}), \mathbf{h}(\mathbf{x}) \in \mathbb{R}^n$ such that $\deg \mathbf{h}(\mathbf{x}) > \deg \mathbf{v}(\mathbf{x}) > (1, \ldots, 1)$. (If $\mathbf{v}(\mathbf{x}) \equiv \mathbf{0}$, then $\deg \mathbf{h}(\mathbf{x}) > (1, \ldots, 1)$.)

Let us construct for system (4.1) a positive definite function $V(x_1, \ldots, x_n)$ (3.7). Then, we will have

$$\dot{V}_t(x_1, \ldots, x_n) = 0.5 \cdot (\mathbf{h}^T(\mathbf{x})(B\mathbf{x} + \mathbf{v}(\mathbf{x})) + (\mathbf{x}^T B^T + \mathbf{v}^T(\mathbf{x}))\mathbf{h}(\mathbf{x})) - r \cdot \mathbf{h}^T(\mathbf{x})\mathbf{h}(\mathbf{x}).$$

We denote by $\mathbb{W}$ the set of all points from $\mathbb{R}^n$ such that $\dot{V}_t(x_1, \ldots, x_n) \leq 0$. Let also $\mathbb{L} \subset \mathbb{W}$ be the set of all points in $\mathbb{W}$ such that $\dot{V}_t(x_1, \ldots, x_n) = 0$. We also denote by $\mathbb{X} \subset \mathbb{W}$ an open set in $\mathbb{W}$ such that $\forall \mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{X}$ $\dot{V}_t(x_1, \ldots, x_n) < 0$. (Thus, $\mathbb{X} \cup \mathbb{L} = \mathbb{W}$.)

Theorem 4.1. *Let the point* $\mathbf{0}$ *be a unique equilibrium point for system (4.1) in* $\mathbb{W}$*. Suppose also that:*

1) point $\mathbf{0}$ *is a saddle point;*

2) there exists a value of parameter $r > 0$ *such that for an arbitrary vector of initial data* $(x_{10}, ..., x_{n0})^T \in \mathbb{X}$ *the solutions* $x_1(x_{10}, ..., x_{n0}, t), ..., x_n(x_{10}, ..., x_{n0}, t)$ *of system (4.1) satisfy to equality*

$$\liminf_{t\to\infty} V(x_1(x_{10}, \ldots, x_{n0}, t), \ldots, x_n(x_{10}, \ldots, x_{n0}, t)) = 0.$$

Then under the conditions of Theorem 3.1 in system (4.1) there exists a chaotic dynamics.

Proof. According to Theorem 3.1, there exists the value $r = r_0$ such that all solutions of system (4.1) are bounded. Therefore, for $r = r_0$ the set $\mathbb{W}$ is a compact positively invariant set with respect to system (4.1).

According to LaSalla's theorem every solution of system (4.1) starting in $\mathbb{W}$ approaches to the largest invariant set $\mathbb{M} \subset \mathbb{L}$ as $t \to \infty$ (see [25,31]). In our case, by assumption $A^T + A = 0$ and condition 1) of Theorem 4.1, the role of the set $\mathbb{M}$ can be played either by a limit cycle or by a homoclinic trajectory connected at $\mathbf{0}$ (see Theorem 3.2). If both conditions 1) and 2) of Theorem 4.1 are satisfied, then there exists a sequence of values $r_{01} > r_{02} > \cdots > r_{0k} > \cdots$ of parameter r such that $\lim\limits_{k\to\infty} r_{0k} = r_c \geq 0$ and the set $\mathbb{M}(r_c)$ at the critical value r_c is the homoclinic orbit. (Indeed, let $\mathbb{N}_s$ and $\mathbb{N}_u$ be stable and unstable manifolds of the point $\mathbf{0}$ [21, 34]. Let's denote by $\mathbf{x}_0 = (x_{01}, \ldots, x_{0n})^T \in \mathbb{N}_u$ the starting point. Since at the point $\mathbf{x}_0$ we have $\dot{V}_t(x_{01}, \ldots, x_{0n}) \geq 0$, then the solution $\mathbf{x}(\mathbf{x}_0, t)$ of system (4.1) should be attracted to a certain limit cycle in $\mathbb{L}$. According to condition 2) of Theorem 4.1, this limit cycle for some value of parameter r will pass arbitrarily close to the origin (to the manifold $\mathbb{N}_s$). In other words, near point $\mathbf{0}$ on trajectory $\mathbf{x}(\mathbf{x}_0, t)$ there will be point $\mathbf{x}_1 = (x_{11}, \ldots, x_{1n})^T$ such that $\dot{V}_t(x_{11}, \ldots, x_{1n}) \leq 0$. Therefore, there must be the value r_c of parameter r for which $\mathbb{N}_s \cap \mathbb{N}_u \neq \emptyset$. This means the existence of homoclinic orbit.)

The last statement enables us to construct a discrete mapping for system (4.1). Theorems 3.2 and 3.3 allow us to assert that a limit cycle can exist in perturbed system (4.1) for some values of parameter r. Let T_0 be the period of this cycle. In this case, the continuous relation (3.10) for the positive definite function $V(t) \equiv V(x_1(t), \ldots, x_n(t))$ can be rewritten as

$$V(x_1(t_{k+1}), \ldots, x_n(t_{k+1})) \leq V(x_1(t_k), \ldots, x_n(t_k))$$

$$\times \exp \int\limits_{t_k}^{t_{k+1}} \left[\frac{W(x_1(\tau), \ldots, x_n(\tau)) - b \cdot (h_1(x_1(\tau), 2\gamma_1) + ... + h_n(x_n(\tau), 2\gamma_n)}{V(x_1(\tau, \ldots, x_n(\tau))} \right] d\tau,$$

where $t_{k+1} - t_k = T_0; k = 0, 1, \ldots$

A discrete analogue of this relation according to the technique described in [32] – [34] can be represented in the following form:

$$V_{k+1} = V_k \exp(p + \phi(V_k) - rV_k^q); k = 0, 1, 2, \dots \tag{4.2}$$

Here $p \geq 0$; $\phi(u)$ is a linear combination of power functions u^{q_i} of one variable $u > 0$ (with $q_i > 0$) and $0 < \deg \phi(u) = \max q_i < q = \min(\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n)$, $i = 1, \dots, l$.

In [32–34] it is shown that for some $r = r_c$ mapping (4.2) generates a chaotic dynamics. Therefore, system (4.1) at $r = r_c$ will also exhibit chaotic behavior. □

Comment 4.1. Theorems 3.2 and 3.3 guarantee the existence of periodic trajectories for unperturbed systems. This means that for small perturbations, periodic motions (limit cycles) will appear in system (4.1). Namely the existence of limit cycle in system (4.1) allows starting a cascade of bifurcations leading to the appearance of a chaotic attractor.

Comment 4.2. Note that the condition $A^T + A = 0$ cannot be omitted. Indeed, if this condition is satisfied, then $\forall r > 0$ $A^T + A - 2rI = -2rI < 0$. Otherwise (at $A^T + A \neq 0$), the inequality $A^T + A - 2rI < 0$ does not hold $\forall r > 0$. Therefore, if $A^T + A \neq 0$ condition 2) of Theorem 4.1 generally speaking cannot be achieved.

Nevertheless, the results obtained for system (4.1) can be strengthened if matrix A is replaced by a similar non-antisymmetric matrix $A_n = H^{-1}AH$ $(A_n^T + A_n \neq 0)$:

$$\dot{\mathbf{x}}(t) = B\mathbf{x} + \mathbf{v}(\mathbf{x}) + (A_n - rI)\mathbf{h}(\mathbf{x}). \tag{4.3}$$

In order to check condition 2) of Theorem 4.1 we can proceed in the following way.

1. In system (4.1) set parameter r large enough. 2. Decrease the parameter r until the limit cycle appears in system (4.1). 3. Continue slowly decreasing the parameter r until limit cycles of any multiplicity appear in system (4.1). (This means that a non-negative function $V(x_1(t), \dots, x_n(t))$ will have any number of minimums.) If parameter r is close enough to value r_c (which is unknown), then the minimums of function $V(x_1(t), \dots, x_n(t))$ will take arbitrarily small (not zero!) values. This will mean that $r \approx r_c$ (see Fig. 4.5).

4.1. Homoclinic chaos

A large number of chaotic attractors arise when so-called homoclinic orbits exist in a dynamical system (see [18–21,32–37], and references to papers on chaotic dynamics cited elsewhere).

Definition 4.1. ([35]). A bounded trajectory $\mathbf{x}(t, \mathbf{x}_0) \in \mathbb{R}^n$ of system (3.6) is called a homoclinic orbit if the trajectory converges to the same equilibrium point as $t \to \pm\infty$.

Let $A = (a_1, \dots, a_n) \in \mathbb{R}^{n\times n}$ be the antisymmetric matrix composed of columns $a_i \in \mathbb{R}^n; i = 1, \dots, n$.

Let us $\forall s \in \{1, \dots, n\}$ denote by symbol $A_s \in \mathbb{R}^{n\times n}$ ($A^s \in \mathbb{R}^{n\times n}$) the matrix obtained from matrix A by replacing the column a_s (row $-a_s^T$) with the column $-a_s$ (row a_s^T).

Definition 4.2. Matrix A_s (or A^s) will be called partially antisymmetric.

Let $p < n$ be a positive integer. If columns $a_{s_1}, \dots, a_{s_p}, 1 \le s_1 < \dots < s_p \le n$ of the matrix A are replaced by columns $-a_{s_1}, \dots, -a_{s_p}$, then the resulting matrix will also be denoted as $A_{s_1 \dots s_p}$ and called partially antisymmetric. A similar designation $A^{s_1 \cdots s_p}$ is retained for the rows.

Let $\mathbf{h}(\mathbf{x}) = (h_1(x_1, \alpha_1 \vee \beta_1), ..., h_n(x_n, \alpha_n \vee \beta_n))^T$ and $\mathbf{f}(\mathbf{x}) = (f_1(x_1, \gamma_1 \vee \delta_1), ..., f_n(x_n, \gamma_n \vee \delta_n))^T$ be two vectors of power functions.

Theorem 4.2. *Let ψ and $\phi \neq 0$ be arbitrary real numbers and let $\mathbf{h}(\mathbf{x})$ be odd function. Consider the following system of ordinary differential equations*

$$\dot{\mathbf{x}}(t) = \psi A_{s_1 \dots s_p} \mathbf{f}(\mathbf{x}) + \phi A \mathbf{h}(\mathbf{x}), \tag{4.4}$$

in which it is assumed that $\deg \mathbf{h}(\mathbf{x}) > \deg \mathbf{f}(\mathbf{x})$. Then any solution $\mathbf{x}(t, \mathbf{x}_0)$ of system (4.4) is periodic. In addition, if the function $\mathbf{f}(\mathbf{x})$ is odd, then there exists a vector of initial conditions $\mathbf{x}_0^*$ such that trajectory $\mathbf{x}(t, \mathbf{x}_0^*)$ is the homoclinic orbit connected at equilibrium point $\mathbf{0}$.

Proof. (a1) Let $\psi = 0$. Then the assertion of Theorem 4.2 follows from Theorem 3.2.

Let $p = 1$ and $s_1 = s$. Introduce also the following vector $\mathbf{f}_s(\mathbf{x}) = (f_1(x_1), \dots, -f_s(x_s), \dots, f_n(x_n))^T$, where $s \in \{1, \dots, n\}$.

(a2) Let $\psi \cdot \phi \neq 0$.

Consider two scalar products:

$\mathbf{h}^T(\mathbf{x}) \cdot \dot{\mathbf{x}} = \psi \mathbf{h}^T(\mathbf{x}) \cdot A_s \mathbf{f}(\mathbf{x}) + \phi \mathbf{h}^T(\mathbf{x}) \cdot A\mathbf{h}(\mathbf{x}) = \psi \mathbf{h}^T(\mathbf{x}) A \mathbf{f}_s(\mathbf{x})$

and

$\mathbf{f}_s^T(\mathbf{x}) \cdot \dot{\mathbf{x}} = \psi \mathbf{f}_s^T(\mathbf{x}) \cdot A_s \mathbf{f}(\mathbf{x}) + \phi \mathbf{f}_s^T(\mathbf{x}) \cdot A\mathbf{h}(\mathbf{x}) = \psi \mathbf{f}^T(\mathbf{x}) \cdot A_s^s \mathbf{f}(\mathbf{x}) + \phi \mathbf{f}_s^T(\mathbf{x}) \cdot A\mathbf{h}(\mathbf{x}) = \phi \mathbf{h}^T(\mathbf{x}) A^T \mathbf{f}_s(\mathbf{x})$. (Here we have used the obvious equality: $A_s^s + (A_s^s)^T = 0$.)

From here it follows that $\phi \mathbf{h}^T(\mathbf{x}) \cdot \dot{\mathbf{x}} + \psi \mathbf{f}_s^T(\mathbf{x}) \cdot \dot{\mathbf{x}} = 0$. Therefore, if we define the derivative $\dot{V}(\mathbf{x})$ of some function $V(\mathbf{x})$ by formula $\dot{V}(\mathbf{x}) \equiv \phi \mathbf{h}^T(\mathbf{x}) \cdot \dot{\mathbf{x}} + \psi \mathbf{f}_s^T(\mathbf{x}) \cdot \dot{\mathbf{x}}$, then we will have

$$\begin{aligned} V(\mathbf{x}) \equiv \phi \frac{h_1(x_1, \gamma_1 + 1)}{\gamma_1 + 1} + \dots + \phi \frac{h_n(x_n, \gamma_n + 1)}{\gamma_n + 1} + \psi \frac{f_1(x_1, \xi_1 + 1)}{\xi_1 + 1} \\ + \dots + \psi \frac{f_{s-1}(x_{s-1}, \xi_{s-1} + 1)}{\xi_{s-1} + 1} - \psi \frac{f_s(x_s, \xi_s + 1)}{\xi_s + 1} + \dots \\ + \psi \frac{f_n(x_n, \xi_n + 1)}{\xi_n + 1} = C(\mathbf{x}_0) = const, \end{aligned}$$

where

$$\forall i \in \{1, ..., n\} \ \gamma_i = \begin{cases} \beta_i, \text{if } x_i < 0, \\ \alpha_i, \text{if } x_i \geq 0, \end{cases} \xi_i = \begin{cases} \delta_i, \text{if } x_i < 0, \\ \gamma_i, \text{if } x_i \geq 0. \end{cases}$$

Without loss of generality, we can assume that $\phi > 0$. Then from the conditions of Theorem 4.2 it follows that for a sufficiently large norm $\|\mathbf{x}\|$ of the vector $\mathbf{x}$, we have $V(\mathbf{x}) > 0$. This means that if $V(\mathbf{x}_0) = C(\mathbf{x}_0) > 0$, then the function $V(\mathbf{x}) = C(\mathbf{x}_0) > 0$ is bounded. Consequently, any solution of system (4.4) is closed and therefore periodic.

Now let the vector of initial conditions $\mathbf{x}_0^*$ be such that $V(\mathbf{x}_0^*) = 0$. Therefore, if the function $\mathbf{f}(\mathbf{x})$ is odd, then the function $V(\mathbf{x})$ is even. From here it follows that $V(\mathbf{x}_0^*) = V(-\mathbf{x}_0^*) = 0$ and in addition, $V(\mathbf{0}) = 0$. Consequently, there is a trajectory of the system (4.4), which for $t = 0$ leaves some point $\mathbf{x}_0 \neq \mathbf{0}$ arbitrarily close to $\mathbf{x}_0^* \approx \mathbf{0}$ and at $t \to \pm\infty$ approaches to the point $\mathbf{0}$.

Now, in system (4.4), we will change the sign of time: $t \to -\tau$. Then we will have

$$\dot{\mathbf{x}}(\tau) = -\psi A_{s_1 \ldots s_p}\mathbf{f}(\mathbf{x}) - \phi A\mathbf{h}(\mathbf{x}). \tag{4.5}$$

Further, we apply to the study of solutions of equation (4.5) the same technique as for the study of solutions of equation (4.4). As a result, instead of analyzing of the equation $V(\mathbf{x}(t)) = C(\mathbf{x}_{t=0}) > 0$, we come to an analysis of the equation $-V(\mathbf{x}(\tau)) = -C(\mathbf{x}_{\tau=0}) < 0$. Thus, based on the structure of even function $V(\mathbf{x}(t))$, we have $V(\mathbf{x}(\tau)) = V(\mathbf{x}(-t)) = V(\mathbf{x}(t)) = C(\mathbf{x}_0) > 0$. This means the existence of a homoclinic orbit.

(a3) Now let $p > 1$. Introduce the vector

$$\mathbf{f}_{s_1 \ldots s_p}(\mathbf{x}) = (f_1(x_1), \ldots, f_{s_1-1}(x_{s_1-1}), \\ -f_{s_1}(x_{s_1}), \ldots, -f_{s_p}(x_{s_p}), f_{s_p+1}(x_{s_p+1}), \ldots, f_n(x_n))^T.$$

Then the proof of case $p > 1$ completely repeats the proof of case $p = 1$ if we take into account the obvious equality: $A^{s_1 \ldots s_p}_{s_1 \ldots s_p} + (A^{s_1 \ldots s_p}_{s_1 \ldots s_p})^T = 0$. □

We replace the odd activation vector-function $\mathbf{h}(\mathbf{x})$ with the odd activation vector-function $\mathcal{H}(\mathbf{x}, \mathbf{r}_1, \ldots, \mathbf{r}_n)$. Similarly, we make the change of variables $\mathbf{f}(\mathbf{x}) \to \mathcal{F}(\mathbf{x}, \mathbf{q}_1, \ldots, \mathbf{q}_n)$ (see (3.12)). Then, the following corollary is obvious.

Corollary of Theorem 4.2. *The statements of Theorem 4.2 remain valid if equation (4.4) is replaced by the equation*

$$\dot{\mathbf{x}}(t) = \psi A_{s_1 \ldots s_p}\mathcal{F}(\mathbf{x}, \mathbf{q}_1, \ldots, \mathbf{q}_n) + \phi A\mathcal{H}(\mathbf{x}, \mathbf{r}_1, \ldots, \mathbf{r}_n),$$

and keep the conditions presented for functions $\mathbf{h}(\mathbf{x})$ *and* $\mathbf{f}(\mathbf{x})$ *also for functions* $\mathcal{H}(\mathbf{x}, \mathbf{r}_1, \ldots, \mathbf{r}_n)$ *and* $\mathcal{F}(\mathbf{x}, \mathbf{q}_1, \ldots, \mathbf{q}_n)$.

Since $\phi \neq 0$, we can assume that $\phi = 1$. Consider the following special case of system (4.1):

$$\dot{\mathbf{x}}(t) = B\mathbf{x} + \psi A_{s_1 \ldots s_p}\mathbf{f}(\mathbf{x}) + (A - rI)\mathbf{h}(\mathbf{x}), r \geq 0. \tag{4.6}$$

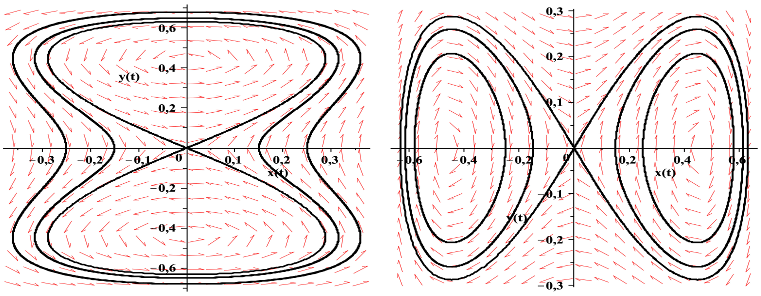
Thus, if conditions of Theorems 4.1 and 4.2 are satisfied for system (4.6), then we will get some chaotic behavior of this system. (Note that if the vectors $\mathbf{h}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ in system (4.4) is replaced by the vectors $\mathcal{H}(\mathbf{x}, \mathbf{s}_1, \dots, \mathbf{s}_n)$ and $\mathcal{F}(\mathbf{x}, \mathbf{s}_1, \dots, \mathbf{s}_n)$, then by virtue of Corollary of Theorem 3.1, the statement of Theorem 4.2 is preserved.)

Definition 4.3. Chaos arising in system (4.6) will be called homoclinic.

1. Suppose that in system (4.4) we have $n = 2$, $\phi = 1$, $\psi = 1$, and $\mathbf{f}(\mathbf{x}) = (x, y)^T$, $\mathbf{h}(\mathbf{x}) = (x^3, y^3)^T$,

$$A_s = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

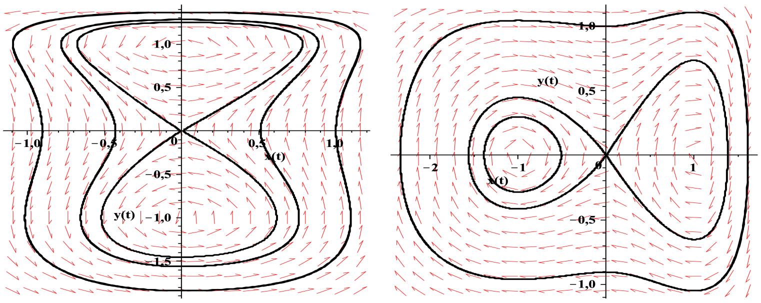
Then, depending on parameter ϕ, we obtain the phase portraits (see Fig. 4.1):



(a1) (a2)

Fig. 4.1. Two types of homoclinic orbits in system (4.4) at $\phi = 5$ (a1) and $\phi = -5$ (a2). The homoclinic orbit resulting from the transformation (at $\|\mathbf{x}_0\| \to 0$): (a1) one periodic trajectory and its self-intersection at point $\mathbf{0}$; (a2) two different periodic trajectories lying to the left and right of the vertical axis and their merging at point $\mathbf{0}$. The point $\mathbf{0}$ is saddle; the eigenvalues of the Jacobi matrix at the point $\mathbf{0}$ are ± 1.

2. Suppose that in system (4.4) we have $n = 2$, $\mathbf{f}(\mathbf{x}) = (x, y)^T$, $\mathbf{h}(\mathbf{x}) = (\mathit{piecewise}(x < 0, -(-x)^{1.5}, x^{3.5}; \mathit{piecewise}(y < 0, -(-y)^{2.5}, y^{7.5})^T$, and matrices A_s and A are the same as in the previous example. Then, depending on parameters ψ and ϕ, we obtain the following phase portraits (see Fig. 4.2):



(a1) (a2)

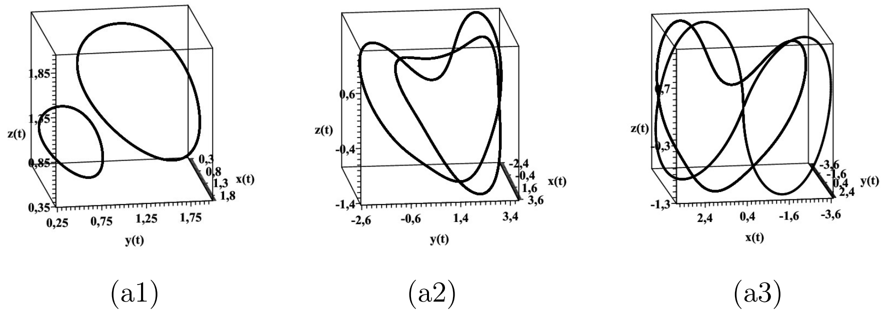
Fig. 4.2. Orbits of system (4.4): (a1) $\psi = 1$, $\phi = 1$; (a2) $\psi = 1$, $\phi = -1$.

3. Now suppose that in system (4.6) we have $n = 3$, $r = 0.01$, and
$\mathbf{f}(\mathbf{x}) = (x^{1.2\vee 1,2}, y^{1.2\vee 1.2}, z^{1.2\vee 1.2})^T$, $\mathbf{h}(\mathbf{x}) = (x^{2\vee 2}, y^{2\vee 2}, z^{2\vee 2})^T$,

$$B = 0 \text{ or } B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -0.25 & -0.46 \\ 0 & 0.46 & 0.26 \end{pmatrix}, A_s = \begin{pmatrix} 0 & -0.8 & -0.675 \\ 0.8 & 0 & 0.7 \\ -0.675 & 0.7 & 0 \end{pmatrix},$$

$$A = \begin{pmatrix} -0.01 & -0.8 & 0.675 \\ 0.8 & -0.01 & -0.7 \\ -0.675 & 0.7 & -0.01 \end{pmatrix}.$$

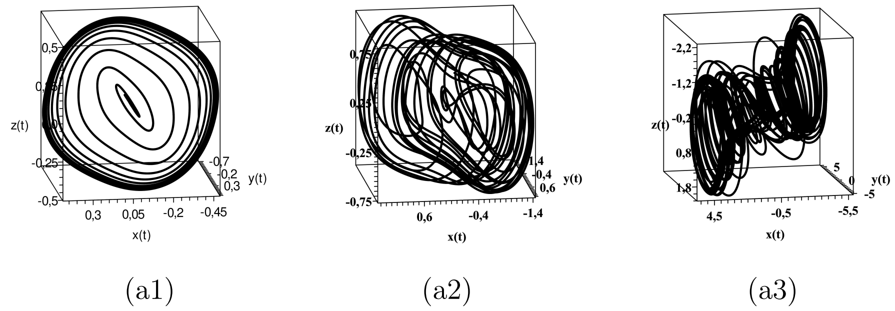
Let us first verify Theorem 4.2 (see Fig.4.3).



(a1) (a2) (a3)

Fig. 4.3. Verification of Theorem 4.2 for $n = 3$ and $B = 0$: (a1) $\psi = 0$, there are only periodic trajectories; (a2) $\psi = -2$, if the starting points are far from the origin, then there are only periodic trajectories; (a3) $\psi = -2$, if the starting point is near the origin ($x_0 = 0, y_0 = 0.001, z_0 = 0.001$), then there is a homoclinic trajectory.

Now let $B \neq 0$. Then, we obtain the following phase portraits (see Fig.4.4):



(a1) (a2) (a3)

Fig. 4.4. The birth of homoclinic chaos in system (4.6), depending on the change in parameter ψ: (a1) $\psi = -0.5$; (a2) $\psi = -1$; (a3) $\psi = -3$.

The result of Theorem 4.2 can be generalized in the following way.

Let $P \in \mathbb{R}^{n\times n}$ be a real matrix. We introduce the vector

$$\mathbf{h}(P\mathbf{x}) = \Big(h_1(\sum_{j=1}^{n} p_{1j}x_j), ..., h_n(\sum_{j=1}^{n} p_{nj}x_j)\Big)^T$$

(see (3.14)).

Introduce also the following $l = 2^n$ matrices:

$$G_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, G_2 = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \ldots,$$

$$G_{l-1} = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, G_l = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Then, equation (4.6) can be rewritten as

$$\dot{\mathbf{x}}(t) = B\mathbf{x} + \psi AG_i\mathbf{f}(\mathbf{x}) + (A - rI)\mathbf{h}(\mathbf{x}); i \in \{2, \ldots, l-1\}. \tag{4.7}$$

(Matrices G_1 and G_l do not change the evenness or oddness of the vector function $\mathbf{f}(\mathbf{x})$. Therefore, they are excluded from further study.)

Equation (4.7) can be used in two directions.

(d1) Simulation of systems containing a limit cycle.

In this case, $\psi = 0$ and instead of (4.7), you can use the system

$$\dot{\mathbf{x}}(t) = \mathbf{c} + B\mathbf{x} + (A - rI)\mathbf{h}(\mathbf{x}), \mathbf{c} \in \mathbb{R}^n, r > 0$$

or the system

$$\dot{\mathbf{x}}(t) = \mathbf{c} + B\mathbf{x} + \mathbf{h}((A - rI)\mathbf{x}), \mathbf{c} \in \mathbb{R}^n, r > 0. \tag{4.8}$$

(Here, the shift vector $\mathbf{c}$ was added to the right side of system (4.7).)

(d2) Simulation of systems containing a homoclinic orbit (for example, Lorenz-like systems [37]).

Clearly, if $\psi \neq 0$, then among the solutions of equation (4.7) there exists a homoclinic orbit.

In addition, the homoclinic orbit can be obtained in the following way.

Let $\psi = -1$, $\mathbf{h}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, $\mathbf{c}, \mathbf{c}_0 \in \mathbb{R}^n$, $r \geq 0$, and instead of (4.7), you can use the system

$$\dot{\mathbf{x}}(t) = \mathbf{c} + B\mathbf{x} + \mathbf{h}((A - rD_0)G_i\mathbf{x} + \mathbf{c}_0) - \mathbf{h}((A - rD_0)G_j\mathbf{x} + \mathbf{c}_0), \tag{4.9}$$

where $D_0 = $diag$(d_{11}, \ldots, d_{nn}) \in \mathbb{R}^{n\times n}; i, j \in \{2, \ldots, l-1\}$. (The case $D_0 = I$ is not excluded.)

Then, for certain values of the parameters, the homoclinic orbit will exist among the solutions of system (4.9). (Note that in system (4.9) the components $h_i(v_i), i = 1, \ldots, n$, of vector $\mathbf{h}(\mathbf{v})$ can be either even or odd activation functions.)

For example, for the Lorenz system

$$\dot{x} = \sigma(y - x), \dot{y} = ax - y - xz, \dot{z} = -bz + xy, \tag{4.10}$$

we have: $\mathbf{h}(\mathbf{x}) = \mathbf{h}(x, y, z) = (h_1(x), h_2(y), h_3(z))^T = (x^2, y^2, z^2)^T, \mathbf{c} = \mathbf{c}_0 = 0, r = 0$,

$$B = \begin{pmatrix} -\sigma & \sigma & 0 \\ a & -1 & 0 \\ 0 & 0 & -b \end{pmatrix},$$

$$A = AG_1 = 0.5 \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}, AG_2 = 0.5 \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix},$$

$$\mathbf{h}(AG_1\mathbf{x}) - \mathbf{h}(AG_2\mathbf{x}) = \begin{pmatrix} 0.25(y+z)^2 \\ 0.25(-x+z)^2 \\ 0.25(-x-y)^2 \end{pmatrix} - \begin{pmatrix} 0.25(y+z)^2 \\ 0.25(x+z)^2 \\ 0.25(x-y)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ -xz \\ xy \end{pmatrix}.$$

The use of representation (4.9) to approximate derivatives can be motivated by the following considerations.

Let $h(u)$ be the even activation function (3.2). Then the function $\phi(u) = h(au + c) - h(-au + c); a, c \in \mathbb{R}$, is the extended odd activation function [25].

Indeed, without loss of generality, we can assume that $a \geq 0$ and $c \geq 0$. Let u^* be the root of function $\phi(u)$ (it is easy to check that this root is unique).

It is clear that if $u \geq u^*$ ($u < u^*$), then $\phi(u) \geq 0$ ($\phi(u) < 0$). Thus, if there are numbers k, i, j such that $h_k((A - rI)G_i\mathbf{x}) - h_k((A - rI)G_j\mathbf{x}) \equiv 0$, then the k-th equation of system (4.9) is a composition of only odd activation functions (linear and $\phi(u_k)$, where u_k is a function of $\mathbf{x}$); $k \in \{1, \ldots, n\}; i \neq j$. Since any of these functions separates points (see Definition 2.1), the k-th equation of system (4.9) satisfies the approximation Theorem 2.2. (Let $\mathbf{c}_0 \neq 0$. Obviously, if numbers $i, j \in \{2, \ldots, l-1\}$, and $i \neq j$ are such that $G_i + G_j = 0$, then all nonzero components of vector $\mathbf{h}((A - rI)G_i\mathbf{x} + \mathbf{c}_0) - \mathbf{h}((A - rI)G_j\mathbf{x} + \mathbf{c}_0)$ are odd activation functions.)

4.2. Examples of attractors of 3D power systems

1. Strange chaotic attractor. Consider a specific system (4.1) for $n = 3$. Assume that in this system

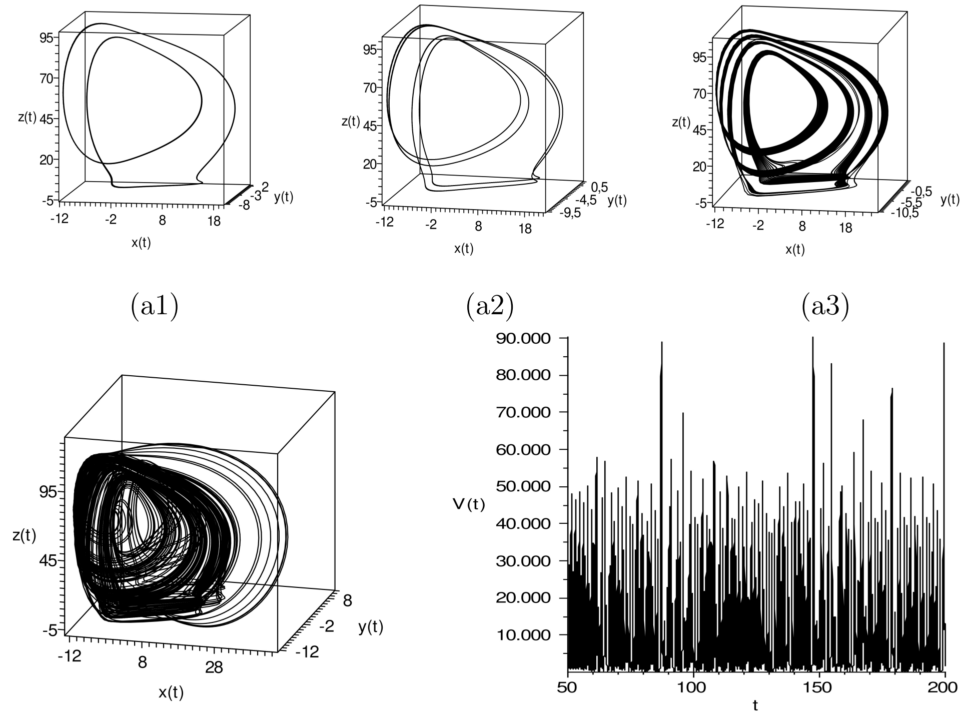
$$\mathbf{v}(\mathbf{x}) \equiv \mathbf{0}, \; B = \begin{pmatrix} 0.1 & -0.2 & 4.8 \\ -2.9 & -1.4 & 2.0 \\ -0.1 & -1.7 & 1.9 \end{pmatrix}, \; A - rI = \begin{pmatrix} -r & -0.1 & -0.7 \\ 0.1 & -r & -0.2 \\ 0.7 & 0.2 & -r \end{pmatrix}.$$

The eigenvalues of the matrix B are $\lambda_{1,2} = -1.1645 \pm 2.4811i, \lambda_3 = 2.9291$. Thus, the origin is the saddle focus.

Let's define vectors: $\mathbf{x} = (x, y, z)^T$, $\mathbf{h}(\mathbf{x}) = (h_1(x), h_2(y), h_3(z))^T$, where

$$
\begin{aligned}
h_1(x) &= -(-x)^{3.0} \text{ if}(x<0) \text{ and } x^{1.8} \text{ if}(x \geq 0)\\
h_2(y) &= -(-y)^{3.0} \text{ if}(y<0) \text{ and } y^{1.8} \text{ if}(y \geq 0)\\
h_3(z) &= -(-z)^{3.0} \text{ if}(z<0) \text{ and } z^{1.5} \text{ if}(z \geq 0).
\end{aligned}
$$

The following Fig. 4.5 shows the transition of system (4.1) from regular regime to chaotic behavior:



(a1) (a2) (a3)

(a4) (a5)

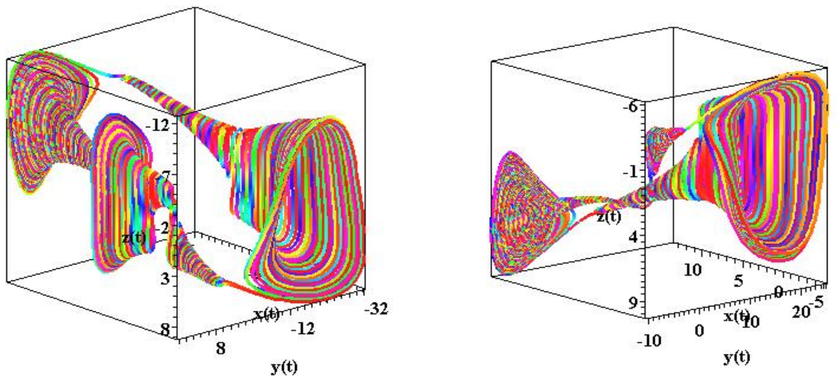
Fig. 4.5. A cascade of bifurcations of the limit cycle in system (4.1) for different values of r (transition to chaos): (a1) $r = 0.024$, (a2) $r = 0.019$, (a3) $r = 0.0183$, (a4) $r = r_c \approx 0.0142$ (it is a chaotic dynamic); the verification of condition 2) of Theorem 4.1 is shown in the graph (a5).

Thus, we have shown that systems (3.5) can simulate a chaotic processes.

2. Strange non-chaotic attractor. Theorems 3.1 and 4.1 are valid if $r > 0$. For $r = 0$ the statements of these theorems have not been proved. In this regard, consider 3D system (4.3) in which $r = 0$ (compare with system (3.17)): $\mathbf{v}(\mathbf{x}) \equiv \mathbf{0}$,

$$
B = \begin{pmatrix} 0 & 0.01 & 0 \\ -0.01 & 0 & 1 \\ 1 & -2 & -0.05 \end{pmatrix}, A_n = H^{-1}AH = \begin{pmatrix} 0 & 0 & 0 \\ 0.01 & 0 & -0.1 \\ 0 & 0.005 & 0 \end{pmatrix}.
$$

In Fig. 4.6 presents a new type of attractors that can be generated by system (4.3) with odd activation functions $h_1(x) = x^{\gamma \vee \delta}, h_2(y) = y^{\gamma \vee \delta}, h_3(z) = z^{\gamma \vee \delta}$:



(a1) (a2)

Fig. 4.6. New strange attractors in system (4.3): (a1) $\gamma = 3.5714, \delta = 3$ and (a2) $\gamma = 3.1428, \delta = 4.3314$

Note that matrix $A_n = H^{-1}AH$ is similar to antisymmetric matrix

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\sqrt{5}/100 \\ 0 & \sqrt{5}/100 & 0 \end{pmatrix}.$$

Thus, system (4.3) allows simulating more complex processes than system (4.1).

5. Generalization of the concept of power activation function

Formulas (3.1) and (3.2), which introduce power activation functions, have two drawbacks:

1) if $0 < \alpha \leq 1$ or $0 < \beta \leq 1$, then the functions (3.1) and (3.2) are non-differentiable;

2) functions (3.1) and (3.2) do not take into account the shift of the argument.

In this connection, we introduce the following function (see Fig.7.1, Fig.7.2):

$$w(u,\alpha,\beta,b,c) = \mathit{piecewise}\Big[u+\frac{b}{c} < -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{1}{\beta}\Big(-\Big(u+\frac{b}{c}\Big)\Big)^{\beta},$$
$$u+\frac{b}{c} \leq c^{\frac{1}{\alpha-1}}, c\cdot\Big(u+\frac{b}{c}\Big), \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha}\Big(u+\frac{b}{c}\Big)^{\alpha}\Big]. \quad (5.1)$$

Here $\alpha > 0$, $\beta > 0$, $\alpha \neq 1$, and $\beta \neq 1$ are degrees; $c > 0$ is the tangent of angle of inclination of a straight line $w = cu + b$; b a given bias of argument.

We put in formula (5.1) $b = 0$. Then we will have

$w(u,\alpha,\beta,c) = \mathit{piecewise}$

$$\Big[u < -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{(-u)^{\beta}}{\beta}, u \leq c^{\frac{1}{\alpha-1}}, cu, \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{u^{\alpha}}{\alpha}\Big]. \quad (5.2)$$

(Formula (5.2) can be obtained from formula (5.1) by introducing a new variable $z := u + b/c$, which in (5.2) is denoted again as $u := z$.)

In the optimization problem using gradient methods, it is necessary to use the derivative of the function $w(u, \alpha, \beta, c)$. In the case of $\alpha > 0$, $\beta > 0$, $\alpha \neq 1, \beta \neq 1$, and $c \geq 0$ this formula is as follows:

$$\dot{w}_u(u, \alpha, \beta, c) = \mathit{piecewise}\Big[u < -c^{\frac{1}{\beta-1}}, (-u)^{\beta-1}, u \leq c^{\frac{1}{\alpha-1}}, c, u^{\alpha-1}\Big]. \tag{5.3}$$

If $\lim \beta \to 1$, then

$$w(u, \alpha, \beta, c) \to \mathit{piecewise}\Big[u \leq c^{\frac{1}{\alpha-1}}, cu, \frac{\alpha-1}{\alpha} c^{\frac{\alpha}{\alpha-1}} + \frac{u^\alpha}{\alpha}\Big],$$

$$\dot{w}_u(u, \alpha, \beta, c) \to \mathit{piecewise}\Big[u \leq c^{\frac{1}{\alpha-1}}, c, u^{\alpha-1}\Big]; \tag{5.4}$$

if $\lim \alpha \to 1$, then

$$w(u, \alpha, \beta, c) \to \mathit{piecewise}\Big[u < -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta} c^{\frac{\beta}{\beta-1}} - \frac{(-u)^\beta}{\beta}, cu\Big],$$

$$\dot{w}_u(u, \alpha, \beta, c) \to \mathit{piecewise}\Big[u < -c^{\frac{1}{\beta-1}}, (-u)^{\beta-1}, c\Big]; \tag{5.5}$$

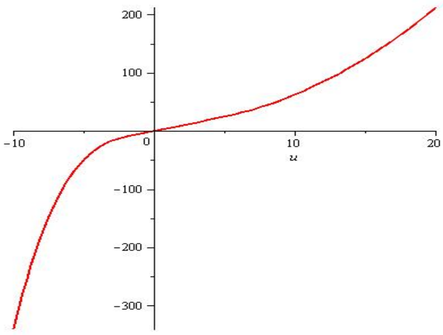
if $\lim \alpha \to 1$ and $\lim \beta \to 1$, then $w(u, \alpha, \beta, c) \to cu$ and $\dot{w}_u(u, \alpha, \beta, c) \to c$.

(Note that formula (5.2) is transformed into formula (5.1) if we put in (5.2) $u := u + b/c$. Thus, we have $w(u + b/c, \alpha, \beta, c) \equiv w(u, \alpha, \beta, b, c)$.)

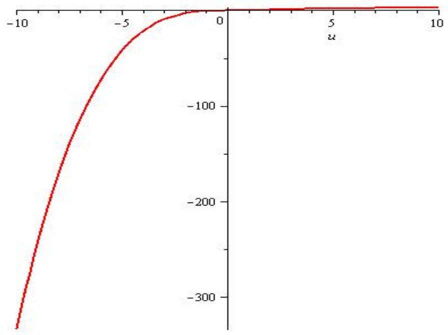
Finally, if we put $c = 0$ in formula (5.2) , then we obtain (with insignificant additions) function (3.1) :

$$w(u, \alpha, \beta) = \mathit{piecewise}\Big[u < 0, -\frac{(-u)^\beta}{\beta}, \frac{u^\alpha}{\alpha}\Big], \tag{5.6}$$

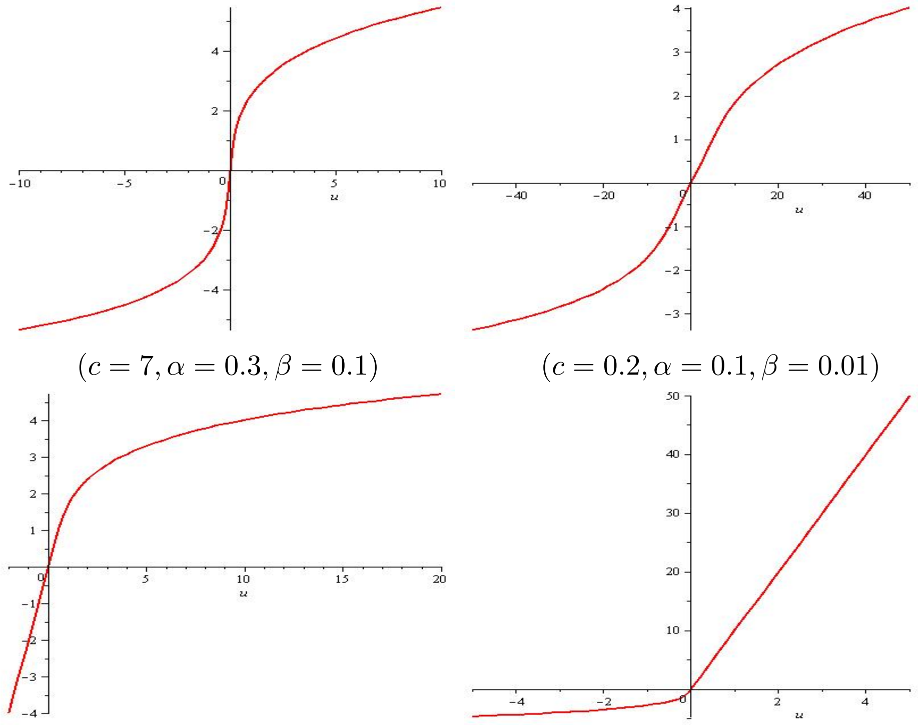
$$\dot{w}_u(u, \alpha, \beta) = \mathit{piecewise}\Big[u < 0, (-u)^{\beta-1}, u^{\alpha-1}\Big]; \alpha > 1, \beta > 1.$$



$(c = 5, \alpha = 2, \beta = 3)$



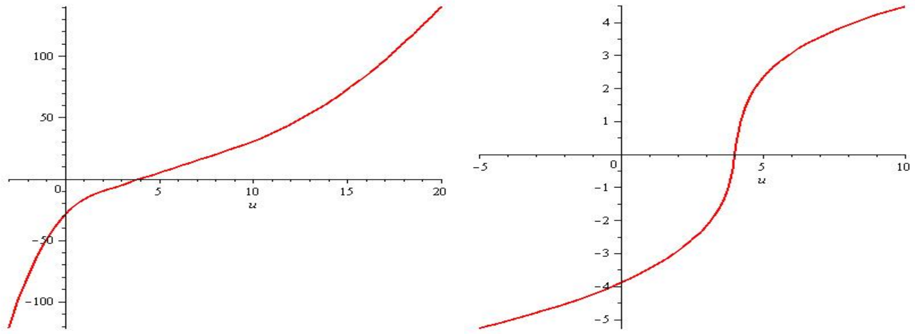
$(c = 0.5, \alpha = 0.2, \beta = 3)$



$(c = 7, \alpha = 0.3, \beta = 0.1)$ $(c = 0.2, \alpha = 0.1, \beta = 0.01)$

$(c = 2, \alpha = 0.01, \beta = 2)$ $(c = 10, \alpha = 1.5, \beta = 0.02)$

Fig.5.1. The activation differentiable power function
$$w(u) = piecewise\Big[u \le -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{(-u)^\beta}{\beta}, u \le c^{\frac{1}{\alpha-1}}, cu, \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{u^\alpha}{\alpha}\Big]$$
for different values of the parameters α, β, and c.



$(c = 5, b = -20, \alpha = 2, \beta = 3)$ $(c = 5, b = 20, \alpha = 0.2, \beta = 0.3)$

Fig.5.2. The activation differentiable power function $w(u) = piecewise\Big[u + \frac{b}{c} \le -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{1}{\beta}(-u - \frac{b}{c})^\beta, u + \frac{b}{c} \le c^{\frac{1}{\alpha-1}}, cu + b, \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha}(u + \frac{b}{c})^\alpha\Big]$ with a given bias $b \neq 0$ for different values of the parameters α, β, and c.

Note that the functions (5.1) and (5.2) are differentiable on the whole interval $(-\infty, \infty)$ for any $\alpha > 0, \alpha \neq 1$ and $\beta > 0, \beta \neq 1$. At the same time, function (5.6) is non-differentiable for $0 < \alpha \leq 1$ or $0 < \beta \leq 1$, at point $u = 0$. (If $\alpha = \beta = 1$, then we get the linear function $w(u) = u$, which is useless for modeling with the help of neural networks.)

Thus, functions (5.1) and (5.2) are by a generalization of the power odd activation function (3.1) (or (5.6)). This generalization is that function (5.2) (unlike function (3.1)) is differentiable. Therefore, it becomes possible to use these functions in the gradient methods of search algorithms (for example, in Algorithms 1 and 2 or in the backpropagation method [1–3, 9, 16]).

6. Algorithms for adjusting the weight coefficients of neural ODEs with power activation functions

Suppose that we study the behavior $\mathbf{x}(t)$ of some dynamical system and we can determine the dimension n of the space in which this system operates.

We introduce real numbers $\Delta t > 0$ and $0 < T < \infty$ such that $T >> n\Delta t$. Assume that for any $t \in [0, T]$ vectors $\mathbf{x}(t + k\Delta t)$ and $\dot{\mathbf{x}}(t + k\Delta t) \in \mathbb{R}^n$ can be measured; $k = 0, 1, \ldots.$ (If the measurement of vector $\dot{\mathbf{x}}(t)$ is impossible, then the standard approximation of derivative

$$\dot{\mathbf{x}}(t) \approx \frac{1}{\Delta t}(\mathbf{x}(t + \Delta t) - \mathbf{x}(t))$$

is used to find it.)

Algorithms are based on the least squares method [1] and the fact that we know sufficient precision the components of $\mathbf{x}(t)$ and its derivative $\dot{\mathbf{x}}(t)$.

Suppose that n time series

$$\begin{array}{l} g_{10}, g_{11}, g_{12}, \cdots, g_{1N}, \\ g_{20}, g_{21}, g_{22}, \cdots, g_{2N}, \\ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot\ \ \cdot \\ g_{n0}, g_{n1}, g_{n2}, \cdots, g_{nN} \end{array}$$

are given on the same time interval T in equally spaced N nodes: $0, \Delta t, \ldots, k\Delta t, \ldots, N\Delta t = T$. Thus, $\Delta t = T/N$.

The objective is to determine the dynamical system described by the equation $\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x})$ from the known time series; here $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is realized by a multilayer neural network $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t \cdot \mathbf{F}(\mathbf{x}_k); k = 0, 1, \ldots$ (see [9–11, 17, 25]). The difference between the input $\mathbf{x}$ and the output $\mathbf{F}(\mathbf{x})$ is compared with $\dot{\mathbf{x}}(t)$ to generate an error $\mathbf{e}(t) \in \mathbb{R}^n$. This error is used to adjust the network parameters so that $\mathbf{e}(t) = \dot{\mathbf{x}}(t) - \mathbf{F}(\mathbf{x}) \to 0$. The function $\mathbf{F}(\mathbf{x})$ is the right-hand side of system (3.5) with unknown parameters. It is necessary to minimize the error $\mathbf{e}(t)$ (in any norm) by adjusting some of the parameters included in (3.5).

6.1. Algorithm 1: architecture of neural ODEs does not use antisymmetric matrices

The algorithm implements a search procedure for approximating a time series by solutions of a system of differential equations.

In this system: 1. There is a linear part. 2. The nonlinear part contains one fixed even and one odd activation function that must be adjusted. 3. All matrices of the system are matrices of general form (antisymmetric matrices are not used).

The model describing the process should be presented in the form of the following system of differential equations of order n:

$$\begin{cases} \dot{x}_1(t) = c_{10} + c_{11}x_1 + \cdots + c_{1n}x_n + b_{11}u(x_1) + b_{12}u(x_2) + \cdots + b_{1n}u(x_n) \\ \qquad\quad +d_{11}h(x_1) + d_{12}h(x_2) + \cdots + d_{1n}h(x_n), \\ \dot{x}_2(t) = c_{20} + c_{21}x_1 + \cdots + c_{2n}x_n + b_{21}u(x_1) + b_{22}u(x_2) + \cdots + b_{2n}u(x_n) \\ \qquad\quad +d_{21}h(x_1) + d_{22}h(x_2) + \cdots + d_{2n}h(x_n), \\ \qquad\quad . \; , \\ \dot{x}_n(t) = c_{n0} + c_{n1}x_1 + \cdots + c_{nn}x_n + b_{n1}u(x_1) + b_{n2}u(x_2) + \cdots + b_{nn}u(x_n) \\ \qquad\quad +d_{n1}h(x_1) + d_{n2}h(x_2) + \cdots + d_{nn}h(x_n). \end{cases} \tag{6.1}$$

Here $u(x_i) = piecewise(x_i < 0, (-x_i)^\delta, x_i^\gamma)$; $\delta > 1$, $\gamma > 1$ (even functions with fix degrees); $h(x_i) = piecewise(x_i < 0, -(-x_i)^\beta, x_i^\alpha)$; $\alpha > \gamma$, $\beta > \delta$ (odd function with adjustable degrees); $i = 1, \ldots, n$.

The purpose of the algorithm is to determine vector $c_0 = (c_{10}, \ldots, c_{n0})^T \in \mathbb{R}^n$, matrices $C = \{c_{ij}\} \in \mathbb{R}^{n\times n}, B = \{b_{ij}\} \in \mathbb{R}^{n\times n}$, $D = \{d_{ij}\} \in \mathbb{R}^{n\times n}$, and degrees α, β of system (6.1); $i, j = 1, \ldots, n$. The answer is presented in the form of matrix $Y = (c_0, C, B, D) \in \mathbb{R}^{n\times(3n+1)}$.

1. Fix a learning selections

$$\begin{array}{c} g_{10}, g_{11}, g_{12}, \ldots, g_{1m}, \\ g_{20}, g_{21}, g_{22}, \ldots, g_{2m}, \\ . \; . \; . \; . \; . \; . \; . \; . \; . \; . \\ g_{n0}, g_{n1}, g_{n2}, \ldots, g_{nm}, \end{array}$$

where $0 < 1 + 3n \leq m \leq N$. (The number of training elements m in the time series should be more than the number of unknown coefficients $1 + 3n$ in any of the equations of the system.)

1.1. Introduce positive numbers $step, \gamma, \delta$, and integers $M_a > 0$, $M_b > 0$. (Let, for example, be: $step := 0.1$, $\gamma := 1.5$, $\delta := 1.5$, and $M_a := 20$, $M_b := 20$.)

1.2. Construct the matrix of numerical derivatives

$$DER := \frac{1}{\Delta t}\begin{pmatrix} g_{11} - g_{10} & g_{21} - g_{20} & \cdots & g_{n1} - g_{n0} \\ g_{12} - g_{11} & g_{22} - g_{21} & \cdots & g_{n2} - g_{n1} \\ \vdots & \vdots & \vdots & \vdots \\ g_{1,m} - g_{1,m-1} & g_{2,m} - g_{2,m-1} & \cdots & g_{n,m} - g_{n,m-1} \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

1.3. Construct the matrix

$$U := \begin{pmatrix} u(g_{10}) & u(g_{20}) & \dots & u(g_{n0}) \\ u(g_{11}) & u(g_{21}) & \dots & u(g_{n1}) \\ \vdots & \vdots & \vdots & \vdots \\ u(g_{1,m-1}) & u(g_{2,m-1}) & \dots & u(g_{n,m-1}) \end{pmatrix} \in \mathbb{R}^{m\times n}.$$

2. Put $k_a := 1$, $k_b := 1$. (Start a double cycle for integer variables k_a and k_b.)

2.1. Fix the number $\alpha := \gamma + k_a \cdot step$.

2.2. Fix the number $\beta := \delta + k_b \cdot step$.

2.3. Construct a perturbation matrix

$$H := \begin{pmatrix} h(g_{10}) & h(g_{20}) & \dots & h(g_{n0}) \\ h(g_{11}) & h(g_{21}) & \dots & h(g_{n1}) \\ \vdots & \vdots & \vdots & \vdots \\ h(g_{1,m-1}) & h(g_{2,m-1}) & \dots & h(g_{n,m-1}) \end{pmatrix} \in \mathbb{R}^{m\times n},$$

and the Jacobi matrix for system (6.1):

$$W := \left(\begin{array}{cccc|c|c} 1 & g_{10} & \cdots & g_{n0} & & \\ 1 & g_{11} & \cdots & g_{n1} & & \\ \vdots & \vdots & \vdots & \vdots & U & H \\ 1 & g_{1,m-1} & \cdots & g_{n,m-1} & & \end{array}\right) \in \mathbb{R}^{m\times(3n+1)}.$$

2.4. Check the condition $\operatorname{rank} W = 3n + 1$. If $\operatorname{rank} W < 3n + 1$ then fix some number $\mu \in (0, 1]$ else $\mu := 0$.

2.5. Find the matrix Y using the least squares method:

$$Y^T = (W^T W + \mu I)^{-1} W^T \cdot DER \in \mathbb{R}^{(3n+1)\times n}.$$

(Here $I \in \mathbb{R}^{(3n+1)\times(3n+1)}$ is the identity matrix.)

3. Compute the matrix of errors

$$ERR := (e_{ij}) = DER - W \cdot Y^T \in \mathbb{R}^{m\times n}$$

and the total computational error at iterative steps k_a and k_b:

$$E_{k_a,k_b} := tr(ERR^T \cdot ERR) = \sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}^2.$$

4. If $k_b < M_b$ then $k_b := k_b + 1$ and go to item 2.2.

4.1. If $k_a = M_a$ then go to item **5** else $k_b := 1$, $k_a := k_a + 1$ and go to item 2.1.

5. Compute the integer numbers $l_a \in \{1, ..., M_a\}$ and $l_b \in \{1, ..., M_b\}$ such that

$$E_{l_a,l_b} = \min(E_{1,1}, \dots, E_{1,M_b}, E_{2,1}, \dots, E_{2,M_b}, \dots, E_{M_a,1}, \dots, E_{M_a,M_b}).$$

(End of the double cycle for integer variables k_a and k_b.)
6. Print the matrix $Y \equiv Y_{l_a,l_b} \in \mathbb{R}^{n\times(3n+1)}$, the numbers $\alpha > 1$, $\beta > 1$ and stop the algorithm.
7. Solve system (6.1). If the solutions of system (6.1) diverge, then to increase the values γ and δ by a small number $\Delta > 0$ ($\gamma := \gamma + \Delta > 0$, $\delta := \delta + \Delta$) and go to item **2**. If for several values γ and δ the solutions of system (6.1) still diverge, then stop the algorithm.

Comment 6.1. If there are no matrices B or D in system (6.1), then there are no fixed degrees of activation functions. Indeed, let $B = 0$, then Algorithm 1 works directly. If $D = 0$ holds, then the following redesignations of variables $B \to D, \gamma \to \alpha$, and $\delta \to \beta$ should be introduced into Algorithm 1.
Comment 6.2. If the matrix $D + D^T$ is negative definite, then the resulting model of neural ODE will generate bounded solutions (see [25]).

6.2. Algorithm 2: antisymmetric matrices are used in the architecture of neural ODEs

It is known that in most real dynamical systems, chaotic phenomena arise as a result of the development of certain periodic processes. In this regard, in order to model the chaotic behavior of such systems, it is desirable to include a mechanism generating a limit cycle in the architecture of the neural network. In turn, the cascade of bifurcations of the limit cycle will lead to the appearance of chaos in the modeled system. The following algorithm is designed to simulate limit cycles in real dynamic processes.

As an object of modeling, we will choose some generalization of system (4.8). Consider the following matrix

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_{n-1} \\ S_n \end{pmatrix} = \begin{pmatrix} -d_{11} & d_{12} & \dots & d_{1,n-1} & d_{1n} \\ -d_{12} & -d_{22} & \dots & d_{2,n-1} & d_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -d_{1,n-1} & -d_{2,n-1} & \dots & -d_{n-1,n-1} & d_{n-1,n} \\ -d_{1n} & -d_{2n} & \dots & d_{n-1,n} & -d_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n} \tag{6.2}$$

such that $(S + D_0) + (S + D_0)^T = 0$. (Thus, $(S + D_0)$ is antisymmetric.)

Now, we introduce the following system

$$\begin{cases} \dot{x}_1(t) = f_1(c_{10} + \sum\limits_{i=1}^{n} c_{1i}x_i) + h_1(S_1 \cdot \mathbf{x} + b_1), \\ \dot{x}_2(t) = f_2(c_{20} + \sum\limits_{i=1}^{n} c_{2i}x_i) + h_2(S_2 \cdot \mathbf{x} + b_2), \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \dot{x}_n(t) = f_n(c_{n0} + \sum\limits_{i=1}^{n} c_{ni}x_i) + h_n(S_n \cdot \mathbf{x} + b_n). \end{cases} \tag{6.3}$$

Here $f_i(u) = piecewise(u < 0, \pm(-u)^{\beta_i}, u^{\alpha_i})$ are odd or even power functions, $h_i(v) = piecewise(v < 0, -(-v)^{\gamma_i},\ v^{\gamma_i})$ are odd power functions and $\gamma_i > 1$;

$i = 1, \ldots, n$. (System (4.8) follows from system (6.3) if we put $\deg f_i(u) = 1$, $D_0 = I$, and $b_i = 0$ in the last system.)

It is known that the basis of search algorithms is the Jacobi matrix Jac. Let's form this matrix for system (6.3) using representation (6.2). In this case, the vector of parameters $\mathbf{z} \in \mathbb{R}^{C_n^2+n(n+2)}$ will be constructed as follows :

$$\mathbf{z} = (c_{10}, \ldots, c_{1n}, \ldots, c_{n0}, \ldots, c_{nn}, d_{11}, \ldots, d_{1n}, d_{22}, \ldots, d_{2n}, \ldots, d_{nn}, b_1, \ldots, b_n)^T.$$

Let $m = 1$. We introduce the following auxiliary matrices that will be needed to construct the Jacobi matrix for system (6.3):

$$J_1 = Diag_1 \cdot \begin{pmatrix} (1, x_1, \ldots, x_n) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (1, x_1, \ldots, x_n) \end{pmatrix} \in \mathbb{R}^{n \times (n+1)n},$$

$$J_2 = Diag_2 \cdot (-\mathbf{x}, P_1, P_2, \ldots, P_{n-1}, I_n) \in \mathbb{R}^{n \times (C_n^2+n+1)},$$

where

$$Diag_1 = \begin{pmatrix} f_1'(c_{10} + \sum\limits_{i=1}^{n} c_{1i}x_i) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & f_n'(c_{n0} + \sum\limits_{i=1}^{n} c_{ni}x_i) \end{pmatrix} \in \mathbb{R}^{n \times n},$$

$$Diag_2 = \begin{pmatrix} h_1'(S_1 \cdot \mathbf{x} + b_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & h_n'(S_n \cdot \mathbf{x} + b_n) \end{pmatrix} \in \mathbb{R}^{n \times n},$$

and

$$P_i = \begin{pmatrix} O_{(i-1)\times(n-i)} \\ ------ \\ x_{i+1}, \ldots, x_n \\ ----- \\ -x_i I_{n-i} \end{pmatrix} \in \mathbb{R}^{n \times (n-i)};$$

$O_{(i-1)\times(n-i)} \in \mathbb{R}^{(i-1)\times(n-i)}$ is the zero matrix; $I_{n-i} \in \mathbb{R}^{(n-i)\times(n-i)}$ and $I_n \in \mathbb{R}^{n\times n}$ are the identity matrices; $i = 1, ..., n-1$.

Now, we briefly describe the algorithm for adjusting the weights for system (6.3).

1. Introducing objects that do not change in the iterative process.

1.1. Fix a learning selections (they are formed from time series):

$$g_{10}, g_{11}, g_{12}, \ldots, g_{1m} \equiv \mathbf{g}_1^T$$

$$g_{20}, g_{21}, g_{22}, \ldots, g_{2m} \equiv \mathbf{g}_2^T$$

$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$

$$g_{n0}, g_{n1}, g_{n2}, \ldots, g_{nm} \equiv \mathbf{g}_n^T.$$

1.2. Fix triplets of real numbers $\alpha_i \geq 1$, $\beta_i \geq 1$ and $\gamma_i > \alpha_i, \gamma_i > \beta_i$; $i = 1, \ldots, n$.

1.3. Introduce the matrix

$$G = \begin{pmatrix} 1 & g_{10} & g_{20} & \cdots & g_{n0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & g_{1,m-1} & g_{2,m-1} & \cdots & g_{n,m-1} \end{pmatrix} = (\mathbf{1}, \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n) \in \mathbb{R}^{m\times(n+1)}.$$

1.4. Construct the columns of numerical derivatives:

$$D_i = \frac{1}{\Delta t} \begin{pmatrix} g_{i1} - g_{i0} \\ g_{i2} - g_{i1} \\ \vdots \\ g_{i,m} - g_{i,m-1} \end{pmatrix} \in \mathbb{R}^m; i = 1, \ldots, n.$$

2. $k = 0$. (The beginning of a global cycle.) Introduce the nonzero vector $Y_k \in \mathbb{R}^{C_n^2+n(n+2)}$ of initial approximations:

$$Y_k := (c_{10}, \ldots, c_{1n}, \ldots, c_{n0}, \ldots, c_{nn}, d_{11}, \ldots, d_{1n}, d_{22}, \ldots, d_{2n}, \ldots, d_{nn}, b_1, \ldots, b_n)^T;$$

$$\text{(for example)} Y_0 := (1, \ldots, 1, d_{11} = 0, ..., d_{nn} = 0, 0, \ldots, 0) \in \mathbb{R}^{C_n^2+n(n+2)},$$

the previous error $T_{k-1} = T_{-1} := 10^{10}$, and the number $\epsilon := 0.001$.

2.1. Construct the matrices

$$F_i = \begin{pmatrix} f_{11}^{(i)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{mm}^{(i)} \end{pmatrix} \in \mathbb{R}^{m\times m},$$

where $f_{jj}^{(i)} = piecewise(c_{i0} + \sum\limits_{s=1}^{n} c_{is}g_{sj} < 0, -\beta_i(-c_{i0} - \sum\limits_{s=1}^{n} c_{is}g_{sj})^{\beta_i - 1}, \alpha_i(c_{i0} + \sum\limits_{s=1}^{n} c_{is}g_{sj})^{\alpha_i-1})$, and

$$H_i = \begin{pmatrix} h_{11}^{(i)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & h_{mm}^{(i)} \end{pmatrix} \in \mathbb{R}^{m\times m},$$

where $h_{jj}^{(i)} = piecewise(\sum\limits_{l=1}^{n} S_{il}g_{lj} + b_i < 0, \gamma_i \cdot (-\sum\limits_{l=1}^{n} S_{il}g_{lj} - b_i)^{\gamma_i - 1}, \gamma_i \cdot (\sum\limits_{l=1}^{n} S_{il}g_{lj} + b_i)^{\gamma_i-1})$; $i = 1, \ldots, n$; $j = 0, 1, \ldots, m-1$.

2.2. Now we will use auxiliary matrices J_1 and J_2 obtained for $m = 1$ in order to construct the Jacobi matrix of system (6.3) in case $m > 1$. To do this, we need to replace scalar variables $1, x_1, \dots, x_n$, respectively, by vectors vectors $\mathbf{1}, \mathbf{g}_1, \dots, \mathbf{g}_n \in \mathbb{R}^m$. Then, we have

$$Jac = (J_1(\mathbf{1}, \mathbf{g}_1, \dots, \mathbf{g}_n) | J_2(\mathbf{1}, \mathbf{g}_1, \dots, \mathbf{g}_n)) = W_k$$

$$= \left(\begin{array}{cccc|cccc|} F_1 \cdot G & 0 & \dots & 0 & -H_1 \cdot \mathbf{g}_1 & H_1 \cdot \mathbf{g}_2 & \dots & H_1 \cdot \mathbf{g}_n \\ 0 & F_2 \cdot G & \dots & 0 & 0 & -H_2 \cdot \mathbf{g}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & F_n \cdot G & 0 & 0 & \dots & -H_n \cdot \mathbf{g}_1 \end{array}\right.$$

$$\left.\begin{array}{|cccc|c|c|cccc} 0 & 0 & \dots & 0 & & 0 & H_1 \cdot \mathbf{1} & 0 & \dots & 0 \\ -H_2 \cdot \mathbf{g}_2 & H_2 \cdot \mathbf{g}_3 & \dots & H_2 \cdot \mathbf{g}_n & & 0 & 0 & H_2 \cdot \mathbf{1} & \dots & 0 \\ 0 & -H_3 \cdot \mathbf{g}_2 & \dots & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -H_n \cdot \mathbf{g}_2 & & -H_n \cdot \mathbf{g}_n & 0 & 0 & \dots & H_n \cdot \mathbf{1} \end{array}\right).$$

Here W_k is $(nm \times (C_n^2 + n(n+2)))$-matrix and $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^m$.

2.3. Check condition $\operatorname{rank} W_k = C_n^2 + n(n+2)$. If $\operatorname{rank} W_k < C_n^2 + n(n+2)$ then fix some number $\mu \in (0, 1]$ else $\mu := 0$.

2.4. Calculation of the right side of system (6.3):

$$R_{ij} = piecewise(c_{i0} + \sum_{s=1}^{n} c_{is} g_{sj} < 0, -(-c_{i0} - \sum_{s=1}^{n} c_{is} g_{sj})^{\beta_i}, (c_{i0} + \sum_{s=1}^{n} c_{is} g_{sj})^{\alpha_i})$$

$$+piecewise(\sum_{l=1}^{n} S_{il} g_{lj} + b_i < 0, -(-\sum_{l=1}^{n} S_{il} g_{lj} - b_i)^{\gamma_i}, (\sum_{l=1}^{n} S_{il} g_{lj} + b_i)^{\gamma_i});$$

where $i = 1, ..., n$; $j = 0, ..., m-1$.

2.5. Form vectors:

$$R_i = \begin{pmatrix} R_{i0} \\ R_{i1} \\ \vdots \\ R_{i,m-1} \end{pmatrix} \in \mathbb{R}^m; i = 1, \dots, n.$$

3. Compute the vector of errors

$$E_k = (e_1, \dots, e_m, e_{m+1}, \dots, e_{2m}, e_{2m+1}, \dots, e_{3m}, \dots, e_{n-1,m}, \dots, e_{nm})^T :=$$

$$\begin{pmatrix} D_1 - R_1 \\ D_2 - R_2 \\ \vdots \\ D_n - R_n \end{pmatrix} \in \mathbb{R}^{nm}$$

and the total computation error at iteration step k: $T_k := \sum_{j=1}^{nm} e_j^2$.

4. If

$$T_{k-1} > T_k + \epsilon,$$

then compute the vector

$$Y_{k+1}^T := Y_k^T + (W_k^T \cdot W_k + \mu I)^{-1} \cdot W_k^T \cdot E_k \in \mathbb{R}^{C_n^2+n(n+2)}$$

else go to item 6. (Here $I \in \mathbb{R}^{(C_n^2+n(n+2))\times(C_n^2+n(n+2))}$ is the identity matrix.)

5. Put $k := k+1$,

$$Y_k := (c_{10}, \dots, c_{1n}, \dots, c_{n0}, \dots, c_{nn}, d_{11}, \dots, d_{1n}, d_{22}, \dots, d_{2n}, \dots, d_{nn}, b_1, \dots, b_n)^T$$

and go to item 2.1.

6. Print a graph of error changes $T_0, T_1, \dots, T_{k-1}$.

7. Print the vector $Y_{k-1} \in \mathbb{R}^{C_n^2+n(n+2)}$ and stop the algorithm.

8. Solve system (6.3). (If the solutions of system (6.3) diverge, then in the initial conditions replace the value $d_{11} = 0$ with a sufficiently small value $|d_{11}|$ and go to item 2.)

Comment 6.3. A chaotic dynamic process modeling should begin with building a limit cycle. As Theorems 3.2 and 3.3 show, for a periodic trajectory to appear in system (6.3), it is sufficient that $b_1 = \dots = b_n = 0$. Therefore, Algorithm 2 can be started by putting $b_1 = \dots = b_n = 0$.

Comment 6.4. If the constructed model does not satisfy the necessary requirements, then it is possible to increase the number of neurons in the equation in accordance with the architecture of system (3.5).

7. Modeling

7.1. Applications of Algorithm 1

To test the operability of Algorithm 1, the following system [36] will be used:

$$\begin{cases} \dot{x}(t) = y(t), \\ \dot{y}(t) = -x(t) + z(t), \\ \dot{z}(t) = 2 - 0.8x^2(t) + z^2(t). \end{cases} \tag{7.1}$$

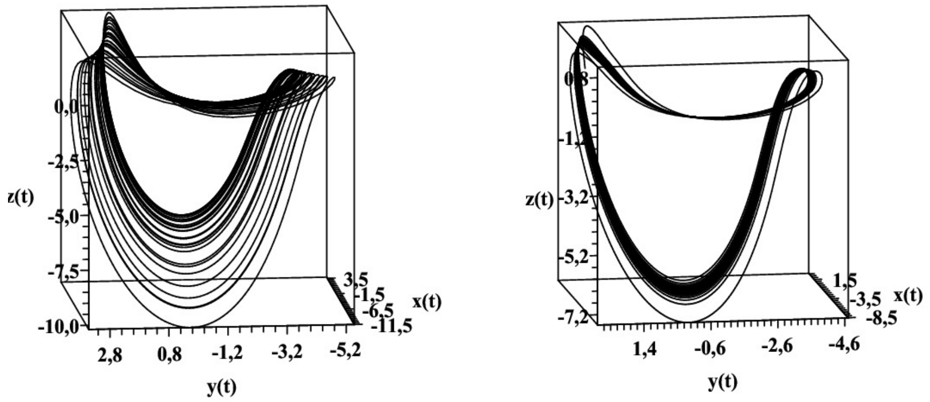
For system (7.1), we solve the Cauchy problem with initial data $x_0 = 0, y_0 = 2.3, z_0 = 0$. From the obtained continuous $x(t), y(t), z(t)$, we form time series with step $\Delta t = 1$. To denote these series, we use the notations $x_1(t) = x(t), x_2(t) = y(t), x_3(t) = z(t); t = 0, 1, 2, \dots, 300$. Thus, we have $m = 300$.

We assume that $B = 0$. Then the use of Algorithm 1 with odd activation functions leads to such coefficients of system (6.1):

$$c_0 = \begin{pmatrix} 0.00020 \\ -0.01967 \\ 1.96685 \end{pmatrix}, C = \begin{pmatrix} 0.00955 & 0.99992 & -0.00951 \\ -0.95541 & 0.00848 & 0.95059 \\ -4.44895, & 0.15144 & 4.93128 \end{pmatrix},$$

$$D = \begin{pmatrix} 0.00000 & 0.00000 & 0.00000 \\ -0.03484 & 0.00131 & 0.04126 \\ 3.48409 & -0.13084 & -4.12689 \end{pmatrix};\ \alpha = 0.66666, \beta = 1.55555. \tag{7.2}$$

Phase portraits of system (7.1) and (7.2) are presented in Fig. 7.1:



(a1) (a2)

Fig. 7.1. Phase portraits of system (7.1)(a1) and (7.2)(a2)

Now, in order to simulate processes in system (7.1), we assume $D = 0$ and again use Algorithm 1, but with even activation functions. Then we get such coefficients of system (6.1):

$$c_0 = \begin{pmatrix} 0.00021 \\ -0.02067 \\ 2.06700 \end{pmatrix}, C = \begin{pmatrix} 0.01000 & 0.99990 & -0.01000 \\ -0.99964 & 0.01018 & 0.99956 \\ -0.02591 & -0.01766 & 0.03368 \end{pmatrix},$$

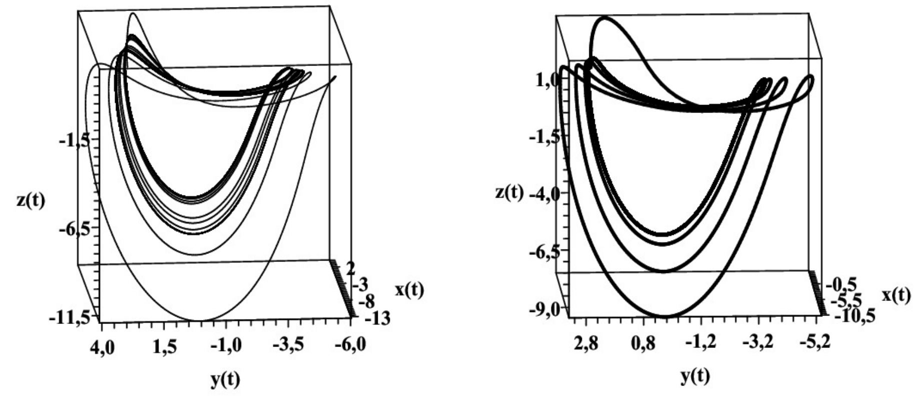
$$B = \begin{pmatrix} 0.00000 & 0.00000 & 0.00000 \\ 0.00735 & 0.00000 & 0.00926 \\ -0.73552 & -0.00296 & 0.92592 \end{pmatrix}; \gamma = 2.05263, \delta = 2.05263 \tag{7.3}$$

and

$$c_0 = \begin{pmatrix} 0.00006 \\ -0.00619 \\ 0.61919 \end{pmatrix}, C = \begin{pmatrix} 0.00989 & 0.99993 & -0.00021 \\ -0.98893 & 0.00679 & 0.98823 \\ -1.09730 & 0.32107 & 1.16668 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.00000 & 0.00000 & 0.00000 \\ 0.02060 & -0.00296 & -0.02438 \\ -2.06034 & 0.29609 & 2.43823 \end{pmatrix}; \gamma = 0.66666, \delta = 1.55555. \tag{7.4}$$

Phase portraits of system (7.3) and (7.4) are presented in Fig.7.2:



(a1) (a2)

Fig. 7.2. Phase portraits of system (7.3)(a1) and (7.4)(a2)

Thus, we can assert that the use of even activation functions leads to more accurate modeling (especially, system (7.3)) than use of odd activation functions.

7.2. Applications of Algorithm 2

Consider the following special case of system (4.9):

$$\begin{cases} \dot{x}_1(t) = a_{10} + \sum\limits_{j=1}^{3} a_{1j}x_j + h(-d_{11}x_1 + d_{12}x_2 + d_{13}x_3) \\ \qquad\quad +\nu h(-l_1d_{11}x_1 + l_2d_{12}x_2 + l_3d_{13}x_3) \\ \dot{x}_2(t) = a_{20} + \sum\limits_{j=1}^{3} a_{2j}x_j + h(-d_{12}x_1 - d_{22}x_2 + d_{23}x_3) \\ \qquad\quad +\nu h(-l_1d_{12}x_1 - l_2d_{22}x_2 + l_3d_{23}x_3) \\ \dot{x}_3(t) = a_{30} + \sum\limits_{j=1}^{3} a_{3j}x_j + h(-d_{13}x_1 - d_{23}x_2 - d_{33}x_3) \\ \qquad\quad +\nu h(-l_1d_{13}x_1 - l_2d_{23}x_2 - l_3d_{33}x_3). \end{cases} \tag{7.5}$$

Here if $\nu = 0(\nu = -1)$, then there is a limit cycle (a homoclinic orbit).

In addition, if $\nu = 0$, then $h(v_i) = piecewise(v_i < 0, -(-v_i)^\gamma, v_i^\gamma)$ is an odd function; if $\nu = -1$, then $h(v_i) = piecewise(v_i < 0, (-v_i)^\gamma, v_i^\gamma)$ is an even function. Here $\gamma > 1$; $d_{11}, d_{12}, d_{13}, d_{23} \in \mathbb{R}$, where $d_{11} \geq 0$; $i = 1, \ldots, 3$.

A triple of integers (l_1, l_2, l_3) accepts only one set of $2^n - 2 = 6$ $(n = 3)$ possible: $(-1, -1, 1); (-1, 1, -1); (1, -1, -1); (1, 1, -1); (1, -1, 1); (-1, 1, 1)$. The triplet (l_1, l_2, l_3) indicates one of the possible types of homoclinic orbits that can be realized in system (7.5).

In this subsection, Algorithm 2 will only be used to simulate systems with possible homoclinic orbits $(\nu = -1)$. Therefore, instead of the nonlinear part

$\mathbf{h}(S\mathbf{x}+\mathbf{b})$ in the system (6.3), the nonlinear part $\mathbf{h}(SG_1\mathbf{x}+\mathbf{b})-\mathbf{h}(SG_2\mathbf{x}+\mathbf{b})$ should be used.

In order to check the performance of Algorithm 2, several well-known systems were used. From the obtained continuous $x(t), y(t), z(t)$, we form time series with step $\Delta t = 1$, where $t = 0, 1, 2, \ldots, 200$. Thus, we have $m = 200$.

1. Lorenz system (4.10) at $x_0 = 1, y_0 = 1, z_0 = 8$. For the Lorentz approximation, we need to take $(l_1, l_2, l_3) = (-1, 1, 1)$.

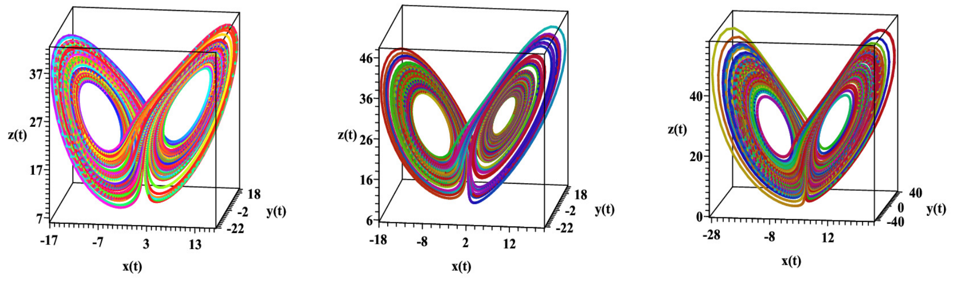
1.1. For the classical Lorentz attractor, we have for system (4.10) $\sigma = 10, a = 27, b = 2.7$ and $h(v_i) = \mathit{piecewise}(v_i < 0, (-v_i)^2, v_i^2); i = 1, \ldots, 3$. A graph of this system is shown in Fig. 7.3 (a1).

1.2. We form time series for system (4.10) and assign $h(v_i) = \mathit{piecewise}(v_i < 0, (-v_i)^{1.5}, v_i^{1.5})$; $i = 1, \ldots, 3$. Then Algorithm 2 applied to system (7.5) leads to the following results:

$$A_0 = \begin{pmatrix} 0.71269 \\ 0.69584 \\ 3.92098 \end{pmatrix}, A = \begin{pmatrix} -10.01641 & 9.94437 & -0.00899 \\ 54.83057 & -2.42695 & -0.02138 \\ -0.44238 & 0.27646 & -3.32041 \end{pmatrix},$$

$d_{11} = 0.00607, d_{12} = 2.63824, d_{13} = 1.08829, d_{22} = d_{11}, d_{23} = 1.68198, d_{33} = d_{11}$. The graph of this system is shown in Fig. 7.3(a2).

1.3. Now we add the following terms to the corresponding right-hand sides of each of equations (7.5):



(a1) (a2) (a3)

Fig. 7.3. Phase portraits of system (4.10)(a1), (7.5)(a2), and [(7.5)+(7.6)](a3)

$$h_1(a_{10} + \sum_{j=1}^{3} a_{1j}x_j), h_1(a_{20} + \sum_{j=1}^{3} a_{2j}x_j), h_1(a_{30} + \sum_{j=1}^{3} a_{3j}x_j), \tag{7.6}$$

where $h_1(v_i) = \mathit{piecewise}(v_i < 0, -(-v_i)^{1.1}, v_i^{1.1}); i = 1, \ldots, 3$. In this case, Algorithm 2 leads to the following results:

$$A_0 = \begin{pmatrix} -0.09276 \\ -0.37044 \\ -6.40899 \end{pmatrix}, A = \begin{pmatrix} -4.25589 & 4.15270 & 0.00343 \\ 27.12403 & -2.14111 & 0.00926 \\ 0.00761 & -0.00245 & -1.53054 \end{pmatrix},$$

$d_{11} = 0.00081, d_{12} = 2.75407, d_{13} = 0.83019, d_{22} = d_{11}, d_{23} = 2.63333, d_{33} = d_{11}$. The graph of this system is shown in Fig.7.3 (a3).

Thus, in the case 1.3 (see Fig. 7.3(a3)), there was no improvement in the quality of approximation.

2. Consider the following system without equilibrium points [38]:

$$\begin{cases} \dot{x}(t) = 10y(t), \\ \dot{y}(t) = -x(t) + 3z(t) + x(t)z(t), \\ \dot{z}(t) = 1 + x(t) - z(t) - x(t)y(t) + 0.25x(t)z(t), \end{cases} \tag{7.7}$$

where $x_0 = y_0 = z_0 = 1$. The graph of this system is shown in Fig. 7.4 (a1).

2.1. From the obtained continuous $x(t), y(t), z(t)$ of system (7.7), we form time series with step $\Delta t = 1$, where $t = 0, 1, 2, \ldots, 5000$. Thus, we have $m = 5000$. We assign $h(v_i) = piecewise(v_i < 0, (-v_i)^2, v_i^2)$; $i = 1, \ldots, 3$. Then Algorithm 2 applied to system (7.5) leads to the following results:

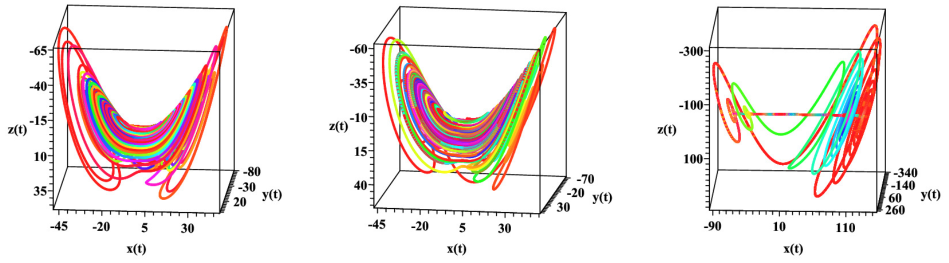
$$A_0 = \begin{pmatrix} 0.00017 \\ -0.05053 \\ 2.22924 \end{pmatrix}, A = \begin{pmatrix} -0.01871 & 9.98680 & 0.02484 \\ -0.99742 & -0.36465 & 3.00322 \\ 0.99120 & 0.01164 & -0.52040 \end{pmatrix},$$

$d_{11} = -0.1000, d_{12} = -0.00573, d_{13} = -0.00573, d_{23} = 43.63002, d_{22} = -0.1182, d_{33} = 10.81835$. The graph of this system is shown in Fig. 7.4 (a2).

2.2. Now we will assign $h(v_i) = piecewise(v_i < 0, (-v_i)^{1.7}, v_i^{1.7})$; $i = 1, \ldots, 3$. In this case, Algorithm 2 leads to the following results:

$$A_0 = \begin{pmatrix} 0.05826 \\ 1.49581 \\ 13.51462 \end{pmatrix}, A = \begin{pmatrix} -0.22723 & 9.75617 & 0.24837 \\ -1.93969 & -3.59190 & 2.94801 \\ 0.92691 & 0.18019 & -6.16120 \end{pmatrix},$$

$d_{11} = 0.01850, d_{12} = 1.30116, d_{13} = 1.30705, d_{23} = -0.71111, d_{22} = 0.01463, d_{33} = -0.16564$. The graph of this system is shown in Fig. 7.4 (a3).



(a1) (a2) (a3)

Fig. 7.4. Phase portraits of system (7.7)(a1), system (7.5) at $\deg h(v) = 2$ (a2), and system (7.5) at $\deg h(v) = 1.7$ (a3)

Thus, the quality of the approximation deteriorates if $\deg h(v) < 2$.

7.3. Modernization of Algorithm 2 for use of activation functions (5.2)

Assume for simplicity that $n = 3$. Now we will demonstrate Algorithm 2 for modeling a dynamic process for given time series. We also suppose that the result of such modeling should be the following system of neural ODEs:

$$\begin{cases} \dot{x}_1(t) = c_{10} + c_{11}x_1 + c_{12}x_2 + c_{13}x_3 + h_1(-d_{11}x_1 + d_{12}x_2 + d_{13}x_3), \\ \dot{x}_2(t) = c_{20} + c_{21}x_1 + c_{22}x_2 + c_{23}x_3 + h_2(-d_{12}x_1 - d_{22}x_2 + d_{23}x_3), \\ \dot{x}_3(t) = c_{30} + c_{31}x_1 + c_{32}x_2 + c_{33}x_3 + h_3(-d_{13}x_1 - d_{23}x_2 - d_{33}x_3). \end{cases} \tag{7.8}$$

(System (7.8) is a special case of system (6.3), in which we set $f_i(u) = u$, $b_i = 0; i = 1, \dots, 3$.)

Now we will assume that in system (7.8) : $h_i(u) = w_i(u, \alpha_i, \beta_i, c_i), i = 1, \dots, 3$ (see (5.2)). The main goal of the modernization of Algorithm 2 is to replace the Jacobi matrix of system (6.3) with the Jacobi matrix of system (7.8).

In system (7.8) the space of parameters (weights) is defined by the vector

$$Y := (c_{10}, \dots, c_{13}, c_{20}, \dots, c_{23}, c_{30}, \dots, c_{33}, d_{11}, d_{12}, d_{13}, d_{22}, d_{23}, d_{33})^T \in \mathbb{R}^{18}.$$

Therefore, the Jacobi matrix in this case is

$$W = \left(\begin{array}{ccc|ccc|cc|c} G & 0 & 0 & -h_1^{'} \cdot \mathbf{g}_1 & h_1^{'} \cdot \mathbf{g}_2 & h_1^{'} \cdot \mathbf{g}_3 & 0 & 0 & 0 \\ 0 & G & 0 & 0 & -h_2^{'} \cdot \mathbf{g}_1 & 0 & -h_2^{'} \cdot \mathbf{g}_2 & h_2^{'} \cdot \mathbf{g}_3 & 0 \\ 0 & 0 & G & 0 & 0 & -h_3^{'} \cdot \mathbf{g}_1 & 0 & -h_3^{'} \cdot \mathbf{g}_2 & -h_3^{'} \cdot \mathbf{g}_3 \end{array} \right).$$

Here the derivative $h_i^{'} := h_{iu}^{'}(u) = h_{iu}^{'}(r_1x_1 + r_2x_2 + r_3x_3)$ is calculated by formula (5.3). (For example, if $h(u) = (x_1 + x_2 + x_3)^2$, then $h^{'}(u) = 2(x_1 + x_2 + x_3)$).

Quite often in system (7.8) (or (6.3)) it is assumed that $d_{11} = d_{22} = d_{33}$. In this case, the space of parameters is defined by the vector

$$Y := (c_{10}, \dots, c_{13}, c_{20}, \dots, c_{23}, c_{30}, \dots, c_{33}, d_{11}, d_{12}, d_{13}, d_{23})^T \in \mathbb{R}^{16}$$

and the Jacobi matrix takes the form

$$W = \begin{pmatrix} G & 0 & 0 & -h_1^{'} \cdot \mathbf{g}_1 & h_1^{'} \cdot \mathbf{g}_2 & h_1^{'} \cdot \mathbf{g}_3 & 0 \\ 0 & G & 0 & -h_2^{'} \cdot \mathbf{g}_2 & -h_2^{'} \cdot \mathbf{g}_1 & 0 & h_2^{'} \cdot \mathbf{g}_3 \\ 0 & 0 & G & -h_3^{'} \cdot \mathbf{g}_3 & 0 & -h_3^{'} \cdot \mathbf{g}_1 & -h_3^{'} \cdot \mathbf{g}_2 \end{pmatrix} \in \mathbb{R}^{3m \times 16}.$$

Further, we form the vectors of the right parts of system (7.8): $R_{1j} = c_{10} + c_{11}g_{1j} + c_{12}g_{2j} + c_{13}g_{3j} + h_1(-d_{11}g_{1j} + d_{12}g_{2j} + d_{13}g_{3j})$, $R_{2j} = c_{20} + c_{21}g_{1j} + c_{22}g_{2j} + c_{23}g_{3j} + h_2(-d_{12}g_{1j} - d_{11}g_{2j} + d_{23}g_{3j})$, $R_{3j} = c_{30} + c_{31}g_{1j} + c_{32}g_{2j} + c_{33}g_{3j} + h_3(-d_{13}g_{1j} - d_{23}g_{2j} - d_{11}g_{3j})$; $j = 0, \dots, m-1$.

Finally we compose vectors

$$R_i = \begin{pmatrix} R_{i0} \\ R_{i1} \\ \vdots \\ R_{i,m-1} \end{pmatrix} \in \mathbb{R}^m; i = 1, \dots, 3.$$

System (7.8) is written for arbitrary activation functions $h_1(u), \dots, h_3(u)$. If these functions are given by formula (5.2), then system (7.8) takes the following form:

$$\left\{\begin{aligned}
&\dot{x}_1(t) = c_{10} + c_{11}x_1 + c_{12}x_2 + c_{13}x_3\\
&\quad +piecewise\Big[-d_{11}x_1 + d_{12}x_2 + d_{13}x_3 < -c_1^{\frac{1}{\beta_1-1}}, -\tfrac{\beta_1-1}{\beta_1}c_1^{\frac{\beta_1}{\beta_1-1}}\\
&\quad -\frac{(d_{11}x_1 - d_{12}x_2 - d_{13}x_3)^{\beta_1}}{\beta_1}, -d_{11}x_1 + d_{12}x_2 + d_{13}x_3 \le c_1^{\frac{1}{\alpha_1-1}},\\
&\quad c_1\cdot(-d_{11}x_1 + d_{12}x_2 + d_{13}x_3), \tfrac{\alpha_1-1}{\alpha_1}c_1^{\frac{\alpha_1}{\alpha_1-1}} + \frac{(-d_{11}x_1 + d_{12}x_2 + d_{13}x_3)^{\alpha_1}}{\alpha_1}\Big],\\
&\dot{x}_2(t) = c_{20} + c_{21}x_1 + c_{22}x_2 + c_{23}x_3\\
&\quad +piecewise\Big[-d_{12}x_1 - d_{22}x_2 + d_{23}x_3 < -c_2^{\frac{1}{\beta_2-1}}, -\tfrac{\beta_2-1}{\beta_2}c_2^{\frac{\beta_2}{\beta_2-1}}\\
&\quad -\frac{(d_{12}x_1 + d_{22}x_2 + d_{23}x_3)^{\beta_2}}{\beta_2}, -d_{12}x_1 - d_{22}x_2 + d_{23}x_3 \le c_2^{\frac{1}{\alpha_2-1}},\\
&\quad c_2\cdot(-d_{12}x_1 - d_{22}x_2 + d_{23}x_3), \tfrac{\alpha_2-1}{\alpha_2}c_2^{\frac{\alpha_2}{\alpha_2-1}} + \frac{(-d_{12}x_1 - d_{22}x_2 + d_{23}x_3)^{\alpha_2}}{\alpha_2}\Big],\\
&\dot{x}_3(t) = c_{30} + c_{31}x_1 + c_{32}x_2 + c_{33}x_3\\
&\quad +piecewise\Big[-d_{13}x_1 - d_{23}x_2 - d_{33}x_3 < -c_3^{\frac{1}{\beta_3-1}}, -\tfrac{\beta_3-1}{\beta_3}c_3^{\frac{\beta_3}{\beta_3-1}}\\
&\quad -\frac{(d_{13}x_1 + d_{23}x_2 + d_{33}x_3)^{\beta_3}}{\beta_3}, -d_{13}x_1 - d_{23}x_2 - d_{33}x_3 \le c_3^{\frac{1}{\alpha_3-1}},\\
&\quad c_3\cdot(-d_{13}x_1 - d_{23}x_2 - d_{33}x_3), \tfrac{\alpha_3-1}{\alpha_3}c_3^{\frac{\alpha_3}{\alpha_3-1}} + \frac{(-d_{13}x_1 - d_{23}x_2 - d_{33}x_3)^{\alpha_3}}{\alpha_3}\Big].
\end{aligned}\right. \tag{7.9}$$

Now everything is ready to use Algorithm 2 when modeling 3D systems with activation functions (5.2).

8. Conclusion

As can be seen from the above examples, Algorithms 1 and 2 correctly indicate the tendency of the behavior of dynamic processes described by certain time series. Now, based on the obtained models, we can describe them more precisely: for example, if approximate the derivatives by finite differences of the second or third order. Various combinations of odd and even power functions can also be used in the simulation. (Corollary of Theorem 2.2 guarantees the achievement of the required approximation accuracy only for odd activation functions. Nevertheless, the introduction of even activation functions sometimes makes it possible to achieve the required accuracy with a smaller number of terms.) Moreover, adding new power-law nonlinearities to the model equations also improves the quality of the approximation.

The main results of this article are as follows:

1. Due to the concept of odd activation function introduced in [25], a simplified version of the classical approximation Theorem 2.2 (this is Corollary of Theorem

2.2) was obtained. As a result, representation (2.1), where generally speaking $\xi_j \neq 1$, was replaced by representation (3.4) with $\xi_j = 1$; $j = 1, \ldots, m$.

2. Conditions for the existence of periodic solutions in neural systems of ordinary differential equations with power nonlinearities were found.

3. Theorems 3.1 and 4.1 represent a constructive approach to solving the problem of the existence of chaos in dynamical systems with power-law nonlinearities. (This approach is more general than the approaches indicated in [32–34].)

In addition, it was found that neural ODEs with power-law activation functions can generate strange non-chaotic attractors (see Fig. 4.6).

4. Algorithms for approximating time series using antisymmetric neural ODE, the architecture of which allows the possibility of modeling limit cycles or homoclinic orbits, have been developed (Algorithm 2 and its modifications associated with modeling systems (4.9) and (7.5)).

5. Let $\mathbf{g}(A\mathbf{x}+\mathbf{b}) := (g_1(a_{11}x_1 + \cdots + a_{1n}x_n + b_1), ..., g_n(a_{n1}x_1 + \cdots + a_{nn}x_n + b_n))^T$. Algorithm 2 can also be applied to simulate the following systems, which generalize system (6.3):

$$\dot{\mathbf{x}}(t) = \mathbf{f}_1(A_1\mathbf{x} + \mathbf{b}_1) + ... + \mathbf{f}_k(A_k\mathbf{x} + \mathbf{b}_k) + \mathbf{h}(S\mathbf{x} + \mathbf{c}).$$

Here $\mathbf{f}_1(\mathbf{x}), \ldots, \mathbf{f}_k(\mathbf{x}), \mathbf{h}(\mathbf{x})$ are power vector functions, and the components of function $\mathbf{h}(\mathbf{x})$ are odd; $\deg \mathbf{h}(\mathbf{x}) > \deg \mathbf{f}_i(\mathbf{x}), i = 1, \ldots, k$ (see Section 4); $A_1, \ldots, A_k, S \in \mathbb{R}^{n\times n}$, and the matrix S is the same as in (6.2); $\mathbf{b}_1, \ldots, \mathbf{b}_k, \mathbf{c} \in \mathbb{R}^n$. In addition, the first block of the Jacobi matrix W_k in item 2.2 of Algorithm 2 must be replaced with the following k blocks:

$$\left(\begin{array}{cccc|c|cccc} F_{11}\cdot G & 0 & \ldots & 0 & & F_{k1}\cdot G & 0 & \ldots & 0 \\ 0 & F_{12}\cdot G & \ldots & 0 & & 0 & F_{k2}\cdot G & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & F_{1n}\cdot G & & 0 & 0 & \ldots & F_{kn}\cdot G \end{array} \right|,$$

where the matrix G and matrices of derivatives $F_{i1}, \ldots, F_{in}; i = 1, \ldots, k$ are defined in items 1.3 and 2.1 of Algoritm 2.

6. The developed methods for adjusting the coefficients of antisymmetric neural ODEs for solving the approximation problem make it possible to use the obtained coefficients as initial weights for deep learning of recurrent neural networks for which the mentioned neural ODEs were designed.

Indeed, let us replace the reconstructed neural ODE $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$ with its difference analogue $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)\cdot\Delta t \in \mathbb{R}^n$, where $k = 0, 1, \ldots, K < \infty$, and K is the number of layers. Then, by adjusting the value of step $\Delta t > 0$, we can get RNN ($\Delta t \to 0$), which adequately simulates the process under study for some $(\Delta t)^*$ and K^*.

References

1. S. Haykin, *Neural Networks. A Comprehensive Foundation*, Second Edition, Pearson Education, Prentice Hall, 2005.

2. E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, The MIT Press Cambridge, Massachusetts, London, England, 2007.
3. B. Cessac, *A view of neural networks as dynamical systems*, arXiv preprint *arXiv: 0901.2203v2 [nlin.AO]*, (2019), 1–62.
4. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press Cambridge, Massachusetts, London, England, 2017.
5. D. Krotov, J. Hopfield, *Large associative memory problem in neurobiology and machine learning*, arXiv preprint *arXiv:2008.06996v1[q-bio. NC]*, (2020), 1–8.
6. Q. Li, *Dynamical Systems and Machine Learning*, Summer School, Peking University, 2020.
7. C. - L. Yang, C. S. Suh, *A general framework for dynamic complex networks*, Journal of Vibration Testing and System Dynamics, **5** (1), (2021), 87–111.
8. S. Sonoda, N. Murata, *Neural network with unbounded activation functions is universal approximator*, Applied and Computational Harmonic Analysis, **43**, 2017, 233–268.
9. R. T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, *Neural ordinary differential equations*, arXiv preprint *arXiv:1806.07366v5[cs.LG]*, (2019), 1–18.
10. R. T. Q. Chen, D. Duvenaud, *Neural networks with cheap differential operators*, arXiv preprint *arXiv:1912.03579v1[cs.LG]*, (2019), 1–11.
11. J. Kelly, J. Bettencourt, M. Johnson, D. Duvenaud, *Learning differential equations that are easy to solve*, arXiv preprint *arXiv:2007.04504v2[cs.LG]*, (2020), 1–18.
12. J. Jia, A. R. Benson, *Neural jump stochastic differential equations*, arXiv preprint *arXiv:1905.10403v3[cs.LG]*, (2020), 1–14.
13. B. Chang, M. Chen, E. Haber, E.D. Chi, *Antisymmetric RNN: A Dynamical System Viev on Recurrent Neural Networks*, In conference ICLR, May 6 - May 9, New Orleans, Louisiana, USA, (2019), 1–15.
14. J. Lorraine, P. Vicol, D. Duvenaud, *Optimizing millions of hyperparameters by implicit differentiation*, arXiv preprint *arXiv:1911.02590v1[cs.LG]*, (2019), 1–18.
15. S. Massaroli, M. Poli, J. Park, A. Yamashita, H. Asama, *Dissecting neural ODEs*, arXiv preprint *arXiv:2002.08071v3[cs.LG]*, (2020), 1–23.
16. A. F. Queiruga, N. B. Erichson, D. Taylor, M.W. Mahoney, *Continuous-in-depth neural networks*, arXiv preprint *arXiv:2008.02389v1[cs.LG]*, (2020), 1–29.
17. V. Ye. Belozyorov, D. V. Dantsev, S. A. Volkova. *On Equivalence of Linear Control Systems and Its Usage to Verification of the Adequacy of Different Models for A Real Dynamic Process*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **28**(1)(2020), 43–97.
18. N. A. Magnitskii, *Universal theory of dynamical chaos in nonlinear dissipative systems of differential equations*, Commun. Nonlinear Sci. Numer. Simul., **13**, (2008), 416–433.
19. N. A. Magnitskii, *Universality of transition to chaos in all kinds of nonlinear differential equations*, Nonlinearity, Bifurcation and Chaos – Theory and Applications, Chapter 6. Intech, (2012), 133–174.
20. N. A. Magnitskii, *Bifurcation theory of dynamical chaos*, Chaos Theory, Chapter 11. Intech, (2018), 197–215.
21. V. Ye. Belozyorov, *Universal approach to the problem of emergence of chaos in autonomous dynamical systems*, Nonlinear Dynamics, **95** (1), (2019), 579–595.

22. N.-E. Tatar, *Hopfield neural networks with unbounded monotone activation functions*, *Advances in Artificial Neural Systems*, Hindawi Publishing Corporation, **2012**, 2012, 571358-1–5.
23. S. M. Richards, F. Berkenkamp, A. Krause, *The Lyapunov neural network: adaptive stability certification for safe learning of dynamical systems*, *arXiv preprint arXiv:1808.009241v2[cs.SY]*, (2018), 1–11.
24. E. Haber, L. Ruthotto, *Stable Architectures for Deep Neural Networks*, *arXiv preprint arXiv: 1705.03341v1[cs.LG]*, (2019), 1–23.
25. V. Ye. Belozyorov, D. V. Dantsev, *Stability of neural ordinary differential equations with power nonlinearities*, *Journal of optimization, differential equations and their applications (JODEA)*, **28** (2), (2020), 21–46.
26. V. Ye. Belozyorov, Ye. M. Kosariev, M. M. Pulin, V. G. Sychenko, V. G. Zaytsev. *A new mathematical model of dynamic process in direct current traction power supply system*, *Journal of Optimization, Differential Equations and Their Applications (JODEA)*, **27**(1) (2019), 21–55.
27. Z. Wang, J. Liang, Y. Liu, *Mathematical problems for complex networks*, *Mathematical Problems in Engineering*, Hindawi Publishing Corporation, **2012**, (2012), ID 934680-1–5.
28. A. N. Gorban, V. L. Dunin-Barkovsky, A. N. Kirdin, E. M. Mirkes, A.Yu. Novokhodko, D. A. Rossiev, S. A. Terekhov, M. Yu. Senashova, V. G. Tsaregorodtsev, *Neuroinformatics*, Nauka, Novosibirsk, 1998 (in rus).
29. H. Zhang, X. Gao, J. Unterman, T. Arodz, *Approximation capabilities of neural ordinary differential equations*, *arXiv preprint arXiv: 1907.12998v1[cs.LG]*, (2019), 1–11.
30. Q. Li, T. Lin, Z. Shen, *Deep learning via dynamical systems: an approximation perspective*, *arXiv preprint arXiv:1912.10382v1[cs.LG]*, (2019), 1–30.
31. H. K. Khalil, *Nonlinear Systems – 2nd Edition*, (Prentice Hall/New-Jersy), 1996.
32. V. Ye. Belozyorov, *A novel search method of chaotic autonomous quadratic dynamical systems without equilibrium points*, *Nonlinear Dyn.*, **86**, (2016), 835–860.
33. V. Ye. Belozyorov, S. A. Volkova, *Role of logistic and Ricker's maps in appearance of chaos in autonomous quadratic dynamical systems*, *Nonlinear Dyn.*, **83**, (2016), 719–729.
34. V. Ye. Belozyorov, *On novel conditions of chaotic attractors existence in autonomous polynomial dynamical systems*, *Nonlinear Dyn.*, **91**, (2018), 2435–2452.
35. V. Ye. Belozyorov, *On existence of homoclinic orbits for some types of autonomous quadratic systems of differential equations*, *Applied Mathematics and Computation*, **217**(9), (2011), 4582–4595.
36. S. Jafary, J. C. Sprott, S. Mohammad Reza Hashemi Golpayegani, *Elementary quadratic chaotic flows no equilibria*, *Physics Letters A*, **377**, (2013), 699–702.
37. X. Wang, G. Chen, *A gallery Lorenz-like and Chen-like attractors*, *Int. J. Bifurc. Chaos*, **23**(4), (2013), ID 1330011-1–20.
38. D. Dantsev, *A novel type of chaotic attractor for quadratic systems without equilibriums*, *Int. J. Bifurc. Chaos*, **28**(1), (2018), ID 1850001-1–7 .

Received 05.05.2022

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **30**, Issue **1**, June **2022**, pp. 42–70, DOI 10.15421/142202

ISSN (print) 2617–0108
ISSN (on-line) 2663–6824

FICTITIOUS CONTROLS AND APPROXIMATION OF AN OPTIMAL CONTROL PROBLEM FOR PERONA-MALIK EQUATION

Peter Kogut[*], Yaroslav Kohut[†], Rosanna Manzo[‡]

Abstract. We discuss the existence of solutions to an optimal control problem for the Cauchy-Neumann boundary value problem for the evolutionary Perona-Malik equations. The control variable v is taken as a distributed control. The optimal control problem is to minimize the discrepancy between a given distribution $u_d \in L^2(\Omega)$ and the current system state. We deal with such case of non-linearity when we cannot expect to have a solution of the original boundary value problem for each admissible control. Instead of this we make use of a variant of its approximation using the model with fictitious control in coefficients of the principle elliptic operator. We introduce a special family of regularized optimization problems for linear parabolic equations and show that each of these problems is consistent, well-posed, and their solutions allow to attain (in the limit) an optimal solution of the original problem as the parameter of regularization tends to zero.

Key words: Perona-Malik equation, optimal control problem, fictitious control, control in coefficients, approximation approach..

2010 Mathematics Subject Classification: 49Q20, 47J35, 49J45, 93C20.

Communicated by Prof. O. M. Stanzhytskyi

1. Introduction

Recently, in the context of time interpolation of satellite multi-spectral images, the following model has been proposed (see [8])

$$u_t - div\left(f\left(|\nabla u|\right)\nabla u\right) + (\nabla u, \boldsymbol{b}) = v \quad \text{in} \quad Q = (0,T)\times\Omega, \tag{1.1}$$

$$u(0,x) = u_0(x) \quad \text{in} \quad \Omega, \tag{1.2}$$

$$\partial_\nu u(t,x) = 0 \quad \text{on} \quad \Sigma = (0,T)\times\partial\Omega, \tag{1.3}$$

where $\Omega \subset \mathbb{R}^2$ is a Lipschitz domain, $\boldsymbol{b} \in \mathfrak{B}_{ad}$ and $v \in \mathfrak{V}_{ad}$ are the control functions with

$$\mathfrak{B}_{ad} = \left\{\boldsymbol{b} \in L^\infty(Q)^2 \cap BV(Q)^2 \ : \ \|\boldsymbol{b}\|_{L^\infty(Q)^2} \le \kappa\right\}, \tag{1.4}$$

$$\mathfrak{V}_{ad} = \left\{v \in L^2(0,T;L^2(\Omega))\right\}, \tag{1.5}$$

[*] Department of Differential Equations, Oles Honchar Dnipro National University, 72, Gagarin av., Dnipro, 49010, Ukraine; EOS Data Analytics Ukraine, Gagarin av., 103a, Dnipro, Ukraine p.kogut@i.ua, peter.kogut@eosda.com

[†] Department of Mathematical Analysis and Theory of Functions, Oles Honchar Dnipro National University, Gagarin av., 72, Dnipro, 49010, Ukraine y.p.kohut@gmail.com

[‡] Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084, SA, Italy (rmanzo@unisa.it)

∂_ν stands for the outward normal derivative, $f \in C^{1,1}(\mathbb{R}_+)$ is a non-increasing real function such that $f(s) \to 0$ when $s \to +\infty$ and $f(s) \to 1$ when $s \to +0$. In particular,

$$f\left(|\nabla u|\right) = \frac{1}{1 + |\nabla u|^2}. \tag{1.6}$$

In fact, the Cauchy-Neumann problem (1.1)–(1.3) can be viewed as some improvement of the Perona-Malik model [23] that was proposed in order to avoid the blurring in images and to reduce the diffusivity at those locations which have a larger likelihood to be edges. This likelihood is measured by $|\nabla u|^2$.

However, the indicated problem is ill-posed due to the degenerate behavior of the multiplayer $f(|\nabla u|)$, $f(|\nabla u|) \longrightarrow 0$ as the gradient $|\nabla u|$ tends to infinity. So, equation (1.1) acts like a standard convection-diffusion equation inside the regions where the magnitude of the gradient of u is weak, whereas at those points where the magnitude of the gradient is large enough, the diffusion is 'stopped'.

Moreover, it can be shown that the equation (1.1), as an example of the nonlinear equation of the porous medium type, combines forward-backforward diffusion flow with the convection (or drift) of the function u in accordance with the velocity field b. In particular, the operator $div\left(f\left(|\nabla u|\right)\nabla u\right)$ implies the forward diffusion in the regions where the squared gradient magnitude of the function u is less than 1, whereas the backward diffusion appears in the area where absolute values of the gradient are larger than 1.

Thus, the model (1.1) is an ill-posed problem from the mathematical point of view and can produce many unexpected phenomena (see [13]). In particular, we have no results of existence and consistency of the initial-boundary value problem (1.1)–(1.3). To overcome this problem, many authors have been looking for some regularizations of the equation (1.1) which inherit its usefulness in image restoration but have better mathematical behavior (see, for instance, [1,3,7,14,15, 21] and the references therein). In order to guarantee the existence and uniqueness of solution to the initial-boundary value problem (1.1)–(1.3), the authors in [8] proposed to specify the equation (1.1) as follows

$$u_t - div\left(K(t,x)\nabla u\right) + (\nabla u, \boldsymbol{b}) = v \quad \text{in} \;\; Q = (0,T) \times \Omega \tag{1.7}$$

with $K(t,x) = f\left(|\nabla Y^*_\sigma|\right)$, where $\nabla Y^*_\sigma = \nabla G_\sigma * Y^*$ is the spatially regularized gradient of Y^*, G_σ denotes the two-dimensional Gaussian filter kernel,

$$\begin{aligned} G_\sigma(x) &= \frac{1}{2\pi\sigma^2} e^{-\frac{|x|^2}{2\sigma^2}}, \quad x \in \mathbb{R}^2, \\ \left(\nabla G_\sigma * Y^*\right)(x) &:= \int_\Omega \nabla G_\sigma(x-y) Y^*(y)\, dy, \quad \forall\, x \in \Omega, \end{aligned}$$

and $Y^* \in C([0,T]; L^2(\Omega))$ is a special function which describes the simplest model of image evolution over the interval $[0,T]$, and this function is defined as a solution of the following optimization problem

$$\int_\Omega \Big[\Big(\frac{\partial Y}{\partial t}\Big|_{t=(T_0+T_1)/2} - div\left(f\left(|\nabla Y_\sigma|\right)\nabla Y\right)|_{t=(T_0+T_1)/2} \\ + \Big(\nabla Y|_{t=(T_0+T_1)/2}, \boldsymbol{b}\Big) - v\Big)^2\Big]\,dx \\ + \int_\Omega \left[\lambda_1^2|\nabla v|^2 + \lambda_2^2\left(|\nabla \boldsymbol{b}_1|^2 + |\nabla \boldsymbol{b}_2|^2\right)\right]\,dx \to \inf_{\substack{v\in H^1(\Omega)\\ \boldsymbol{b}\in H^1(\Omega;\mathbb{R}^2)}}. \quad (1.8)$$

However, it is well-known that the Perona–Malik model with the spatially regularized gradient has several serious practical and theoretical difficulties. The first one is that the spatial regularization of gradient in the form $f\left(|\nabla G_\sigma * u|\right)$ leads to the loss of accuracy in the case when the signal is noisy, with white noise, for instance [7]. Then the noise introduces very large, in theory unbounded, oscillations of the gradient ∇u. As a result, the conditional smoothing introduced by the model will not help, since all these noise edges will be kept.

The second drawback of the Perona–Malik model with the regularized gradient (see also the model (1.7), (1.2), (1.3)) is the fact that the space-invariant Gaussian smoothing inside the divergent term tends to push the edges in u away from their original locations. We refer to [26] where this issue is studied in details. This effect, known as edge dislocation, can be detrimental especially in the context of the boundary detection problem and its application to the remote sensing and monitoring.

In view of this, our prime interest in this paper is to study the equation (1.1) and the corresponding PDE-constrained optimization problem without the space-invariant Gaussian smoothing inside the divergent term. With that in mind we consider the following optimal control problem

$$(\mathcal{R})\quad \text{Minimize } J(v,u) = \int_{Q_T}\left|D\left(\frac{1}{1+|\nabla u|^2}\right)\right| \\ + \frac{1}{2}\int_\Omega |u(T)-u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla u|^2\,dxdt + \frac{\gamma}{2}\int_0^T\int_\omega |v|^2\,dxdt \quad (1.9)$$

subject to the constraints

$$u_t - div\left(\frac{\nabla u}{1+|\nabla u|^2}\right) = v\chi_\omega \quad \text{in } \ Q_T := (0,T)\times\Omega, \quad (1.10)$$

$$\partial_\nu u = 0 \quad \text{on } \ (0,T)\times\partial\Omega, \quad (1.11)$$

$$u(0,\cdot) = u_0 \quad \text{in } \ \Omega, \quad (1.12)$$

$$v \in \mathfrak{V}_{ad} := L^2(0,T;L^2(\omega)), \quad (1.13)$$

where $T>0$, Ω is a bounded open subset of $\mathbb{R}^N$ with a Lipschitz boundary, $N\ge 2$, ω is an open nonempty subset of Ω, $\chi_\omega = \left\{\begin{array}{l} 1,\ x\in\omega, \\ 0,\ x\in\Omega\setminus\omega \end{array}\right\}$ is the characteristic function of the set ω, ∂_ν stands for the outward normal derivative, $u_0, u_d \in L^2(\Omega)$ are given functions, λ, γ are given positive constants, and $v : \omega \to \mathbb{R}$ is a control.

Let us mention that control problems for the non-smoothed Perona-Malik equation have received very little attention in the literature. Formulating the control problem (1.9)–(1.13) for the nonlinear equation of the porous medium type is mainly motivated by the observation that this statement can be successfully applied to the image processing, in particular, to the reduction of mixture of Gaussian and impulse noise with keeping safe the image contours and texture (see, for instance, [2] and the references therein). On the other hand, the novelty of this problem is that we involve into optimization the nonlinear equation with rather special type (non-convex and non-coercive) of non-linearity. Because of this the situation is even more delicate since (1.10) is not well-posed for the given type of non-linearity.

As was mentioned before, the operator $div\,(f\,(|\nabla u|)\,\nabla u)$ with a function f given by (1.6) provides an example of a non-linear operator in divergence form with a so-called degenerate nonlinearity. Moreover, since the function $\mathbb{R}^N \ni s \mapsto \frac{s}{1+|s|^2} \in \mathbb{R}^N$ is neither monotone nor coercive, we have no existence result for the initial-boundary value problem (IBVP) (1.10)–(1.12) and its uniqueness. With that in mind, we say that (v, u) is a feasible pair to the problem (1.9)–(1.13) if

$$v \in \mathfrak{V}_{ad} := L^2(0,T;L^2(\omega)), \quad u \in L^2(0,T;H^1(\Omega)), \quad J(v,u) < +\infty, \tag{1.14}$$

and the following integral identity

$$\int_0^T \int_\Omega \left(-u\frac{\partial\varphi}{\partial t} + \frac{(\nabla u, \nabla\varphi)}{1+|\nabla u|^2} \right) dxdt = \int_0^T \int_\omega v\varphi\, dxdt + \int_\Omega u_0(x)\varphi(0,x)\, dx \tag{1.15}$$

holds for any function $\varphi \in \Phi$, where

$$\Phi = \left\{ \varphi \in C^1(\overline{Q_T}) \ : \ \varphi(T,\cdot) = 0 \text{ in } \Omega \text{ and } \partial_\nu\varphi = 0 \text{ on } \ (0,T)\times\partial\Omega \right\}.$$

In order to find out in what sense the solution takes the initial value $u(0,\cdot) = u_0$, we give the following result.

Proposition 1.1. Let (v,u) be a feasible pair to the problem (1.9)–(1.13). Then, for any $\eta \in C_0^\infty(\Omega)$, the scalar function $h(t) = \displaystyle\int_\Omega u(t,x)\eta(x)\, dx$ belongs to $W^{1,1}(0,T)$ and $h(0) = \displaystyle\int_\Omega u_0(x)\eta(x)\, dx$.

Proof. We set $\varphi(t,x) = \eta(x)\zeta(t)$ where $\zeta(\cdot)$ is a smooth function on $[0,T]$ and $\zeta(T) = 0$. Then it is clear that $\varphi \in \Phi$ and, therefore, the integral identity (1.15) yields the equality

$$\int_0^T \left[-h(t)\zeta'(t) + \underbrace{\left(\int_\Omega \frac{(\nabla u, \nabla\eta)}{1+|\nabla u|^2}\, dx - \int_\omega \eta v\, dx \right)}_{H(t)} \zeta(t) \right] dt = \underbrace{\left(\int_\Omega u_0\eta\, dx \right)}_{k} \zeta(0). \tag{1.16}$$

Since $h \in L^1(0,T)$ and $H \in L^1(0,T)$, it follows from (1.16) that $h \in W^{1,1}(0,T)$, i.e., the function $h(t)$ is absolutely continuous on $[0,T]$. Moreover, from (1.16) we deduce that $h(0) = k$. □

For further convenience we denote the set of all feasible solutions to the problem (1.9)–(1.13) by Ξ. Because of the degenerate behavior of multiplier $f(|\nabla u|)$, the structure of the set Ξ and its main topological properties are unknown in general.

The main focus in this paper consists in providing an approximation framework which in spite of the technical difficulties leads to an implementable scheme, namely, to the so-called indirect approach proving the existence of optimal solutions and giving the procedure of their efficient approximation. With that in mind, we show that the original optimal control problem (1.9)–(1.13) can be approximated efficiently by a special family of optimal control problems for linear parabolic equations with the fictitious BV-control in the principle part of elliptic operator $div\,(\rho \nabla u)$. In spite of the fact that the concept of fictitious controls is not new in the literature, in this paper we utilize it in a new manner combining it with the pointwise convergence of the gradients of solutions to some parabolic equations.

The paper is organized as follows. In the next section, we give some preliminaries and notions that will be needed in the sequel. Section 3 contains a few technical results concerning the almost everywhere convergence of the gradients of solutions to linear parabolic equations with BV-coefficients in the main part of the elliptic operator. These results were obtained in the spirit of Bocardo and Murat approach (see Theorems 4.1 and 4.3 in [6]). In Section 4 we give a precise statement of the fictitious optimal control problems for linear parabolic equations with the constrained BV-controls in the coefficients. We also discuss in this section the existence issues for the proposed control problems. The announced approximation framework is the subject of Section 5, where we provide an asymptotic analysis of a family of approximated optimal control problems and show that some optimal pairs to the original problem (1.9)–(1.13) can be attained (in an appropriate topology) by optimal solutions to the approximated problems.

2. Preliminaries and Basic Definitions

We begin with some notation. For vectors $\xi \in \mathbb{R}^N$ and $\eta \in \mathbb{R}^N$, $(\xi, \eta) = \xi^t \eta$ denotes the standard vector inner product in $\mathbb{R}^N$, where ^{t} denotes the transpose operator. The norm $|\xi|$ is the Euclidean norm given by $|\xi| = \sqrt{(\xi, \xi)}$.

Let Ω be a given bounded open subset of R^N ($N \geq 2$) with a sufficiently smooth boundary. We suppose that the unit outward normal $\nu = \nu(x)$ is well-defined for $\mathcal{H}^{N-1}$-a.a. $x \in \partial\Omega$, where a.a. it means here with respect to the $(N-1)$-dimensional Hausdorff measure $\mathcal{H}^{N-1}$. For any subset $D \subset \Omega$ we denote by $|D|$ its N-dimensional Lebesgue measure $\mathcal{L}^N(D)$. For a subset $D \subseteq \Omega$ let $\overline{D}$ denote its closure and ∂D its boundary. We define the characteristic function χ_D

of D by

$$\chi_D(x) := \begin{cases} 1, & \text{for } x \in D, \\ 0, & \text{otherwise.} \end{cases}$$

Let X denote a real Banach space with norm $\|\cdot\|_X$, and let X' be its dual. Let $\langle\cdot,\cdot\rangle_{X';X}$ be the duality form on $X' \times X$. By $\rightharpoonup$ and $\overset{*}{\rightharpoonup}$ we denote the weak and weak* convergence in normed spaces, respectively.

For given $1 \le p \le +\infty$, the space $L^p(\Omega;\mathbb{R}^N)$ is defined by

$$L^p(\Omega;\mathbb{R}^N) = \left\{ f : \Omega \to \mathbb{R}^N \ : \ \|f\|_{L^p(\Omega;\mathbb{R}^N)} < +\infty \right\},$$

where $\|f\|_{L^p(\Omega;\mathbb{R}^N)} = \left(\int_\Omega |f(x)|^p\,dx\right)^{1/p}$ for $1 \le p < +\infty$. The inner product of two functions f and g in $L^p(\Omega;\mathbb{R}^N)$ with $p \in [1,\infty)$ is given by

$$(f,g)_{L^p(\Omega;\mathbb{R}^N)} = \int_\Omega (f(x),g(x))\ dx = \int_\Omega \sum_{k=1}^N f_k(x)g_k(x)\,dx.$$

We denote by $C_c^\infty(\mathbb{R}^N)$ a locally convex space of all infinitely differentiable functions with compact support. We recall here some functional spaces that will be used throughout this paper. We define the Banach space $H^1(\Omega)$ as the closure of $C_c^\infty(\mathbb{R}^N)$ with respect to the norm

$$\|y\|_{H^1(\Omega)} = \left(\int_\Omega \left(y^2 + |\nabla y|^2\right)\,dx\right)^{1/2}.$$

We denote by $\left(H^1(\Omega)\right)'$ the dual space of $H^1(\Omega)$. We also set $H^1(\Omega;\partial\Omega) = \left\{u \in H^1(\Omega)\ :\ \frac{\partial u}{\partial \nu} = 0\right\}$.

Let $k > 0$. In what follows, we will often use composition of functions in Sobolev space $H^1(\Omega)$ with the Lipschitz continuous function

$$T_k(s) = \max\left\{-k, \min\left\{s,k\right\}\right\}.$$

We recall the well-know result on Sobolev spaces about composition with regular functions.

Theorem 2.1. *Let $G : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function such that $G(0) = 0$. If u belongs to $H^1(\Omega)$, then $G(u)$ belongs to $H^1(\Omega)$ as well, and*

$$\nabla G(u) = G'(u)\nabla u \quad \textit{almost everywhere in } \Omega.$$

As a result, we have

$$\nabla T_k(u) = \nabla u \chi_D\{|u| \le k\} \quad \text{almost everywhere in } \Omega. \tag{2.1}$$

Weak and Strong Convergence in $L^1(\Omega)$. Throughout the paper we will often use the concepts of the weak and strong convergence in $L^1(\Omega)$. Hereinafter, ε

denotes a small parameter which varies within a strictly decreasing sequence of positive numbers converging to 0. When we write $\varepsilon > 0$, we consider only the elements of this sequence, in the case $\varepsilon \geq 0$ we also consider its limit $\varepsilon = 0$. Let $\{a_\varepsilon\}_{\varepsilon>0}$ be a sequence in $L^1(\Omega)$. We recall that $\{a_\varepsilon\}_{\varepsilon>0}$ is called equi-integrable if for any $\delta > 0$ there is $\tau = \tau(\delta)$ such that $\int_S |a_\varepsilon|\, dx < \delta$ for all a_ε and for every measurable subset $S \subset \Omega$ of Lebesgue measure $|S| < \tau$. A sufficient condition for the sequence $\{a_\varepsilon\}_{\varepsilon>0}$ to be equi-integrable is that there exists a constant $C > 0$ such that

$$\sup_{\varepsilon>0} \int_\Omega |a_\varepsilon|^{1+\theta}\, dx \leq C \tag{2.2}$$

for some $\theta > 0$.

Theorem 2.2 (Dunford–Pettis). *Let $\{a_\varepsilon\}_{\varepsilon>0}$ be a sequence in $L^1(\Omega)$. Then this sequence is relatively compact with respect to the weak convergence in $L^1(\Omega)$ if and only if $\{a_\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^1(\Omega)$, i.e., $\sup_{\varepsilon>0} \|u_\varepsilon\|_{L^1(\Omega)} < +\infty$, and $\{a_\varepsilon\}_{\varepsilon>0}$ is equi-integrable.*

Theorem 2.3 (Lebesgue–Vitali). *If a sequence $\{a_\varepsilon\}_{\varepsilon>0} \subset L^1(\Omega)$ is equi-integrable and there exists a function $a \in L^1(\Omega)$ such that $a_\varepsilon(x) \to a(x)$ almost everywhere in Ω then $a_\varepsilon \to a$ in $L^1(\Omega)$.*

A typical application of VitaliвЂ™s theorem is provided by the next simple lemma.

Lemma 2.1. *Let $\{a_\varepsilon\}_{\varepsilon>0}$ be a sequence in $L^1(\Omega)$ such that $a_\varepsilon(x) \to a(x)$ almost everywhere in Ω, and this sequence is uniformly bounded in $L^p(\Omega)$ for some $p > 1$. Then*

$$a_\varepsilon \to a \quad \text{in } L^r(\Omega) \text{ for all } 1 \leq r < p. \tag{2.3}$$

The next lemma is useful in many applications.

Lemma 2.2. *Let $\{a_\varepsilon\}_{\varepsilon>0}$, $\{b_\varepsilon\}_{\varepsilon>0}$, a, and b be a measurable functions such that*

$$a_\varepsilon(x) \to a(x) \quad \text{a.e. in } \Omega, \quad \sup_{\varepsilon>0} \|a_\varepsilon\|_{L^\infty(\Omega)} < \infty, \tag{2.4}$$

$$b_\varepsilon \rightharpoonup b \quad \text{in } L^1(\Omega). \tag{2.5}$$

Then

$$ab \in L^1(\Omega) \quad \text{and} \quad a_\varepsilon b_\varepsilon \rightharpoonup ab \quad \text{in } L^1(\Omega). \tag{2.6}$$

Functions with Bounded Variation. Let $f : \Omega \to \mathbb{R}$ be a function of $L^1(\Omega)$. Define

$$\int_\Omega |Df| = \sup\Big\{ \int_\Omega f\, div\, \varphi\, dx \ : \\ \varphi = (\varphi_1, \dots, \varphi_N) \in C_0^1(\Omega; \mathbb{R}^N),\ |\varphi(x)| \leq 1 \text{ for } x \in \Omega \Big\},$$

where $div\, \varphi = \sum_{i=1}^{N} \frac{\partial \varphi_i}{\partial x_i}$. According to the Radon-Nikodym theorem, if $\int_\Omega |Df| < +\infty$ then the distribution Df is a measure and there exist a vector-valued function $\nabla f \in [L^1(\Omega)]^N$ and a measure $D_s f$, singular with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N \lfloor \Omega$ restricted to Ω, such that

$$Df = \nabla f \mathcal{L}^N \lfloor \Omega + D_s f.$$

Definition 2.1. A function $f \in L^1(\Omega)$ is said to have a bounded variation in Ω if $\int_\Omega |Df| < +\infty$. By $BV(\Omega)$ we denote the space of all functions in $L^1(\Omega)$ with bounded variation.

Under the norm $\|f\|_{BV(\Omega)} = \|f\|_{L^1(\Omega)} + \int_\Omega |Df|$, $BV(\Omega)$ is a Banach space. The following compactness result for BV-functions is well-known:

Proposition 2.1. The uniformly bounded sets in BV-norm are relatively compact in $L^1(\Omega)$.

Definition 2.2. A sequence $\{f_k\}_{k=1}^{\infty} \subset BV(\Omega)$ weakly-$*$ converges to some $f \in BV(\Omega)$, and we write $f_k \overset{*}{\rightharpoonup} f$ if and only if the two following conditions hold: $f_k \to f$ strongly in $L^1(\Omega)$, and $Df_k \rightharpoonup Df$ weakly-$*$ in $\mathcal{M}(\Omega; \mathbb{R}^N)$, where $\mathcal{M}(\Omega; \mathbb{R}^N)$ stands for the space of all vector-valued Borel measures which is, according to the Riesz theory, the dual of the space $C(\Omega; \mathbb{R}^N)$ of all continuous vector-valued functions φ vanishing at infinity.

In the proposition below we give a compactness result related to this convergence, together with the lower semicontinuity property (see [4]):

Proposition 2.2. Let $\{f_k\}_{k=1}^{\infty}$ be a sequence in $BV(\Omega)$ strongly converging to some f in $L^1(\Omega)$ and satisfying $\sup_{k\in\mathbb{N}} \int_\Omega |Df_k| < +\infty$. Then

(i) $f \in BV(\Omega)$ and $\int_\Omega |Df| \le \liminf_{k\to\infty} \int_\Omega |Df_k|$;

(ii) $f_k \overset{*}{\rightharpoonup} f$ in $BV(\Omega)$.

The following embedding results for BV-function is useful in many applications (see [5, p.378]).

Proposition 2.3. Let Ω be an open bounded subset of $\mathbb{R}^N$ with a Lipschitz boundary. Then the embedding $BV(\Omega) \hookrightarrow L^{\frac{N}{N-1}}(\Omega)$ is continuous and the embeddings $BV(\Omega) \hookrightarrow L^p(\Omega)$ are compact for all p such that $1 \le p < \frac{N}{N-1}$. Moreover, there exists a constant $C_{em} > 0$ which depends only on Ω and p such that for all u in $BV(\Omega)$,

$$\left(\int_\Omega |u|^p \, dx \right)^{1/p} \le C_{em} \|u\|_{BV(\Omega)}, \quad \forall\, p \in \left[1, \frac{N}{N-1}\right].$$

3. Some Auxiliaries

In this section we give a few technical results that can be viewed as some specification of the well-known results of Bocardo and Murat (see Theorems 4.1 and 4.3 in [6]).

Proposition 3.1. Let $\{u_k\}_{k\in\mathbb{N}}$ be a weakly convergent sequence in $L^2(0,T;H^1(\Omega))$, and

$$u_k \rightharpoonup u \quad \text{weakly in } \; L^2(0,T;H^1(\Omega)). \tag{3.1}$$

Assume that

$$\frac{\partial u_k}{\partial t} = h_k \quad \text{in } \; \mathcal{D}'((0,T)\times\Omega) \quad \forall\, k\in\mathbb{N}, \tag{3.2}$$

where $\{h_k\}_{k\in\mathbb{N}}$ is a bounded sequence in $L^2(0,T;H^{-1}(\Omega))$. Then

$$u_k \to u \quad \text{strongly in } \; L^2_{loc}(0,T;L^2_{loc}(\Omega)). \tag{3.3}$$

Proof. For arbitrary test functions $\psi\in C_0^\infty(\Omega)$ and $\eta\in C_0^\infty(0,T)$, we set

$$\phi(t,x)=\eta(t)\psi(x), \quad z_k=\phi u_k, \quad \alpha_k=\phi h_k+\frac{\partial\phi}{\partial t}u_k.$$

Then, in view of the dense embeddings $H^1(\Omega)\subset L^2(\Omega)\subset H^{-1}(\Omega)$, we see that, for any bounded open subset S such that $\operatorname{supp}(\psi)\subset S\subset\Omega$,

$$\begin{aligned} z_k(t,\cdot)\in H_0^1(S) \text{ and } \frac{\partial\phi(t,\cdot)}{\partial t}u_k(t,\cdot)&\in H^{-1}(S) \;\text{ a.e. } \; t\in(0,T),\\ \frac{\partial z_k}{\partial t}&=\alpha_k \quad \text{in } \; \mathcal{D}'((0,T)\times S), \quad \forall\, k\in\mathbb{N},\\ \sup_{k\in\mathbb{N}}\|z_k\|_{L^2(0,T;H_0^1(S))}\le C, \; \sup_{k\in\mathbb{N}}\|\alpha_k\|_{L^2(0,T;H^{-1}(S))}&\le C \text{ with some } \; C>0. \end{aligned} \tag{3.4}$$

Moreover, all these functions have their support included in the same compact subset of $(0,T)\times S$.

Since the embeddings $H_0^1(S)\hookrightarrow L^2(S)$ and $L^2(S)\hookrightarrow H^{-1}(S)$ are compact, the brilliant Aubin's Lemma (see [24, Section 8, Corollary 4]) and conditions (3.4) ensure that the sequence $\{z_k\}_{k\in\mathbb{N}}$ is compact in $L^2(0,T;L^2(S))$. This implies (3.3). □

Proposition 3.2. Let $\varepsilon\in(0,1)$ and $K\in(0,\infty)$ be given values. Assume that the sequences

$$\begin{gathered} \{u_k\}_{k=1}^\infty\subset L^2(0,T;H^1(\Omega)), \quad \{v_k\}_{k=1}^\infty\subset L^2(0,T;L^2(\Omega)),\\ \text{and} \quad \{\rho_k\}_{k=1}^\infty\subset BV(Q_T)\cap L^\infty(Q_T) \end{gathered} \tag{3.5}$$

are bounded and such that

$$u_k \rightharpoonup u \text{ weakly in } L^2(0,T;H^1(\Omega)), \tag{3.6}$$

$$v_k \rightharpoonup v \text{ weakly in } L^2(0,T;L^2(\Omega)), \tag{3.7}$$

$$\rho_k \rightharpoonup \rho \text{ weakly-}* \text{ in } BV(Q_T) \text{ and a.e. in } Q_T, \tag{3.8}$$

$$\rho_k \geq \varepsilon \quad \text{a.e. in } \ Q_T, \quad \forall\, k \in \mathbb{N}, \tag{3.9}$$

$$\frac{\partial u_k}{\partial t} - div\,(\rho_k \nabla u_k) = v_k \quad \text{in } \ \mathcal{D}'(Q_T), \quad \forall\, k \in \mathbb{N}. \tag{3.10}$$

Then

$$\nabla T_K(u_k) \to \nabla T_K(u) \text{ strongly in } L^2_{loc}(0,T;L^2_{loc}(\Omega))^N, \tag{3.11}$$

where $T_K : \mathbb{R} \to \mathbb{R}$ is the truncation at height K.

Proof. Let us denote the duality pairing between

$$L^2(0,T;H^{-1}(\Omega)) \quad \text{ and } \quad L^2(0,T;H^1_0(\Omega))$$

by $< \cdot,\cdot >_{Q_T}$. We also set $S_K(u) = \displaystyle\int_0^u T_K(s)\,ds$. Then, using the trick with approximation by convolution, it is easy to show that:

For any $\phi \in C_0^\infty(0,T;C_0^\infty(\Omega))$ and any $u \in L^2(0,T;H^1(\Omega))$ with $\dfrac{\partial u}{\partial t} \in L^2(0,T;H^{-1}(\Omega))$, we have

$$\left\langle \frac{\partial u}{\partial t}, \phi T_K(u) \right\rangle_{Q_T} = - \iint_{Q_T} \frac{\partial \phi}{\partial t} S_K(u)\, dxdt. \tag{3.12}$$

With an arbitrary compact subset $A \subset Q_T = (0,T)\times\Omega$ we associate a function $\phi_A \in C_0^\infty(0,T;C_0^\infty(\Omega))$ such that $0 \leq \phi_A(t,x) \leq 1$ in Q_T and $\phi_A(t,x) = 1$ on A. Then using in (3.10) the test function

$$z_k = [T_K(u_k) - T_K(u)]\,\phi_A,$$

we obtain

$$\left\langle \frac{\partial u_k}{\partial t}, \phi_A T_K(u_k) \right\rangle_{Q_T} \overset{\text{by (3.12)}}{=} - \iint_{Q_T} \frac{\partial \phi_A}{\partial t} S_K(u_k)\, dxdt$$

and, therefore, (3.10) yields

$$\begin{aligned} - \iint_{Q_T} \frac{\partial \phi_A}{\partial t} S_K(u_k)\, dxdt - \left\langle \frac{\partial u_k}{\partial t}, \phi_A T_K(u) \right\rangle_{Q_T} & \\ + \iint_{Q_T} \phi_A \rho_k \left(\nabla u_k, \nabla T_K(u_k) - \nabla T_K(u)\right)\, dxdt & \\ + \iint_{Q_T} [T_K(u_k) - T_K(u)]\, \rho_k \left(\nabla u_k, \nabla \phi_A\right)\, dxdt & \\ = \int_0^T \langle v_k, [T_K(u_k) - T_K(u)]\,\phi_A \rangle_{H^{-1}(\Omega),H^1_0(\Omega)}\, dt. & \end{aligned} \tag{3.13}$$

As follows from the initial assumptions (3.5)–(3.8), the sequence $\{h_k\}_{k\in\mathbb{N}}$ with

$$h_k = div\,(\rho_k \nabla u_k) + v_k$$

is bounded in $L^2(0,T;H^{-1}(\Omega))$. Then, Proposition 3.1 implies that, up to a subsequence, the following assertion holds

$$\begin{gathered} T_K(u_k) - T_K(u) \rightharpoonup 0 \quad \text{weakly in } L^2(0,T;H^1(\Omega)), \\ T_K(u_k) - T_K(u) \to 0 \quad \text{strongly in } L^2_{loc}(Q_T), \text{ and a.e. in } Q_T. \end{gathered} \tag{3.14}$$

Therefore, the last term in (3.13) tends to zero as $k\to\infty$.

Moreover, using the fact that $\rho_k(x) - \rho(x) \to 0$ a.e. in Q_T and the sequence $\{(\nabla u_k, \nabla\phi_A)\}_{k\in\mathbb{N}}$ is bounded in $L^2(Q_T)$, we deduce that

$$\iint_{Q_T} [T_K(u_k) - T_K(u)]\,\rho_k\,(\nabla u_k, \nabla\phi_A)\,dxdt \to 0 \quad \text{as } k\to\infty.$$

Since

$$\frac{\partial}{\partial t} S_K(u_k) = T_K(u_k)\frac{\partial u_k}{\partial t} \quad \text{in } \mathcal{D}'((0,T)\times\Omega) \quad \forall\, k\in\mathbb{N},$$

it follows from Proposition 3.1 that $S_K(u_k)\to S_K(u)$ strongly in $L^2_{loc}(Q_T)$, which yields

$$\lim_{k\to\infty}\iint_{Q_T}\frac{\partial\phi_A}{\partial t} S_K(u_k)\,dxdt = \iint_{Q_T}\frac{\partial\phi_A}{\partial t}S_K(u)\,dxdt.$$

As for the second term in (3.13), we see that $\phi_A T_K(u)\in L^2(0,T;H^1_0(\Omega))$ and $\frac{\partial u_k}{\partial t}$ is a bounded term in $L^2(0,T;H^{-1}(\Omega))$. Hence,

$$\left\langle \frac{\partial u_k}{\partial t}, \phi_A T_K(u)\right\rangle_{Q_T} \to \left\langle \frac{\partial u}{\partial t}, \phi_A T_K(u)\right\rangle_{Q_T} \overset{\text{by (3.12)}}{=} \iint_{Q_T}\frac{\partial\phi_A}{\partial t}S_K(u)\,dxdt$$

as $k\to\infty$.

Thus, we have shown that

$$\lim_{k\to\infty}\iint_{Q_T}\phi_A\rho_k\,(\nabla u_k, \nabla T_K(u_k) - \nabla T_K(u))\,dxdt = 0. \tag{3.15}$$

Taking this fact into account, we observe that

$$\begin{aligned}
&\iint_{Q_T} \phi_A \rho \big| \nabla T_K(u_k) - \nabla T_K(u) \big|^2 \, dxdt \\
&\qquad = \iint_{Q_T} \phi_A \left(\rho - \rho_k\right) \left(\nabla T_K(u_k), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad\quad + \iint_{Q_T} \phi_A \left(\rho_k \nabla T_K(u_k) - \rho \nabla T_K(u), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad = \iint_{Q_T} \phi_A \left(\rho - \rho_k\right) \left(\nabla T_K(u_k), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad\quad + \iint_{Q_T} \phi_A \left(\rho_k \nabla T_K(u_k), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad\quad - \iint_{Q_T} \phi_A \left(\rho \nabla T_K(u), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad = \iint_{Q_T} \phi_A \left(\rho - \rho_k\right) \left(\nabla T_K(u_k), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad\quad + \iint_{Q_T} \phi_A \left(\rho_k \nabla u_k, \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \\
&\qquad\quad - \iint_{Q_T} \phi_A \left(\rho_k \nabla u, \nabla T_K(u_k) - \nabla T_K(u)\right) \chi_{\Lambda_k} \, dxdt \\
&\qquad\quad - \iint_{Q_T} \phi_A \left(\rho \nabla T_K(u), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt, \qquad (3.16)
\end{aligned}$$

where χ_{Λ_K} stands for the characteristic function of the set

$$\Lambda_k := \left\{(t,x) \in Q_T \ : \ |u_k(t,x)| > K\right\}.$$

In view of (3.8), (3.14), and (3.15), we have:

$$\begin{gathered}
\iint_{Q_T} \phi_A \left(\rho - \rho_k\right) \left(\nabla T_K(u_k), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \overset{\text{by Lemma 2.2}}{\to} 0, \\
\iint_{Q_T} \phi_A \left(\rho_k \nabla u_k, \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \overset{\text{by (3.15)}}{\to} 0, \\
\iint_{Q_T} \phi_A \left(\rho \nabla T_K(u), \nabla T_K(u_k) - \nabla T_K(u)\right) dxdt \overset{\text{by (3.14)}}{\to} 0.
\end{gathered}$$

As a result, it follows from (3.16) that

$$\begin{aligned}
&\lim_{k\to\infty} \iint_{Q_T} \phi_A \rho \big| \nabla T_K(u_k) - \nabla T_K(u) \big|^2 \, dxdt \\
&\qquad\qquad = - \lim_{k\to\infty} \iint_{Q_T} \phi_A \left(\rho_k \nabla u, \nabla T_K(u_k) - \nabla T_K(u)\right) \chi_{\Lambda_k} \, dxdt.
\end{aligned}$$

Utilizing the fact that $\chi_{\Lambda_K}\nabla T_K(u_k)=0$ almost everywhere in Q_T, we see that

$$\lim_{k\to\infty}\iint_{Q_T}\phi_A\rho\left|\nabla T_K(u_k)-\nabla T_K(u)\right|^2dxdt \\ =\lim_{k\to\infty}\iint_{Q_T}\phi_A\left(\rho_k\nabla u,\nabla T_K(u)\right)\chi_{\Lambda_k}\,dxdt.$$

Moreover, in view of the weak convergence (3.6) and the Lebesgue dominated Theorem, we have

$$\phi_A\nabla T_K(u)\chi_{\Lambda_k}\to 0\quad\text{strongly in } L^2(Q_T)^N.$$

Hence,

$$0=\lim_{k\to\infty}\iint_{Q_T}\phi_A\rho\left|\nabla T_K(u_k)-\nabla T_K(u)\right|^2dxdt\geq\varepsilon\left\|T_K(u_k)-\nabla T_K(u)\right\|^2,$$

and we arrive at the announced convergence (3.11). □

In fact, the main result of Proposition 3.2 can be specified as follows.

Theorem 3.1. *Let $\varepsilon\in(0,1)$ be a given value and let*

$$\begin{gathered}\{u_k\}_{k=1}^\infty\subset L^2(0,T;H^1(\Omega)),\quad \{v_k\}_{k=1}^\infty\subset L^2(0,T;L^2(\Omega)),\\ \textit{and}\quad \{\rho_k\}_{k=1}^\infty\subset BV(Q_T)\cap L^\infty(Q_T)\end{gathered}\tag{3.17}$$

be bounded sequences satisfying conditions (3.6)–(3.10)*. Then*

$$\nabla u_k\to\nabla u \textit{ strongly in } L^q(0,T;L^q(\Omega))^N \textit{ for any } q\in[1,2).\tag{3.18}$$

Proof. We fix an arbitrary compact subset $A\subset Q_T=(0,T)\times\Omega$ and associate with it a smooth function $\phi_A\in C_0^\infty(0,T;C_0^\infty(\Omega))$ such that $0\leq\phi_A(t,x)\leq 1$ in Q_T and $\phi_A(t,x)=1$ on A. In accordance with the initial assumptions, the functions $\{v_k\}_{k=1}^\infty$ and v belong to the space $L^2(0,T;H^{-1}(\Omega))$. Hence,

$$\frac{\partial u}{\partial t}\in L^2(0,T;H^{-1}(\Omega))\quad\text{and}\quad\frac{\partial u_k}{\partial t}\in L^2(0,T;H^{-1}(\Omega)),\quad\forall\,k\in\mathbb{N}.$$

Therefore, in order to perform the usual integration by parts in the variational equality (3.10), we can use for this $T_K(u_k-u)\phi_A$ as a test function. Taking into account the representation (3.12) and using the fact that

$$\frac{\partial\left(u_k-u\right)}{\partial t}-div\left(\rho_k\nabla u_k-\rho\nabla u\right)=v_k-v\quad\text{in }\mathcal{D}'(Q_T),\quad\forall\,k\in\mathbb{N},$$

we obtain

$$\begin{aligned}-\iint_{Q_T}\frac{\partial\phi_A}{\partial t}S_K(u_k-u)\,dxdt+\iint_{Q_T}\phi_A\left(\rho_k\nabla u_k-\rho\nabla u,\nabla T_K(u_k-u)\right)dxdt\\ +\iint_{Q_T}\left[T_K(u_k-u)\right]\left(\rho_k\nabla u_k-\rho\nabla u,\nabla\phi_A\right)dxdt\\ =\int_0^T\left\langle v_k-v,\left[T_K(u_k-u)\right]\phi_A\right\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}dt.\end{aligned}\tag{3.19}$$

Due to Proposition 3.1, we have

$$T_K(u_k - u) \rightharpoonup 0 \text{ weakly in } L^2(0,T;H^1(\Omega)), \tag{3.20}$$

$$T_K(u_k - u) \to 0 \ \text{ strongly in } L^2_{loc}(Q_T), \text{ and a.e. in } Q_T, \tag{3.21}$$

$$S_K(u_k - u) \to 0 \text{ strongly in } L^2_{loc}(Q_T). \tag{3.22}$$

Then, in view of (3.7), the first and last terms in (3.19) tend to zero as $k \to \infty$. Moreover, using the fact that $\{\rho_k\}_{k=1}^{\infty} \subset L^\infty(Q_T)$, $\rho_k(x) - \rho(x) \to 0$ a.e. in Q_T, and the sequence $\{(\nabla u_k - \nabla u, \nabla \phi_A)\}_{k\in\mathbb{N}}$ is bounded in $L^2(Q_T)$, by the Lebesgue dominated theorem we deduce that

$$\begin{aligned}\iint_{Q_T} & \left[T_K(u_k - u)\right] (\rho_k \nabla u_k - \rho \nabla u, \nabla \phi_A)\, dxdt \\ &= \iint_{Q_T} \left[T_K(u_k - u)\right] \rho_k \left(\nabla u_k - \nabla u, \nabla \phi_A\right) dxdt \\ &\quad + \iint_{Q_T} \left[T_K(u_k - u)\right] (\rho_k - \rho) \left(\nabla u, \nabla \phi_A\right) dxdt \to 0 \quad \text{as } k \to \infty. \end{aligned} \tag{3.23}$$

Thus, passing to the limit in (3.19) when k tends to infinity, we obtain

$$\begin{aligned}\lim_{k\to\infty} \iint_{Q_T} & \phi_A \left(\rho_k \nabla u_k - \rho \nabla u, \nabla T_K(u_k - u)\right) dxdt \\ &= \lim_{k\to\infty} \iint_{Q_T} \phi_A \rho \left(\nabla u_k - \nabla u, \nabla T_K(u_k - u)\right) dxdt \\ &\quad + \lim_{k\to\infty} \iint_{Q_T} \phi_A (\rho_k - \rho) \left(\nabla u_k, \nabla T_K(u_k - u)\right) dxdt \\ &= \lim_{k\to\infty} \iint_{Q_T} \phi_A \rho \left(\nabla (u_k - u), \nabla T_K(u_k - u)\right) dxdt = 0, \end{aligned} \tag{3.24}$$

where

$$\lim_{k\to\infty} \iint_{Q_T} \phi_A (\rho_k - \rho) \left(\nabla u_k, \nabla T_K(u_k - u)\right) dxdt = 0$$

by Lemma 2.2. Setting

$$E_k := \phi_A \rho |\nabla (u_k - u)|^2 \quad \text{ in } Q_T$$

and splitting the set A onto

$$\begin{aligned} B_k^K &= \left\{(t,x) \in A \ : |u_k(t,x) - u(t,x)| \le K\right\}, \\ G_k^K &= \left\{(t,x) \in A \ : |u_k(t,x) - u(t,x)| > K\right\}, \end{aligned}$$

we see that

$$\begin{aligned}\iint_A E_k^\theta \, dxdt &= \iint_{B_k^K} E_k^\theta \, dxdt + \iint_{G_k^K} E_k^\theta \, dxdt \\ &\le \left(\iint_{B_k^K} E_k \, dxdt\right)^\theta |B_k^K|^{1-\theta} + \left(\iint_{G_k^K} E_k \, dxdt\right)^\theta |G_k^K|^{1-\theta}\end{aligned}$$

by Hólder inequality with some $\theta \in (0,1)$. Since, for K fixed, we have $|G_k^K| \to 0$ as $k \to \infty$, and since the sequence $\{\rho\nabla(u_k - u)\}_{k=1}^{\infty}$ is bounded in $L^2(0,T;L^2(\Omega)^N)$, it follows that $\sup_{k\in\mathbb{N}} \|E_k\|_{L^1(Q_T)} < \infty$, and, therefore,

$$\lim_{k\to\infty} \left(\iint_{G_k^K} E_k\, dxdt \right)^{\theta} |G_k^K|^{1-\theta} = 0.$$

Hence,

$$\begin{aligned} 0 &\le \lim_{k\to\infty} \iint_A E_k^\theta\, dxdt \\ &\le \lim_{k\to\infty} \left[\left(\iint_{B_k^K} E_k\, dxdt \right)^{\theta} |B_k^K|^{1-\theta} \right] \\ &= \left(\lim_{k\to\infty} \iint_{Q_T} \phi_A \rho \left(\nabla(u_k - u), \nabla T_K(u_k - u) \right) dxdt \right)^{\theta} \\ &\quad \times \lim_{k\to\infty} |B_k^K|^{1-\theta} \overset{\text{by } (3.24)}{=} 0. \end{aligned} \tag{3.25}$$

As a result, we deduce from (3.25) that $E_k^\theta \to 0$ strongly in $L^1(A)$. So, using a sequence of compact sets $A \subset Q_T$, there exists a subsequence of $\{E_k\}_{k\in\mathbb{N}}$ such that

$$E_{k_n}(t,x) \to 0 \quad \text{for almost each } (t,x) \in Q_T.$$

Then the estimate (3.9) implies that

$$\nabla u_{k_n}(t,x) \to \nabla u(t,x) \quad \text{for almost each } (t,x) \in Q_T \text{ as } n \to \infty.$$

To conclude the proof, it remains to notice that since the sequence $\{\nabla u_k\}_{k=1}^{\infty}$ is bounded in the space $L^2(0,T;L^2(\Omega)^N)$, it follows from Vitaly's theorem (see Lemma 2.1) that

$$\nabla u_k \to \nabla u \quad \text{strongly in } L^q(Q_T).$$

□

4. Regularization of the Original Optimal Control Problem

We introduce the following family of approximating control problems

$$\begin{aligned} (\mathcal{R}_\varepsilon) \quad \text{Minimize } J_\varepsilon(\rho,v,u) &= \frac{1}{2}\int_\Omega |u(T) - u_d|^2\, dx + \frac{\lambda}{2}\int_0^T \int_\Omega |\nabla u|^2\, dxdt \\ &+ \frac{\gamma}{2}\int_0^T \int_\omega |v|^2\, dxdt + \int_{Q_T} |D\rho| + \frac{1}{\varepsilon}\int_0^T \int_\Omega |\rho - \frac{1}{1+|\nabla u|^2}|^2\, dxdt \end{aligned} \tag{4.1}$$

subject to the constraints

$$u_t - div\,(\rho\nabla u) = v\chi_\omega \quad \text{in} \;\; Q_T := (0,T)\times\Omega, \tag{4.2}$$

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on} \;\; (0,T)\times\partial\Omega, \tag{4.3}$$

$$u(0,\cdot) = u_0 \quad \text{in} \;\; \Omega, \tag{4.4}$$

$$v \in \mathfrak{V}_{ad} := L^2(0,T;L^2(\omega)), \tag{4.5}$$

$$\rho \in \mathfrak{R}_{ad} := \{h \in BV(Q_T)\cap L^\infty(Q_T) \;:\; 0 \le h(t,x) \le 1 \text{ a.e. in } \;Q_T\}. \tag{4.6}$$

We say that a tuple (ρ, v, u) is a feasible solution to the problem (4.1)–(4.6) if

$$\rho \in \mathfrak{R}_{ad}, \quad v \in \mathfrak{V}_{ad}, \quad u \in L^2(0,T;H^1(\Omega)), \tag{4.7}$$

$$\rho(t,x) \ge \max\left\{\frac{\varepsilon^2}{1+\varepsilon^2}, \frac{1}{1+|\nabla u(t,x)|^2}\right\} \text{ a.e. in } \;Q_T, \tag{4.8}$$

and this triplet satisfies the following integral identity

$$\int_0^T\int_\Omega (-\varphi_t u + \rho\,(\nabla u,\nabla\varphi))\,dxdt = \int_0^T\int_\omega v\varphi\,dxdt + \int_\Omega u_0(x)\varphi(0,x)\,dx \tag{4.9}$$

for each $\varphi \in \Psi$, where

$$\Psi = \left\{\varphi \in C^1(\overline{Q_T}) \;:\; \varphi(T,\cdot) = 0 \text{ in } \Omega \text{ and } \partial_\nu\varphi = 0 \text{ on } \;(0,T)\times\partial\Omega\right\}.$$

The set of all feasible solution is denoted by Ξ_ε.

Remark 4.1. Let us show that $\Xi_\varepsilon \ne \emptyset$ for each $\varepsilon > 0$. Indeed, taking $z = e^{-\alpha t}u$, we obtain the following IBVP for z:

$$z_t + \alpha z - div\,\widehat{A} = e^{-\alpha t}v\chi_\omega, \quad z\Big|_{i=0} = u_0, \tag{4.10}$$

where the vector function $\widehat{A} = \rho e^{\alpha t}\nabla z$ possesses the following monotonicity, coercivity, and boundedness conditions

$$\left(\widehat{A}(t,x,\xi) - \widehat{A}(t,x,\eta), \xi - \eta\right) \ge 0,$$

$$\left(\widehat{A}(t,x,\xi),\xi\right) \ge \frac{\varepsilon^2}{1+\varepsilon^2}|\xi|^2, \left(\widehat{A}(t,x,\xi),\xi\right) \le e^{\alpha T}|\xi|^2,$$

and the operator $Bz = \alpha z - div\,\widehat{A}$ is coercive in the space $L^2(0,T;H^1(\Omega))$, i.e.

$$\begin{aligned}\langle Bz, z\rangle_{L^2\left(0,T;(H^1(\Omega))^*\right);L^2(0,T;H^1(\Omega))} &\ge \alpha\|z\|^2_{L^2(Q_T)} + \frac{\varepsilon^2}{1+\varepsilon^2}\|\nabla z\|^2_{L^2(Q_T;\mathbb{R}^N)}\\ &\ge c_0\|z\|^2_{L^2(0,T;H^1(\Omega))}.\end{aligned}$$

Hence, the problem (4.10) has a unique solution for each $v \in \mathfrak{V}_{ad}$ [20]. As for the original IBVP, the same result follows by multiplying of z by $e^{\alpha t}$. Moreover, in this case the integral identity (4.9) holds for any function $\varphi \in \Psi$ and the energy equality

$$\int_\Omega u^2(t,x)\,dx + 2\int_0^t\int_\Omega \rho|\nabla u|^2\,dxdt = 2\int_0^t\int_\omega vu\,dxdt + \int_\Omega u_0^2\,dx, \quad 0\le t\le T, \quad (4.11)$$

is valid.

Our next step deals with the study of topological properties of the set of feasible solutions Ξ_ε to the problem (4.1)–(4.6).

Definition 4.1. A sequence $\{(\rho_k, v_k, u_k)\in\Xi_\varepsilon\}_{k\in\mathbb{N}}$ is called bounded if

$$\sup_{k\in\mathbb{N}}\left[\|\rho_k\|_{BV(Q_T)} + \|v_k\|_{L^2(0,T;L^2(\omega))} + \|u_k\|_{L^2(0,T;H^1(\Omega))}\right] < +\infty.$$

Definition 4.2. We say that a bounded sequence $\{(\rho_k, v_k, u_k)\in\Xi_\varepsilon\}_{k\in\mathbb{N}}$ of feasible solutions τ-converges to a triplet

$$(\rho, v, u)\in BV(Q_T)\times L^2(0,T;L^2(\omega))\times L^2(0,T;H^1(\Omega))$$

if conditions

$$u_k \rightharpoonup u \text{ weakly in } L^2(0,T;H^1(\Omega)), \quad (4.12)$$

$$v_k \rightharpoonup v \text{ weakly in } L^2(0,T;L^2(\omega)), \quad (4.13)$$

$$\rho_k \rightharpoonup \rho \text{ weakly-}* \text{ in } BV(Q_T) \text{ and a.e. in } Q_T \quad (4.14)$$

hold true.

Remark 4.2. As follows from Theorem 3.1, if $\{(\rho_k, v_k, u_k)\in\Xi_\varepsilon\}_{k\in\mathbb{N}}$ is a τ-convergent sequence of feasible solutions and $(\rho_k, v_k, u_k)\xrightarrow{\tau}(\rho, v, u)$, then $\nabla u_k\to\nabla u$ strongly in $L^q(0,T;L^q(\Omega))^N$ for any $q\in[1,2)$ and, passing to a subsequence if necessary, we can assert that $\nabla u_k(t,x)\to\nabla u(t,x)$ a.e. in $Q_T=(0,T)\times\Omega$.

Remark 4.3. As immediately follows from (4.9), if (ρ, v, u) is a feasible solution to the problem (4.1)–(4.6), then the equality

$$\frac{\partial u_k}{\partial t} - div\,(\rho_k\nabla u_k) = \chi_\omega v_k \quad \text{in } \mathcal{D}'(Q_T)$$

holds in the sense of distributions for each $k\in\mathbb{N}$. Moreover, if a sequence $\{(\rho_k, v_k, u_k)\in\Xi_\varepsilon\}_{k\in\mathbb{N}}$ is bounded in the sense of Definition 4.1, then $div\,(\rho_k\nabla u_k)+\chi_\omega v_k\in L^2(0,T;H^{-1}(\Omega))$. Therefore, $u_k\in C([0,T];L^2(\Omega))$ for all $k\in\mathbb{N}$ (see [25, Proposition III.1.2]) and due to J.L. Lions [22, Chapitre 1, Theorem 5.1] (we refer also to [24] for some generalizations), the Banach space

$$W=\left\{\varphi\ :\varphi\in L^2(0,T;H^1(\Omega)), \frac{\partial\varphi}{\partial t}\in L^2(0,T;H^{-1}(\Omega))\right\}$$

with the norm of the graph

$$\|\varphi\|_W = \|\varphi\|_{L^2(0,T;H^1(\Omega))} + \left\|\frac{\partial\varphi}{\partial t}\right\|_{L^2(0,T;H^{-1}(\Omega))},$$

is compactly embedded into $L^2(0,T;L^2(\Omega))$.

Thus, the first term in the objective functional (4.1) is well defined onto the set of feasible solutions. So, if $\{u_k\}_{k\in\mathbb{N}}$ is a bounded sequence in W and $u_k \rightharpoonup u$ weakly in $L^2(0,T;H^1(\Omega))$, then $u_k \to u$ strongly in $L^2(0,T;L^2(\Omega))$ and, as a consequence, $u_k(T,\cdot) \to u(T,\cdot)$ strongly in $L^2(\Omega)$.

Before proceeding further, we establish the following important property.

Proposition 4.1. For every $\varepsilon \in (0,1)$ the set Ξ_ε is sequentially closed with respect to the τ-convergence.

Proof. Let $\{(\rho_k, v_k, u_k)\}_{k\in\mathbb{N}} \subset \Xi_\varepsilon$ be a τ-convergent sequence of feasible solutions to the optimal control problem (4.1)–(4.6). Let (ρ, v, u) be its τ-limit. Our aim is to show that $(\rho, v, u) \in \Xi_\varepsilon$.

Since the inclusions $\chi_\omega v \in \mathfrak{V}_{ad} := L^2(0,T;L^2(\Omega))$ and $u \in L^2(0,T;H^1(\Omega))$ are obvious, let us show that the condition (3.9) is valid for some $\varepsilon > 0$. Indeed, in view of Remark 4.2, we can suppose that, up to a subsequence,

$$u_k(t,x) \to u(t,x) \quad \text{and} \quad \frac{1}{1+|\nabla u_k(t,x)|^2} \to \frac{1}{1+|\nabla u(t,x)|^2} \text{ a.e. in } Q_T.$$

Hence, in view of the definition of τ-convergence, the limit passage in the relation

$$\rho_k(t,x) \geq \max\left\{\frac{\varepsilon^2}{1+\varepsilon^2}, \frac{1}{1+|\nabla u_k(t,x)|^2}\right\} \text{ a.e. in } Q_T$$

immediately leads us to the inequality (3.9) with $\widehat{\varepsilon} = \frac{\varepsilon^2}{1+\varepsilon^2}$. As for the inclusion $\rho \in \mathfrak{R}_{ad}$, it is a direct consequence of the weak-$*$ compactness of bounded set $\mathfrak{R}_{ad}$ in $BV(Q_T)$.

It remains to show that the limit triplet (ρ, v, u) is related by the integral identity (4.9). To do so, it is enough to fix an arbitrary test function $\varphi \in \Psi$ and pass to the limit in relation

$$\int_0^T \int_\Omega \left(-\varphi_t u_k + \rho_k \left(\nabla u_k, \nabla\varphi\right)\right) dxdt \\ = \int_0^T \int_\omega v_k \varphi \, dxdt + \int_\Omega u_0(x)\varphi(0,x)\, dx. \quad (4.15)$$

Since $\rho_k \nabla u_k \to \rho\nabla u$ strongly in $L^q(Q_T)$ for $q \in [1,2)$ by Lemma 2.1, it follows that the limit passage in (4.15) leads to the integral identity (4.9). Thus, (ρ, v, u) is a feasible solution to optimal control problem (4.1)–(4.6). □

We are now in a position to state the existence of optimal solutions to the problem (4.1)–(4.6).

Theorem 4.1. *Let $u_d \in L^\infty(\Omega)$ be a given function, and let λ and γ be given constants. Then, for each $\varepsilon \in (0,1)$, the optimal control problem (4.1)–(4.6) admits at least one solution $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon$.*

Proof. Let $\varepsilon \in (0,1)$ be a fixed value. Then, as it was indicated in Remark 4.1, the optimal control problem (4.1)–(4.6) is consistent, that is, $\Xi_\varepsilon \neq \emptyset$.

Let $\{(\rho_k, v_k, u_k) \in \Xi_\varepsilon\}_{k\in\mathbb{N}}$ be a minimizing sequence to the problem (4.1)–(4.6). Then the relation

$$\inf_{(\rho,v,u)\in\Xi_\varepsilon} J_\varepsilon(\rho,v,u) = \lim_{k\to\infty} \Big[\frac{1}{2}\int_\Omega |u_k(T)-u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla u_k|^2\,dxdt$$
$$+ \frac{\gamma}{2}\int_0^T\int_\omega |v_k|^2\,dxdt \quad + \int_{Q_T}|D\rho_k| + \frac{1}{\varepsilon}\int_0^T\int_\Omega |\rho_k - \frac{1}{1+|\nabla u_k|^2}|^2\,dxdt\Big] < +\infty$$

and definition of the set $\mathfrak{R}_{ad}$ imply existence of a constant $C>0$ such that

$$\begin{aligned} \sup_{k\in\mathbb{N}} \|\nabla u_k\|_{L^2(0,T;L^2(\Omega)^N)} &\le C,\\ \sup_{k\in\mathbb{N}} \|v_k\|_{L^2(0,T;L^2(\omega))} &\le C,\\ \text{and} \sup_{k\in\mathbb{N}} \|\rho_k\|_{BV(Q_T)} &\le C. \end{aligned} \tag{4.16}$$

Then, from the energy equality (4.11), we deduce that

$$\begin{aligned}\int_0^T\int_\Omega u_k^2(t,x)\,dxdt &\le 2T\int_0^T\int_\omega v_k u_k\,dxdt + T\int_\Omega u_0^2\,dx\\ &\le 2T^2\int_0^T\int_\omega v_k^2\,dxdt + \frac{1}{2}\int_0^T\int_\Omega u_k^2\,dxdt + T\int_\Omega u_0^2\,dx.\end{aligned}$$

Hence,

$$\sup_{k\in\mathbb{N}} \|u_k\|_{L^2(0,T;L^2(\Omega))} \le 4T^2C^2 + 2T\|u_0\|^2_{L^2(\Omega)}.$$

Utilizing this fact together with (4.16), we see that $\{(\rho_k, v_k, u_k) \in \Xi_\varepsilon\}_{k\in\mathbb{N}}$ is a bounded sequence in the sense of Definition 4.1. As a result, there exist functions $\rho^0_\varepsilon \in BV(Q_T)$, $v^0_\varepsilon \in L^2(0,T;L^2(\omega))$, and $u^0_\varepsilon \in L^2(0,T;H^1(\Omega))$ such that, up to a subsequence, $(\rho_k, v_k, u_k) \xrightarrow{\tau} (\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon)$ as $k\to\infty$. Since the set Ξ_ε is sequentially closed with respect to the τ-convergence (see Proposition 4.1), it follows that the τ-limit tuple $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon)$ is a feasible solution to optimal control problem (4.1)–(4.6) (i.e., $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon$). To conclude the proof, we observe that $\nabla u_k(t,x) \to \nabla u^0_\varepsilon(t,x)$ a.e. in Q_T (see Remark 4.2) and, therefore,

$$\rho_k(t,x) - \frac{1}{1+|\nabla u_k(t,x)|^2} \to \rho^0_\varepsilon(t,x) - \frac{1}{1+|\nabla u^0_\varepsilon(t,x)|^2} \text{ a.e. in } Q_T.$$

Since

$$\left\|\rho_k - \frac{1}{1+|\nabla u_k|^2}\right\|_{L^\infty(Q_T)} \le 2 \text{ for all } k \in \mathbb{N},$$

it follows that the sequence $\left\{\rho_k - \dfrac{1}{1+|\nabla u_k|^2}\right\}_{k\in\mathbb{N}}$ is equi-integrable. Hence, Vitaly's theorem implies that

$$\rho_k - \frac{1}{1+|\nabla u_k|^2} \to \rho^0_\varepsilon - \frac{1}{1+|\nabla u^0_\varepsilon|^2} \quad \text{strongly in } L^2(Q_T) \tag{4.17}$$

(see Lemma 2.1). Taking this fact into account and observing that

$$\begin{aligned}
\liminf_{k\to\infty} \int_0^T \int_\Omega |\rho_k - \frac{1}{1+|\nabla u_k|^2}|^2\,dxdt &\overset{\text{by (4.17)}}{=} \int_0^T \int_\Omega |\rho^0_\varepsilon - \frac{1}{1+|\nabla u^0_\varepsilon|^2}|^2\,dxdt,\\
\lim_{k\to\infty} \int_\Omega |u_k(T) - u_d|^2\,dx &\overset{\text{by Remark (4.3)}}{\ge} \int_\Omega |u^0_\varepsilon(T) - u_d|^2\,dx,\\
\lim_{k\to\infty} \int_0^T \int_\Omega |\nabla u_k|^2\,dxdt &\overset{\text{by (4.12)}}{=} \int_0^T \int_\Omega |\nabla u^0_\varepsilon|^2\,dxdt,\\
\liminf_{k\to\infty} \int_0^T \int_\omega |v_k|^2\,dxdt &\overset{\text{by (4.13)}}{\ge} \int_0^T \int_\Omega |v^0_\varepsilon|^2\,dxdt,\\
\liminf_{k\to\infty} \int_{Q_T} |D\rho_k| &\overset{\text{by (4.14)}}{\ge} \int_{Q_T} |D\rho^0_\varepsilon|,
\end{aligned}$$

we see that the cost functional J_ε is sequentially lower τ-semicontinuous. Thus

$$J_\varepsilon(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \le \liminf_{k\to\infty} J_\varepsilon(\rho_k, v_k, u_k) \le \lim_{k\to\infty} J_\varepsilon(\rho_k, v_k, u_k) = \inf_{(\rho,v,u)\in\Xi_\varepsilon} J_\varepsilon(\rho, v, u),$$

and, therefore, $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon)$ is an optimal triplet. □

5. Asymptotic Analysis of the Approximated OCP $(\mathcal{R}_\varepsilon)$

The main goal of this section is to show that the original OCP $(\mathcal{R})$ is solvable and some solutions can be attained (in an appropriate topology) by optimal solutions to the approximated problems $(\mathcal{R}_\varepsilon)$. With that in mind, we make use of the concept of variational convergence of constrained minimization problems (see [9, 17, 18]) and study the asymptotic behavior of a family of OCPs $(\mathcal{R}_\varepsilon)$ as $\varepsilon \to 0$.

Before proceeding further, we adopt the following concept.

Definition 5.1. Let

$$\{(\rho_\varepsilon, v_\varepsilon, u_\varepsilon)\}_{\varepsilon>0} \subset BV(Q_T) \times L^2(0,T;L^2(\omega)) \times L^2(0,T;H^1(\Omega))$$

be an arbitrary sequence. We say that this sequence is bounded if

$$\sup_{\varepsilon>0} \left[\|\rho_\varepsilon\|_{BV(Q_T)} + \|v_\varepsilon\|_{L^2(0,T;L^2(\omega))} + \|u_\varepsilon\|_{L^2(0,T;H^1(\Omega))}\right] < +\infty.$$

Definition 5.2. We say that a bounded sequence

$$\{(\rho_\varepsilon, v_\varepsilon, u_\varepsilon)\}_{\varepsilon>0} \subset BV(Q_T) \times L^2(0,T;L^2(\omega)) \times L^2(0,T;H^1(\Omega))$$

is w-convergent as $\varepsilon \to 0$ and $(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \overset{w}{\to} (\rho, v, u)$ if $(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \overset{\tau}{\to} (\rho, v, u)$ as $\varepsilon \to 0$, i.e.,

$$u_\varepsilon \rightharpoonup u \text{ weakly in } L^2(0,T;H^1(\Omega)), \tag{5.1}$$

$$v_\varepsilon \rightharpoonup v \text{ weakly in } L^2(0,T;L^2(\omega)), \tag{5.2}$$

$$\rho_\varepsilon \rightharpoonup \rho \text{ weakly-}* \text{ in } BV(Q_T) \text{ and a.e. in } Q_T; \tag{5.3}$$

and $\nabla u_\varepsilon \to \nabla u$ strongly in $L^1(0,T;L^1(\Omega)^N)$.

The following technical result will play a significant role in the sequel.

Lemma 5.1. *Let $\{(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ be a τ-convergent sequence of feasible solutions to OCPs* (4.1)–(4.6)*, and let*

$$(\rho, v, u) \in BV(Q_T) \times L^2(0,T;L^2(\omega)) \times L^2(0,T;H^1(\Omega))$$

be its τ-limit. Then $(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \overset{w}{\to} (\rho, v, u)$ as $\varepsilon \to 0$, and (ρ, v, u) is subjected to the constrains

$$\rho \in \mathfrak{R}_{ad}, \quad v \in \mathfrak{V}_{ad}, \quad u \in L^2(0,T;H^1(\Omega)), \tag{5.4}$$

$$\rho(t,x) \geq \frac{1}{1+|\nabla u(t,x)|^2} \ a.e. \ in \ \ Q_T, \tag{5.5}$$

$$\int_0^T \int_\Omega \left(-\varphi_t u + \rho\left(\nabla u, \nabla \varphi\right)\right) dxdt \\ = \int_0^T \int_\omega v\varphi \, dxdt + \int_\Omega u_0(x)\varphi(0,x)\, dx, \quad \forall \varphi \in \Psi. \tag{5.6}$$

Proof. Since $\{(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ is a sequence of feasible solutions, it implies that the equality

$$\int_0^T \int_\Omega \left(-\varphi_t u_\varepsilon + \rho_\varepsilon\left(\nabla u_\varepsilon, \nabla \varphi\right)\right) dxdt \\ = \int_0^T \int_\omega v_\varepsilon\varphi \, dxdt + \int_\Omega u_0(x)\varphi(0,x)\, dx, \quad \forall \varphi \in \Psi \tag{5.7}$$

holds true for all $\varepsilon > 0$. Then the limit passage in (5.7) leads to the relation (5.6). Setting in this relation the test function φ as an element of $C_c^\infty(Q_T) \subset \Psi$, we see that the τ-limit (ρ, v, u) satisfies the equation

$$\frac{\partial u}{\partial t} - div\left(\rho \nabla u\right) = \chi_\omega v$$

in the sense of distributions $\mathcal{D}'(Q_T)$. So, in view of Remark 4.3, we can suppose that, for each $\varepsilon > 0$, we have the equalities

$$\frac{\partial\,(u_\varepsilon - u)}{\partial t} - div\,(\rho_k \nabla u_\varepsilon - \rho\nabla u) = (v_\varepsilon - v)\chi_\omega \quad \text{in } \mathcal{D}'(Q_T). \tag{5.8}$$

Therefore, arguing as in the proof of Theorem 3.1, we use for (5.8) the test function $T_K(u_\varepsilon - u)\phi_A$, where A is a compact subset of Q_T, and the function $\phi_A \in C_0^\infty(0,T; C_0^\infty(\Omega))$ is such that $0 \le \phi_A(t,x) \le 1$ in Q_T and $\phi_A(t,x) = 1$ on A. After integration by parts, we obtain

$$\begin{aligned}\iint_{Q_T} \phi_A \rho_\varepsilon\,(\nabla u_\varepsilon - \nabla u, \nabla T_K(u_\varepsilon - u))\;dxdt &= \iint_{Q_T} \frac{\partial \phi_A}{\partial t} S_K(u_\varepsilon - u)\,dxdt \\ &- \iint_{Q_T} \phi_A(\rho_\varepsilon - \rho)\,(\nabla u, \nabla T_K(u_\varepsilon - u))\;dxdt \\ &- \iint_{Q_T} \phi_A \rho\,(\nabla u_\varepsilon - \nabla u, \nabla T_K(u_\varepsilon - u))\;dxdt \\ &- \iint_{Q_T} \phi_A(\rho_\varepsilon - \rho)\,(\nabla u_\varepsilon, \nabla T_K(u_\varepsilon - u))\;dxdt \\ &+ \int_0^T \langle (v_\varepsilon - v)\chi_\omega, [T_K(u_\varepsilon - u)]\,\phi_A\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}\;dt.\end{aligned} \tag{5.9}$$

Since, by Proposition 3.1,

$$\begin{aligned}&T_K(u_\varepsilon - u) \rightharpoonup 0 \quad \text{weakly in } L^2(0,T;H^1(\Omega)),\ \text{strongly in } L^2_{loc}(Q_T),\ \text{and a.e. in } Q_T,\\ &S_K(u_\varepsilon - u) \to 0 \quad \text{strongly in } L^2_{loc}(Q_T) \text{ as } \varepsilon \to 0,\end{aligned}$$

it follows from (5.1)–(5.3) and the Lebesgue dominated theorem that the right hand side of (5.9) tends to zero as $\varepsilon \to 0$. Hence, passing to the limit in (5.9), we deduce:

$$\lim_{\varepsilon\to 0} \iint_{Q_T} \phi_A \rho_\varepsilon\,(\nabla u_\varepsilon - \nabla u, \nabla T_K(u_\varepsilon - u))\;dxdt = 0. \tag{5.10}$$

Setting

$$E_\varepsilon := \phi_A \rho_\varepsilon |\nabla(u_\varepsilon - u)|^2 \quad \text{in } Q_T$$

and aligning the set A into

$$\begin{aligned}B_\varepsilon &= \{(t,x) \in A\ : |u_\varepsilon(t,x) - u(t,x)| \le K\},\\ G_\varepsilon &= \{(t,x) \in A\ : |u_\varepsilon(t,x) - u(t,x)| > K\},\end{aligned}$$

we see that

$$\iint_A E_\varepsilon^\theta\,dxdt \le \left(\iint_{B_\varepsilon} E_\varepsilon\,dxdt\right)^\theta |B_\varepsilon|^{1-\theta} + \left(\iint_{G_\varepsilon} E_\varepsilon\,dxdt\right)^\theta |G_\varepsilon|^{1-\theta}$$

by Hölder inequality with some $\theta \in (0,1)$. Since, for K fixed, we have $|G_\varepsilon| \to 0$ as $\varepsilon \to 0$, and the sequence $\{\rho_\varepsilon \nabla(u_\varepsilon - u)\}_{\varepsilon>0}$ is bounded in $L^2(0,T;L^2(\Omega)^N)$, it follows that $\sup_{\varepsilon>0} \|E_\varepsilon\|_{L^1(Q_T)} < \infty$, and, therefore,

$$\lim_{\varepsilon\to 0} \left(\iint_{G_\varepsilon} E_\varepsilon \, dxdt \right)^\theta |G_\varepsilon|^{1-\theta} = 0.$$

Hence,

$$\begin{aligned} 0 &\le \lim_{\varepsilon\to 0} \iint_A E_\varepsilon^\theta \, dxdt \\ &\le \lim_{\varepsilon\to 0} \left[\left(\iint_{B_\varepsilon} E_\varepsilon \, dxdt \right)^\theta |B_\varepsilon|^{1-\theta} \right] \\ &= \left(\lim_{\varepsilon\to 0} \iint_{Q_T} \phi_A \rho \left(\nabla(u_\varepsilon - u), \nabla T_K(u_\varepsilon - u) \right) \, dxdt \right)^\theta \lim_{\varepsilon\to 0} |B_\varepsilon|^{1-\theta} \\ &\overset{\text{by } (3.24)}{=} 0. \end{aligned} \tag{5.11}$$

As a result, we deduce from (5.11) that $E_\varepsilon^\theta \to 0$ strongly in $L^1(A)$. So, using a sequence of compact sets $A \subset Q_T$ converging in an appropriate sense to Q_T, there exists a subsequence of $\{E_\varepsilon\}_{\varepsilon>0}$ (still denoted by the same index) such that

$$E_\varepsilon(t,x) \to 0 \quad \text{for almost each } (t,x) \in Q_T \text{ as } \varepsilon \to 0.$$

Thus,

$$\rho_\varepsilon(t,x) \left| \nabla u_\varepsilon(t,x) - \nabla u(t,x) \right|^2 \to 0 \quad \text{for a.e. } (t,x) \in Q_T \text{ as } \varepsilon_n \to 0. \tag{5.12}$$

Utilizing the fact that $(\rho_\varepsilon, v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon$ for each $\varepsilon > 0$ and observing that $\dfrac{\varepsilon^2}{1+\varepsilon^2} \to 0$ as $\varepsilon \to 0$, we see that

$$\rho_\varepsilon(t,x) \ge \max\left\{ \frac{\varepsilon^2}{1+\varepsilon^2}, \frac{1}{1+|\nabla u_\varepsilon(t,x)|^2} \right\} \ge \frac{1}{1+|\nabla u_\varepsilon(t,x)|^2} \text{ a.e. in } Q_T \tag{5.13}$$

for $\varepsilon > 0$ small enough. Hence, from (5.13) and (5.11) we deduce:

$$\begin{aligned} 0 &\le \lim_{\varepsilon\to 0} \iint_{Q_T} \frac{1}{1+|\nabla u_\varepsilon|^2} \left| \nabla u_\varepsilon - \nabla u \right|^2 \, dxdt \\ &\le \lim_{\varepsilon\to 0} \iint_{Q_T} \rho_\varepsilon \left| \nabla u_\varepsilon - \nabla u \right|^2 \, dxdt = 0. \end{aligned} \tag{5.14}$$

Since

$$
\begin{aligned}
&\|\nabla u_\varepsilon - \nabla u\|^2_{L^1(0,T;L^1(\Omega)^N)} = \left(\int_0^T \int_\Omega |\nabla u_\varepsilon - \nabla u|\, dxdt\right)^2 \\
&\le \left(\int_0^T \left(\int_\Omega \frac{1}{1+|\nabla u_\varepsilon(x)|^2} \left|\nabla u_\varepsilon - \nabla u\right|^2\, dx\right)^{1/2} \left(\int_\Omega \left(1+|\nabla u_\varepsilon(x)|^2\right)\, dx\right)^{1/2} dt\right)^2 \\
&\quad \le \int_0^T \int_\Omega \frac{1}{1+|\nabla u_\varepsilon(x)|^2} \left|\nabla u_\varepsilon - \nabla u\right|^2\, dxdt \int_0^T \int_\Omega \left(1+|\nabla u_\varepsilon(x)|^2\right)\, dx\, dt \\
&\qquad \le \left(|Q_T| + \sup_{\varepsilon>0} \|u_\varepsilon\|^2_{L^2(0,T;H^1(\Omega))}\right) \int_\Omega \frac{1}{1+|\nabla u_\varepsilon(x)|^2} \left|\nabla u_\varepsilon - \nabla u\right|^2\, dx,
\end{aligned}
$$

it follows from (5.14) that

$$
\begin{aligned}
&\lim_{\varepsilon\to 0} \|\nabla u_\varepsilon - \nabla u\|^2_{L^1(0,T;L^1(\Omega)^N)} \\
&\qquad\qquad \le C \lim_{\varepsilon\to 0} \iint_{Q_T} \frac{1}{1+|\nabla u_\varepsilon|^2} \left|\nabla u_\varepsilon - \nabla u\right|^2\, dxdt = 0. \qquad (5.15)
\end{aligned}
$$

Thus, we can specify the τ-convergence properties (5.1)–(5.3) as follows: in addition to (5.1) $\nabla u_\varepsilon \to \nabla u$ strongly in $L^1(0,T;L^1(\Omega)^N)$, and there exists a subsequence $\{\varepsilon'\}$ such that

$$
\nabla u_{\varepsilon'}(t,x) \to \nabla u(t,x) \quad \text{a.e. in } Q_T. \qquad (5.16)
$$

To conclude the proof, it remains to show that

$$
\rho(t,x) \ge \frac{1}{1+|\nabla u(t,x)|^2} \text{ a.e. in } \ Q_T. \qquad (5.17)
$$

To do so, it is enough to observe that

$$
\rho_\varepsilon(t,x) \ge \max\left\{\frac{\varepsilon^2}{1+\varepsilon^2}, \frac{1}{1+|\nabla u_\varepsilon(t,x)|^2}\right\} \ge \frac{1}{1+|\nabla u_\varepsilon(t,x)|^2} \text{ a.e. in } \ Q_T \quad (5.18)
$$

for $\varepsilon > 0$ small enough. Using the pointwise convergence (5.16) and (5.3) and passing to the limit in (5.18) as $\varepsilon \to 0$, we arrive to the announced property (5.5). □

Our next step is to discuss the issue related to the existence of solutions to the original optimal control problem (1.9)–(1.13) and their attainability by optimal solutions of the approximated problems $(\mathcal{R}_\varepsilon)$. Before we go on, we assume that the set of feasible solution Ξ to the problem (1.9)–(1.13) is non-empty. In the case when the initial state u_0 is sufficiently smooth and supp $(u_0) \subset \omega$, this assumption can be easily verified. Indeed, let $\varphi \in C^\infty([0,T]; C^\infty_c(\omega))$ be an arbitrary function such that $\varphi(0,x) = u_0(x)$ in Ω. Then it is easy to check that the pair

$$
(v,u) := \left(\left[\varphi_t - div\left(\frac{\nabla\varphi}{1+|\nabla\varphi|^2}\right)\right] \Big\lfloor_{x\in\omega}, \varphi\right)
$$

belongs to the set Ξ. Thus, $\Xi \neq \emptyset$.

We begin with the following result that can be viewed as a direct consequence of Lemma 5.1 and Theorem 4.1.

Proposition 5.1. Let $u_d \in L^\infty(\Omega)$ be a given function, and λ and γ be given constants. Let $\left\{(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ be a bounded sequence of optimal solutions to the approximated problems (4.1)–(4.6) when the small parameter ε varies within a strictly decreasing sequence of positive numbers converging to zero. Then there is a subsequence of $\left\{(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$, still denoted by the suffix ε, and distributions $\rho^0 \in \mathfrak{R}_{ad} \subset BV(Q_T)$, $v^0 \in \mathfrak{V}_{ad}$, and $u^0 \in L^2(0,T;H^1(\Omega))$ such that they satisfy conditions (5.5)–(5.6), and $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \overset{w}{\to} (\rho^0, v^0, u^0)$ as $\varepsilon \to 0$.

The key point in Proposition 5.1 is the assumption that a given sequence of optimal solutions to the approximated problems (4.1)–(4.6) is bounded. Let us show that this assumption can be omitted if only the original optimal control problem is consistent, i.e. $\Xi \neq \emptyset$.

Proposition 5.2. Assume that $\Xi \neq \emptyset$. Let $\left\{(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ be a sequence of optimal solutions to the approximated problems (4.1)–(4.6). Then there exists a constant $C > 0$ independent of $\varepsilon > 0$ such that

$$\sup_{\varepsilon>0} \left[\|\rho^0_\varepsilon\|_{BV(Q_T)} + \|v^0_\varepsilon\|_{L^2(0,T;L^2(\omega))} + \|u^0_\varepsilon\|_{L^2(0,T;H^1(\Omega))}\right] \le C. \tag{5.19}$$

Proof. Let $(\widehat{v}, \widehat{u}) \in \Xi$ be a feasible solution to optimal control problem (1.9)–(1.13). Hence, this pair satisfies conditions (1.14)–(1.15). Setting $\widehat{\rho} := (1+|\nabla \widehat{u}|^2)^{-1}$ in Q_T, we see that

$$0 \le \widehat{\rho}(t,x) \le 1 \text{ a.e. in } Q_T \quad \text{and} \quad \widehat{\rho} \in BV(Q_T) \cap L^\infty(Q_T),$$

and the pair $(\widehat{\rho}, \widehat{u})$ satisfies inequalities (4.8) for $\varepsilon > 0$ small enough. Hence, $\widehat{\rho} \in \mathfrak{R}_{ad}$ and, as a consequence, we deduce: $(\widehat{\rho}, \widehat{v}, \widehat{u}) \in \Xi_\varepsilon$ for $\varepsilon > 0$ small enough. Therefore,

$$\begin{aligned}\inf_{(\rho,v,u)\in\Xi_\varepsilon} J_\varepsilon(\rho,v,u) &= J_\varepsilon(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon) \le J_\varepsilon\left(\widehat{\rho}, \widehat{v}, \widehat{u}\right) \\ &= \frac{1}{2}\int_\Omega |\widehat{u}(T) - u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla \widehat{u}|^2\,dxdt \\ &\quad + \frac{\gamma}{2}\int_0^T\int_\omega |\widehat{v}|^2\,dxdt + \int_{Q_T} |D\widehat{\rho}| = C < +\infty.\end{aligned}$$

From this and definition of the set $\mathfrak{R}_{ad}$, we deduce that

$$\|\nabla u^0_\varepsilon\|^2_{L^2(0,T;L^2(\Omega)^N)} \le \frac{2}{\lambda}C, \quad \|v^0_\varepsilon\|^2_{L^2(0,T;L^2(\Omega))} \le \frac{2}{\gamma}C, \tag{5.20}$$

$$\int_{Q_T} |D\rho^0_\varepsilon| \le C, \quad \|\rho^0_\varepsilon\|_{BV(\Omega)} \le |Q_T| + C, \tag{5.21}$$

$$\int_0^T\int_\Omega \left|\rho^0_\varepsilon - \frac{1}{1+|\nabla u^0_\varepsilon|^2}\right|^2 dxdt \le C\varepsilon \tag{5.22}$$

for all $\varepsilon > 0$ small enough. Then energy equality (4.11) implies that

$$\int_0^T \int_\Omega \left[u_\varepsilon^0\right]^2 \, dxdt \leq 2T \int_0^T \int_\omega v_\varepsilon^0 u_\varepsilon^0 \, dxdt + T \int_\Omega u_0^2 \, dx$$
$$\leq 2T^2 \int_0^T \int_\omega \left[v_\varepsilon^0\right]^2 \, dxdt + \frac{1}{2} \int_0^T \int_\Omega \left[u_\varepsilon^0\right]^2 \, dxdt + T \int_\Omega u_0^2 \, dx.$$

Therefore,

$$\sup_{\varepsilon>0} \|u_\varepsilon^0\|_{L^2(0,T;L^2(\Omega))} \leq 8T^2 \frac{C}{\gamma} + 2T\|u_0\|^2_{L^2(\Omega)}. \tag{5.23}$$

Thus, the sequence $\left\{(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ is bounded in

$$BV(Q_T) \times L^2(0,T;L^2(\omega)) \times L^2(0,T;H^1(\Omega)).$$

□

The next step of our analysis is to show that the pair (v^0, u^0) is optimal to the original OCP $(\mathcal{R})$ provided (ρ^0, v^0, u^0) is a cluster tuple of a given sequence of optimal solutions $\left\{(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$. To do so, we will utilize some hints from the recent papers [10, 16] where the so-called indirect approach to the existence problem of optimal solutions has been proposed.

Theorem 5.1. *Assume that $\Xi \neq \emptyset$. Let $\left\{(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ be a sequence of optimal solutions to the approximated problems* (4.1)–(4.6)*. Let $(\rho^0, v^0, u^0) \in BV(Q_T) \times L^2(0,T;L^2(\omega)) \times L^2(0,T;H^1(\Omega))$ be a w-cluster tuple (in the sense of Definition 5.2) of a given sequence of optimal solutions Then*

$$(v^0, u^0) \in \Xi, \quad \rho^0(t,x) = \frac{1}{1 + |\nabla u^0(t,x)|^2} \ \textit{a.e. in} \ Q_T, \tag{5.24}$$

$$\lim_{\varepsilon\to 0} \inf_{(\rho,v,u)\in\Xi_\varepsilon} J_\varepsilon(\rho, v, u) = \lim_{\varepsilon\to 0} J_\varepsilon(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) = J(v^0, u^0) = \inf_{(v,u)\in\Xi} J(v, u). \tag{5.25}$$

Proof. Arguing as in the proof of Proposition 5.2, it can be shown that there exists a constant $C > 0$ such that estimates (5.20)–(5.23) hold true. Hence, the sequence $\left\{(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ is compact with respect to the τ-convergence. Moreover, in view of Proposition 5.1 and the Lebesgue Dominated Theorem, we can suppose that, up to a subsequence,

$$(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \overset{w}{\to} (\rho^0, v^0, u^0) \tag{5.26}$$

$$\frac{1}{1 + |\nabla u_\varepsilon^0|^2} \to \frac{1}{1 + |\nabla u^0|^2} \quad \text{strongly in } L^2(Q_T) \text{ as } \varepsilon \to 0, \tag{5.27}$$

$$\rho_\varepsilon^0(t,x) - \frac{1}{1 + |\nabla u_\varepsilon^0(t,x)|^2} \to \rho^0(t,x) - \frac{1}{1 + |\nabla u^0(t,x)|^2} \ \text{a.e. in} \ Q_T, \tag{5.28}$$

and $\left(\rho_\varepsilon^0 - \left(1 + |\nabla u_\varepsilon^0|^2\right)^{-1}\right) \in L^\infty(\Omega)$.

Then it follows from Vitaly's theorem (see Lemma 2.1) that

$$\rho_\varepsilon^0 - \left(1+|\nabla u_\varepsilon^0|^2\right)^{-1} \to \rho^0 - \frac{1}{1+|\nabla u^0|^2} \text{ strongly in } L^2(\Omega).$$

However, as follows from the third estimate in (5.22), the L^2-limit of the sequence $\left\{\rho_\varepsilon^0 - \frac{1}{1+|\nabla u_\varepsilon^0|^2}\right\}_{\varepsilon>0}$ is equal to zero. Hence, we obtain

$$\rho^0(t,x) = \frac{1}{1+|\nabla u^0(t,x)|^2} \quad \text{a.e. in } \ Q_T.$$

Thus,

$$(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) \overset{w}{\to} \left(\frac{1}{1+|\nabla u^0|^2}, v^0, u^0\right) \qquad \text{as } \varepsilon \to 0.$$

Taking into account Proposition 5.1, we see that (v^0,u^0) is a feasible solution to the original OCP $(\mathcal{R})$. Moreover, as a direct consequence of the properties (5.27), we have the following estimate

$$\begin{aligned}\liminf_{\varepsilon\to 0} J_\varepsilon(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) &\geq \frac{1}{2}\int_\Omega |u^0(T)-u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla u^0|^2\,dxdt \\ &+ \frac{\gamma}{2}\int_0^T\int_\Omega |v^0|^2\,dxdt + \int_{Q_T}\left|D\left(\frac{1}{1+|\nabla u^0|^2}\right)\right| = J(v^0,u^0). \quad (5.29)\end{aligned}$$

Let us assume for a moment that the pair (v^0,u^0) is not optimal for $(\mathcal{R})$-problem. Then there exists another pair $(v^*,u^*)\in\Xi$ such that

$$J(v^*,u^*) < J(v^0,u^0) < +\infty. \qquad (5.30)$$

Setting $\rho^* = \left(1+|\nabla u^*|^2\right)^{-1}$, we deduce from condition $(v^*,u^*)\in\Xi$ that the tuple (ρ^*,v^*,u^*) is a feasible solution to each approximate problem $(\mathcal{R}_\varepsilon)$, i.e.,

$$(\rho^*,v^*,u^*)\in\Xi_\varepsilon, \quad \forall\,\varepsilon\in(0,1). \qquad (5.31)$$

Taking this fact into account, we get

$$\begin{aligned}J(v^0,u^0) &= \frac{1}{2}\int_\Omega |u^0(T)-u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla u^0|^2\,dxdt \\ &+ \frac{\gamma}{2}\int_0^T\int_\Omega |v^0|^2\,dxdt + \int_{Q_T}\left|D\left(\frac{1}{1+|\nabla u^0|^2}\right)\right| \\ &\overset{\text{by (5.29)}}{\leq} \liminf_{\varepsilon\to 0} J_\varepsilon(\rho_\varepsilon^0, v_\varepsilon^0, u_\varepsilon^0) = \liminf_{\varepsilon\to 0}\inf_{(\rho,v,u)\in\Xi_\varepsilon} J_\varepsilon(\rho,v,u) \\ &\leq \lim_{\varepsilon\to 0} J_\varepsilon(\rho^*,v^*,u^*) = \frac{1}{2}\int_\Omega |u^*(T)-u_d|^2\,dx + \frac{\lambda}{2}\int_0^T\int_\Omega |\nabla u^*|^2\,dxdt \\ &+ \frac{\gamma}{2}\int_0^T\int_\Omega |v^*|^2\,dxdt + \int_{Q_T}\left|D\left(\frac{1}{1+|\nabla u^*|^2}\right)\right| \\ &+ \frac{1}{\varepsilon}\int_0^T\int_\Omega \left|\rho^* - \frac{1}{1+|\nabla u^* \ |^2}\right|^2 dxdt = J(v^*,u^*).\end{aligned}$$

Thus, $J(v^0,u^0) \le J(v^*,u^*)$ and we come into a conflict with condition (5.30). Hence, the limit pair (v^0,u^0) is optimal for the original OCP $(\mathcal{R})$. □

As follows from Theorem 5.1, the optimal solutions to the approximated problems $(\rho^0_\varepsilon, v^0_\varepsilon, u^0_\varepsilon)$ can be considered as a basis for the construction of suboptimal controls to the original problem $(\mathcal{R})$ (for the details we refer to [9,11,12,19])

References

1. L. Afraites, A. Atlas, F. Karami, D. Meskine, *Some class of parabolic systems applied to image processing*, Discrete and Continuous Dynamical Systems, Series B, **21** (6) (2016), 1671–1687.
2. L. Afraites, A. Hadri, A. Laghrib, M. Nachaoui, *A non-convex denoising model for impulse and Gaussian noise mixture removing using bi-level parameter identification*, Inverse Problems and Imaging, (2022), doi: 10.3934/ipi.2022001, 1–44.
3. L. Alvarez, P.-L. Lions, J.-M. Morel, *Image selective smoothing and edge detection by nonlinear diffusion*, *SIAM J. Numer. Anal,*, **29** (1992), 845–866.
4. L. Ambrosio, N. Fusco, D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford University Press, New York, 2000.
5. H. Attouch, G. Buttazzo, G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, *SIAM*, Philadelphia, 2006.
6. L. Boccardo, F. Murat, *Almost everywhere convergence of the gradients of solutions to elliptic and parabolic equations*, Nonlinear Analysis, Theory, Methods and Applications, **19** (6) (1992), 581–597.
7. F. Catté, P.L. Lions, J-M. Morel, T. Coll, *Image Selective Smoothing and Edge Detection by Nonlinear Diffusion*, *SIAM Journal on Numerical Analysis*, **29** (1) (1992), 182–193.
8. C. D'Apice, P.I. Kogut, R. Manzo, *On coupled two-level variational problem in Sobolev-Orlicz space*, submitted for publication.
9. C. D'Apice, U. De Maio, P.I. Kogut, *Suboptimal boundary control for elliptic equations in critically perforated domains*, *Ann. Inst. H. Poincaré Anal. Non Lineaire*, **25**(2008), 1073–1101.
10. C. D'Apice, U. De Maio, P. Kogut, *An indirect approach to the existence of quasi-optimal controls in coefficients for multi-dimensional thermistor problem*, in "Contemporary Approaches and Methods in Fundamental Mathematics and Mechanics", Editors: Sadovnichiy, Victor A., Zgurovsky, Michael (Eds.). Springer. Chapter 24, (2020), 489–522.
11. U. De Maio, P.I. Kogut, R. Manzo, *Asymptotic Analysis of an Optimal Boundary Control Problem for Ill-Posed Elliptic Equation in Domains with Rugous Boundary*, Asymptotic Analysis, **118** (3) (2020), 209–234.
12. T. Horsin, P.I. Kogut, *Optimal L^2-Control Problem in Coefficients for a Linear Elliptic Equation. I. Existence Result*, Mathematical Control and Related Fields, **5** (1) (2015), 73–96.
13. L. Evans, M. Portilheiro, *Irreversibility and hysteresis for a forward-backward dffusion equation*, Mathematical Models and Methods in Applied Sciences, **14** (11) (2004), 1599–1620.
14. P. Guidotti, *Anisotropic diffusions of image processing from PeronaвЂ“Malik on*, *Advanced Studies in Pure Mathematics*, **67** (2015), 131–156.

15. F. Karami, L. Ziad, K. Sadik, *A splitting algorithm for a novel regularization of Perona-Malik and application to image restoration*, EURASIP Journal on Advances in Signal Processing, **2017** (46) (2017), 1–9.
16. P. Kogut, *On optimal and quasi-optimal controls in coefficients for multi-dimensional thermistor problem with mixed Dirichlet-Neumann boundary conditions*, Control and Cybernetics, **48**(1) (2019), 31–68.
17. P. Kogut, G. Leugering, *On S-homogenization of an optimal control problem with control and state constraints*, Zeitschrift fuer Analysis und ihre Anwendungen, **20** (2001), 395–429.
18. P. Kogut, G. Leugering *Optimal Control Problems for Partial Differential Equations on Reticulated Domains. Approximation and Asymptotic Analysis*, Series: Systems and Control, *Birkhäuser Verlag*, Boston, 2011.
19. P.I. Kogut, R. Manzo, *On Vector-Valued Approximation of State Constrained Optimal Control Problems for Nonlinear Hyperbolic Conservation Laws*, Journal of Dynamical and Control Systems, **19** (2) (2013), 381–404.
20. *O.A. Ladyzhenskaja, V.A. Solonnikov, N.N. UralвЂ™ceva*, *Linear and quasilinear equations of parabolic type*, Translations of the American Mathematical Society, American Mathematical Society, Providence, 1968.
21. S. Lecheheb, M. Maouni, H. Lakhal, *Image Restoration Using a Novel Model Combining the Perona-Malik Equation and the Heat Equation*, International Journal of Analysis and Applications, **19** (2) (2021), 228–238.
22. J.-L. Lions, *Contrôlabilité exacte. Perturbations et stabilisation de systèmes distribués*, Masson, Paris, 1988.
23. P. Perona, J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intelligence, **12** (1990), 161–192.
24. J Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. pura Appl., **146** (1987), 65–96.
25. R.E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, vol. 49 of Mathematical Surveys and Monographs, *American Mathematical Society*, Providence, RI, 1997.
26. V.B Surya Prasath, J.M. Urbano, D. Vorotnikov, *Analysis of adaptive forward-backward diffusion flows with applications in image processing*, Inverse Problems, **31** (2015), Id 105008, 1–30.

Received 25.03.2022

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **30**, Issue **1**, June **2022**, pp. 71–87, DOI 10.15421/142203

ISSN (print) 2617–0108
ISSN (on-line) 2663–6824

NONLINEAR EVOLUTIONARY PROBLEM OF FILTRATION CONSOLIDATION WITH THE NON-CLASSICAL CONJUGATION CONDITION

Olha R. Michuta[*], Petro M. Martyniuk[†]

Abstract. Finite-element solutions of the initial-boundary value problem for a nonlinear parabolic equation in an inhomogeneous domain with the conjugation condition of a non-ideal contact were found. The initial boundary value problem is a mathematical model of an important technical problem of filtration consolidation of inhomogeneous soils. Inhomogeneity is considered in terms of the presence of thin inclusions, physico-chemical characteristics of which differ from those of the main soil. The problem of long-term consolidation is especially pronounced in soils with low filtration coefficient. Low permeability of the porous medium causes deviation from the linear relationship between the pressure gradient and the filtration rate. Weak formulation of the problem is suggested, and the accuracy of the approximate finite element solution, its existence and uniqueness are substantiated for the case of Darcy's nonlinear law. A test example and the effect of the nonlinear filtration law for thin inclusion on the dynamics of scattering of excess pressures in the entire area of the problem are considered.

Key words: nonlinear initial-boundary value problem, finite element method, consolidation, threshold gradient, nonlinear filtration law, conjugation condition.

2010 Mathematics Subject Classification: 49J20, 49K20, 58J37.

Communicated by Prof. V. O. Kapustyan

1. Introduction

A nonlinear initial-boundary value problem is investigated for the parabolic equation in the inhomogeneous domain $\Omega = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, where Ω_1, Ω_2 are some given domains. By inhomogeneity we mean the presence of a contact interface $\omega = \overline{\Omega_1} \cap \overline{\Omega_2}$ which from a physical view point means a thin inclusion of the third material. Differences in the physical characteristics of the materials of the inclusion ω and regions Ω_i, $i = 1, 2$, can lead to the discontinuity of the solutions of the initial boundary value problem at the inclusion. Regarding the study of problems in inhomogeneous environments of this type, we will focus on the methodology where the study of processes at the thin inclusion itself is taken outside the general initial-boundary value problem. The physical characteristics of the thin inclusion material and the inclusion thickness are taken into account. The presence of a thin inclusion is taken into account in the general initial-boundary

[*] Department of Computer Science and Applied Mathematics , National University of Water and Environmental Engineering, 11, Soborna st., Rivne, 33028, Ukraine, o.r.michuta@nuwm.edu.ua

[†] Department of Computer Science and Applied Mathematics , National University of Water and Environmental Engineering, 11, Soborna st., Rivne, 33028, Ukraine, p.m.martyniuk@nuwm.edu.ua

value problem by the so-called conjugation conditions for an unknown function. This approach, when using numerical methods, avoids solving the problem in the inclusion itself and thus simplifies the solution process.

The above approach to simulating inhomogeneities in soils began to develop in the work of I.I. Lyashko, I. V. Sergienko, V. V. Skopetskii, V. S. Deineka and is quite fully described in review monographs [6, 15, 16]. The works [2, 11, 14–17] are also worth mentioning in this direction which develop both the solution methods and the qualitative theory of initial-boundary value problems with possible discontinuous solutions.

As noted above, the initial-boundary value problem, more specifically the conjugation with non-ideal contact, include physical characteristics of the material of the thin inclusion (inclusions themselves may be both of natural origin and artificial). The parameters of the material of thin soil inclusions (filtration coefficient, porosity, thermal conductivity, etc.) are nonlinearly dependent on the effect of external factors. Considering the initial-boundary value problems as mathematical models of physico-chemical processes in porous soil media, the presence of such dependences requires modification of the conjugation conditions. The above-mentioned works [2,6,11,14–17] assumed the parameters of inclusions to be constant, which is reflected in the conjugation conditions with non-ideal contact. Mathematical models for the distribution of inorganic chemicals in porous media and modification of conjugation conditions taking into account nonlinear dependences of material parameters of thin geobarriers on the effect of physicochemical factors, including chemical suffusion [23] were developed in [18, 19].

The influence of organic substances on the development of microorganisms and the effect of bioclogging processes on the value of pressure jumps at a thin geobarrier were studied in [20,21]. Modified conjugation conditions and mathematical models of moisture transfer in inhomogeneous porous media are presented in [5, 12]. The method of modification of conjugation conditions for the partial case of Darcy's nonlinear law is shown in [13].

Here, we investigate the initial-boundary value problem for a quasilinear parabolic equation as a mathematical model of soil consolidation. The problem of consolidation (compaction) of soils is especially relevant for water-saturated clays. This is caused by the low filtration coefficient of clay soils and, as a result, the long time for the transition of clay bases of civil and industrial buildings to a stable state. Weak permeability of clay soils raises questions about the limits of Darcy's filtration law in its classical form [9]. It will be recalled that Darcy's classical law mathematically records the linear relationship between the filtration rate and the pressure gradient. The linearity of Darcy's basic filtration law has its physically determined limits. The linearity is violated both for highly permeable porous media and for weakly permeable ones. In particular, for weakly permeable porous media, this is manifested in the presence of the so-called "threshold gradient", below which the relationship between the filtration rate and the pressure gradient becomes nonlinear.

The nonlinearity of Darcy's law for consolidation problems is taken into account in e.g. [10, 24, 25]. However, only homogeneous media without thin inclusions are considered there. Additionally, power laws were considered nonlinear (dependence of the filtration rate on the pressure gradient raised to a certain degree other than unity). Quasilinear filtration processes were studied in [3, 4] where the linear dependence of the filtration rate on the pressure gradient is preserved, but the filtration coefficient nonlinearly depends on the physico-chemical parameters of the porous medium.

Thus, the objectives of this work are: 1) modification of the conjugation condition on a thin inclusion under Darcy's nonlinear law; 2) formation of a mathematical model of filtration consolidation of inhomogeneous soil in the presence of the threshold gradient; 3) investigation of finite element solutions of the corresponding boundary value problem, numerical experiments and analysis of the significance of the nonlinearity of Darcy's law on the value of excess pressures and their jumps.

2. The problem of nonlinear filtration through a thin inclusion

It is suggested in [9] to generalize nonlinear filtration laws in the form (in one-dimensional case)

$$u = -k\left(\frac{\partial h}{\partial x} - \frac{I}{\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha}, \left(\frac{i}{I^*}\right)^{\alpha}\right)\operatorname{sgn}\left(\frac{\partial h}{\partial x}\right)\right), \tag{2.1}$$

where i is the absolute value of the pressure gradient, i.e. in the one-dimensional case $i = \left|\frac{\partial h}{\partial x}\right|$; k is the filtration coefficient of the porous medium; u is the filtration rate; I is the absolute value of the pressure gradient below which the linearity of Darcy's law is violated (the so-called threshold gradient); α is an empirical parameter;

$$I^* = \frac{\alpha}{\gamma\left(\frac{1}{\alpha}\right)}I,$$

$$\gamma\left(a, x\right) = \int_0^x s^{a-1}e^{-s}ds,$$

$$\gamma\left(a\right) = \int_0^\infty s^{a-1}e^{-s}ds.$$

Here $\gamma\left(a\right)$, $\gamma\left(a, x\right)$ are the so-called gamma function and the lower incomplete gamma function.

As noted in [9], Eq. (2.1) includes previously proposed nonlinear filtration laws for permeable soils, i.e. Hansbo's law (1960), Swartzendruber's law (1961), Zou's law (1996).

The use of the law in the form of (2.1) is quite inconvenient in terms of applying the finite element method. Given that $\frac{\partial h}{\partial x} = i\operatorname{sgn}\left(\frac{\partial h}{\partial x}\right)$, it follows from (2.1) that

$$u = -k(n)\left(1 - \frac{I}{\left(i+\varepsilon^2\right)\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha}, \left(\frac{i}{I^*}\right)^{\alpha}\right)\right)\frac{\partial h}{\partial x},$$

or

$$u = -k^*(n, I)\frac{\partial h}{\partial x}, \tag{2.2}$$

where

$$k^*(n, I) = k(n)\left(1 - \frac{I}{(i + \varepsilon^2)\,\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha}, \left(\frac{i}{I^*}\right)^{\alpha}\right)\right).$$

Here n is the soil porosity; $\varepsilon > 0$ is a small constant. Since in this work we consider the problem of soil compaction, the dependence of the filtration coefficient on porosity should be taken into account.

We assume (due to thinness of the inclusion ω, Fig. 1) that the filtration processes in the cross-section of this inclusion are stationary (or at least quasi-stationary). Thus, similarly to [5, 12, 13, 18–21], for the inclusion of thickness d, we consider the following filtration problem:

$$\frac{d}{d}\left(-k^*_\omega(n_\omega, I_\omega)\frac{dh}{d\xi}\right) = 0, 0 < \xi < d, \tag{2.3}$$

$$h(0) = h^-, h(d) = h^+. \tag{2.4}$$

Here, h^-, h^+ are the known values of pressures, and the sub-script ω means the corresponding characteristic for the inclusion ω (Fig. 4.1). Repeating the reasoning of [5, 12, 13, 18–21], we have

$$h(\xi) = \frac{\displaystyle\int_0^\xi \frac{dx}{k^*_\omega(n_\omega, I_\omega)}}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(n_\omega, I_\omega)}}\left(h^+ - h^-\right) + h^-.$$

However, hereafter we are more interested not in the pressure itself, but in its gradient. As a result

$$\frac{dh}{d\xi} = \frac{1}{k^*_\omega(n_\omega, I_\omega)\displaystyle\int_0^d \frac{dx}{k^*_\omega(n_\omega, I_\omega)}}\left(h^+ - h^-\right). \tag{2.5}$$

3. Conjugation condition for nonlinear filtration law

According to [16], similarly to [5, 13] the conjugation condition is derived on the basis of the law of conservation of fluid flow through the cross-section area of the inclusion surface along the normal in time Δt. As the flow

$$q = -k^*_\omega(n_\omega, I_\omega)\frac{dh}{d}\Delta t = u\Delta t$$

and

$$q = q^+ = q^-,$$

then

$$u^{\pm}\big|_{x=\xi} = -k^*_\omega(n_\omega, I_\omega)\frac{dh}{d\xi}. \tag{3.1}$$

From (2.5) and (3.1) we have the final formula of the conjugation condition with non-ideal contact for pressures at the inclusion, with nonlinear filtration law in the form (2.2)

$$u^{\pm}\big|_{x=\xi} = -\frac{[h]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(n_\omega, I_\omega)}}. \tag{3.2}$$

Here $[h] = h^+ - h^-$ is the pressure jump at the inclusion.

Before formulating the mathematical model of the problem, we note that the porosity n (as well as n_ω) of the soil in the regions Ω_i, $i = 1, 2$, is related to the void ratio e as $n = \frac{e}{1+e}$. In its turn, and it is shown in the section of numerical experiment results, e depends on pressures $h(x,t)$. Thus $n = n(h)$, and in subsequent calculations $k^* = k^*(h, I)$, $k^*_\omega = k^*_\omega(h, I_\omega)$.

4. Nonlinear mathematical model of filtration consolidation of porous medium with thin inclusion

Here we consider the process of filtration consolidation of the soil layer of total thickness l with a thin inclusion ω of thickness d which is located at the depth $x = \xi$ (Fig. 4.1). The material of the thin inclusion differs in its physico-chemical characteristics from those of the main soil.

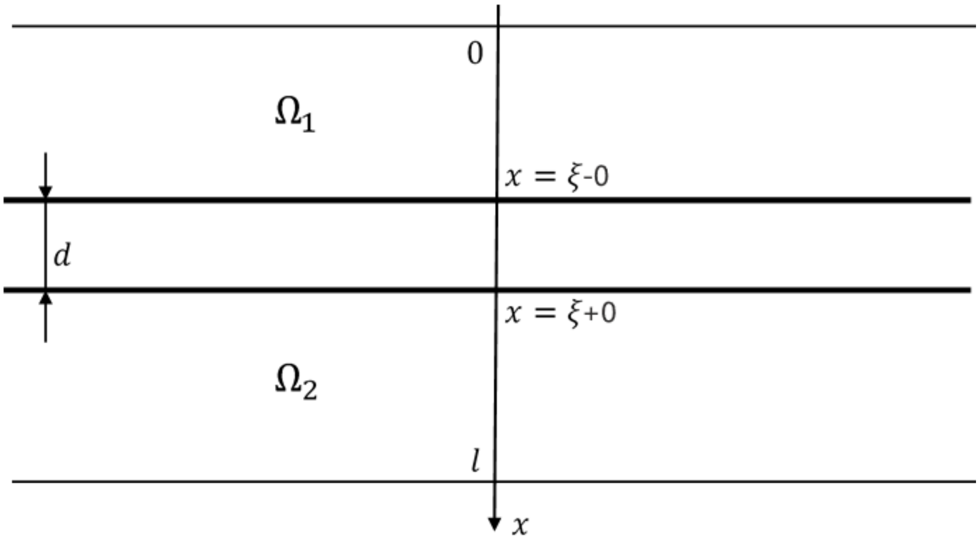


Fig. 4.1. A layer of soil of thickness l with a thin inclusion ω of thickness d ($d \ll l$).

The formulation of the mathematical model of the described problem will partially use the work of scientists reviewed in Introduction. The mathematical model will include the equations of filtration consolidation [8,22,23]. As a result, we have the following boundary value problem:

$$\frac{\partial h}{\partial t}=\frac{1+e}{\gamma a}\frac{\partial}{\partial x}\left(k^*(h,I)\frac{\partial h}{\partial x}\right),\ x\in\Omega_1\cup\Omega_2,\ t\in(0,T],\tag{4.1}$$

$$\left.h\left(x,t\right)\right|_{x=0}=0, t\in[0,T],\tag{4.2}$$

$$\left.u\left(x,t\right)\right|_{x=l}=\left.\left(-k^*(h,I)\frac{\partial h}{\partial x}\right)\right|_{x=l}=0,\quad t\in[0,T],\tag{4.3}$$

$$h\left(x,0\right)=h_0\left(x\right),\quad x\in\overline{\Omega}_1\cup\overline{\Omega}_2,\tag{4.4}$$

$$\left.u^{\pm}\right|_{x=\xi}=\left.\left(-k^*(h,I)\frac{\partial h}{\partial x}\right)^{\pm}\right|_{x=\xi}=-\frac{[h]}{\displaystyle\int_0^d\frac{dx}{k^*_\omega(h,I_\omega)}}.\tag{4.5}$$

Here,

$$k^*(h,I)=k\left(h\right)\left(1-\frac{I}{\left(i+\varepsilon^2\right)\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha},\left(\frac{i}{I^*}\right)^{\alpha}\right)\right),$$

$$\Omega_1=\left(0;\xi\right),\Omega_2=\left(\xi;l\right),0<\xi<l;\Omega=\Omega_1\cup\Omega_2;$$

$T>0$ is the specified time duration; $h_0\left(x\right)$ is a known function; a is the soil compressibility coefficient; $n,\ \ n_\omega$ are the porosity of soil and inclusion material, respectively; $e=\frac{n}{1-n}$ is the soil void ratio; γ is the specific weight of the pore fluid; h is the pressure; $k,\ \ k_\omega$ are the filtration coefficients of the main soil and soil inclusion, respectively; u is the filtration rate which is determined according to (2.2); $u^{\pm}$ are the filtration rates at $x=\xi-0$ and $x=\xi+0$, respectively; $[h]=h^+-h^-$ is the pressure jump at the thin inclusion.

We shall show that k^* is always positive. Let us turn to the starting relations from [9] from which the generalized Darcy's law (2.1) is derived. Particularly, it is based on ([9, formula (1.20)])

$$\frac{dq}{di}=k\left(1-e^{-i\left(\frac{i}{I^*}\right)^{\alpha}}\right)\tag{4.6}$$

from which we have [9, formula (1.21)]

$$q=k\left(i-\frac{I}{\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha},\left(\frac{i}{I^*}\right)^{\alpha}\right)\right).\tag{4.7}$$

From (4.7), $\left.q\right|_{i=0}=0$, from (4.6), $\left.\frac{dq}{di}\right|_{i>0}>0$. Therefore, the function q as the function of the absolute value of i is increasing for $i\in(0;+\infty)$, and is equal to zero for $i=0$. That is, $\left.q\right|_{i>0}>0$. It follows that

$$i-\frac{I}{\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha},\left(\frac{i}{I^*}\right)^{\alpha}\right)>0$$

or

$$1-\frac{I}{i\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha},\left(\frac{i}{I^*}\right)^{\alpha}\right)>0.$$

Then it further reinforces the above inequality that when $i>0$

$$1-\frac{I}{(i+\varepsilon^2)\gamma\left(\frac{1}{\alpha}\right)}\gamma\left(\frac{1}{\alpha},\left(\frac{i}{I^*}\right)^{\alpha}\right)>0.$$

Since $\gamma\left(\frac{1}{\alpha},0\right)=0$, then for $i=0$ from the formula for $k^*(h,I)$ we obtain $k^*(h,I)=k(h)>0$. Therefore

$$k^*(h,I)>0,\ i\in[0;+\infty)$$

and similarly

$$k^*_\omega(h,I_\omega)>0,\ i\in[0;+\infty).$$

Similarly to [6] we introduce the following notation: $Q_T=\Omega\times(0;T]$, $Q_T^1=\Omega_1\times(0;T]$, $Q_T^2=\Omega_2\times(0;T]$.

Assume that the function $h_0(x)$ is continuous on each of the closures $\overline{\Omega}_1$, $\overline{\Omega}_2$. Also with respect to the coefficients k^*, k^*_ω, assume that

1)

$$0<k^*_{min}\le k^*(s_1,s_2)\le k^*_{max}<\infty,$$

$$0<k^*_{\omega,min}\le k^*_\omega(s_1,s_2)\le k^*_{\omega,max},$$

$\forall s_1\in(-\infty;+\infty)$, $\forall s_2\in[0;+\infty)$; k^*_{min}, k^*_{max}, $k^*_{\omega,min}$, $k^*_{\omega,max}$ are positive constants;

2)

$$\left|k^*\left(p_1,s_1\right)-k^*\left(p_2,s_2\right)\right|\le k^*_L\left|p_1-p_2\right|,\ 0<k^*_L<\infty;$$

$$\left|k^*_\omega\left(p_1,s_1\right)-k^*_\omega\left(p_2,s_2\right)\right|\le k^*_{\omega,L}\left|p_1-p_2\right|,\ 0<k^*_{\omega,L}<\infty.$$

Also, the function $k^*=k^*(h,I)$ must be continuous on $\overline{\Omega}_1$, $\overline{\Omega}_2$ and continuously differentiated on Ω_1, Ω_2.

Definition 4.1. The classical solution of the initial-boundary value problem (4.1)–(4.5) which allows a discontinuity of the first kind at the point $x=\xi$ is a function $h(x,t)\in\Psi$ that satisfies $\forall(x,t)\in\overline{Q}_T$ equation (4.1) and the initial condition (4.4).

Here, Ψ is a set of functions $\psi(x,t)$ which, together with $\frac{\partial\psi}{\partial x}$, are continuous on each of the closures $\overline{Q}_T^1$, $\overline{Q}_T^2$, have bounded continuous partial derivatives $\frac{\partial\psi}{\partial t}$, $\frac{\partial^2\psi}{\partial x^2}$ on Q_T^1, Q_T^2, and satisfy conditions (4.2), (4.3), (4.5).

For further calculations we will note one more aspect. The conjugation condition with non-ideal contact (see [16, page 291, formula (7.4)]) which can be called classic

$$\left.\left(\varkappa\left(x,u\right)\frac{\partial u}{\partial x}\right)\right|_{x=\xi}=r\left[u\right]$$

includes r, a known constant, and $0 < r_0 \leq r < \infty$. Consider condition (4.5). When the second of conditions 1) for the coefficient in the right part of the conjugation condition (4.5)

$$\left(-k^*(h,I)\frac{\partial h}{\partial x}\right)^{\pm}\Bigg|_{x=\xi} = -\frac{[h]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(h,I_\omega)}}$$

is satisfied, we have

$$\frac{d}{k^*_{\omega,max}} \leq \int_0^d \frac{dx}{k^*_\omega(h,I_\omega)} \leq \frac{d}{k^*_{\omega,min}}.$$

Further,

$$k^*_{\omega,min}\frac{[h]}{d} \leq \frac{[h]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(h,I_\omega)}} \leq k^*_{\omega,max}\frac{[h]}{d}.$$

Thus, in the case of a modified conjugation condition (4.5), we have

$$0 < \frac{k^*_{\omega,min}}{d} \leq r \leq \frac{k^*_{\omega,max}}{d} < \infty. \tag{4.8}$$

Estimate (4.8) allows us to generalize the theorems proved in [6,16] for problems with the classical conjugation condition with non-ideal contact, for the case of a modified conjugation condition (4.5).

5. Generalized solution of problem (4.1)-(4.5)

Similarly to [6] let H_0 be the space of functions $s\,(x)$ that in each of the regions Ω_i belong to the Sobolev space $W_2^1\,(\Omega_i)$, $i = 1, 2$, and they acquire zero values at the ends of the segment $[0; l]$ where the function $h\,(x,t)$ is set the boundary conditions of the first kind.

Let $h(x,t) \in \Psi$ be the classical solution of the initial-boundary value problem (4.1)–(4.5). Take $s\,(x) \in H_0$. Multiply equation (4.1) and initial condition (4.4) by $s\,(x)$. Integrating them on the segment $[0; l]$ and taking into account the conjugation conditions (4.5), we obtain

$$\int_0^l \frac{\gamma a}{1+e}\frac{\partial h}{\partial t}s\,(x)\,dx + \int_0^l k^*(h,I)\frac{\partial h}{\partial x}\frac{ds}{dx}dx + \frac{[h]\,[s]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(h,I_\omega)}} = 0, \tag{5.1}$$

$$\int_0^l h\,(x,0)\,s\,(x)\,dx = \int_0^l h_0\,(x)\,s\,(x)\,dx. \tag{5.2}$$

Thus, if $h(x,t) \in \Psi$ is a classical solution of the initial-boundary value problem (4.1)-(4.5), then $h(x,t)$ is a solution of problem (5.1), (5.2) in a weak formulation.

Let H be the space of functions $v\,(x,t)$ that are square-integrable together with their first derivatives $\frac{\partial v}{\partial t}$, $\frac{\partial v}{\partial x}$ on each of the intervals $(0;\xi)$, $(\xi;l)$, $\forall t \in (0\,;T]$, $T>0$, and they satisfy the same boundary conditions of the first kind as the function $h\,(x,t)$.

Definition 5.1. Function $h\,(x,t) \in H$ that for any $s\,(x) \in H_0$ satisfies integral relations (5.1), (5.2) is called a generalized solution of the initial-boundary value problem (4.1)–(4.5).

An approximate generalized solution of the initial-boundary value problem (4.1)-(4.5) will be sought in the form

$$\hat{h}\,(x,t) = \sum_{i=1}^{N} h_i\,(t)\,\varphi_i\,(x), \tag{5.3}$$

where $\{\varphi_i(x)\}_{i=1}^{N}$ is the basis of finite-dimensional subspace $M_0 \subset H_0$; $h_i\,(t)$, $i=\overline{1,N}$ are unknown coefficients that depend only on time.

The set of functions that can be represented in the form (5.3) generate a finite-dimensional subspace $M_1 \subset H_1$.

Definition 5.2. An approximate generalized solution of the initial-boundary value problem (4.1)–(4.5) is a function $\hat{h}(x,t) \in M_1$ that for an arbitrary function $S(x) \in M_0$ satisfies integral relations

$$\int_0^l \frac{\gamma a}{1+e}\frac{\partial \hat{h}}{\partial t} S\,(x)\,dx + \int_0^l k^*(\hat{h},I)\frac{\partial \hat{h}}{\partial x}\frac{dS}{dx}dx + \frac{\left[\hat{h}\right]\,[S]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(\hat{h},I_\omega)}} = 0, \tag{5.4}$$

$$\int_0^l \hat{h}\,(x,0)\,S\,(x)\,dx = \int_0^l h_0\,(x)\,S\,(x)\,dx. \tag{5.5}$$

Next, from the weak formulation of (5.4), (5.5) we obtain (assuming the function $S(x)$ equal to each basis function $\varphi_i(x)$, $i=\overline{1,N}$) the Cauchy problem for a system of nonlinear differential equations

$$\mathbf{M}\,(\mathbf{H})\,\frac{d\mathbf{H}}{dt} + \mathbf{L}\,(\mathbf{H})\,\mathbf{H}\,(\boldsymbol{t}) = \mathbf{0}, \tag{5.6}$$

$$\widetilde{\mathbf{M}}\mathbf{H}^{(\mathbf{0})} = \widetilde{\mathbf{F}}, \tag{5.7}$$

where

$$\widetilde{\mathbf{F}} = \left(\tilde{f}_i\right)_{i=1}^{N}, \quad \widetilde{\mathbf{M}} = (\tilde{m}_{ij})_{i,j=1}^{N}, \quad \tilde{m}_{ij} = \int_0^l \varphi_i\varphi_j dx, \quad \mathbf{M} = (m_{ij})_{i,j=1}^{N},$$
$$\mathbf{L} = (l_{ij})_{i,j=1}^{N}, \quad \mathbf{H} = (h_i\,(t))_{i=1}^{N}, \quad \mathbf{H}^{(\mathbf{0})} = (h_i\,(0))_{i=1}^{N},$$

$$m_{ij} = \int_0^l \frac{\gamma a}{1+e} \varphi_i \varphi_j dx,$$

$$l_{ij} = \int_0^l k^*(\hat{h}, I) \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} dx + \frac{[\varphi_i]\,[\varphi_j]}{\displaystyle\int_0^d \frac{dx}{k^*_\omega(\hat{h}, I_\omega)}}.$$

The square matrix $\mathbf{M}\,(\mathbf{H})$ is symmetric and positive definite because $\frac{\gamma a}{1+e} > 0$, $\forall (x,t) \in \overline{Q}_T$. Given the proven positivity of the coefficients k^*, k^*_ω, as well as the assumptions 1) made regarding their limitations, the matrix $\mathbf{L}\,(\mathbf{H})$ will also be symmetric and positively definite [6, page 417]. Next, similarly to [6, problem (3.14) of Chapter 8], we write the system (5.7) in the form

$$\frac{d\mathbf{H}}{dt} = \mathbf{\Phi}\,(\mathbf{H})\,, \tag{5.8}$$

where $\mathbf{\Phi}\,(\mathbf{H}) = -\mathbf{M}^{-1}\mathbf{L}\,(\mathbf{H})\,\mathbf{H}\,(\boldsymbol{t})$**.** The functions $\mathbf{\Phi}\,(\mathbf{H})$, $\partial\mathbf{\Phi}/\partial\mathbf{H}$ are continuous. Thus, there is a single approximate generalized solution $\hat{h}\,(x,t) \in M_1$ of the initial-boundary value problem (4.1)–(4.5).

We will introduce the following norms [6, page 380]:

$$\begin{aligned}
&\|u\|^2_{L_2} = \int_0^l u^2(x,t)dx,\\
&\|u\|^2_{H_0^1} = \left\|\frac{\partial u}{\partial x}\right\|^2_{L_2},\\
&\|u\|^2_{L_2\times L_2} = \|u\|^2_{L_2(Q_T)} = \int_0^T\int_0^l u^2 dxdt,\\
&\|u\|^2_{H_0^1\times L_2} = \int_0^T \|u\|^2_{H_0^1} dt = \int_0^T\int_0^l \left(\frac{\partial u}{\partial x}\right)^2 dxdt,\\
&\|u\|_{L_2\times L_\infty} = \sup_{t\in(0,T]} \|u(\cdot,t)\|_{L_2},\\
&\|\nabla_x u\|_{L_\infty\times L_\infty} = \sup_{(x,t)\in Q_T} \left|\frac{\partial u(x,t)}{\partial x}\right|,\\
&\|u\|^2_{W_2^1\times L_2} = \int_0^T\int_0^l \left(u^2 + \left(\frac{\partial u}{\partial x}\right)^2\right) dxdt,\\
&\|[u]\|^2_{L_2} = \int_0^T [u]^2 dt = \int_0^T (u(\xi+0,t) - u(\xi-0,t))^2 dt.
\end{aligned}$$

Similarly to [6, page 380, Theorem 1] we can prove the following result.

Theorem 5.1. *Let $h(x,t)$ be a classical solution of the initial-boundary value problem (4.1)–(4.5), and $\hat{h}(x,t)$ be a generalized solution of this problem from space M_1. Then, under the conditions 1), 2) imposed on k^*, k^*_ω, taking into account*

(4.8)*, there are such positive constants c,* δ_1*,* δ_2*, that for an arbitrary function* $\tilde{h}(x,t) \in M_1$ *the following inequality holds:*

$$\left\|h-\hat{h}\right\|^2_{L_2\times L_\infty}+\delta_1\left\|h-\hat{h}\right\|^2_{H_0^1\times L_2}+\delta_2\left\|\left[h-\hat{h}\right]\right\|^2_{L_2} \\ \leq c\Big\{\left\|h-\tilde{h}\right\|^2_{L_2\times L_\infty}+\left\|h-\tilde{h}\right\|^2_{H_0^1\times L_2} \\ +\left\|\left[h-\tilde{h}\right]\right\|^2_{L_2}+\left\|\frac{\partial(h-\tilde{h})}{\partial t}\right\|^2_{L_2\times L_2}\Big\}, \quad \forall\tilde{h}\in M_1. \quad (5.9)$$

Dependence (5.9) is used in estimating the accuracy of the finite element method.

6. Finite element method

We will cover the closure $\Omega = \overline{\Omega}_1 \cup \overline{\Omega}_2$ with a finite element grid with the total number of nodes N. The point $x = \xi$ should be double numbered, the node on the left $x = \xi - 0$ and the node on the right $x = \xi + 0$. Let in (4.8) $\varphi_i(x)$ be the basis functions of the finite element method which allow a discontinuity of the first kind at the point $x = \xi$ and are polynomials of m-th degree. Then the space of functions $\hat{h}(x,t)$ of the form (5.3) with the specified basic functions is denoted H_m^N.

Theorem 6.1. *Let the classical solution* $h(x,t)$ *of the boundary value problem* (4.1)–(4.5) *have partial derivatives* $\frac{\partial^{m+1}(\cdot)}{\partial x^{m+1}}$*,* $\frac{\partial^{m+2}(\cdot)}{\partial x^m \partial t}$ *limited on* Q_T^i*,* $i = 1, 2$*. Then the approximate generalized solution* $\hat{h}(x,t) \in H_m^N$ *has an estimate*

$$\left\|h-\hat{h}\right\|_{W_2^1\times L_2} \leq c\cdot h_{max}^m,$$

where m is the degree of FEM polynomials, $c = const > 0$*,*

$$h_{max} = \max_{i=\overline{0,N-1}} \left(x_{i+1} - x_i\right),$$

$[x_{i+1}; x_i]$ *are finite elements.*

Proof. The validity of the theorem follows from the estimate (5.9) of the previous theorem taking into account the interpolation estimates [6, page 387, Theorem 2]. □

7. Time discretization methods

Problem (5.6), (5.7) is a Cauchy problem for a system of nonlinear differential equations of the first order. Finding its solution also requires the use of appropriate discretization methods. The application of the Crank-Nicolson method is subrstantiated in [15]:

$$\mathbf{M}\left(\frac{1}{2}\left(\mathbf{H}^{(j+1)}+\mathbf{H}^{(j)}\right)\right)\frac{\mathbf{H}^{(j+1)}-\mathbf{H}^{(j)}}{\tau}$$
$$+\mathbf{L}\left(\frac{1}{2}\left(\mathbf{H}^{(j+1)}+\mathbf{H}^{(j)}\right)\right)\cdot\frac{1}{2}\left(\mathbf{H}^{(j+1)}+\mathbf{H}^{(j)}\right)=\mathbf{0}, \quad j=0,1,2,...,m_\tau-1.$$

Here the time segment $[0,T]$ is divided into m_τ equal parts with the step $\tau=\frac{T}{m_\tau}$; $\mathbf{H}^{(j)}$ is the approximate solution of the Cauchy problem (5.6), (5.7) for $t=j\tau$. Let also introduce the following notation: h_j is the classical solution of the initial-boundary value problem (4.1)–(4.5) for $t=j\tau$; $\hat{h}_j$ is the approximate generalized solution of the initial-boundary value problem (4.1)-(4.5) for $t=j\tau$; $\phi_{j+1/2}=\frac{1}{2}(\phi_{j+1}+\phi_j)$; $z_j=h_j-\hat{h}_j$.

Given (4.8), similarly to Theorem 5 [6, Chapter 8] it is also valid the following result.

Theorem 7.1. *Let $h(x,t)$ be a classical solution of the initial-boundary value problem (4.1)–(4.5). Let the functions $\frac{\partial h}{\partial t}$, $\frac{\partial h}{\partial x}$ be twice continuously differentiable over time on $\overline{Q}_T^i$, $i=1,2$. Let also assume that the derivatives $\frac{\partial^3 h}{\partial t^3}$, $\frac{\partial^3 h}{\partial t^2 \partial x}$ are uniformly limited in modulus by a constant c_1, $\forall(x,t)\in\overline{Q}_T$. If conditions 1), 2) are satisfied, then there are positive constants c, δ_1, r_0, τ_0, that depend on the constants of conditions 1), 2), as well as T, l, such that for $\forall\tau\le\tau_0$ the classical solution $h(x,t)$ and the approximate generalized solution obtained using the Crank-Nicolson method, $\hat{h}(x,t)\in M_1$, of the problems (4.1)–(4.5) and (5.6), (5.7), respectively, satisfy the inequality*

$$\|z_{m_\tau}\|^2_{L_2}+\delta_1\sum_{j=0}^{m_\tau-1}\|z_{j+1/2}\|^2_{H_0^1}\tau+r_0\sum_{j=0}^{m_\tau-1}\left[z_{j+1/2}\right]^2\tau$$
$$\le c\left(\sum_{j=0}^{m_\tau-1}\left\|(h-\tilde{h})_{j+1/2}\right\|^2_{H_0^1}\tau++\sum_{j=1}^{m_\tau-1}\left\|\frac{(h-\tilde{h})_{j+1/2}-(h-\tilde{h})_{j-1/2}}{\tau}\right\|^2_{L_2}\tau\right.$$
$$+\sum_{j=0}^{m_\tau-1}\left[(h-\tilde{h})_{j+1/2}\right]^2\tau+\left\|(h-\tilde{h})_0\right\|^2_{L_2}+\left\|(h-\tilde{h})_{m_\tau-1/2}\right\|^2_{L_2}$$
$$\left.+\left\|(h-\tilde{h})_{1/2}\right\|^2_{L_2}+O(\tau^4)\right),\ \forall\tilde{h}\in M_1. \quad (7.1)$$

Similarly to Theorem 6 [6, Chapter 8] and taking into account estimate (7.1), we have:

Theorem 7.2. *Let the classical solution $h(x,t)$ of the problem (4.1)–(4.5) satisfy the conditions of Theorem 7.1. Then for the errors z of the approximate generalized* *solution* $\hat{h}(x,t) \in H_m^N$ *of the problem* (5.4), (5.5) *obtained using the Crank-Nicolson method, the following estimate is valid:*

$$\|z_{m_\tau}\|_{L_2}^2 + \delta_1 \tau \sum_{j=0}^{m_\tau - 1} \|z_{j+1/2}(h)\|_{H_0^1}^2 \le c \cdot \left(h_{max}^{2m} + \tau^4\right).$$

However, the practical implementation of the Crank-Nicolson method for the nonlinear Cauchy problem (5.6), (5.7) requires the use of iterations. Instead of the Crank-Nicolson method, one can use the predictor-corrector method [16], which for the system of equations (5.6) has the following form:

$$\mathbf{M}\left(\mathbf{H}^{(j)}\right)\frac{\mathbf{W}^{(j+1)} - \mathbf{H}^{(j)}}{\tau} + \mathbf{L}\left(\mathbf{H}^{(j)}\right)\frac{1}{2}\left(\mathbf{W}^{(j+1)} + \mathbf{H}^{(j)}\right) = \mathbf{0},$$

$$\mathbf{M}\left(\frac{1}{2}\left(\mathbf{W}^{(j+1)} + \mathbf{H}^{(j)}\right)\right)\frac{\mathbf{H}^{(j+1)} - \mathbf{H}^{(j)}}{\tau} +$$

$$+\mathbf{L}\left(\frac{1}{2}\left(\mathbf{W}^{(j+1)} + \mathbf{H}^{(j)}\right)\right) \cdot \frac{1}{2}\left(\mathbf{H}^{(j+1)} + \mathbf{H}^{(j)}\right) = \mathbf{0},\ j = 0, 1, 2, ..., m_\tau - 1,$$

where $\mathbf{W}^{(j+1)}$ are auxiliary vector functions.

From the view point of the simplicity of practical implementation, a fully implicit linearized difference scheme has proved itself well [8,21,23]. For the system (5.6), it has the form

$$\mathbf{M}\left(\mathbf{H}^{(j)}\right)\frac{\mathbf{H}^{(j+1)} - \mathbf{H}^{(j)}}{\tau} + \mathbf{L}\left(\mathbf{H}^{(j)}\right) \cdot \mathbf{H}^{(j+1)} = \mathbf{0},\ j = 0, 1, 2, ..., m_\tau - 1.$$

8. Results of numerical experiments and their analysis

According to [9, formula (1.24)],

$$I = A\tilde{k}^B,$$

where $A = 4.0 \times 10^{-12}$ and $B = -0.78$ are empirical parameters; $\tilde{k}$ ($\left[\tilde{k}\right] = m^2$) is the soil permeability coefficient, i.e. $k = \frac{\tilde{k}\rho g}{\mu}$, ρ is the density of pore fluid, μ is its viscosity, g is the acceleration of free fall. Since these studies do not yet take into account non-isothermal conditions, the dynamic viscosity of water at constant temperature $25^\circ C$ is used which is

$$\mu = 1.03 \cdot 10^{-8}\ Pa \cdot day.$$

Soil parameters for the numerical experiments were taken from the Hydrus-1D freeware. Specifically, Sandy Clay was considered as the main soil, with $k_0 = 0.0288\ m/day$, $n_0 = 0.38$. Then for the main soil $\tilde{k} = 2.98 \cdot 10^{-14}\ m^2$, and $I =$

0.142. Silty Clay was taken as the inclusion soil, with $k_{0\omega} = 0.0048\ m/day$, $n_{0\omega} = 0.46$, where index "0"denotes the initial values. Then for the inclusion soil $\tilde{k} = 4.97 \cdot 10^{-15}\ m^2$, $I = 0.574$.

The parameter α is also important. According to [9, formula (1.24)] $\alpha \geq 0$. The authors state that the parameter α mainly characterizes the smoothness of the transition from nonlinear to linear part of the curve of dependence $u = u(i)$ for the filtration rate and depends on the distribution of pore sizes in the porous medium. An increase in α means a sharper transition. A larger distribution range of pore sizes means a smoother transition between the linear and nonlinear parts and thus a smaller α value. For instance, for $\alpha \to \infty$ we have from the generalized law (2.1) [1]

$$u = \begin{cases} 0, & i \leq I; \\ -k\,(i - I)\,\mathrm{sgn}(i), & i \geq I. \end{cases}$$

When $\alpha \to 0$ we obtain the transition to Darcy's linear law.

We used $\alpha = 2$ for the main soil, and $\alpha = 5$ hor the thin inclusion soil in the following numerical experiments.

According to the linear compression dependence for soils,

$$e = -a\sigma + const.$$

Here σ are vertical stresses in the soil skeleton (in one-dimensional case). Further,

$$\frac{\partial e}{\partial t} = -a\frac{\partial \sigma}{\partial t}.$$

Also, according to Terzaghi's effective stress principle [22, 23]

$$\frac{\partial \sigma}{\partial t} = -\gamma\frac{\partial h}{\partial t},$$

and

$$\frac{\partial e}{\partial t} = a\gamma\frac{\partial h}{\partial t}.$$

From the last ratio we obtain

$$\frac{e^{(j+1)} - e^{(j)}}{\tau} = a\gamma\frac{h^{(j+1)} - h^{(j)}}{\tau}, \quad j = 0, 1, 2, ..., m_\tau - 1,$$

or

$$e^{(j+1)} = a\gamma\left(h^{(j+1)} - h^{(j)}\right) + e^{(j)}, \quad j = 0, 1, 2, ..., m_\tau - 1.$$

Obtained ratio was used to determine the variable filtration coefficient in the void ratio according to the Kozeny-Carman equation [7]

$$k = k_0\frac{1 + e_0}{1 + e}\left(\frac{e}{e_0}\right)^3,$$

where k_0, e_0 are the initial values of filtration coefficient and void ratio; k, e are their variable values over time.

In equation (4.1), the soil compressibility coefficient $a = 5.12 \times 10^{-7} \frac{m^2}{H}$, specific gravity of pore fluid $\gamma_c = 10^4 \frac{H}{m^2}$. Initial pressure distribution $h_0(x) = 20\ m$ is corresponding to the application of the respective load to the soil surface. Unobstructed outflow of pore fluid is provided at the upper limit, and there is no drainage at the lower limit.

The model problem considered a soil layer of $l = 25\ m$ thickness. The depth of inclusion $\xi = 10\ m$, and its thickness $d = 0.2\ m$. The x variable step was $0.04\ m$, the time step $\tau = 10\ day$. Piece-square functions were used as FEM basis. The results of numerical experiments are plotted in Figs. 8.2, 8.3.

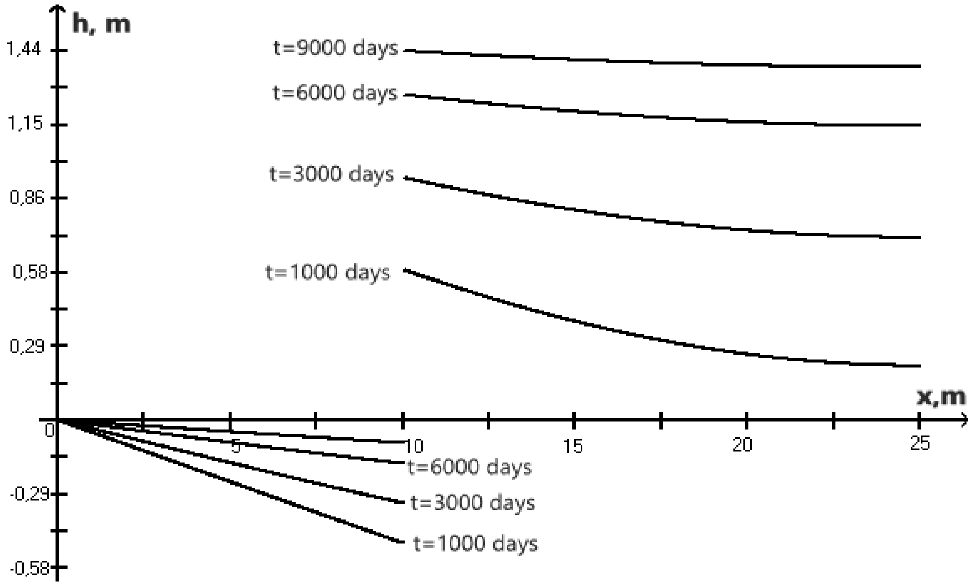


Fig. 8.2. Difference of the distribution of the pressure fields for cases of nonlinear and linear Darcy's laws.

The nonlinearity in Darcy's law has virtually no effect on the distribution of excess pressure during the first 1000 days from the beginning of the study process (Figs. 8.2, 8.3). However, as of the 3000th day, the relative difference in the pressure jumps on the thin inclusion in the linear and nonlinear laws reached 8.4%. Then such relative differences continue to increase and reach 42% on the 9000th day (about 1.5 meters in absolute terms). Thus, the nonlinearity in Darcy's law and the presence of the threshold gradient can introduce significant changes in the distribution of pressures, particularly in the long run. This is important both in terms of natural heterogeneous soils and in terms of hydraulic structures with fine inclusions.

References

1. J. Bear, *Hydraulics of groundwater*, *McGraw-Hill Inc.*, New York, 1979.

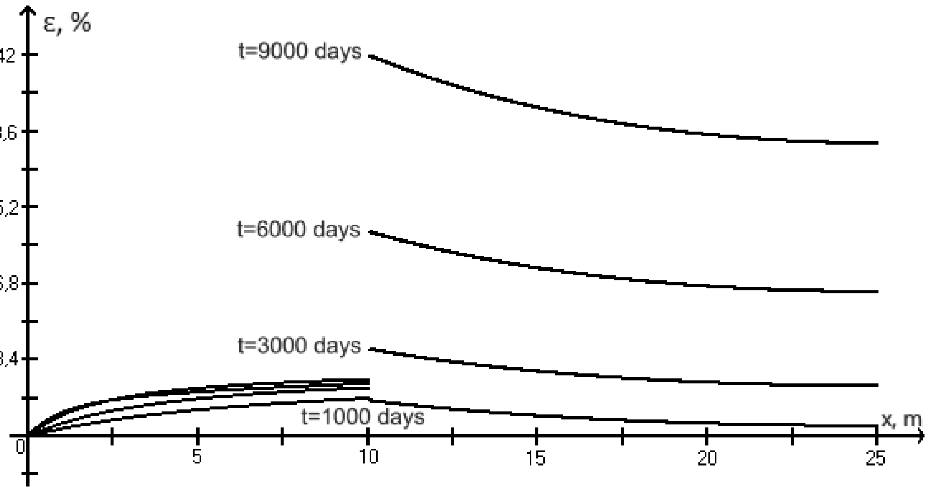


Fig. 8.3. Relative difference of the distribution of the pressure fields for the cases of nonlinear and linear Darcy's laws (in relation to the linear case).

2. V. I. Bilenko, K. V. Bozhonok, S. Yu. Dzyadyk, O. B. Stelya, *Piecewise Polynomial Algorithms for the Analysis of Processes in Inhomogeneous Media*, Cybernetics and Systems Analysis, **54** (4) (2018), 636–642.
3. A. Ya. Bomba, A. P. Safonyk, E. A. Fursachik, *Identification of Mass Transfer Distribution Factor and Its Account for Magnetic Filtration Process Modeling*, Journal of Automation and Information Sciences, **45** (4) (2013), 16–22.
4. A. Ya. Bomba, S. V. Yaroshchak, *Complex approach to modeling of two-phase filtration processes under control conditions*, Journal of Mathematical Sciences, **184** (1) (2012), 56–68.
5. Y. Chui, P. Martyniuk, M. Kuzlo, O. Ulianchuk–Martyniuk, *The conditions of conjugation in the tasks of moisture transfer on a thin clay inclusion taing into account salt solutions and themperature*, Journal of Theoretical and Applied Mechanics(Bulgaria), **49** (1) (2019), 28–38.
6. V. S. Deineka, I. V. Sergienko, V. V. Skopetsky, *Models and methods for solving problems with conjugation conditions*, Naukova dumka (Kiev), 1998.
7. F. M. Francisca, D. A. Glatstein, *Long term hydraulic conductivity of compacted soils permeated with landfill leachate*, Applied Clay Science, **49** (2010), 187–193.
8. V. A. Herus, N. V. Ivanchuk, P. M. Martyniuk, *A System Approach to Mathematical and Computer Modeling of Geomigration Processes Using Freefem++ and Parallelization of Computations*, Cybernetics and Systems Analysis, **54** (2) (2019), 284–294.
9. H.-H. Liu, *Fluid Flow in the Subsurface: History, Generalization and Applications of Physical Laws*,Springer, Switzerland, 2017.
10. L. Chuan-Xun, W. Chang-Jian, L. Meng-Meng, L. Jian-Fei, X. Kang-He, *One-dimensional large-strain consolidation of soft clay with non-Darcian flow and nonlinear compression and non-Darcian flow and nonlinear compression and permeability of soil*, Journal of Central South University, **44** (4) (2017), 967–976.

11. S. I. Lyashko, D. A. Nomirovskii, *The Generalized Solvability and Optimization of Parabolic Systems in Domains with Thin Low-Permeable Inclusions*, Cybernetics and Systems Analysis, **39** (5) (2003), 737–745.
12. P.M. Martyniuk, O.R. Michuta, O.V. Ulianchuk–Martyniuk, M.T. Kuzlo, *Numerical investigation of pressure head jump values on a thin inclusion in one-dimensional non-linear soil moisture transport problem*, Int. J. of Appl. Mathem., **31** (4) (2018), 649–660.
13. O. Michuta, N. Ivanchuk, P. Martyniuk, O. Ostapchuk, *A finite element study of elastic filtration in soils with thin inclusions*, Eastern-European Journal of Enterprise Technologies, **5** (5-107) (2020), 41–48.
14. V. V. Semenov, *Optimization of parabolic systems with conjugation conditions*, *Dop. NANU*, **5** (2003), 66–72.
15. I. V. Sergienko, V. S. Deineka, *Models with conjugation conditions and high-accuracy methods of their discretization*, Cybernetics and Systems Analysis, **36** (1) (2000), 83–101.
16. I. V. Sergienko, V. V. Skopetsky, V. S. Deineka, *Mathematical modeling and research of processes in inhomogeneous environments*, Naukova dumka(Kiev), 1991.
17. I. B. Tymchyshyn, D. A. Nomirovskii, *Generalized Solvability of a Parabolic Model Describing Transfer Processes in Domains with Thin Inclusions, Differential Equations*, **57** (8) (2021), 1053–1062.
18. O. V. Ulianchuk–Martyniuk, *Numerical simulation of the effect of semi-permeable properties of clay on the value of concentration jumps of contaminants in a thin geochemical barrier*, Eurasian Journal of Mathematical and Computer Applications, **8** (1) (2020), 91–104.
19. O. Ulianchuk–Martyniuk, O. Michuta, *Conjugation conditions in the problem of filtering chemical solutions in the case of structural changes to the material and chemical suffusion in the geobarrier*, JP Journal of Heat and Mass Transfer, **19** (1) (2020), 141–154.
20. O. Ulianchuk–Martyniuk, O. Michuta, N. Ivanchuk, *Biocolmatation and the finite element modeling of its influence on changes in the head drop in a geobarrier*, Eastern-European Journal of Enterprise Technologies, **4** (10-106) (2020), 18–26.
21. O. V. Ulianchuk–Martyniuk, O. R. Michuta, N. V. Ivanchuk, *Finite element analysis of the diffusion model of the bioclogging of the geobarrier*, Eurasian Journal Of Mathematical and Computer Applications, **9** (4) (2021), 100–114.
22. A. P. Vlasyuk, P. M. Martynyuk, *Numerical solution of three-dimensional problems of filtration consolidation with regard for the influence of technogenic factors by the method of radial basis functions*, Journal of Mathematical Sciences, **171** (5) (2010), 632–648.
23. A. P. Vlasyuk, P. M. Martynyuk, O. R. Fursovych, *Numerical solution of a one-dimensional problem of filtration consolidation of saline soils in a nonisothermal regime*, Journal of Mathematical Sciences, **160** (4) (2009), 525–535.
24. Xu-dong Zhao, Wen-hui Gong, *Numerical solution of nonlinear large strain consolidation based on non-Darcian flow*, Mathematical Problems in Engineering, **2019)** (2019).
25. Z. Liu, Ya. Xia, M. Shi, J. Zhang, X. Zhu, *Numerical simulation and experiment study on the characteristics of non-Darcian flow and rheological consolidation of saturated clay*, Water, **11** (2019), 1385.

Received 20.04.2022

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **30**, Issue **1**, June **2022**, pp. 88–97, DOI 10.15421/142204

ISSN (print) 2617–0108
ISSN (on-line) 2663–6824

HIGHER-ORDER OPTIMALITY CONDITIONS FOR DEGENERATE UNCONSTRAINED OPTIMIZATION PROBLEMS

Viktor Zadachyn*

Abstract. In this paper necessary and sufficient conditions of a minimum for the unconstrained degenerate optimization problem are presented. These conditions generalize the well-known optimality conditions. The new optimality conditions are presented in terms of polylinear forms and Hesse's pseudoinverse matrix. The results are illustrated by examples.The formulation and appearance of these conditions differ from high-order optimality conditions by other authors. The suggested representation of high-order optimality conditions makes them convenient for the evaluation of the convergence rate for unconstrained optimization methods in the case of a singular minimum point, for example, for the analysis of Newton's and quasi-Newton's methods.

Key words: unconditional optimization, degenerate minimum point, optimality conditions, multilinear form.

2010 Mathematics Subject Classification: 49K10, 65K10, 90C30, 90C46.

Communicated by Prof. O. M. Kiselyova

1. Introduction

Unconstrained optimization is the aim of many papers, and it has a variety of applications (see, for example, [1–5]).Nevertheless, the existing numerical methods for solving the general unconstrained optimization problem up to the second order have a very low convergence rate in the case of degenerate problems [6–17] since for increasing the convergence rate, it is necessary to use derivatives of orders greater than two [6, 7]. At the same time, using derivatives of the 3^{rd} and 4^{th} orders makes a numerical method very time-consuming.

For the analysis of the convergence rate for unconstrained optimization methods in the case of a singular minimum point, it is necessary to have appropriate high-order optimality conditions.

A broad literature review on optimality conditions is presented in [18], therefore we will not go into details in this paper. In addition to the above review, many papers have been devoted to high-order optimality conditions (for example, [19, 20]). For the unconstrained optimization degenerate problem, high-order optimality conditions are formulated in [19], but the form of these conditions is not

*Viktor Zadachyn
Department of Information Systems, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine Email: zadachinvm@gmail.com

convenient for application. In [20] high-order optimality conditions for unconstrained optimization have been also considered, but they are not convenient for usage as well.

This paper aims to represent generalized necessary and sufficient conditions of a minimum for unconstrained optimization degenerate problems, which improve to some extent mixed order necessary and sufficient conditions for a minimum proposed in [19] , but also differ from high-order optimality conditions suggested in [20]. The developed optimality conditions should be more convenient for evaluating the convergence rate of unconstrained optimization methods in the case of a singular minimum point.

2. Higher-order optimality conditions

Remind that the degenerate problem of unconstrained optimization is to find

$$\min f(x), \quad x \in R^n, \tag{2.1}$$

where $f(x)$ is of class C^p $(p \geq 4)$ under the assumption that a point $x^* \in R^n$ of local minimum of $f(x)$ is such that the Hessian matrix $f^{(2)}(x^*)$ is degenerate but is not zero identically.

Introduce the following notations:

$$R_1 = Ker(f^{(2)}(x^*)) = \{x \in R^n | \; f^{(2)}(x^*) \cdot x = 0\}; \quad R_2 = \{y \in R^n \mid y \bot x\}$$

is the orthogonal complement of R_1 (i.e., $R^n = R_1 \oplus R_2$); P is an orthogonal projector onto the subspace R_1; $P^{\perp}$ is an orthogonal projector onto the subspace R_2; $f^{(l)}(x^*)$ is an l-th derivative of $f(x)$ at $f^{(l)}\,(x^*) \cdot \left[u^i, v^{l-i}\right]$ is a multilinear form of l arguments u, $v \in R^n$ (the superscripts i and l-i indicate the multiplicity of occurrences of the corresponding argument). Notice that the value of symmetric multilinear form is invariant concerning various permutations of arguments.

Denote by $R^{(n)^{p/2}}$ the space of $\left(\frac{p}{2}\right)$-dimensional arrays of the dimension $n \times n \times \ldots \times n$. Then $f^{\left(\frac{p}{2}+1\right)}\,(x^*)$ can be considered as a linear mapping $f^{\left(\frac{p}{2}+1\right)}\,(x^*) : R^{(n)^{p/2}} \to R^n$; the mapping $(f^{\left(\frac{p}{2}+1\right)}\,(x^*))^T : R^n \to R^{(n)^{p/2}}$ is a conjugate to $f^{\left(\frac{p}{2}+1\right)}\,(x^*)$ linear mapping. The mapping $f^{(p)}\,(x^*)$ can be considered as a linear mapping $f^{(p)}\,(x^*) : R^{(n)^{p/2}} \to R^{(n)^{p/2}}$, i.e. the value of the multilinear form $f^{(p)} \cdot (x^*)\,[u^p] = U^T f^{(p)}\,(x^*)\,U$, where $U \in R^{(n)^{p/2}}$, is a $\left(\frac{p}{2}\right)$ — dimensional matrix with entries $U_{i,j,\ldots,k} = u_i u_j \cdots u_k$.

Remark also that if $(f^{(2)}\,(x^*))^+$ is a pseudoinverse matrix [21] to $f^{(2)}(x^*)$, then

$$P = I - \; (f^{(2)}\,(x^*))^+ f^{(2)}\,(x^*)\,, \quad P^{\perp} \; = (f^{(2)}\,(x^*))^+ f^{(2)}(x^*). \tag{2.2}$$

Theorem 2.1. *(generalized necessary condition for minimum). Let $f(x)$ be a function such that*

- *f attains at point $x^* \in R^n$ a local minimum;*

- *f is p times ($p \geq 4$, p is even) continuously differentiable in the neighborhood $V(x^*)$ of x^*;*
- *for all $u \in R^n$*
$$f^{(2l)}(x^*)\left[\,(Pu)^{2l}\right] = 0, \tag{2.3}$$
for $l = 1, \dots, \left(\frac{p}{2} - 1\right)$.

Then for all $u \in R^n$

$$f^{(1)}(x^*) = 0, \quad f^{(2)}(x^*)\left[u^2\right] \geq 0; \tag{2.4}$$

$$f^{(2)}(x^*)\left[\,\left(P^{\perp}u\right)^2\right] \geq m_2||P^{\perp}u||^2; \tag{2.5}$$

$$\begin{gathered} f^{(2l+1)}(x^*)\left[\,(Pu)^{2l+1}\right] = 0, \\ \textit{for}\; l = 1, \dots, \left(\tfrac{p}{2} - 1\right),\; f^{(p)}(x^*)\left[(Pu)^p\right] \geq 0; \end{gathered} \tag{2.6}$$

$$f^{(l+1)}(x^*)\left[\,\left(P^{\perp}u\right),\; (Pu)^l\right] = 0, \textit{for}\; l = 1, \dots, \left(\frac{p}{2} - 1\right); \tag{2.7}$$

$$\left(f^{(p)}(x^*) - \frac{p!}{2\left(\left(\frac{p}{2}\right)!\right)^2}(f^{\left(\frac{p}{2}+1\right)}(x^*))^T(f^{(2)}(x^*))^+(f^{\left(\frac{p}{2}+1\right)}(x^*))\right)\left[(Pu)^p\right] \geq 0, \tag{2.8}$$

where $m_2 > 0$.

Proof. The conditions (2.4) are well known. The condition (2.5) means that the matrix $f^{(2)}(x^*)$ is not identically zero. Moreover, $m_2 > 0$ is equal to a minimal nonzero eigenvalue of the matrix $f^{(2)}(x^*)$. The conditions (2.6) and (2.7) follow from Theorem 2.1 [19] that was proved for the case of Hilbert space. Additionally, in Theorem 2.1 [19] it was proved that under condition (2.3) the following inequality

$$\begin{aligned} F_0(x^*, u) \equiv {} & \frac{1}{2}f^{(2)}(x^*)\left[\,\left(P^{\perp}u\right)^2\right] \\ & + \frac{1}{\left(\frac{p}{2}\right)!}f^{\left(\frac{p}{2}+1\right)}(x^*)\left[\left(P^{\perp}u\right), (Pu)^{p/2}\right] + \frac{1}{p!}f^{(p)}(x^*)\left[(Pu)^p\right] \geq 0 \end{aligned} \tag{2.9}$$

holds for all $u \in R^n$.

Taking into account (2.2), we can rewrite (2.9) as follows

$$\begin{aligned} F_0(x^*, u) = {} & \frac{1}{2}f^{(2)}(x^*)\left[\left(P^{\perp}u + \frac{1}{\left(\frac{p}{2}\right)!}(f^{(2)}(x^*))^+(f^{\left(\frac{p}{2}+1\right)}(x^*))\left[(Pu)^{p/2}\right]\right)^2\right] \\ & + \frac{1}{p!}\left(f^{(p)}(x^*) - \frac{p!}{2\left(\left(\frac{p}{2}\right)!\right)^2}(f^{\left(\frac{p}{2}+1\right)}(x^*))^T(f^{(2)}(x^*))^+(f^{\left(\frac{p}{2}+1\right)}(x^*))\right)\left[(Pu)^p\right]. \end{aligned} \tag{2.10}$$

Since $F_0(x^*, u) \geq 0$ for all $u \in R^n$, consider decomposition $u = u_1 + u_2$, where $u_1 = Pu \in R_1$, $u_2 = P^{\perp}u = -\frac{1}{(\frac{p}{2})!}(f^{(2)}(x^*))^+(f^{\left(\frac{p}{2}+1\right)}(x^*))\left[(Pu)^{p/2}\right] \in R_2$. Then (2.8) follows from (2.4), (2.9), (2.10). □

Corollary 2.1. *(generalized necessary conditions for a minimum of* 4^{th} *order). Let* $f(x)$ *be a function such that it attains at point* $x^* \in R^n$ *a local minimum and is four times continuously differentiable in the neighborhood of* x^*.

Then, for all $u \in R^n$

$$f^{(1)}(x^*) = 0,\ f^{(2)}(x^*)\left[u^2\right] = f^{(2)}(x^*)\left[\left(P^{\perp}u\right)^2\right] \geq 0; \tag{2.11}$$

$$f^{(2)}(x^*)\left[\left(P^{\perp}u\right)^2\right] \geq m_2||P\perp u||^2;$$

$$f^{(3)}(x^*)\left[(Pu)^3\right] = 0, f^{(4)}(x^*)\left[(Pu)^4\right] \geq 0; \tag{2.12}$$

$$\left(f^{(4)}(x^*) - 3\ (f^{(3)}(x^*))^T(f^{(2)}(x^*))^+(f^{(3)}(x^*))\right)[(Pu)^4] \geq 0, \tag{2.13}$$

where $m_2 > 0$.

Theorem 2.2. *(generalized sufficient minimum condition). Let* $f(x)$ *be a* p *times* $(p \geq 4,\ p$ *is even) continuously differentiable function in the neighborhood* $V(x^*)$ *of point* x^**at which the conditions (2.3)–(2.7) are satisfied and for all* $u \in R^n$

$$\left(f^{(p)}(x^*) - \frac{p!}{2\left((\frac{p}{2})!\right)^2}(f^{\left(\frac{p}{2}+1\right)}(x^*))^T(f^{(2)}(x^*))^+(f^{\left(\frac{p}{2}+1\right)}(x^*))\right) \times [(Pu)^p] \geq m_p||Pu||^p, \tag{2.14}$$

where $m_p > 0$.

Then x^* *is a point of strict local minimum of the function* $f(x)$ *and for all* x *from the sufficiently small neighborhood* $V(x^*)$ *the following inequality*

$$f(x) - f(x^*) \geq m_0 \cdot (||P\perp v||^2 + ||Pv||^p), \tag{2.15}$$

where $v = x - x^*$ *and* $m_0 > 0$*, holds.*

Proof. Since the function $f(x)$ is p times continuously differentiable in the neighborhood $V(x^*)$, according to the Taylor series expansion we have the following:

$$\begin{aligned} f(x) - f(x^*) &= \sum_{l=1}^{p} \frac{1}{l!} f^{(l)}(x^*)\left[(v)^l\right] + O(||v||^{p+1}) \\ &= f^{(1)}(x^*)[v] + \frac{1}{2} f^{(2)}(x^*)\left[\left(P^{\perp}v\right)^2\right] \\ &\quad + \sum_{l=3}^{p} \frac{1}{l!} \sum_{i=0}^{l} C_l^i f^{(l)}(x^*)\left[\left(P^{\perp}v\right)^{l-i}, (Pv)^i\right] + O(||v||^{p+1}), \end{aligned}$$

for all $x \in V(x^*)$, where $v = x - x^*$. Taking into account the conditions (2.3)–(2.7), in a sufficiently small neighborhood $V(x^*)$ the following equality

$$\begin{aligned} f(x) - f(x^*) = {} & \frac{1}{2} f^{(2)}(x^*) \left[\left(P^{\perp} v\right)^2 \right] \\ & + \frac{1}{\left(\frac{p}{2}\right)!} f^{\left(\frac{p}{2}+1\right)}(x^*) \left[\left(P^{\perp} v\right),\ (Pv)^{\frac{p}{2}} \right] \\ & + \frac{1}{p!} f^{(p)}(x^*) \left[(Pv)^p \right] \\ & + O(||P^{\perp} v||^3) + O(||P^{\perp} v|| \cdot ||Pv||^{\frac{p}{2}+1}) \\ & + O(||P^{\perp} v||^2 \cdot ||Pv||^{\frac{p}{2}}) + O(||v||^{p+1}) \end{aligned}$$

holds. Because of (2.9) and (2.10), we have

$$\begin{aligned} f(x) - f(x^*) = {} & \frac{1}{2} f^{(2)}(x^*) \left[\left(P^{\perp} v + \frac{1}{\left(\frac{p}{2}\right)!} (f^{(2)}(x^*))^+ (f^{\left(\frac{p}{2}+1\right)}(x^*)) \left[(Pv)^{\frac{p}{2}} \right] \right)^2 \right] \\ & + \frac{1}{p!} \left(f^{(p)}(x^*) - \frac{p!}{2\left(\left(\frac{p}{2}\right)!\right)^2} (f^{\left(\frac{p}{2}+1\right)}(x^*))^T (f^{(2)}(x^*))^+ (f^{\left(\frac{p}{2}+1\right)}(x^*)) \right) \left[(Pv)^p \right] \\ & + O(||P^{\perp} v||^3) + O(||P^{\perp} v|| \cdot ||Pv||^{\frac{p}{2}+1}) + O(||P^{\perp} v||^2 \cdot ||Pv||^{\frac{p}{2}}) + O(||v||^{p+1}). \end{aligned}$$

Thus, from (2.5) and (2.14) it follows that

$$\begin{aligned} f(x) - f(x^*) \geq {} & \frac{m_2}{2} \left\| P^{\perp} v + \frac{1}{\left(\frac{p}{2}\right)!} (f^{(2)}(x^*)) + (f^{\left(\frac{p}{2}+1\right)}(x^*)) \left[(Pv)^{\frac{p}{2}} \right] \right\|^2 \\ & + \frac{m_p}{p!} \|Pv\|^p - N_1 \left\| P^{\perp} v \right\|^3 - N_2 \left\| P^{\perp} v \right\| \cdot \|Pv\|^{\frac{p}{2}+1} \\ & - N_3 \left\| P^{\perp} v \right\|^2 \cdot \|Pv\|^{\frac{p}{2}} - N_4 \|v\|^{p+1}, \end{aligned} \tag{2.16}$$

where N_1, N_2, N_3, and N_4 are some positive constants.

Consider $x \in V(x^*)$ such that

$$\left\| P^{\perp} v \right\| \geq \frac{2}{\left(\frac{p}{2}\right)!} \left\| (f^{(2)}(x^*))^+ \right\| \cdot \left\| f^{\left(\frac{p}{2}+1\right)}(x^*) \right\| \cdot \|Pv\|^{\frac{p}{2}}.$$

Then

$$\begin{aligned} & \left\| P^{\perp} v + \frac{1}{\left(\frac{p}{2}\right)!} (f^{(2)}(x^*))^+ (f^{\left(\frac{p}{2}+1\right)}(x^*)) \left[(Pv)^{\frac{p}{2}} \right] \right\| \geq \left\| P^{\perp} v \right\| \\ & - \left\| \frac{1}{\left(\frac{p}{2}\right)!} (f^{(2)}(x^*))^+ (f^{\left(\frac{p}{2}+1\right)}(x^*)) \left[(Pv)^{\frac{p}{2}} \right] \right\| \geq \frac{1}{2} \left\| P^{\perp} v \right\|, \end{aligned}$$

and hence (2.16) implies

$$f(x) - f(x^*) \geq \frac{1}{2}\left(\frac{m_2}{8}\left\|P^{\perp}v\right\|^2 + \frac{m_p}{p!}\left\|Pv\right\|^p\right)$$
$$\geq min(\frac{m_2}{16}, \frac{m_p}{2\cdot p!})\cdot\left(\left\|P^{\perp}v\right\|^2 + \left\|Pv\right\|^p\right) \tag{2.17}$$

in a sufficiently small neighborhood $V(x^*)$.

Let $x \in V(x^*)$ is such that

$$\left\|P^{\perp}v\right\| < \frac{2}{\left(\frac{p}{2}\right)!}\left\|(f^{(2)}(x^*))^+\right\|\cdot\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|\cdot\left\|Pv\right\|^{\frac{p}{2}}.$$

Then

$$\left\|Pv\right\|^{\frac{p}{2}} > \frac{\left(\frac{p}{2}\right)!}{2}\left\|(f^{(2)}(x^*))^+\right\|^{-1}\cdot\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|^{-1}\cdot\left\|P^{\perp}v\right\|.$$

The inequality (2.5) implies that $\left\|(f^{(2)}(x^*))^+\right\| \leq \frac{1}{m_2}$ and, hence,

$$\left\|Pv\right\|^{\frac{p}{2}} > \frac{\left(\frac{p}{2}\right)!}{2}\left\|(f^{(2)}(x^*))^+\right\|^{-1}\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|^{-1}\left\|P^{\perp}v\right\|$$
$$\geq \frac{\left(\frac{p}{2}\right)!}{2}m_2\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|^{-1}\left\|P^{\perp}v\right\|.$$

Then from (2.16) we obtain

$$f(x) - f(x^*) \geq \frac{1}{2}\frac{m_p}{p!}\left\|Pv\right\|^p$$
$$\geq \frac{1}{4}\frac{m_p}{p!}\left\|Pv\right\|^p + \frac{1}{4}\frac{m_p}{p!}\left(\frac{\left(\frac{p}{2}\right)!}{2}m_2\right)^2\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|^{-2}\left\|P^{\perp}v\right\|^2$$
$$\geq min\left(\frac{m_p}{4\cdot p!}, \frac{1}{4}\frac{m_p}{p!}\left(\frac{\left(\frac{p}{2}\right)!}{2}m_2\right)^2\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\|^{-2}\right)$$
$$\times\left(\left\|P^{\perp}v\right\|^2 + \left\|Pv\right\|^p\right) \tag{2.18}$$

in a sufficiently small neighborhood $V(x^*)$.

It is worth to emphasize that in our calculations given above (see, for instance, (2.18)), we implicitly assumed that $\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\| > 0$. In the case when

$$\left\|f^{\left(\frac{p}{2}+1\right)}(x^*)\right\| = 0$$

(i.e., $f^{\left(\frac{p}{2}+1\right)}(x^*) \equiv 0$), from (2.16) we conclude that

$$\begin{aligned} f\left(x\right)-f\left(x^{*}\right) &\geq \frac{1}{2}\left(\frac{m_2}{8}\left\|P^{\perp}v\right\|^{2}+\frac{m_p}{p!}\left\|Pv\right\|^{p}\right) \\ &\geq min(\frac{m_2}{4},\frac{m_p}{2\cdot p!})\cdot\left(\left\|P^{\perp}v\right\|^{2}+\left\|Pv\right\|^{p}\right) \qquad (2.19)\end{aligned}$$

in a sufficiently small neighborhood $V(x^*)$,

Thus, according to (2.17) – (2.19) , for all $x \in V(x^*)$ different from x^*, there is a positive constant m_0 such that inequality (2.15) is fulfilled, i.e. x^* is a point of a strict local minimum of the function $f(x)$. □

Corollary 2.2. *(generalized sufficient condition for a minimum of the 4^{th} order). Let $f(x)$ be a four times continuously differentiable function in some neighborhood $V(x^*)$ of the point x^*, at which conditions (2.11) and (2.12) are satisfied, and for all $u \in R^n$*

$$\left(f^{(4)}\left(x^{*}\right)-3\ (f^{(3)}\left(x^{*}\right))^{T}(f^{(2)}\left(x^{*}\right))^{+}(f^{(3)}\left(x^{*}\right))\right)[(Pu)^4]\geq m_4\left\|Pu\right\|^{4}, \quad (2.20)$$

where $m_4 > 0$.

Then x^* is a point of the strict local minimum of the function $f(x)$ and for all x from a sufficiently small neighborhood $V(x^*)$ the following inequality

$$f\left(x\right)-f\left(x^{*}\right)\geq m_0\cdot\left(\left\|P^{\perp}v\right\|^{2}+\left\|Pv\right\|^{4}\right), \qquad (2.21)$$

where $v = x - x^*$ and $m_0 > 0$, holds.

The given above generalized necessary and sufficient minimum conditions provide constructive optimality criteria for degenerate problem (2.1). We illustrate the naturalness of conditions (2.13) and (2.20) with the following examples.

Example 2.1. Consider the function $f\left(x\right)=\left(x_1+x_2^2\right)^2$, $x \in R^2$. This function attains its minimum at the points of a set $X=\left\{x\in R^2 |\ x_1=-x_2^2\right\}$.

Consider the point $x^*=(0,\ 0)^T \in X$. Then,

$$f^{(2)}\left(x^{*}\right)=\begin{pmatrix}2&0\\0&0\end{pmatrix},\quad \left(f^{(2)}\left(x^{*}\right)\right)^{+}=\begin{pmatrix}\frac{1}{2}&0\\0&0\end{pmatrix},$$

$$f^{(3)}\left(x^{*}\right)=\left(A\mid B\right),\quad f^{(4)}\left(x^{*}\right)=\begin{pmatrix}(C\mid C)\\(C\mid D)\end{pmatrix},$$

where $A=\begin{pmatrix}0&0\\0&4\end{pmatrix}, B=\begin{pmatrix}0&4\\4&4\end{pmatrix}, C=\begin{pmatrix}0&0\\0&0\end{pmatrix}, D=\begin{pmatrix}0&0\\0&24\end{pmatrix}$. The orthogonal projector $P=\begin{pmatrix}0&0\\0&1\end{pmatrix}$ and the orthogonal projector $P^{\perp}=\begin{pmatrix}1&0\\0&0\end{pmatrix}$.

The point x^* is not a point of strict local minimum, although the following condition

$$f^{(2)}(x^*)\left[\left(P^{\perp}u\right)^2\right] = 2\left\|P^{\perp}v\right\|^2, \ f^{(4)}(x^*)\,[(Pu)^4] = 24\cdot\|Pv\|^4, \ \forall\, u\in R^2$$

is satisfied. In addition,

$$\left(f^{(4)}(x^*) - 3\,(f^{(3)}(x^*))^T(f^{(2)}(x^*))^+(f^{(3)}(x^*))\right)[(Pu)^4] = 0, \ \forall\, u\in R^2.$$

Example 2.2. Consider the function $f(x) = x_1^2 + x_1x_2^2 + x_2^4$, $x\in R^2$. This function attains the minimal value at $x^* = (0,0)^T$, which is a point of the strict local minimum.

Then

$$f^{(2)}(x^*) = \begin{pmatrix}2 & 0\\ 0 & 0\end{pmatrix}, \quad \left(f^{(2)}(x^*)\right)^+ = \begin{pmatrix}\frac{1}{2} & 0\\ 0 & 0\end{pmatrix},$$

$$f^{(3)}(x^*) = (A \mid B), \quad f^{(4)}(x^*) = \begin{pmatrix}(C \mid C)\\ (C \mid D)\end{pmatrix},$$

where $A = \begin{pmatrix}0 & 0\\ 0 & 2\end{pmatrix}, B = \begin{pmatrix}0 & 2\\ 2 & 0\end{pmatrix}, C = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}, D = \begin{pmatrix}0 & 0\\ 0 & 24\end{pmatrix}$. The orthogonal projector $P = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$ and the orthogonal projector $P^{\perp} = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}$.

For x^* the following conditions

$$f^{(2)}(x^*)\left[\left(P^{\perp}u\right)^2\right] = 2\left\|P^{\perp}u\right\|^2, \ f^{(4)}(x^*)\,[(Pu)^4] = 24\cdot\|Pu\|^4, \ \forall\, u\in R^2$$

$$\left(f^{(4)}(x^*) - 3\,(f^{(3)}(x^*))^T(f^{(2)}(x^*))^+(f^{(3)}(x^*))\right)[(Pu)^4]] = 18\cdot\|Pu\|^4, \ \forall\, u\in R^2$$

are satisfied.

Therefore, the condition (2.20) provides strictness of the minimum, while the condition $f^{(4)}(x^*)\,[(Pu)^4] \geq m\ \|Pu\|^4, \forall\, u\in R^2, m>0$, is not sufficient for this.

3. Conclusion

The suggested necessary and sufficient conditions of a minimum for unconstrained optimization degenerate problems generalize the known optimality conditions. The formulation and appearance of these conditions differ from the high-order optimality conditions proposed by other authors. Owing to the results obtained, the suggested optimality conditions can be used for the analysis of the convergence rate of unconstrained optimization methods in the case of a singular minimum point, for example, Newton's method and quasi-Newton's methods. These issues will be considered in future papers.

4. Acknowledgement

The author would like to thank Alexander L. Yampolsky for his valuable comments.

References

1. W. Ring, B. Wirth, *Optimization methods on Riemannian manifolds and their application to shape space*, SIAM Journal on Optimization, **22** (2) (2012), 596–627.
2. N.G. Maratos, M.A. Moraitis, *Some results on the Sign recurrent neural network for unconstrained minimization*, Neurocomputing, **287** (2018), 1–25.
3. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.
4. A.R. Sankar, V.N. Balasubramanian, *Are saddles good enough for deep learning?*, arXiv preprint arXiv:1706.02052 (2017).
5. D. Mehta, T. Chen, T. Tang, J.D. Hauenstein, *The Loss Surface Of Deep Linear Networks Viewed Through The Algebraic Geometry Lens*, arXiv preprint arXiv: 1810.07716 (2018).
6. K.N. Belash, A.A. Tret'yakov, *Methods for solving degenerate problems*, USSR Comput. Math. And Math. Phys, **28** (4) (1988), 90–94.
7. E. Szczepanik, A. Prusinska, A. Tret'yakov, *The p-Factor Method for Nonlinear Optimization*, Schedae Informaticae, **21** (2012), 141–157.
8. M. Avriel, *Nonlinear Programming: Analysis and Methods*, Dover Publishing, New York, 2003.
9. D.H. Li, M. Fukushima, L. Qi, N. Yamashita, *Regularized newton methods for convex minimization problems with singular solutions*, Comput. Optim. Appl., **28** (2004), 131–147.
10. C. Shen, X. Chen, Y. Liang, *A regularized Newton method for degenerate unconstrained optimization problems*, Optimization Letters, **6** (2012), 1913–1933.
11. Q. Li, *A Modified Fletcher-Reeves-Type Method for Nonsmooth Convex Minimization*, Stat., Optim. & Inf. Comput., **2** (3) (2014), 200–210.
12. T.N. Graspa, *A modified Newton direction for unconstrained optimization*, Optimization, **63** (7) (2014), 983–1004.
13. X. Han, J. Zhang, J. Chen, *A new hybrid conjugate gradient algorithm for unconstrained optimization*, Bulletin of the Iranian Mathematical Society, **43** (6) (2017), 2067–2084.
14. X. Li, B. Wang, W. Hu, *A modified nonmonotone BFGS algorithm for unconstrained optimization*, Journal of Inequalities and Applications, **183** (2017), 1–18.
15. K. Ghazali, J. Sulaiman, Y. Dasril, D. Gabda, *Newton-SOR Iteration for Solving Large-Scale Unconstrained Optimization Problems with an Arrowhead Hessian Matrices*, Journal of Physics: Conference Series, **1358:1** (2019), 1–10.
16. S. Taheri, M. Mammadov, S. Seifollahi, *Globally convergent algorithms for solving unconstrained optimization problems*, Optimization, **64** (2) (2015), 249–263.
17. W. Quapp, *Searching Minima of an N-Dimensional Surface: A Robust Valley Following Method*, Computers and Mathematics with Applications, **41** (2001), 407–414.
18. Jean-Paul Penot, *Higher-order optimality conditions and higher-order tangents sets*, SIAM Journal on Optimization, **27** (4) (2017), 2508–2527.

19. V. M. Zadachin, *Necessary and sufficient conditions for a minimum of mixed order*, Ukrainian Mathematical Journal, **41**(1989), 367-370. DOI: 10.1007/BF01060332.

20. B. Jiménez, V. Novo, *Higher-order optimality conditions for strict local minima. Annals of Operations Research*, **157** (2008), 183–192.

21. G. H. Golub, C.F. Van Loan, *Matrix Computations.* 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

Received 28.01.2022

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **30**, Issue **1**, June **2022**, pp. 98–121, DOI 10.15421/142205

ISSN (print) 2617–0108
ISSN (on-line) 2663–6824

INITIAL-BOUNDARY VALUE PROBLEMS FOR ANISOTROPIC PARABOLIC EQUATIONS WITH VARIABLE EXPONENTS OF THE NONLINEARITY IN UNBOUNDED DOMAINS WITH CONDITIONS AT INFINITY

Mykola Bokalo*

Abstract. We deal with the initial-boundary value problems with some restrictions at infinity for linear and nonlinear anisotropic parabolic second-order equations in unbounded domains with respect to the spatial variables. The weak solutions of our problem in Lebesgue and Sobolev spaces with variable exponents is considered. We prove theorems on the existence and uniqueness of the weak solutions using the method based on Saint-Venant principle, and the monotonicity method. Moreover, we obtain estimate of the weak solutions.

Key words: parabolic equation, variable exponent of nonlinearity, Lebesgue space with variable exponent, Sobolev space with variable exponent, unbounded domain, Saint-Venant principle, monotonicity method.

2010 Mathematics Subject Classification: 35D30, 35K10, 35K20, 35K55.

Communicated by Prof. O. Kupenko

1. Introduction

Let n be a natural number, and $\mathbb{R}^n$ be the linear space of ordered collections $x=(x_1,...,x_n)$ of real numbers with a norm $|x|:=(|x_1|^2+...+|x_n|^2)^{1/2}$. Suppose that Ω is an unbounded domain in $\mathbb{R}^n$, and $\partial\Omega$ (boundary of the domain Ω) is a piecewise-smooth surface. Let $\nu=(\nu_1,...,\nu_n)$ be a outward-pointing normal unit vector on $\partial\Omega$. Suppose $\partial\Omega=\Gamma_0\cup\Gamma_1$, where Γ_0 is a closure of an open set on $\partial\Omega$ (in particular, $\Gamma_0=\emptyset$ or $\Gamma_0=\partial\Omega$), $\Gamma_1:=\partial\Omega\setminus\Gamma_0$. Put $Q:=\Omega\times(0,T)$, $\Sigma_0:=\Gamma_0\times(0,T)$, $\Sigma_1:=\Gamma_1\times(0,T)$, where $T>0$. Denote by $Bd(\Omega)$ the set of all bounded subdomains of Ω.

We consider the problem: *to find the function* $u:\overline{Q}\to\mathbb{R}$ *that satisfies (in some sense) the parabolic equation*

$$u_t-\sum_{i=1}^{n}\frac{d}{dx_i}a_i(x,t,u,\nabla u)+a_0(x,t,u,\nabla u)=f(x,t),\quad (x,t)\in Q\,, \tag{1.1}$$

[*] Department of Mathematical Statistics and Differential Equations, Ivan Franko National University of Lviv, 1, Universytetska St., Lviv, 79000, Ukraine, mm.bokalo@gmail.com

the boundary conditions

$$u\Big|_{\Sigma_0} = 0, \qquad \frac{\partial u}{\partial \nu_a}\Big|_{\Sigma_1} = 0, \tag{1.2}$$

and the initial condition

$$u(x,0) = u_0(x)\,, \quad x \in \Omega, \tag{1.3}$$

where $a_i : Q \times \mathbb{R}^{1+n} \to \mathbb{R}$, $i = \overline{0,n}$, $f : Q \to \mathbb{R}$, $u_0 : \Omega \to \mathbb{R}$ *are given real-valued functions,* $\dfrac{\partial u(x,t)}{\partial \nu_a} := \sum_{i=1}^{n} a_i(x,t,u,\nabla u)\,\nu_i(x)$ *is an exterior conormal derivative of* u *in point* $(x,t) \in \Sigma_1$.

Remark 1.1. An simpler example of the equations of type (1.1) considered here is

$$u_t - \sum_{i,j=1}^{n} (\widehat{a}_{ij}(x,t)u_{x_j})_{x_i} + \sum_{j=1}^{n} \widehat{a}_j(x,t)u_{x_j} + \widehat{a}_0(x,t)u = f(x,t), \quad (x,t) \in Q, \tag{1.4}$$

where $\widehat{a}_{ij} = \widehat{a}_{ji} \in L_\infty(Q)$, $i,j = \overline{1,n}$, are functions such that for a.e. $(x,t) \in Q$ we have

$$\sum_{i,j=1}^{n} \widehat{a}_{ij}(x,t)\eta_i\eta_j \geqslant \omega \sum_{l=1}^{n} |\eta_l|^2, \quad (\eta_1,...,\eta_n) \in \mathbb{R}^n, \qquad \omega = \text{const} > 0,$$

and $\widehat{a}_j \in L_\infty(Q)$, $j = \overline{0,n}$, $f : Q \to \mathbb{R}$ is such that $f \in L_2(\Omega' \times (0,T))$ for all $\Omega' \in Bd(\Omega)$.

In Remark 3.4, we have given additional conditions for the coefficients of equation (1.4), which together with those indicated here guarantee the existence and uniqueness of a weak solution of problem (1.4), (1.2), (1.3) in some class of functions, which have corresponding behavior at infinity. □

Remark 1.2. An more complex example of the equations of type (1.1) considered here is

$$u_t - \sum_{i,j=1}^{k} (\widehat{a}_{ij}(x,t)u_{x_j})_{x_i} - \sum_{i=k+1}^{n} (\widehat{a}_i(x,t)|u_{x_i}|^{p_i(x)-2}u_{x_i})_{x_i} + \widehat{a}_0(x,t)u = f(x,t), \tag{1.5}$$

$(x,t) \in Q$, where $k \in \{1,...,n-1\}$ and Ω such that $\Omega \cap \{x = (x_1,...,x_k,x_{k+1},...,x_n) \in \mathbb{R}^n \,\big|\, |x_1|^2 + ... + |x_k|^2 < \tau^2\}$ is bounded for each $\tau > 0$, for example, $\Omega = \Omega_1 \times \Omega_2$, Ω_1 is an unbounded domain in space $\{(x_1,...,x_k) \,\big|\, x_1,...,x_k \in \mathbb{R}\}$, and Ω_2 is a bounded domain in space $\{(x_{k+1},...,x_n) \,\big|\, x_{k+1},...,x_n \in \mathbb{R}\}$. Also we suppose that 1) $\widehat{a}_{ij} = \widehat{a}_{ji} \in L_\infty(Q)$, $i,j = \overline{1,k}$, are functions such that for a.e. $(x,t) \in Q$ a quadratic form $\sum_{i,j=1}^{k} \widehat{a}_{ij}(x,t)\eta_i\eta_j$, $(\eta_1,...,\eta_k) \in \mathbb{R}^k$, is positive, 2) for every $i \in$

$\{0, k+1, ..., n\}$ a function $\widehat{a}_i : Q \to \mathbb{R}$ is measurable, and $0 < \operatorname{ess\,inf}_{\Omega' \times (0,T)} \widehat{a}_i \leqslant \operatorname{ess\,sup}_{\Omega' \times (0,T)} \widehat{a}_i < +\infty$ for all $\Omega' \in Bd(\Omega)$, 3) for every $i \in \{k+1, ..., n\}$ a function $p_i : \Omega \to \mathbb{R}$ is measurable, and $1 < \operatorname{ess\,inf}_{\Omega'} p_i \leqslant \operatorname{ess\,sup}_{\Omega'} p_i < +\infty$ for all $\Omega' \in Bd(\Omega)$ (the functions $p_i, i = \overline{k+1, n}$, are called *exponents of the nonlinearity*).

In remark 3.5, we have given additional conditions for the coefficients of equation (1.5), which together with those indicated here guarantee the existence and uniqueness of a weak solution of problem (1.5), (1.2), (1.3) in some class of functions, which have corresponding behavior at infinity. □

Initial-boundary value problems for parabolic equations in unbounded domains with respect to the spatial variables were studied by many authors. As is well known, to guarantee the uniqueness of the solution of the initial-boundary value problems for linear parabolic equations in unbounded domains (in particular, these problems can be described by (1.4), (1.2), (1.3)) we need some restrictions on solution's behavior as $|x| \to +\infty$ (for example, solution's growth restriction as $|x| \to +\infty$, or belonging of solution to some functional spaces). Since the uniqueness of solution is the determining condition to the well-posedness of problems for evolutionary equations, then it is naturally to formulate the initial-boundary value problem for equation (1.1) in the following form: to find the solution of this equation that satisfies conditions (1.2), (1.3), and some restrictions on its behavior as $|x| \to +\infty$. Firstly this was obtained in [1]. There it was shown that the classical solution of the Cauchy problem for heat equation

$$u_t - \Delta u = 0, \quad (x,t) \in \mathbb{R}^n \times (0,T], \qquad u|_{t=0} = u_0(x), \ x \in \mathbb{R}^n, \tag{1.6}$$

is a unique in the class of the functions such that

$$|u(x,t)| \leqslant A\, e^{a|x|^2} \quad \text{for all} \quad (x,t) \in \mathbb{R}^n \times [0,T], \tag{1.7}$$

where constants a, A are depending on u, while restriction (1.7) is an essential condition for the uniqueness of the solution of the problem. Or rather, in [1], [2] was proved that problem (1.6) with $u_0 \equiv 0$ has a nontrivial solution with growth $Ae^{a|x|^{2+\varepsilon}}$ as $|x| \to +\infty$ for $\varepsilon > 0$. Remark that restriction (1.7) can be interpreted as an analog of the boundary condition at infinity. Similar results for weak solutions of linear parabolic equations from a wide class were obtained in [3], and to substantiate these results used an analogue of the principle of Saint-Venant known in mechanics. The similar situation is with nonlinear parabolic equations from certain classes (see [4–9], etc).

Note that we need some restrictions on the data-in behavior as $|x| \to +\infty$ to solvability of the initial-boundary value problems for parabolic equations considered above. In particular, in the paper [1] it was shown that a classical solution of a problem (1.6), (1.7) exists if u_0 satisfies the condition: $|u_0(x)| \leqslant B\, e^{b|x|^2}$ for all $x \in \mathbb{R}^n$, where b, B are any constants.

However, there are nonlinear parabolic equations for which the corresponding initial-boundary value problems are unique solvable without any conditions at infinity. First result was proved in [10] for equation (1.5) with $p_0 = \text{const} > 2$, and $p_{k+1} = \ldots = p_n = 2$. Similar results were obtained for nonlinear parabolic equations in [10–20], etc.

Nonlinear differential equations with variable exponents of the nonlinearity (for example, equation (1.5)) appear as mathematical models in various physical processes. In particular, these equations describe electroreological substance flows, image recovering processes, electric current in the conductor with changing temperature field (see [21]). Nonlinear differential equations with variable exponents of the nonlinearity were intensively studied in [22–29], etc. The corresponding generalizations of Lebesgue and Sobolev spaces (see [30]) were used in these investigations.

In this work we consider a class of second order parabolic equations in unbounded domains with respect to the spatial variables, which require for the correct formulation of the initial-boundary value problems of setting conditions for the behavior of the solution at infinity. This class contains both linear (see, for example, (1.4)) and nonlinear equations with variable exponents of the nonlinearity (see, for example, (1.5))). Here we complement and generalize results for linear (see, for example, [3]), and nonlinear parabolic equations with constant exponents of the nonlinearity (see, for example, [6]). As we know from the available sources, nonlinear parabolic equations with variable exponents of the nonlinearity were not previously investigated in the context of the problem under consideration. In our researches, we use an analog of the well-known in mechanics Saint-Venant principle. It was developed in [3, 6, 31, 32], and others. Moreover, to prove the solvability of our problem we use the method of exhaustion for unbounded domains, and the monotonicity method [33].

The article is organized as follows. In Section 2, we describe functional spaces which are used in the sequel. In Section 3, we set the researched problem and formulate the main results. Section 4 contains auxiliary statements that are used in the next section. Finally, Section 5 is devoted to substantiation of the main results.

2. Main notation

Firstly, we introduce some functional spaces. Let $r : \Omega \to \mathbb{R}$ be a measurable function, $r(x) \geqslant 1$ for almost every (a.e.) $x \in \Omega$, and $\operatorname{ess\,sup}_{x\in\Omega'} r(x) < \infty$ for any $\Omega' \in Bd(\Omega)$. For any $\Omega' \in Bd(\Omega)$ we denote by $L_{r(\cdot)}(\Omega')$ the linear space of (classes of) measurable functions $v : \Omega' \to \mathbb{R}$ such that $\rho_{\Omega',r}(v) := \int_{\Omega'} |v(x)|^{r(x)}\,dx < \infty$. This is the Banach space with a norm

$$\|v\|_{L_{r(\cdot)}(\Omega')} := \inf\{\lambda > 0 \ \mid \ \rho_{\Omega',r}(v/\lambda) \leqslant 1\}.$$

Space $L_{r(\cdot)}(\Omega')$ is called *the Lebesgue space with variable exponent* or *generalized Lebesgue space* (see, for example, [30]). If $r(x) > 1$ for a.e. $x \in \Omega$, put by definition $r'(x) := r(x)/(r(x)-1)$ for a.e. $x \in \Omega$. As is well known, the dual space $(L_{r(\cdot)}(\Omega'))'$ can be identified with $L_{r'(\cdot)}(\Omega')$ under the condition $\operatorname{ess\,inf}_{x\in\Omega'} r(x) > 1$. Note also that in the case $r(x) = r = \text{const} \geqslant 1$ for a.e. $x \in \Omega' \in Bd(\Omega)$ we have $L_{r(\cdot)}(\Omega') = L_r(\Omega')$, and $\|\cdot\|_{L_{r(\cdot)}(\Omega')} = \|\cdot\|_{L_r(\Omega')}$.

Denote by $L_{r(\cdot),\mathrm{loc}}(\overline{\Omega})$ the linear space of (classes of) measurable functions $v : \Omega \to \mathbb{R}$ such that their restrictions $v|_{\Omega'}$ belong to the space $L_{r(\cdot)}(\Omega')$ for any set $\Omega' \in Bd(\Omega)$. This space with a family of seminorms $\{\|\cdot\|_{L_{r(\cdot)}(\Omega')} \,|\, \Omega' \in Bd(\Omega)\}$ is complete locally convex. Then a sequence $\{v_l\}_{l=1}^{\infty}$ converges to v in $L_{r(\cdot),\mathrm{loc}}(\overline{\Omega})$ *strongly* (correspondly, *weakly*), if for any domain $\Omega' \in Bd(\Omega)$ the sequence $\{v_l|_{\Omega'}\}_{l=1}^{\infty}$ converges to $v|_{\Omega'}$ in $L_{r(\cdot)}(\Omega')$ *strongly* (correspondly, *weakly*). As above, we introduce the space $L_{r(\cdot)}(Q')$, where $Q' = \Omega' \times (0,T)$, $\Omega' \in Bd(\Omega)$, by using the functional $\rho_{Q',r}(w) := \iint_{Q'} |w(x,t)|^{r(x)}\, dxdt$ instead of $\rho_{\Omega',r}(v)$. Then we define a complete locally convex space $L_{r(\cdot),\mathrm{loc}}(\overline{Q})$ along with a family of seminorms $\{\|\cdot\|_{L_{r(\cdot)}(\Omega'\times(0,T))} \mid \Omega' \in Bd(\Omega)\}$.

Let the following condition holds:

$(\mathbf{P})$ $p = (p_0, p_1, \dots, p_n) : \Omega \to \mathbb{R}^{1+n}$ *is a vector-valued function such that for every* $i \in \{0,1,\dots,n\}$ *the function* $p_i : \Omega \to \mathbb{R}$ *is measurable, and for any* $\Omega' \in Bd(\Omega)$ *we have* $1 < \operatorname{ess\,inf}_{\Omega'} p_i \leqslant \operatorname{ess\,sup}_{\Omega'} p_i < +\infty$.

Let $p' = (p_0', p_1', \dots, p_n')$ be the vector-valued function such that $\frac{1}{p_i(x)} + \frac{1}{p_i'(x)} = 1$ for a.e. $x \in \Omega$, $i = \overline{0,n}$. Obviously, the function p' satisfies condition $(\mathbf{P})$ with p_i' instead of p_i, $i = \overline{0,n}$.

For any domain $\Omega' \in Bd(\Omega)$ we define the space

$$W^1_{p(\cdot)}(\Omega') := \{v \in L_{p_0(\cdot)}(\Omega') \mid v_{x_i} \in L_{p_i(\cdot)}(\Omega'),\ i = \overline{1,n}\}.$$

This is the Banach space with the norm

$$\|v\|_{W^1_{p(\cdot)}(\Omega')} := \|v\|_{L_{p_0(\cdot)}(\Omega')} + \sum_{i=1}^{n} \|v_{x_i}\|_{L_{p_i(\cdot)}(\Omega')}.$$

Space $W^1_{p(\cdot)}(\Omega')$ is called the *Sobolev space with variable exponent* or *generalized Sobolev space* (see, for example, [30]). Denote by $W^1_{p(\cdot),\mathrm{loc}}(\overline{\Omega})$ the complete locally convex space of (classes of) functions $v \in L_{p_0(\cdot),\mathrm{loc}}(\overline{\Omega})$ such that $v_{x_i} \in L_{p_i(\cdot),\mathrm{loc}}(\overline{\Omega})$, $i = \overline{1,n}$, along with a family of seminorms $\bigl\{\|v\|_{W^1_{p(\cdot)}(\Omega'))} \mid \Omega' \in Bd(\Omega)\bigr\}$. Let $\widetilde{W}^1_{p(\cdot),\mathrm{loc}}(\overline{\Omega})$ be the closure of the set $\widetilde{C}^1(\overline{\Omega}) := \{v \in C^1(\overline{\Omega}) \mid v|_{\Gamma_0} = 0\}$ in space $W^1_{p(\cdot),\mathrm{loc}}(\overline{\Omega})$. By $\widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega)$ we denote a subspace of $\widetilde{W}^1_{p(\cdot),\mathrm{loc}}(\overline{\Omega})$ consisting of functions with bounded supports.

For the domain $Q' = \Omega' \times (0,T)$, where $\Omega' \in Bd(\Omega)$, we put

$$W^{1,0}_{p(\cdot)}(Q') := \{w \in L_{p_0(\cdot)}(Q') \ \mid\ w_{x_i} \in L_{p_i(\cdot)}(Q'),\ \ i = \overline{1,n}\}.$$

This is the Banach space with the norm

$$\|w\|_{W^{1,0}_{p(\cdot)}(Q')} := \|w\|_{L_{p_0(\cdot)}(Q')} + \sum_{i=1}^{n} \|w_{x_i}\|_{L_{p_i(\cdot)}(Q')}.$$

Denote by $W^{1,0}_{p(\cdot),\,\mathrm{loc}}(\overline{Q})$ the complete locally convex space of (classes of) functions $w \in L_{p_0(\cdot),\,\mathrm{loc}}(\overline{Q})$ such that $w_{x_i} \in L_{p_i(\cdot),\,\mathrm{loc}}(\overline{Q})$, $i = \overline{1,n}$, along with a family of seminorms $\big\{\|w\|_{W^{1,0}_{p(\cdot)}(\Omega'\times(0,T))} \big|\ \Omega' \in Bd(\Omega)\big\}$. By $\widetilde{W}^{1,0}_{p(\cdot),\,\mathrm{loc}}(\overline{Q})$ we denote a subspace of functions $w \in W^{1,0}_{p(\cdot),\,\mathrm{loc}}(\overline{Q})$ such that $w(\cdot,t)$ belongs to $\widetilde{W}^{1}_{p(\cdot),\,\mathrm{loc}}(\overline{\Omega})$ for a.e. $t \in (0,T)$.

By definition, put

$$C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega})) := \{w : [0,T] \to L_{2,\mathrm{loc}}(\overline{\Omega}) \bigm| w \in C([0,T];L_2(\Omega'))\ \forall\,\Omega' \in Bd(\Omega)\}.$$

This space with the family of seminorms

$$\{\|w\|_{C([0,T];L_2(\Omega'))} := \max_{t\in[0,T]} \|w(\cdot,t)\|_{L_2(\Omega')}\ |\Omega' \in Bd(\Omega)\}$$

is complete locally convex.

Denote by

$$\mathbb{U}_{p,\mathrm{loc}}(\overline{Q}) := \widetilde{W}^{1,0}_{p(\cdot),\mathrm{loc}}(\overline{Q}) \cap C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega})).$$

This space is complete locally convex along with a family of seminorms $\{\|w\|_{W^{1,0}_{p(\cdot)}(\Omega'\times(0,T))} + \|w\|_{C([0,T];L_2(\Omega'))} \mid \Omega' \in Bd(\Omega)\}$.

Finally, let $C^1_c(0,T) \subset C^1(0,T)$ be a set of functions with compact supports on $(0,T)$.

3. Statement of the problem and formulation of main results

We will consider weak solutions of the problem (1.1) – (1.3). To define them, we introduce corresponding data-in classes.

Let $p = (p_0, p_1, \dots, p_n)$ be a vector-valued function that satisfies condition (**P**). By $\mathbb{A}_p$ we denote all ordered collections $(a_0, a_1, \dots, a_n)$ of the real functions satisfying the following conditions:

($\mathbf{A}_1$) for every $i \in \{0,1,\dots,n\}$, function $a_i(x,t,\rho,\xi)$, $(x,t,\rho,\xi) \in Q \times \mathbb{R}^{1+n}$, is a Carathéodory, i.e., function $a_i(x,t,\cdot,\cdot) : \mathbb{R}^{1+n} \to \mathbb{R}$ is a continuous for a.e. $(x,t) \in Q$, and function $a_i(\cdot,\cdot,\rho,\xi) : Q \to \mathbb{R}$ is a measurable for every $(\rho,\xi) \in \mathbb{R}^{1+n}$; in addition, $a_i(x,t,0,0) = 0$ for a.e. $(x,t) \in Q, i = \overline{0,n}$;

($\mathbf{A}_2$) for every $i \in \{0,1,\dots,n\}$, for a.e. $(x,t) \in Q$, and for every $(\rho,\xi) \in \mathbb{R}^{1+n}$ the following inequality holds

$$|a_i(x,t,\rho,\xi)| \leqslant h_{i,1}(x,t)\big(|\rho|^{p_0(x)/p_i'(x)} + \sum_{j=1}^{n} |\xi_j|^{p_j(x)/p_i'(x)}\big) + h_{i,2}(x,t),$$

where $h_{i,1} \in L_{\infty,\,\mathrm{loc}}(\overline{Q})$, $h_{i,2} \in L_{p'_i(\cdot),\,\mathrm{loc}}(\overline{Q})$.

Now we give a definition of a weak solution of problem (1.1) – (1.3). We assume that p satisfies condition (**P**), $(a_0, a_1, ..., a_n) \in \mathbb{A}_p$, $f \in L_{2,\mathrm{loc}}(\overline{Q})$, $u_0 \in L_{2,\mathrm{loc}}(\overline{\Omega})$.

Definition 3.1. A weak solution of problem (1.1) — (1.3) is called a function $u \in \mathbb{U}_{p,\mathrm{loc}}(\overline{Q})$ that satisfies (in the sense of space $C([0,T]; L_{2,\mathrm{loc}}(\overline{\Omega}))$) the initial condition

$$u(\cdot, 0) = u_0(\cdot) \quad \text{a.e. on} \quad \Omega, \tag{3.1}$$

and the integral identity

$$\iint_Q \big[-u\psi\varphi' + \sum_{i=1}^n a_i(x,t,u,\nabla u)\psi_{x_i}\varphi + a_0(x,t,u,\nabla u)\psi\varphi\big]\, dxdt$$
$$= \iint_Q f\psi\varphi\, dxdt \quad \forall\, \psi \in \widetilde{W}^1_{p(\cdot),c}(\Omega),\ \forall\, \varphi \in C^1_c(0,T). \tag{3.2}$$

Suppose $0 \in \Omega$. Let $k \in \{1, \ldots, n\}$ be a number such that for any $\tau > 0$ the set $\widetilde{\Omega}_\tau := \Omega \cap \{x \in \mathbb{R}^n \mid |x_1|^2 + \ldots + |x_k|^2 < \tau^2\}$ is bounded. For any $\tau > 0$ we denote by Ω_τ a connected component of the set $\widetilde{\Omega}_\tau$ that contains 0. For any $\tau > 0$ put $Q_\tau := \Omega_\tau \times (0,T)$. Obviously, $\Omega = \bigcup_{\tau>0} \Omega_\tau$, $Q = \bigcup_{\tau>0} Q_\tau$.

The choice of value k depends on the geometry of the domain Ω (up to the numbering of variables $x_1, ..., x_n$). Obviously, in the general case we can take $k = n$, and, in this case, the class of equations considered below will consist of generalizations of equation (1.4), or rather, of almost linear equations. But in the case of $k < n$ the class of equations to which the following results apply is wider than in the case of $k = n$, and the smaller the value of k the wider the class of these equations (to confirm this, see (1.5)).

Let us illustrate possibilities of the value's k considered two examples.

Example 3.1. Assume $\Omega = \Omega_1 \times \Omega_2$, where Ω_1 is an unbounded domain in $\mathbb{R}^l := \{(x_1, \ldots, x_l) \mid x_i \in \mathbb{R},\ i = \overline{1,l}\}$ for some $l \in \{1, \ldots, n-1\}$, Ω_2 is a bounded domain in $\mathbb{R}^{n-l} := \{(x_{l+1}, \ldots, x_n) \mid x_i \in \mathbb{R},\ i = \overline{l+1,n}\}$, and $0 \in \Omega$. Then we can take arbitrary $k \in \{l, \ldots, n\}$. If $k = l$, then $\Omega_\tau = \Omega_{1,\tau} \times \Omega_2$ for any $\tau > 0$, where $\Omega_{1,\tau}$ is a connected component of the set $\Omega_1 \cap \{(x_1, \ldots, x_l) \in \mathbb{R}^l \mid |x_1|^2 + \ldots + |x_l|^2 < \tau^2\}$ such that $0 \in \Omega_{1,R}$. □

Example 3.2. Suppose

$$\Omega := \{(x_1, x_2) \in \mathbb{R}^2 \mid -\infty < x_1 < +\infty, \quad -\phi_1(x_1) < x_2 < \phi_2(x_1)\},$$

where for each $m \in \{1, 2\}$ a function ϕ_m is continuous on $\mathbb{R}$, and $\phi_m(s) > 0$ for all $s \in \mathbb{R}$. Then we can take either $k = 1$ or $k = 2$. In case $k = 1$, we have $\Omega_\tau = \big\{(x_1, x_2) \in \mathbb{R}^2 \,\big|\, |x_1| < \tau, \quad -\phi_1(x_1) < x_2 < \phi_2(x_1)\big\}$ for any $\tau > 0$. If $k = 2$, then

$$\Omega_\tau = \big\{(x_1, x_2) \in \mathbb{R}^2 \,\big|\, |x_1| < \tau,$$
$$-\min\{\phi_1(x_1), \sqrt{\tau^2 - |x_1|^2}\} < x_2 < \min\{\phi_2(x_1), \sqrt{\tau^2 - |x_1|^2}\}$$

for any $\tau > 0$. □

By definition, put

$$\Gamma_{j,\tau} := \Gamma_j \cap \partial\Omega_\tau,\ j = 0,1, \quad \Gamma_{*,\tau} := \Omega \cap \partial\Omega_\tau,$$

$$\Sigma_{j,\tau} := \Gamma_{j,\tau} \times (0,T),\ j = 0,1, \quad \Sigma_{*,\tau} := \Gamma_{*,\tau} \times (0,T), \quad \tau > 0.$$

We will use a notation

$$\nabla_k v := (v_{x_1}, \dots, v_{x_k}), \quad |\nabla_k v| := (|v_{x_1}|^2 + \ldots + |v_{x_k}|^2)^{1/2}.$$

Everywhere further we will consider that is carried out the following condition:

($\mathbf{P}^*$) *$p = (p_0, p_1, \dots, p_n) : \Omega \to \mathbb{R}^{1+n}$ satisfy condition* ($\mathbf{P}$), *and $p_0(x) = p_1(x) = \ldots = p_k(x) = 2$ for a.e. $x \in \Omega$.*

Suppose $\mathbb{A}_p^*$ is a subset of $\mathbb{A}_p$, which every element satisfies conditions ($\mathbf{A}_1$), ($\mathbf{A}_2$), and the following condition:

($\mathbf{A}_3$) for a.e. $(x,t) \in Q$ and for every (ρ_1, ξ^1), $(\rho_2, \xi^2) \in \mathbb{R}^{1+n}$, we have

$$\sum_{i=1}^{k} |a_i(x,t,\rho_1,\xi^1) - a_i(x,t,\rho_2,\xi^2)| \leqslant g_1(x,t)|\xi'^1 - \xi'^2| + g_2(x,t)|\rho_1 - \rho_2|, \tag{3.3}$$

$$\sum_{i=1}^{n} (a_i(x,t,\rho_1,\xi^1) - a_i(x,t,\rho_2,\xi^2))(\xi_i^1 - \xi_i^2) + (a_0(x,t,\rho_1,\xi^1) - a_0(x,t,\rho_2,\xi^2))(\rho_1 - \rho_2) \geqslant q_1(x,t)|\xi'^1 - \xi'^2|^2 + q_2(x,t)|\rho_1 - \rho_2|^2, \tag{3.4}$$

where $\xi'^j := (\xi_1^j, \dots, \xi_k^j)$, $|\xi'^j| := (|\xi_1^j|^2 + \ldots + |\xi_k^j|^2)^{1/2}$, $j \in \{1,2\}$, and g_1, g_2, q_1, $q_2 : \overline{Q} \to \mathbb{R}$ are continuous functions on $\overline{Q}$ that satisfy the following conditions:

- $g_1(x,t) > 0$, $g_2(x,t) \geqslant 0$, $q_1(x,t) > 0$ for all $(x,t) \in \overline{Q}$, $\inf_{\overline{Q}} q_2 > -\infty$;

- there exist a real number μ, and continuous functions d_1, d_2, λ defined on $[1, +\infty)$ such that

$$q_2(x,t) + \mu > 0 \quad \text{for all } (x,t) \in \overline{Q}, \tag{3.5}$$

$$\text{for all } \ \tau \geqslant 1 : \quad d_1(\tau) \geqslant \max_{\overline{\Sigma_{*,\tau}}} \frac{g_1}{\sqrt{q_1}}, \quad d_2(\tau) \geqslant \max_{\overline{\Sigma_{*,\tau}}} g_2, \tag{3.6}$$

$$\text{for all } \ \tau \geqslant 1 : \quad -\mu < \lambda(\tau) \leqslant \inf_{t,v} \frac{\displaystyle\int_{\Gamma_{*,\tau}} \left[q_1|\nabla_k v|^2 + q_2|v|^2\right] d\Gamma}{\displaystyle\int_{\Gamma_{*,\tau}} |v|^2 \, d\Gamma}, \tag{3.7}$$

where the infimum is taken over all numbers $t \in [0,T]$, and all functions v that are continuously differentiable in the neighborhood of $\overline{\Gamma_{*,\tau}}$, and $v = 0$

on $\partial\Gamma_{*,\tau} \cap \Gamma_0$ (in particular, $-\mu < \lambda(\tau) \leqslant \min_{\overline{\Sigma_{*,\tau}}} q_2$), while

$$\int_1^{+\infty} \frac{d\tau}{A_\mu(\tau)} = +\infty, \tag{3.8}$$

where

$$A_\mu(\tau) := \frac{d_1(\tau)}{\sqrt{\lambda(\tau)+\mu}} + \frac{d_2(\tau)}{\lambda(\tau)+\mu}, \quad \tau \geqslant 1. \tag{3.9}$$

Remark 3.1. If $\sup_{\overline{Q}} \dfrac{g_1}{\sqrt{q_1}} < +\infty$, $\sup_{\overline{Q}} g_2 < +\infty$, then functions d_1, d_2, λ can be chosen as constants. Namely, $d_1(\tau) := d_{1,0}$, $d_2(\tau) := d_{2,0}$, $\lambda(\tau) := \lambda_0$ for all $\tau \geqslant 1$, where $d_{1,0}$, $d_{2,0}$, λ_0 are constants such that

$$d_{1,0} \geqslant \sup_{\overline{Q}} \frac{g_1}{\sqrt{q_1}}, \quad d_{2,0} \geqslant \sup_{\overline{Q}} g_2, \quad \lambda_0 \leqslant \inf_{\overline{Q}} q_2.$$

Then we can take μ such that $\lambda_0 > -\mu$, and

$$A_\mu(\tau) = A_{\mu,0} := \frac{d_{1,0}}{\sqrt{\lambda_0+\mu}} + \frac{d_{2,0}}{\lambda_0+\mu} \quad \text{for all} \ \ \tau \geqslant 1.$$

□

Suppose $\mathbb{A}_p^{**}$, in the case of $k < n$, is a subset of $\mathbb{A}_p^*$, which arbitrary element satisfies the following condition:

($\mathbf{A}_4$) for a.e. $(x,t) \in Q$ and for every $(\rho,\xi) \in \mathbb{R}^{1+n}$, we have

$$\sum_{i=0}^n a_i(x,t,\rho,\xi)\xi_i + a_0(x,t,\rho,\xi)\rho \geqslant q_3(x,t) \sum_{i=k+1}^n |\xi_i|^{p_i(x)} - q_4(x,t)|\rho|^2 - h(x,t), \tag{3.10}$$

where $q_3, q_4 \in C(\overline{Q})$, $q_3(x,t) > 0$ for all $(x,t) \in \overline{Q}$, $0 \leqslant \sup_{\overline{Q}} q_4 < +\infty$, $h \in L_{1,\mathrm{loc}}(\overline{Q})$, $h \geqslant 0$ a.e. on Q.

In the case of $k = n$ we will assume that $\mathbb{A}_p^{**} := \mathbb{A}_p^*$.

It is easy to prove that the initial problem

$$\frac{d\tau}{d\alpha} = A_\mu(\tau), \quad \tau(0) = 1 \tag{3.11}$$

has a unique solution $\tau(\alpha)$, $\alpha \in [0,+\infty)$, and this solution is determined by the equality

$$\int_1^{\tau(\alpha)} \frac{ds}{A_\mu(s)} = \alpha, \quad \alpha \geqslant 0. \tag{3.12}$$

From this and (3.8) it follows that

$$\tau(\alpha) \to +\infty \quad \text{as} \quad \alpha \to +\infty. \tag{3.13}$$

Suppose $\tau(\alpha)$, $\alpha \in [0, +\infty)$, is a solution of problem (3.11), and put

$$\Omega^\alpha := \Omega_{\tau(\alpha)}, \quad \Gamma_j^\alpha := \Gamma_{j,\tau(\alpha)},\, j = 0, 1, \quad \Gamma_*^\alpha := \Gamma_{*,\tau(\alpha)},$$
$$Q^\alpha := Q_{\tau(\alpha)}, \quad \Sigma_j^\alpha := \Sigma_{j,\tau(\alpha)},\, j = 0, 1, \quad \Sigma_*^\alpha := \Sigma_{*,\tau(\alpha)}.$$

Note that in view of (3.13) we have $\Omega = \bigcup_{\alpha>0} \Omega^\alpha$, $Q = \bigcup_{\alpha>0} Q^\alpha$.

Let $\{\Lambda_m\}_{m=1}^\infty$ be a sequence of real numbers such that for all $m \in \mathbb{N}$ we have

$$-\mu < \Lambda_m \leqslant \inf_{t,v} \frac{\displaystyle\int_{\Omega^m} \left[q_1 |\nabla_k v|^2 + q_2 |v|^2\right] dx}{\displaystyle\int_{\Omega^m} |v|^2\, dx}\,, \tag{3.14}$$

where the infimum is taken over all numbers $t \in [0, T]$, and functions $v \in C^1(\overline{\Omega^m})$ such that $v = 0$ on $\partial\Omega^m \setminus \Gamma_1^m$ (in particular, $-\mu < \Lambda_m \leqslant \min_{\overline{Q^m}} q_2$).

Denote

$$\begin{gathered} E_{k,\mu}(w) := q_1 |\nabla_k w|^2 + (q_2 + \mu)|w|^2, \\ \langle w \rangle_\alpha := \left(\iint_{Q^\alpha} E_{k,\mu}(w) e^{-2\mu t}\, dx dt \right)^{1/2}, \quad \alpha \geqslant 0. \end{gathered} \tag{3.15}$$

Now we formulate our main results.

Theorem 3.1 (a uniqueness of the solution). *Let p satisfies condition ($\mathbf{P}^*$), $f \in L_{2,\mathrm{loc}}(\overline{Q})$, $u_0 \in L_{2,\mathrm{loc}}(\overline{\Omega})$, $(a_0, a_1, \ldots, a_n) \in \mathbb{A}_p^*$. Then problem (1.1) – (1.3) has at most one weak solution such that*

$$e^{-R/2} \langle u \rangle_R \to 0 \quad \textit{as} \quad R \to +\infty \tag{3.16}$$

(an analog of the boundary condition at infinity), where $\langle\cdot\rangle_R$ defined in (3.15).

Remark 3.2. Assertion (3.16) is equivalent to the condition

$$e^{-\int_1^r \frac{ds}{A_\mu(s)}} \iint_{Q_r} \left[q_1 |\nabla_k u|^2 + (q_2 + \mu)|u|^2\right] dx dt \to 0 \quad \text{as} \quad r \to +\infty. \tag{3.17}$$

It follows from (3.12), if to remark that $Q^R = Q_r$, if $R = \int_1^r \frac{ds}{A_\mu(s)}$. □

Theorem 3.2 (an existence of the solution). *Let p satisfies condition ($\mathbf{P}^*$), $f \in L_{2,\mathrm{loc}}(\overline{Q})$, $u_0 \in L_{2,\mathrm{loc}}(\overline{\Omega})$, $(a_0, a_1, \ldots, a_n) \in \mathbb{A}_p^{**}$. Also suppose for some number $\varkappa \in (0, 1)$ the following inequality holds*

$$(\Lambda_m + \mu)^{-1} \iint_{Q^m} |f|^2 e^{-2\mu t}\, dx dt + \int_{\Omega^m} |u_0|^2\, dx \leqslant C_1\, e^{(1-\varkappa)m} \quad \forall\, m \in \mathbb{N}, \tag{3.18}$$

where $C_1 > 0$ is a constant.

Then there exists a weak solution of problem (1.1) – (1.3) satisfying condition (3.16). Moreover, for this solution the following estimate is fulfilled:

$$\langle u \rangle_m \leqslant C_2\, e^{(1-\varkappa)m/2} \quad \forall\, m \in \mathbb{N}, \tag{3.19}$$

where $C_2 := [(2 + e^{1/2} - e^{-\varkappa/2})/(1 - e^{-\varkappa/2})]\sqrt{C_1}$, $\langle\cdot\rangle_m$ defined in (3.15).

Remark 3.3. Estimate (3.19) is equivalent to the estimate

$$\iint_{Q_r}\big[q_1|\nabla_k u|^2+(q_2+\mu)|u|^2\big]e^{-2\mu t}\,dxdt\ \leqslant\ C_3\,e^{(1-\varkappa)\int_1^r\frac{ds}{A_\mu(s)}}\quad\forall\, r\geqslant 1,\qquad(3.20)$$

where $C_3>0$ is a constant depending only on $\varkappa$ and C_1.

The statement is substantiated in the same way as (3.17). □

Remark 3.4. For equation (1.4) the conditions of Theorems 1 and 2 are satisfied if functions $\widehat{a}_{ij}$, $i,j=\overline{1,n}$, $\widehat{a}_i$, $i=\overline{0,n}$, are as in Remark 1.1, and for a.e. $(x,t)\in Q$ following hold

$$g_1(x,t)\geqslant\sum_{i=1}^n\Big(\sum_{j=1}^n|\widehat{a}_{ij}(x,t)|^2\Big)^{1/2},\quad g_2(x,t)=0,$$

$$q_1(x,t)=\omega/2,\quad q_2(x,t)\leqslant\Big(\widehat{a}_0(x,t)-\frac{1}{2\omega}\sum_{i=1}^n|\widehat{a}_i(x,t)|^2\Big),$$

where g_1, g_2, q_1, q_2 are as in $(\mathbf{A}_3)$ with $\mu=0$, and f,u_0 satisfy (3.18). □

Remark 3.5. For equation (1.5) the conditions of Theorems 1 and 2 are satisfied if functions $\widehat{a}_{ij}$, $i,j=\overline{1,k}$, $\widehat{a}_i$, $i=\overline{k+1,n}$, $\widehat{a}_0$ are as in Remark 1.2, and for a.e. $(x,t)\in Q$ following inequalities hold

$$\sqrt{k}\sum_{i=1}^k\max_{j\in\{1,\dots,k\}}\big|\widehat{a}_{ij}(x,t)\big|\leqslant g_1(x,t),$$

$$\sum_{i,j=1}^k\widehat{a}_{ij}(x,t)\eta_i\eta_j\geqslant q_1(x,t)\sum_{i=1}^k|\eta_i|^2\quad\forall\,(\eta_1,\dots,\eta_k)\in\mathbb{R}^k,$$

$$\widehat{a}_0(x,t)\geqslant q_2(x,t),\qquad\min_{i\in\{k+1,\dots,n\}}\widehat{a}_i(x,t)\geqslant q_3(x,t),$$

where g_1, q_1, q_2, q_3 are as in $(\mathbf{A}_3)$, $(\mathbf{A}_4)$ together with $g_2=0$, $q_4=0$, $\mu=0$, and f,u_0 satisfy (3.18). □

4. Auxiliary statements

Here we give some auxiliary results which will be used in Section 5. We denote

$$a_i(v)(x,t):=a_i(x,t,v(x,t),\nabla v(x,t)),\quad(x,t)\in Q,\quad i=\overline{0,n},\qquad(4.1)$$

$$\partial_0 v=v,\quad\partial_i v=\partial_i v,\quad i=\overline{1,n}.\qquad(4.2)$$

Recall that $\mathrm{Lip}(\overline{\Omega})$ is the linear space of Lipschitz continuous functions on $\overline{\Omega}$.

Lemma 4.1 (Lemma 1, [24])**.** *Suppose p satisfies condition* $(\mathbf{P})$, $R>0$ *is an arbitrary fixed number, and a function* $w\in \widetilde{W}^{1,0}_{p(\cdot),\mathrm{loc}}(\overline{Q})$ *satisfies the integral identity*

$$\begin{gathered}\iint_{Q_R}\left[-w\psi\varphi'+\sum_{i=0}^{n} g_i\partial_i\psi\varphi\right]dxdt=0\\ \forall\,\psi\in\widetilde{W}^{1}_{p(\cdot),\mathrm{c}}(\Omega),\ supp\,\psi\subset\overline{\Omega_R},\ \forall\,\varphi\in C^1_c(0,T),\end{gathered}\tag{4.3}$$

where $g_i\in L_{p_i'(\cdot),\mathrm{loc}}(\overline{Q})$, $i=\overline{0,n}$, *are given functions.*

Then for arbitrary function $\zeta\in \mathrm{Lip}(\overline{\Omega})$, $supp\,\zeta\subset\overline{\Omega_R}$, $\zeta\geqslant 0$ *we have* $\sqrt{\zeta}w\in C([0,T];L_2(\Omega_R))$ *(hence,* $w\in C([0,T];L_2(\Omega_{R'}))$ *for every* $R'\in(0,R)$*). Moreover, for arbitrary functions* $\theta\in C^1([0,T])$, *and for any numbers* $t_1,t_2\in[0,T]$ $(t_1<t_2)$ *the following equality holds*

$$\frac12\big[\theta(t)\int_{\Omega_R}|w(x,t)|^2\zeta(x)\,dx\big]\Big|_{t=t_1}^{t=t_2}-\frac12\int_{t_1}^{t_2}\int_{\Omega_R}|w|^2\zeta\,\theta'\,dxdt \\ +\int_{t_1}^{t_2}\int_{\Omega_R}\big[\sum_{i=0}^{n}g_i\partial_i(w\zeta)\big]\theta\,dxdt=0.\tag{4.4}$$

If, in addition, it is known that $w|_{\Gamma_{*,R}\times(0,T)}=0$, *then* $w\in C([0,T];L_2(\Omega_R))$, *and we can take* $\zeta(x)=1$, $x\in\overline{\Omega}$, *in (4.4).*

Lemma 4.2 (an analog of Saint-Venant principle)**.** *Assume p satisfies condition* $(\mathbf{P}^*)$, $(a_0,a_1,\dots,a_n)\in\mathbb{A}^*_p$, $f\in L_{2,\mathrm{loc}}(\overline{Q})$, $u_0\in L_{2,\mathrm{loc}}(\overline{\Omega})$. *Suppose* $R>0$ *is an arbitrary number, and* $u_1,u_2\in\mathbb{U}_{p,\mathrm{loc}}(\overline{Q})$ *such that for each* $l\in\{1,2\}$ *we have*

$$u_l(\cdot,0)=u_0(\cdot)\quad a.e.\ on\ \ \Omega^R,\tag{4.5}$$

and

$$\iint_{Q^R}\big[-u_l\psi\varphi'+\sum_{i=0}^{n}a_i(u_l)\partial_i\psi\varphi\big]\,dxdt=\iint_{Q^R}f\psi\varphi\,dxdt,\\ \forall\,\psi\in\widetilde{W}^{1}_{p(\cdot),c}(\Omega),\ supp\,\psi\subset\overline{\Omega^R},\ \ \forall\,\varphi\in C^1_c(0,T).\tag{4.6}$$

Then for every R_1,R_2, $0<R_1<R_2\leqslant R$, *the following inequality holds*

$$\langle u_1-u_2\rangle_{R_1}\leqslant e^{(R_1-R_2)/2}\,\langle u_1-u_2\rangle_{R_2}.\tag{4.7}$$

Remark 4.1. The inequality of type (4.7) has been obtained in [3] for weak solutions from $W^{1,1}_{2,\mathrm{loc}}$ to linear parabolic equations, and in [6], [31], [32] and other works for weak solutions from $W^{1,0}_{2,\mathrm{loc}}$ to quasilinear parabolic equations with constant nonlinearty exponents. This inequality is an analog of the well-known in elasticity theory Saint-Venant principle. □

The proof of Lemma 2. For an arbitrary $x \in \mathbb{R}^n$ we set $x = (x', x'')$, where $x' = (x_1, ..., x_k) \in \mathbb{R}^k$, $x'' = (x_{k+1}, ..., x_n) \in \mathbb{R}^{n-k}$. Let $|x'| = (|x_1|^2 + \ldots + |x_k|^2)^{1/2}$. For any $\delta \in (0,1)$, $\tau \in [1, +\infty)$, $x' \in \mathbb{R}^k$ we denote

$$\psi_\delta(x', \tau) := \begin{cases} 1, & \text{if } |x'| \leqslant \tau - \delta, \\ (\tau - |x'|)/\delta, & \text{if } \tau - \delta < |x'| < \tau, \\ 0, & \text{if } |x'| \geqslant \tau. \end{cases}$$

Obviously, for every $i \in \{1, \ldots, k\}$ we have $\partial_i \psi_\delta(x', \tau) := 0$ if $|x'| < \tau - \delta$ or $|x'| > \tau$, and

$$\partial_i \psi_\delta(x', \tau) = -\frac{x_i}{\delta |x'|} \quad \text{if } \ \tau - \delta < |x'| < \tau. \tag{4.8}$$

By definition, put $w := u_1 - u_2$. Let $\delta \in (0,1)$, $\tau \in (1, \tau(R))$ be arbitrary fixed. We subtract the integral identity (4.6) for $l = 2$ from this identity for $l = 1$. Applying Lemma 1 to their difference with $t_1 = 0$, $t_2 = T$, $\theta(t) := e^{-2\mu t}$, $t \in \mathbb{R}$, $\zeta(x) := \psi_\delta(x', \tau)$, $x = (x', x'') \in \overline{\Omega}$, we obtain

$$\frac{1}{2}\Big[e^{-2\mu t} \int_{\Omega_\tau} |w(x,t)|^2 \psi_\delta(x', \tau)\, dx\Big]\Big|_{t=0}^{t=T} + \mu \iint_{Q_\tau} |w|^2 \psi_\delta e^{-2\mu t}\, dxdt$$
$$+ \iint_{Q_\tau} \Big[\sum_{i=0}^{n} (a_i(u_1) - a_i(u_2)) \partial_i w \psi_\delta\Big] e^{-2\mu t}\, dxdt$$
$$= -\iint_{Q_\tau} \Big[\sum_{i=1}^{k} (a_i(u_1) - a_i(u_2)) w \partial_i \psi_\delta\Big] e^{-2\mu t}\, dxdt. \tag{4.9}$$

Let $\nabla_k w := (\partial_1 w, \ldots, \partial_k w)$, $|\nabla_k w| := (\sum_{i=1}^k |\partial_i w|^2)^{1/2}$. In view of (3.3) we have

$$\sum_{i=1}^{k} |a_i(u_1) - a_i(u_2)| \leqslant g_1 |\nabla_k w| + g_2 |w| \quad \text{a.e. on } \ Q. \tag{4.10}$$

From (4.9), taking into account (3.4), (4.5), (4.8), and (4.10), we deduce

$$\iint_{Q_\tau} \big[q_1 |\nabla_k w|^2 + (q_2 + \mu)|w|^2\big] \psi_\delta e^{-2\mu t}\, dxdt$$
$$\leqslant \frac{1}{\delta} \iint_{Q_\tau \setminus Q_{\tau-\delta}} \big[g_1 |\nabla_k w| + g_2 |w|\big] |w| e^{-2\mu t}\, dxdt. \tag{4.11}$$

Note that for an arbitrary function $P \in L_{1,\text{loc}}(\overline{Q})$ we have

$$\iint_{Q_\tau \setminus Q_{\tau-\delta}} P(x,t)\, dxdt = \int_{\tau-\delta}^{\tau} \Big(\iint_{\Sigma_{*,\sigma}} P(x,t)\, d\Gamma\, dt\Big)\, d\sigma, \quad \tau > 0.$$

Using the latter assertion, we pass to the limit in (4.11) as $\delta \to 0+$. So, we get

$$\iint_{Q_\tau} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]e^{-2\mu t}\,dxdt \\ \leqslant \iint_{\Sigma_{*,\tau}} \big[g_1|\nabla_k w| + g_2|w|\big]|w|e^{-2\mu t}\,d\Gamma\,dt \quad \text{for a.e. } \tau \in (0,\tau(R)). \tag{4.12}$$

From Cauchy-Bunyakovsky-Schvartz inequality it follows that for a.e. $\tau \in (0,\tau(R))$

$$\iint_{\Sigma_{*,\tau}} \big[g_1|\nabla_k w| + g_2|w|\big]|w|e^{-2\mu t}\,d\Gamma\,dt \leqslant \Big(\iint_{\Sigma_{*,\tau}} |g_1|^2|\nabla_k w|^2 e^{-2\mu t}\,d\Gamma\,dt\Big)^{1/2} \\ \times \Big(\iint_{\Sigma_{*,\tau}} |w|^2 e^{-2\mu t}\,d\Gamma\,dt\Big)^{1/2} + \iint_{\Sigma_{*,\tau}} g_2|w|^2 e^{-2\mu t}\,d\Gamma\,dt. \tag{4.13}$$

By virtue of (3.6) and (3.7), we obtain for a.e. $\tau \in (0,\tau(R))$ and for a.e. $t \in (0,T)$

$$\int_{\Gamma_{*,\tau}} |g_1|^2|\nabla_k w|^2\,d\Gamma \leqslant \int_{\Gamma_{*,\tau}} [|g_1|^2/q_1]q_1|\nabla_k w|^2\,d\Gamma \\ \leqslant (d_1(\tau))^2 \int_{\Gamma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]\,d\Gamma, \tag{4.14}$$

$$\int_{\Gamma_{*,\tau}} |w|^2\,d\Gamma \leqslant \int_{\Gamma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]\,d\Gamma \\ \Big/\Big[\int_{\Gamma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]\,d\Gamma\Big/\int_{\Gamma_{*,\tau}} |w|^2\,d\Gamma\Big] \\ \leqslant (\lambda(\tau)+\mu)^{-1} \int_{\Gamma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]\,d\Gamma, \tag{4.15}$$

$$\int_{\Gamma_{*,\tau}} g_2|w|^2\,d\Gamma \leqslant d_2(\tau)\int_{\Gamma_{*,\tau}} |w|^2\,d\Gamma \\ \leqslant d_2(\tau)(\lambda(\tau)+\mu)^{-1} \int_{\Gamma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]\,d\Gamma. \tag{4.16}$$

From (4.12), taking into account (4.13) – (4.16), we infer

$$\iint_{Q_\tau} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]e^{-2\mu t}\,dxdt \\ \leqslant \big[d_1(\tau)(\lambda(\tau)+\mu)^{-1/2} + d_2(\tau)(\lambda(\tau)+\mu)^{-1}\big] \\ \times \iint_{\Sigma_{*,\tau}} \big[q_1|\nabla_k w|^2 + (q_2+\mu)|w|^2\big]e^{-2\mu t}\,d\Gamma\,dt. \tag{4.17}$$

In view of (3.9), (3.15), and (4.17) we establish for a.e. $\tau \in (0, \tau(R))$

$$\iint_{Q_\tau} E_{k,\mu}(w) e^{-2\mu t}\, dxdt \leqslant A_\mu(\tau) \iint_{\Sigma_{*,\tau}} E_{k,\mu}(w) e^{-2\mu t}\, d\Gamma\, dt. \tag{4.18}$$

Denote

$$F(\tau) := \iint_{Q_\tau} E_{k,\mu}(w) e^{-2\mu t}\, dxdt \equiv \int_0^\tau \Big(\int_{\Sigma_{*,\sigma}} E_{k,\mu}(w) e^{-2\mu t}\, d\Gamma\, dt \Big)\, d\sigma, \tag{4.19}$$

for all $\tau \in [1, \tau(R)]$. Then for a.e. $\tau \in (1, \tau(R))$

$$\iint_{\Sigma_{*,\tau}} E_{k,\mu}(w) e^{-2\mu t}\, d\Gamma\, dt = \frac{d}{d\tau} \int_0^\tau \Big(\int_{\Sigma_{*,\sigma}} E_{k,\mu}(w) e^{-2\mu t}\, d\Gamma\, dt \Big)\, d\sigma = \frac{dF(\tau)}{d\tau}. \tag{4.20}$$

From (4.18), using (4.19), and (4.20), we obtain

$$F(\tau) \leqslant A_\mu(\tau) \frac{dF(\tau)}{d\tau} \quad \text{for a.e.} \quad \tau \in [1, \tau(R)]. \tag{4.21}$$

Suppose $\tau = \tau(\alpha)$, $\alpha \in [0, +\infty)$, is a solution of problem (3.11), and R_1, R_2 are arbitrary real numbers such that $0 < R_1 < R_2 \leqslant R$. In view of (3.11) and (4.21) we get

$$F(\tau(\alpha)) \leqslant \frac{dF(\tau(\alpha))}{d\tau} \frac{d\tau(\alpha)}{d\alpha}, \quad \alpha \in [R_1, R_2].$$

It follows that

$$0 \leqslant \frac{dF(\tau(\alpha))}{d\alpha} - F(\tau(\alpha)), \quad \alpha \in [R_1, R_2]. \tag{4.22}$$

Multiplying (4.22) by $e^{-\alpha}$, we deduce $0 \leqslant \dfrac{d}{d\alpha}\Big(e^{-\alpha} F(\tau(\alpha)) \Big)$, $\alpha \in [R_1, R_2]$. Integrating the latter inequality in α from R_1 to R_2, we infer

$$F(\tau(R_1)) \leqslant e^{R_1 - R_2} F(\tau(R_2)). \tag{4.23}$$

From (4.23), taking into account $\langle w \rangle_\alpha = \sqrt{F(\tau(\alpha))}$, we imply (4.7). □

5. Proofs of the main results

The proof of Theorem 1. Let us show that problem (1.1) – (1.3) has no more than one weak solution. Assume the opposite. Let u_1 and u_2 be different weak solutions of problem (1.1) – (1.3), which satisfy condition (3.16). It is clear that for arbitrary $R > 0$ a functional $\langle \cdot \rangle_R$ is a seminorm in space $\mathbb{U}_{p,\mathrm{loc}}(\overline{Q})$. From this fact and (3.16) we deduce

$$e^{-R/2} \langle u_1 - u_2 \rangle_R \leqslant e^{-R/2} (\langle u_1 \rangle_R + \langle u_2 \rangle_R) = e^{-R/2} \langle u_1 \rangle_R + e^{-R/2} \langle u_2 \rangle_R = \beta(R),$$

where $\beta(R) \to 0$ as $R \to +\infty$. Using this assertion and Lemma 2 (see (4.7)) for arbitrary R_1, R_2 such that $R_1 < R_2$, we obtain the estimate

$$\langle u_1 - u_2\rangle_{R_1} \leqslant e^{(R_1-R_2)/2}\langle u_1 - u_2\rangle_{R_2} = e^{R_1/2}\beta(R_2). \tag{5.1}$$

We fix R_1, and tend R_2 to $+\infty$. From (5.1) it follows that $\langle u_1 - u_2\rangle_{R_1} = 0$. Thus $u_1 = u_2$ almost everywhere on Q^{R_1}. As R_1 is arbitrary, we get $u_1 = u_2$ almost everywhere on Q. This contradiction proves Theorem 1. □

The proof of Theorem 2. The proof is in four steps.
Step 1 (the solution's approximations). Let $\alpha > 0$ be an arbitrary number. By $\widehat{W}^1_{p(\cdot)}(\Omega^\alpha)$ define the closure of space $\{v \in C^1\big(\overline{\Omega^\alpha}\big) \mid v\big|_{\partial\Omega^\alpha\setminus\Gamma_1^\alpha} = 0\}$ in $W^1_{p(\cdot)}(\Omega^\alpha)$. By $\widehat{W}^{1,0}_{p(\cdot)}(Q^\alpha)$ denote a space of functions $w \in W^{1,0}_{p(\cdot)}(Q^\alpha)$ such that, for a.e. $t \in (0,T)$, $w(\cdot,t)$ belongs to $\widehat{W}^1_{p(\cdot)}(\Omega^\alpha)$. We set $\widehat{\mathbb{U}}_p(Q^\alpha) := \widehat{W}^{1,0}_{p(\cdot)}(Q^\alpha) \cap C([0,T]; L_2(\Omega^\alpha))$.

For every $l \in \mathbb{N}$ we consider the problem: *to find the function $u_l \in \widehat{\mathbb{U}}_p(Q^l)$ that satisfies (in the sense of space $C([0,T]; L_2(\Omega^l))$) the initial condition*

$$u_l(\cdot,0) = u_0(\cdot) \quad \textit{almost everywhere in } \ \Omega^l, \tag{5.2}$$

and the integral identity

$$\iint_{Q^l} \Big\{-u_l\psi\varphi' + \sum_{i=0}^{n} a_i(u_l)\partial_i\psi\varphi\Big\}\,dxdt = \iint_{Q^l} f\psi\varphi\,dxdt,$$
$$\forall\,\psi \in \widetilde{W}^1_{p(\cdot),c}(\Omega), \ \textit{supp}\,\psi \subset \overline{\Omega^l}, \ \forall\,\varphi \in C^1_c(0,T). \tag{5.3}$$

To prove the existence of the function $u_l \in \widehat{\mathbb{U}}_p(Q^l)$ we use Faedo-Galerkin method (see, for example, [22]). In view of $(\mathbf{A}_3)$ it is easy to show that the function u_l is a unique.

For every $l \in \mathbb{N}$ the function u_l is extended by zero to Q, and the extension denote by u_l again. Obviously, that $u_l \in \mathbb{U}_{p,\mathrm{loc}}(\overline{Q})$. Now we show that there exists a subsequence of the sequence $\{u_l\}_{l=1}^{\infty}$ converging to the weak solution of problem (1.1) – (1.3), (3.16) in some sense. We use an approach from [3], [6], and [33].

Step 2 (the convergence of the sequence of solution's approximations). First we estimate $\langle u_l\rangle_l$ for an arbitrary fixed $l \in \mathbb{N}$. From Lemma 1, putting $w = u_l$, $R = l$, $t_1 = 0$, $t_2 = T$, $\theta(t) = e^{-2\mu t}$, $t \in \mathbb{R}$, $\zeta(x) = 1$, $x \in \overline{\Omega}$, and using (5.3) instead of (4.3), we obtain (see (4.1))

$$\frac{1}{2}e^{-2\mu T}\int_{\Omega^l} |u_l(x,T)|^2\,dx + \iint_{Q^l}\Big[\sum_{i=0}^{n} a_i(u_l)\,\partial_i u_l + \mu|u_l|^2\Big]e^{-2\mu t}\,dxdt$$
$$= \iint_{Q^l} f\,u_l\,e^{-2\mu t}\,dxdt + \frac{1}{2}\int_{\Omega^l}|u_0|^2\,dx. \tag{5.4}$$

From this assertion, taking into account $(\mathbf{A}_1)$ (or rather, the condition $a_i(0)=0,\ i=\overline{0,n}$), $(\mathbf{A}_3)$ (see (3.4)), and Cauchy inequality:

$$ab \leqslant \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2, \quad a,b\in\mathbb{R},\ \varepsilon>0, \tag{5.5}$$

we infer

$$\begin{aligned}&\iint_{Q^l}\big[q_1|\nabla_k u_l|^2+(q_2+\mu)|u_l|^2\big]e^{-2\mu t}\,dxdt\\ &\quad\leqslant \frac{\varepsilon_1}{2}\iint_{Q^l}|u_l|^2e^{-2\mu t}\,dxdt+\frac{1}{2\varepsilon_1}\iint_{Q^l}|f|^2e^{-2\mu t}\,dxdt+\frac{1}{2}\int_{\Omega^l}|u_0|^2\,dxdt,\end{aligned} \tag{5.6}$$

where $\varepsilon_1>0$ is an arbitrary constant.

We have

$$\begin{aligned}\iint_{Q^l}|u_l|^2e^{-2\mu t}\,dxdt &= \int_0^T e^{-2\mu t}\Big(\int_{\Omega^l}|u_l|^2\,dx\Big)\,dt\\ &= \int_0^T e^{-2\mu t}\Big(\int_{\Omega^l}\big[q_1|\nabla_k u_l|^2+(q_2+\mu)|u_l|^2\big]\,dx\Big/\Big[\int_{\Omega^l}\big[q_1|\nabla_k u_l|^2\\ &\quad+(q_2+\mu)|u_l|^2\big]\,dx\Big/\int_{\Omega^l}|u_l|^2\,dx\Big]\Big)\,dt\\ &\leqslant \frac{1}{\Lambda_l+\mu}\iint_{Q^l}\big[q_1|\nabla_k u_l|^2+(q_2+\mu)|u_l|^2\big]e^{-2\mu t}\,dxdt,\end{aligned} \tag{5.7}$$

where Λ_l is defined in (3.14).

From (5.6) and (5.7), putting $\varepsilon_1=\Lambda_l+\mu$, we get

$$\iint_{Q^l}E_{k,\mu}(u_l)e^{-2\mu t}\,dxdt \leqslant (\Lambda_l+\mu)^{-1}\iint_{Q^l}|f|^2e^{-2\mu t}\,dxdt+\int_{\Omega^l}|u_0|^2\,dx.$$

The latter inequality and (3.18) imply the estimate

$$\langle u_l\rangle_l \leqslant \sqrt{C_1}\,e^{(1-\varkappa)l/2}, \quad l\in\mathbb{N}. \tag{5.8}$$

Let $m\in\mathbb{N}$ be an arbitrary fixed number, and let $l,r\in\mathbb{N}$ be arbitrary numbers, while $l\geqslant m$. We have

$$\langle\, u_{l+r}-u_l\rangle_m \leqslant \sum_{i=0}^{r-1}\langle\, u_{l+i+1}-u_{l+i}\rangle_m. \tag{5.9}$$

For each $i\in\{0,\dots,r-1\}$ and the functions u_{l+i+1}, u_{l+i}, using Lemma 2 with $R=l+i$, we obtain

$$\begin{aligned}\langle\, u_{l+i+1}-u_{l+i}\rangle_m &\leqslant e^{-1/2}\langle\, u_{l+i+1}-u_{l+i}\rangle_{m+1}\leqslant\dots\\ &\leqslant e^{-(l+i-m)/2}\langle\, u_{l+i+1}-u_{l+i}\rangle_{l+i}.\end{aligned} \tag{5.10}$$

In view of (5.8), we have

$$\begin{aligned}\langle\, u_{l+i+1} - u_{l+i}\rangle_{l+i} &\leqslant \langle\, u_{l+i+1}\rangle_{l+i+1} + \langle\, u_{l+i}\rangle_{l+i} \\ &\leqslant \sqrt{C_1}\big[e^{(1-\varkappa)(l+i+1)/2} + e^{(1-\varkappa)(l+i)/2}\big] \\ &\leqslant \sqrt{C_1}\big[e^{1/2} + 1\big]e^{(1-\varkappa)(l+i)/2} = C_4\, e^{(1-\varkappa)(l+i)/2}, \qquad (5.11)\end{aligned}$$

where $C_4 := \sqrt{C_1}\big(e^{1/2}+1\big)$.

Using (5.9) – (5.11), we find

$$\begin{aligned}\langle\, u_{l+r} - u_l\rangle_m &\leqslant C_4 \sum_{i=0}^{r-1} e^{-(l+i-m)/2}\, e^{(1-\varkappa)(l+i)/2} \\ &\leqslant C_4 e^{(m-\varkappa l)/2}\sum_{i=0}^{\infty}(e^{-\varkappa/2})^i \leqslant C_5 e^{(m-\varkappa l)/2}, \qquad (5.12)\end{aligned}$$

where

$$C_5 := C_4/(1-e^{-\varkappa/2}) = \sqrt{C_1}(e^{1/2}+1)/(1-e^{-\varkappa/2}). \qquad (5.13)$$

From (5.12) it follows that $\langle\, u_{l+r} - u_l\rangle_m \to 0$ as $l \to +\infty$ uniformly by $r \in \mathbb{N}$, that is, $\{\partial_i u_l\}$, $i = \overline{0,k}$, are Cauchy sequences in space $L_2(Q^m)$, where $m \in \mathbb{N}$ is an arbitrary fixed. Hence, there exists a function $u \in L_{2,\,\mathrm{loc}}(\overline{Q})$ such that $\partial_i u \in L_{2,\mathrm{loc}}(\overline{Q})$, $i = \overline{1,k}$, and

$$\partial_i u_l \underset{l\to\infty}{\longrightarrow} \partial_i u \quad \text{strongly in} \quad L_{2,\,\mathrm{loc}}(\overline{Q}), \quad i = \overline{0,k}. \qquad (5.14)$$

Taking into account ($\mathbf{A}_3$) (see (3.3)), from (5.14) we get

$$a_i(u_l) \underset{l\to\infty}{\longrightarrow} a_i(u) \quad \text{strongly in} \quad L_{2,\,\mathrm{loc}}(\overline{Q}), \quad i = \overline{1,k}. \qquad (5.15)$$

Suppose $m \in \mathbb{N}$ is an arbitrary fixed number, and $l, r \in \mathbb{N}$ are arbitrary numbers such that $l \geqslant m$, $r \geqslant m$. Under the condition $\operatorname{supp}\psi \subset \overline{\Omega^m}$, we subtract the integral identity (5.3) for $l = r$ from this identity for l. Applying Lemma 1 to their difference with $t_1 = 0$, $t_2 = s \in (0, T]$, $\theta(t) := e^{-2\mu t}$, $t \in \mathbb{R}$, $\zeta(x) := \psi_{1/2}(x', \tau(m))$, $x = (x', x'') \in \overline{\Omega}$, we obtain

$$\begin{aligned}&\frac{1}{2}\Big[e^{-2\mu t}\int_{\Omega^m} |u_{lr}(x,t)|^2 \psi_{1/2}(x', \tau(m))\,dx\Big]\Big|_{t=0}^{t=s} \\ &\qquad + \int_0^s\!\!\int_{\Omega^m} \Big[\sum_{i=0}^{n}(a_i(u_l) - a_i(u_r))\partial_i u_{lr} + \mu|u_{lr}|^2\Big]\psi_{1/2} e^{-2\mu t}\,dxdt \\ &\quad = -\int_0^s\!\!\int_{\Omega^m} \Big[\sum_{i=1}^{k}(a_i(u_l) - a_i(u_r))u_{lr}\partial_i\psi_{1/2}\Big] e^{-2\mu t}\,dxdt, \qquad (5.16)\end{aligned}$$

where $u_{lr} := u_l - u_r$.

By virtue of $(\mathbf{A}_3)$ and (4.8), (5.2), from (5.16) for all $s\in[0,T]$ we deduce

$$\int_{\Omega^m}|u_{lr}(x,s)|^2\psi_{1/2}(x',\tau(m))\,dx \leqslant 4e^{2|\mu|T}\int_0^s\int_{\Omega^m}\Big[\sum_{i=1}^k|a_i(u_l)-a_i(u_r)||u_{lr}|\Big]\,dxdt. \quad (5.17)$$

From (5.17), in view of Cauchy-Bunyakovsky-Schvartz inequality, it implies that

$$\begin{aligned}\max_{t\in[0,T]}\int_{\Omega_{\tau(m)-1/2}}&|u_l(x,t)-u_r(x,t)|^2\,dx\\ &\leqslant 4e^{2|\mu|T}\sum_{i=1}^k\Big(\iint_{Q^m}|a_i(u_l)-a_i(u_r)|^2\,dxdt\Big)^{1/2}\\ &\quad\times\Big(\iint_{Q^m}|u_l-u_r|^2\,dxdt\Big)^{1/2}.\end{aligned} \quad (5.18)$$

Using (5.14) and (5.15), from (5.18) we infer that $\{u_l\}$ is the Cauchy sequence in space $C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega}))$. Hence,

$$u\in C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega})) \quad\text{and}\quad u_l\underset{l\to\infty}{\longrightarrow} u \quad\text{in}\quad C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega})). \quad (5.19)$$

Assume $m\in\mathbb{N}$ is an arbitrary fixed number, and $l\in\mathbb{N}$ is an arbitrary number such that $l\geqslant m$. Putting $w=u_l$, $R=\tau(m)$, $t_1=0$, $t_2=T$, $\zeta(x):=\psi_{1/2}(x',\tau(m))$, $x=(x',x'')\in\overline{\Omega}$, $\theta(t):=e^{-2qt}$, $t\in\mathbb{R}$, where $q:=\sup_{\overline{Q}}q_4$ (q_4 from condition $(\mathbf{A}_4)$), and using (5.3) instead of (4.3), from Lemma 1 we obtain

$$\begin{aligned}\frac12 e^{-2qT}&\int_{\Omega^m}|u_l(x,T)|^2\psi_{1/2}(x',\tau(m))\,dx\\ &+\iint_{Q^m}\Big[\sum_{i=0}^n a_i(u_l)\partial_i u_l+q|u_l|^2\Big]\psi_{1/2}e^{-2qt}\,dxdt\\ &=-\iint_{Q^m}\sum_{i=1}^k a_i(u_l)u_l\,\partial_i\psi_{1/2}e^{-2qt}\,dxdt+\iint_{Q^m}fu_l\psi_{1/2}e^{-2qt}\,dxdt\\ &+\frac12\int_{\Omega^m}|u_0|^2\psi_{1/2}(x',\tau(m))\,dx.\end{aligned} \quad (5.20)$$

Estimating the terms of (5.20) with conditions $(\mathbf{A}_1)$, $(\mathbf{A}_3)$ (see (3.3)), $(\mathbf{A}_4)$, (4.8) and Cauchy-Bunyakovsky-Schvartz inequality, we get

$$\begin{aligned}&\iint_{Q_{\tau(m)-1/2}}\Big[q_3\sum_{i=k+1}^n|\partial_i u_l|^{p_i(x)}+(q-q_4)|u_l|^2\Big]e^{-2qt}\,dxdt\\ &\leqslant C_7\Big(\iint_{Q^m}\Big[\sum_{i=0}^k|\partial_i u_l|^2\Big]e^{-2qt}\,dxdt+\iint_{Q^m}\big[|f|^2+h\big]e^{-2qt}\,dxdt+\int_{\Omega^m}|u_0|^2\,dx\Big),\end{aligned} \quad (5.21)$$

where constant $C_7 > 0$ is independent of l, but it may be depended on m.

Using (5.14), from (5.21) we obtain

$$\iint_{Q_{\tau(m)-1/2}} \sum_{i=0}^{n} |\partial_i u_l|^{p_i(x)}\, dxdt \leqslant C_8, \quad m, l \in \mathbb{N},\ l \geqslant m, \tag{5.22}$$

where constant $C_8 > 0$ is independent of l, but it may be depended on m.

By virtue of $(\mathbf{A}_2)$, (5.14), (5.22), and discrete Hölder inequality we deduce that for every $i \in \{0, k+1, \dots, n\}$ and arbitrary $m, l \in \mathbb{N}$, $l \geqslant m$,

$$\iint_{Q_{\tau(m)-1/2}} |a_i(u_l)|^{p_i'(x)}\, dxdt \leqslant C_9 \iint_{Q_{\tau(m)-1/2}} \Big[\sum_{j=0}^{n} |\partial_j u_l|^{p_j(x)}\Big]\, dxdt + C_{10} \leqslant C_{11}, \tag{5.23}$$

where positive constants C_9, C_{10}, C_{11} are independent of l, but they may be depended on m.

In view of (5.22), (5.23), and the reflexivity of spaces $L_{p_i(\,\cdot\,)}(Q_\tau)$, $L_{p_i'(\cdot)}(Q_\tau)$, $i = \overline{k+1, n}$, $\tau > 0$, it follows that there exists a subsequence of the sequence $\{u_l\}_{l=1}^{\infty}$ (without loss of generality we use the notation $\{u_l\}_{l=1}^{\infty}$ for this subsequence), and functions $\chi_0 \in L_{2,\,\mathrm{loc}}(\overline{Q})$, $\chi_i \in L_{p'_i(\cdot),\mathrm{loc}}(\overline{Q})$, $i = \overline{k+1, n}$, such that

$$\partial_i u_l \xrightarrow[l\to\infty]{} \partial_i u \quad \text{weakly in } L_{p_i(\cdot),\mathrm{loc}}(\overline{Q}), \quad i = \overline{k+1, n}, \tag{5.24}$$

$$a_0(u_l) \xrightarrow[l\to\infty]{} \chi_0 \quad \text{weakly in } L_{2,\mathrm{loc}}(\overline{Q}), \tag{5.25}$$

$$a_i(u_l) \xrightarrow[l\to\infty]{} \chi_i \quad \text{weakly in } L_{p'_i(\cdot),\mathrm{loc}}(\overline{Q}), \quad i = \overline{k+1, n}. \tag{5.26}$$

Put

$$\chi_i := a_i(u), \quad i = \overline{1, k}. \tag{5.27}$$

Remark that for every $l \in \mathbb{N}$ (see (5.3)) we have the identity

$$\iint_Q \big[-u_l \psi \varphi' + \sum_{i=0}^{n} a_i(u_l)\partial_i \psi \varphi - f\psi\varphi\big]\, dxdt = 0,$$
$$\forall\, \psi \in \widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega),\ \operatorname{supp}\psi \subset \overline{\Omega^l},\ \forall\, \varphi \in C^1_c(0,T). \tag{5.28}$$

In (5.28) we fix an arbitrary $\psi \in \widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega)$, $\varphi \in C^1_c(0,T)$, and pass to the limit as $l \to \infty$, taking into account (5.14), (5.15), (5.25) – (5.27). So, we get

$$\iint_Q \big[-u\psi\varphi' + \sum_{i=0}^{n} \chi_i \partial_i \psi \varphi - f\psi\varphi\big]\, dxdt = 0. \tag{5.29}$$

To conclude that u is a weak solution of problem (1.1) – (1.3). It remains to show that the following identity holds

$$\iint_Q \sum_{i=0}^{n} \chi_i \partial_i \psi \varphi\, dxdt = \iint_Q \sum_{i=0}^{n} a_i(u)\partial_i \psi \varphi\, dxdt \quad \forall\, \psi \in \widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega),\ \forall\, \varphi \in C^1_c(0,T). \tag{5.30}$$

Indeed, if (5.30) is true, then from this and (5.29) we obtain the integral identity (3.2). In view of (5.14), (5.24) we have $u \in \widetilde{W}^{1,0}_{p(\cdot),\,\mathrm{loc}}(\overline{Q})$. From (5.2), (5.19) we deduce $u \in C([0,T];L_{2,\mathrm{loc}}(\overline{\Omega}))$ (it means that $u \in \mathbb{U}^{b}_{p,\mathrm{loc}}(\overline{Q})$) and the initial condition (1.3) is true. Hence, the function u is a weak solution of problem (1.1) – (1.3).

Step 3 (the correctness of identity (5.30)). To verify the correctness of identity (5.30) we use the monotonicity method [33].

Let $v \in L_{2,\mathrm{loc}}(\overline{Q})$ be an arbitrary function such that $\partial_i v \in L_{p_i(\cdot),\,\mathrm{loc}}(\overline{Q})$, $i=\overline{1,n}$, let $\zeta(x')$, $x'=(x_1,\dots,x_k)\in\mathbb{R}^k$, be a nonnegative continuously differentiable function with bounded support, and let $\theta \in C^1_c(0,T)$, $\theta \geqslant 0$. By virtue of condition $(\mathbf{A}_3)$ (see (3.4)), for every $l\in\mathbb{N}$ we have

$$\iint_Q \Big[\sum_{i=0}^{n}(a_i(u_l)-a_i(v))(\partial_i u_l-\partial_i v)+\mu(u_l-v)^2\Big]\zeta\theta e^{-2\mu t}\,dxdt \geqslant 0. \tag{5.31}$$

We rewrite inequality (5.31) as

$$\iint_Q \Big[\sum_{i=0}^{n} a_i(u_l)\partial_i u_l\Big]\zeta\theta e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=0}^{n}\big(a_i(u_l)\partial_i v + a_i(v)(\partial_i u_l-\partial_i v)\big) + \mu(u_l-v)^2\Big]\zeta\theta e^{-2\mu t}\,dxdt \geqslant 0 \quad \forall\, l\in\mathbb{N}. \tag{5.32}$$

Assume $m\in\mathbb{N}$ such that $\operatorname{supp}\zeta\subset\{x' \,|\, |x'|\leqslant \tau(m)\}$. Using Lemma 1, we obtain from identity (5.28) as $l\geqslant m$

$$\iint_Q \Big[\sum_{i=0}^{n} a_i(u_l)\partial_i u_l\Big]\zeta\theta e^{-2\mu t}\,dxdt = \iint_Q |u_l|^2\zeta(\theta'/2-\mu\theta)e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=1}^{k} a_i(u_l)u_l\partial_i\zeta - f u_l\zeta\Big]\theta e^{-2\mu t}\,dxdt. \tag{5.33}$$

From (5.32) and (5.33) we get

$$\iint_Q |u_l|^2\zeta(\theta'/2-\mu\theta)e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=1}^{k} a_i(u_l)u_l\partial_i\zeta - f u_l\zeta\Big]\theta e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=0}^{n}\big(a_i(u_l)\partial_i v + a_i(v)(\partial_i u_l-\partial_i v)\big) + \mu(u_l-v)^2\Big]\zeta\theta e^{-2\mu t}\,dxdt \geqslant 0. \tag{5.34}$$

In (5.34) we pass to the limit as $l\to\infty$, and by virtue of (5.14), (5.15), (5.25) – (5.27) we infer

$$\iint_Q |u|^2\zeta(\theta'/2-\mu\theta)e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=1}^{k} \chi_i u\partial_i\zeta - f u\zeta\Big]\theta e^{-2\mu t}\,dxdt - \iint_Q \Big[\sum_{i=0}^{n}\big(\chi_i\partial_i v + a_i(v)(\partial_i u-\partial_i v)\big) + \mu(u-v)^2\Big]\zeta\theta e^{-2\mu t}\,dxdt \geqslant 0. \tag{5.35}$$

In view of Lemma 1 it follows from (5.29) next equality

$$\iint_Q \big[\sum_{i=0}^{n} \chi_i \partial_i u\big]\, \zeta\theta e^{-2\mu t}\, dxdt = \iint_Q |u|^2 \zeta(\theta'/2 - \mu\theta) e^{-2\mu t}\, dxdt \\ - \iint_Q \big[\sum_{i=1}^{k} \chi_i u \partial_i \zeta - f u \zeta\big] \theta e^{-2\mu t}\, dxdt. \quad (5.36)$$

Assertions (5.35) and (5.36) imply

$$\iint_Q \big[\sum_{i=0}^{n} \chi_i\, \partial_i u\big] w\theta e^{-2\mu t}\, dxdt - \iint_Q \big[\sum_{i=0}^{n} \big(\chi_i \partial_i v + a_i(v)(\partial_i u - \partial_i v)\big) \\ + \mu(u-v)^2\big] \zeta\theta e^{-2\mu t}\, dxdt \geqslant 0,$$

that is,

$$\iint_Q \big[\sum_{i=0}^{n} (\chi_i - a_i(v))(\partial_i u - \partial_i v) + \mu(u-v)^2\big] \zeta\theta e^{-2\mu t}\, dxdt \geqslant 0. \quad (5.37)$$

In (5.37) we put $v = u - \lambda\psi\varphi$, where λ is an arbitrary number, and $\psi \in \widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega)$, $\varphi \in C^1_c(0,T)$ are arbitrary functions. So, taking into account the arbitrariness of λ, we obtain the equality

$$\iint_Q \big[\sum_{i=0}^{n} \big(\chi_i - a_i(u - \lambda\psi\varphi)\big) \partial_i \psi \varphi + \lambda\mu(\psi\varphi)^2\big] \zeta\theta e^{-2\mu t}\, dxdt = 0.$$

Here we tend λ to 0, using conditions ($\mathbf{A}_1$), ($\mathbf{A}_2$), and Lebesgue dominated convergence theorem. Thus, taking into account the arbitrariness of ζ and θ, we deduce

$$\iint_Q \Big[\sum_{i=0}^{n} \big(\chi_i - a_i(u)\big) \partial_i \psi\Big] \varphi\, dxdt = 0, \quad \psi \in \widetilde{W}^1_{p(\cdot),\mathrm{c}}(\Omega),\ \varphi \in C^1_c(0,T). \quad (5.38)$$

From (5.38) it follows (5.30).

Step 4 (the solution's estimate). Estimate (3.19) is obtained from (5.8), (5.12) and (5.13) by this way: $\langle u\rangle_m \leqslant \langle u - u_m\rangle_m + \langle u_m\rangle_m = \lim_{l\to\infty} \langle u_l - u_m\rangle_m + \langle u_m\rangle_m \leqslant C_2 e^{(1-\varkappa)m/2}$, where $C_2 := \sqrt{C_1} + C_5 = \sqrt{C_1}(2 + e^{1/2} - e^{-\varkappa/2})/(1 - e^{-\varkappa/2})$.

Now it is easy to see that the function u satisfies (3.16). Indeed, let $R > 0$ be an arbitrary number, and m be a natural number such that $m - 1 < R \leqslant m$. Using (3.19), we get

$$\langle u\rangle_R \leqslant \langle u\rangle_m \leqslant C_2 e^{(1-\varkappa)m/2} = C_2 e^{(1-\varkappa)(m-R)/2} e^{(1-\varkappa)R/2} \\ \leqslant C_2 e^{(1-\varkappa)/2} e^{-\varkappa R/2} e^{R/2} = \beta(R) e^{R/2}, \quad R \geqslant 1,$$

where $\beta(R) := C_2 e^{(1-\varkappa)/2} e^{-\varkappa R/2}$. Since $\beta(R) \to 0$ as $R \to +\infty$, then we have (3.16).

So, we have shown that u is a weak solution of problem (1.1) – (1.3) that satisfies (3.16) and (3.19). Theorem 2 is proved. □

References

1. A.N. Tikhonov, *Théoremes d'unicité pour l'équation de la chaleur*, *Mat. Sb.*, **42** (2) (1935), 199–216.
2. S. Täcklind, *Sur les classes quasianalytiques des solutions des équations aux dérivés partielles du type parabolique*, *Nova Acta Regiae Soc. Sci. Upsal. Series 4*, **10** (3) (1936), 3–55.
3. O.A. Oleinik, G.A. Iosifyan, *An analog of Saint-Venant principle and uniqueness of the solutions of the boundary-value problems in unbounded domains for parabolic equations*, *Usp. Mat. Nauk*, **31** (6) (1976), 142–166.
4. O.A. Oleinik, E.V. Radkevich, *The method of introducing a parameter for the investigation of evolution equations*, *Usp. Mat. Nauk*, **33** (5) (1978), 7–72.
5. Ph. Benilan, M.G. Grandall, M. Pierre, *Solutions of the porous medium equations in R^n under optimal conditions on initial values*, Indiana Univ. Math. J., **33** (1) (1984), 51–87.
6. A.E. Shishkov, *The solvability of the boundary-value problems for quasilinear elliptic and parabolic equations in unbounded domains in the classes of functions growing at the infinity*, *Ukr. Math. J.*, **47** (2) (1985), 277–289.
7. A.E. Shishkov, V.F. Akulov, *Analogs of Teklind uniqueness classes for solutions of initial-boundary problems for some quasilinear degenerate parabolic equations*, *Reports of the Academy of Sciences of UkrSSR. Ser. A*, **5** (1989), 23–25.
8. E. Di Benedetto, M.A. Herrero, *On the Cauchy problem and initial traces for a degenerate parabolic equation*, Trans. Amer. Math. Soc., **314** (1) (1989), 187–224.
9. A.L. Gladkov, *The Cauchy problem in classes of increasing functions for the equation of filtration with convection*, *Math. Sbornik*, **186** (6) (1995), 35–56.
10. H. Brézis, *Semilinear equations in $\mathbb{R}^N$ without conditions at infinity*, *Appl. Math. Optim.*, **12** (3) (1984), 271–282.
11. M.A. Herrero, M. Pierre, *The Cauchy problem for $u_t - \Delta u^m = 0$ when $0 < m < 1$*, Trans. Am. Math. Soc., **291** (1) (1985), 145–158.
12. F. Bernis, *Elliptic and parabolic semilinear parabolic problems without conditions at infinity*, Arch. Rational Mech. Anal., **106** (3) (1989), 217–241.
13. E. Di Benedetto, M.A. Herrero, *Non-negative solutions of the evolution p-Laplacian equation. Initial traces and Cauchy problem when $1 < p < 2$*, Arch. Rational Mech. Anal., **111** (3) (1990), 225–290.
14. M.M. Bokalo, *On unique solvability of boundary value problems for semilinear parabolic equations in unbounded domains without conditions at infinity*, Siberian Math. J., **34** (4) (1993), 620–627.
15. M.M. Bokalo, *Boundary value problems for semilinear parabolic equations in unbounded domains without conditions at infinity*, Siberian Math. J., **37** (5) (1996), 860–867.
16. L. Boccardo, Th. Gallouët, J.L. Vazquez, *Solutions of nonlinear parabolic equations without growth restrictions on the data*, Electronic J. Diff. Eq., **60** (2001), 1–20.

17. A. Gladkov, M. Guedda, *Diffusion-absorption equation without growth restrictions on the data at infinity*, *J. Math. Anal. Appl.*, **274** (1) (2002), 16–37.

18. C. Marchi, A. Tesei, *Higher-order parabolic equations without conditions at infinity*, *J. Math. Anal. Appl.*, **269** (2002), 352–368.

19. N.M. Bokalo, *The well-posedness of the first boundary value problem and the Cauchy problem for some quasilinear parabolic systems without conditions at infinity*, *J. Math. Sci.*, **135** (1) (2006), 2625–2636.

20. M. M. Bokalo, O. M. Buhrii, N. Hryadil, *Initial-boundary value problems for nonlinear elliptic-parabolic equations with variable exponents of nonlinearity in unbounded domains without conditions at infinity*, *Nonlinear Analysis. Elsevier. USA*, **192** (2020), 1–17.

21. M. Růžička, *Electroreological fluids: modeling and mathematical theory*, *Springer-Verl., Berlin*, 2000.

22. V. N. Samokhin, *On a class of equations that generalize equations of polytropic filtration*, *Diff. Equat.*, **32** (5) (1996), 648–657.

23. O. Buhrii, S. Lavrenyuk, *Initial boundary-value problem for parabolic equation of polytropic filtration type*, *Visn. Lviv Univ (Herald of Lviv University). Ser. Mech.-Math.*, **56** (2000), 33–43.

24. M.M. Bokalo, I.B. Pauchok, *On the well-posedness of a Fourier problem for nonlinear parabolic equations of higher order with variable exponents of nonlinearity*, *Matematychni Studii*, **26** (1) (2006), 25–48.

25. M. Bokalo, O. Domanska, *On well-posedness of boundary problems for elliptic equations in general anisotropic Lebesgue-Sobolev spaces,* Matematychni Studii, **28** (1) (2007), 77–91.

26. S. Antontsev, S. Shmarev, *Evolution PDEs with nonstandard growth conditions. Existence, uniqueness, localization, blow-up*, *Atlantis Studies in Diff. Eq., Vol. 4, Paris: Atlantis Press*, 2015.

27. V. Rădulescu, D. Repovš, *Partial differential equations with variable exponents: variational methods and qualitative analysis*, *CRC Press, Boca Raton, London, New York*, 2015.

28. M.M. Bokalo, O.M. Buhrii, R.A. Mashiyev, *Unique solvability of initial-boundary-value problems for anisotropic elliptic-parabolic equations with variable exponents of nonlinearity*, *J. Nonl. Evol. Eq. Appl.*, **6** (2013), 67–87.

29. O. Buhrii, N. Buhrii, *Nonlocal in time problem for anisotropic parabolic equations with variable exponents of nonlinearities*, *J. Math. Anal. Appl.*, **473** (2019), 695–711.

30. L. Diening, P. Harjulehto, P. Hästö, M. Růžička, *Lebesgue and Sobolev spaces with variable exponents*, *Springer, Heidelberg*, 2011.

31. N.M. Bokalo, *Energy estimates for solutions and unique solvability of the Fourier problem for linear and quasilinear parabolic equations*, *Diff. Equat.*, **30** (8) (1994), 1226–1234.

32. V.A. Galactionov, A.E. Shishkov, *Saint-Venant's principle in blow-up for higher-order quasilinear parabolic equations,* *Proc. R. Soc. Edinb. Sect. A, Math*, **133** (2003), no. 5, 1075–1119.

33. J.-L. Lions *Quelques méthodes de résolution des problèmes aux limites non linéaires.* Paris: Dunod, 1969.

Received 12.04.2022

Journal of Optimization, Differential Equations and Their Applications

Volume 30 Issue 1 June 2022

For notes

Journal of Optimization, Differential Equations and Their Applications

Volume 30 **Issue 1** **June 2022**

JODEA will publish carefully selected, longer research papers on mathematical aspects of optimal control theory and optimization for partial differential equations and on applications of the mathematic theory to issues arising in the sciences and in engineering. Papers submitted to this journal should be correct, innovative, non-trivial, with a lucid presentation, and of interest to a substantial number of readers. Emphasis will be placed on papers that are judged to be specially timely, and of interest to a substantial number of mathematicians working in this area.

Instruction to Authors:

Manuscripts should be in English and submitted electronically, pdf format to the member of the Editorial Board whose area, in the opinion of author, is most closely related to the topic of the paper and the same time, copy your submission email to the Managing Editor. Submissions can also be made directly to the Managing Editor.

Submission of a manuscript is a representation that the work has not been previously published, has not been copyrighted, is not being submitted for publication elsewhere, and that its submission has been approved by all of the authors and by the institution where the work was carried out. Furthermore, that any person cited as a source of personal communications has approved such citation, and that the authors have agreed that the copyright in the article shall be assigned exclusively to the Publisher upon acceptance of the article.

Manuscript style: Number each page. Page 1 should contain the title, authors names and complete affiliations. Place any footnote to the title at the bottom of Page 1. Each paper requires an abstract not exceeding 200 words summarizing the techniques, methods and main conclusions. AMS subject classification must accompany all articles, placed at Page 1 after Abstract. E-mail addresses of all authors should be placed together with the corresponding affiliations. Each paper requires a running head (abbriviated form of the title) of no more than 40 characters.

Equations should be centered with the number placed in parentheses at the right margin.

Figures must be drafted in high resolution and high contrast on separate pieces of white paper, in the form suitable for photographic reproduction and reduction.

References should be listed alphabetically, typed and punctuated according to the following examples:

1. S. N. Chow, J. K. Hale, *Methods od Bifurcation Theory*, *Springer-Verlad*, New York, 1982.
2. J. Serrin, *Gradient estimates for solutions of nonlinear elliptic and parabolic equations*, in "Contributions to Nonlinear Functional Analysis,"(ed. E.H. Zarantonello), *Academic Press* (1971).
3. S. Smale, *Stable manifolds for differential equations and diffeomorphisms*, *Ann. Scuola Norm. Sup. Pisa Cl.Sci.*, **18** (1963), 97–116.

For journal abbreviations used in bibliographies, consult the list of serials in the latest *Mathematical Reviews* annual index.

Final version of the manuscript should be typeset using LaTeX which can shorten the production process. Files of sample papers can be downloaded from the Journal's home page, where more information on how to prepare TeX files can be found.

Journal of Optimization, Differential Equations and Their Applications

Volume 30 **Issue 1** **June 2022**

ISSN 2617-0108

03001

9 772617 010000

CONTENTS