Journal of Optimization, Differential Equations and Their Applications

ISSN (print)
2617-0108

ISSN (online)
2663-6824

Volume 31

Issue 2

2023

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

Journal of Optimization, Differential Equations and Their Applications

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

p.kogut@i.ua

EDITORIAL BOARD

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 10.15421/14233102.

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 1–23, DOI 10.15421/142308

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

HOMOGENEOUS APPROXIMATION OF ONE-DIMENSIONAL SERIES OF ITERATED INTEGRALS AND TIME OPTIMALITY

Daria M. Andreieva[*], Svetlana Yu. Ignatovich[†]

Abstract. In the paper we consider nonlinear systems depending linearly on control with one-dimensional output. As is well known, under the analyticity requirement, the output can be expressed as a series of iterated integrals of controls with scalar coefficients. Since iterated integrals generate a free associative algebra, algebraic and combinatorial tools can be applied. Developing ideas of the algebraic approach to homogeneous approximation of nonlinear control systems, we propose the definition of a homogeneous approximation for series of iterated integrals with one-dimensional coefficients and study its algebraic properties. In addition, we describe relations between the homogeneous approximation of one-dimensional series of iterated integrals and an approximation in the sense of time optimality. Namely, we give conditions under which the optimal time and optimal controls for the minimal realization of the initial series are approximated by the optimal time and optimal controls for the minimal realization of its homogeneous approximation in a neighborhood of the origin.

Key words: nonlinear control system, series of iterated integrals, free associative algebra, core Lie subalgebra, homogeneous approximation, time-optimal control problem.

2010 Mathematics Subject Classification: 93B15, 93B25, 93C10, 49J15.

Communicated by Prof. V. Belozyorov

1. Introduction

Nonlinear systems depending linearly on the control with output

$$\dot{x} = \sum_{i=1}^{m} X_i(x)u_i, \quad y = h(x), \quad x \in \mathbb{R}^n,\ y \in \mathbb{R}^p,\ u_1, \ldots, u_m \in \mathbb{R}, \tag{1.1}$$

describe many important practical examples including nonholonomic mechanical systems, e.g., control of a unicycle, a flying airplane, a car pulling trailers, etc. [14]. Although the right hand side is nonlinear in x, the linear dependence on controls allows using algebraic tools to study various control problems related to such systems. We briefly explain the approach.

Let us assume that the vector fields $X_1(x), \ldots, X_m(x)$ and the map $h(x)$ are defined and real analytic in a neighborhood of the origin in $\mathbb{R}^n$ and $h(0) = 0$.

[*] Department of Applied Mathematics, V. N. Karazin Kharkiv National University, Svobody sqr., 4, 61022 Kharkiv, Ukraine, andrejeva_darja@ukr.net

[†] Department of Applied Mathematics, V. N. Karazin Kharkiv National University, Svobody sqr., 4, 61022 Kharkiv, Ukraine, s.ignatovich@karazin.ua, ignatovich@ukr.net

We consider a trajectory of the system starting at the origin and corresponding to a certain control $u(t) = (u_1(t), \dots, u_m(t))$; below we denote this trajectory by $x(t; u)$. Then the output $y(t; u) = h(x(t; u))$ can be expressed directly via controls $u_i(t)$. Namely, the following series representation holds,

$$y(t; u) = \sum_{k=1}^{\infty} \sum_{1 \le i_1, \dots, i_k \le m} c_{i_1 \dots i_k} \eta_{i_1 \dots i_k}(t, u), \tag{1.2}$$

where $c_{i_1 \dots i_k} \in \mathbb{R}^p$ are constant vectors depending on the vector fields $X_1, \dots, X_m$ and the map h and their derivatives at $x = 0$ and $\eta_{i_1 \dots i_k}(t, u)$ are nonlinear functionals of u ("iterated integrals") of the form

$$\eta_{i_1 \dots i_k}(t, u) = \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{k-1}} u_{i_1}(\tau_1) u_{i_2}(\tau_2) \cdots u_{i_k}(\tau_k)\, d\tau_k \cdots d\tau_2 d\tau_1, \tag{1.3}$$

which do not depend on $X_1, \dots, X_m$ and h. This representation was first applied to control systems by M. Fliess [6]. One can show that iterated integrals (1.3) considered on a sufficiently wide set of controls are linearly independent. The special form of iterated integrals suggests introducing a concatenation operation; with this operation the linear span of iterated integrals becomes a free associative algebra. As a result, algebraic and combinatorial technique can be applied [12], [18], [11] for various control problems related to systems of the form (1.1).

In particular, if the output is identity, $h(x) = x$, one can study optimal control problems using the direct representation for the trajectory (1.2) (where $y(t, u) = x(t; u)$) as is common for linear control systems. However, in the general case, it is difficult to operate with a series in the right hand side of (1.2). Instead, an approximation can be considered, at least for studying local problems. We emphasize that the linear approximation, where each vector field is replaced by its value at the origin, is definitely inappropriate for the case when $n > m$, since such an approximating system is not controllable even when the initial system is controllable. Therefore, a nonlinear approximation should be introduced; a possible choice is to find a system for which the representation of the trajectory equals a truncation of the series (1.2). The most well-known is the homogeneous approximation studied within the differential geometric approach in [4], [20], [7], [3] and many other papers.

The homogeneous approximation admits a natural and convenient algebraic description [8], [18], which allows finding it explicitly [19]. We recall the main ideas in the next section. The homogeneous approximation approximates the system in the sense of a sub-Riemannian metric [3] as well as in the sense of time optimality [18].

For systems with output, the most interesting case is $p = 1$, when the output is one-dimensional. In [2], we analyzed algebraic properties of the homogeneous approximation of the system, which was a realization of the given one-dimensional series. However, it seems more natural to consider the homogeneous approximation of the series itself. In the present paper we introduce the definition, propose an algebraic description and prove the classification theorem for such homogeneous approximation and explain the relation with the problem of time optimality.

The paper is organized as follows. In Section 2 we recall some definitions and results concerning nonlinear control systems. In particular, we recall the concepts of series of iterated integrals and the corresponding formal series of elements of the abstract free algebra. In Section 3 we introduce the definition of a homogeneous approximation of the series (Definition 3.2) and study its properties. In particular, we show that the minimal realization of the homogeneous approximation has no greater dimension than the minimal realization of the initial series (Theorem 3.1). Also, we prove a classification theorem (Theorem 3.2) which describes all possible core Lie subalgebras for a series and its homogeneous approximation. In Section 4 we consider the time-optimal problem for the system with output, which means the steering the trajectory of the system to a given surface in the minimal time. We show that under some conditions the optimal times and the optimal controls of the minimal realizations of the initial series and of its homogeneous approximation are equivalent in a neighborhood of the origin (Theorems 4.1 and 4.2).

2. Background

2.1. Systems with one-dimensional output

Let us consider the following nonlinear system

$$\dot{x} = \sum_{i=1}^{m} X_i(x)u_i, \tag{2.1}$$

with the one-dimensional output

$$y = h(x), \tag{2.2}$$

assuming that $X_1(x), \ldots, X_m(x)$ and $h(x)$ are real analytic in a neighborhood of the origin and $h(0) = 0$. As above, $x(t; u)$ denotes the trajectory of the system (2.1) starting at the origin, $x(0; u) = 0$, and corresponding to the control $u = u(t) = (u_1(t), \ldots, u_m(t))$, $t \in [0, \theta]$. We assume that controls are bounded; more specifically, the set of admissible controls is

$$B^\theta = \{u(\cdot) \in L_\infty([0, \theta]; \mathbb{R}^m) : \sum_{i=1}^{m} u_i^2(t) \leq 1 \text{ a.e., } t \in [0, \theta]\}. \tag{2.3}$$

Then, for small θ, the output $y(\theta; u) = h(x(\theta; u))$ can be represented as a series (1.2) where $\eta_{i_1 \ldots i_k}(\theta, u)$ are iterated integrals (1.3) and $c_{i_1 \ldots i_k}$ are scalar constant coefficients defied by

$$c_{i_1 \ldots i_k} = X_{i_k} \ldots X_{i_1} h(0). \tag{2.4}$$

Due to analyticity, there exist numbers $C_1, C > 0$ such that coefficients (2.4) satisfy the estimate

$$|c_{i_1\dots i_k}| \le C_1 C^k k! \tag{2.5}$$

On the other hand, since $|u_i(t)| \le 1$, $i = 1, \dots, m$, then $|\eta_{i_1\dots i_k}(\theta, u)| \le \frac{1}{k!}\theta^k$. Hence, the series in the right hand side of (1.2) converges if $\theta < \frac{1}{Cm}$. The representation (1.2) means that the output is directly expressed via inputs. It is especially useful if the sum in the right hand side is finite. Otherwise, the series can be approximated by a finite number of its terms.

2.2. Algebra of iterated integrals

More generally, one can consider series of iterated integrals with arbitrary scalar coefficients; the first question is whether there exists a system of the form (2.1) and the output (2.2) satisfying (1.2) (so-called realizability problem). In other words, one wants to find out whether there exist m analytic vector fields and an analytic scalar function satisfying equalities (2.4).

In order to answer this question as well as many others, it is useful to notice that iterated integrals (1.3) are linearly independent as functionals of $u \in B^\theta$. Therefore, the linear span (over $\mathbb{R}$) of iterated integrals turns into a free associative algebra

$$\mathcal{F}_\theta = \mathrm{Lin}\{\eta_{i_1\dots i_k}(\theta, u) : k \ge 1,\ 1 \le i_1, \dots, i_k \le m\}$$

with the concatenation operation $\eta_{i_1\dots i_k}(\theta, u) \vee \eta_{j_1\dots j_p}(\theta, u) = \eta_{i_1\dots i_k j_1\dots j_p}(\theta, u)$. All algebras $\mathcal{F}_\theta$ with different $\theta > 0$ are isomorphic to each other, so, it is useful to consider an abstract free associative algebra, which is isomorphic to all of them. Namely, let us introduce m abstract independent elements denoted as $\eta_1, \dots, \eta_m$ (letters) and consider all finite sequences of these letters $\eta_{i_1} \cdots \eta_{i_k}$ (words). Then the linear span of all words (over $\mathbb{R}$) with the concatenation operation is a free associative algebra isomorphic to any $\mathcal{F}_\theta$; we denote it by $\mathcal{F}$.

For the sake of brevity, below we use the following notation. Denote by M the set of all multi-indices

$$M = \{I = (i_1, \dots, i_k) : k \ge 1,\ 1 \le i_1, \dots, i_k \le m\}.$$

Then for $I = (i_1, \dots, i_k)$, we write $\eta_{i_1\dots i_k}$ or η_I instead of $\eta_{i_1} \cdots \eta_{i_k}$.

Thus, along with series of iterated integrals (1.2), we consider formal series of elements of $\mathcal{F}$,

$$S = \sum_{I \in M} c_I \eta_I, \tag{2.6}$$

where c_I are scalar coefficients.

Starting with the letters $\eta_1, \dots, \eta_m$, one can form the free Lie algebra $\mathcal{L}$, considering the linear span of all successive commutators of the letters. In other words, $\mathcal{L}$ is the minimal linear subspace including $\eta_1, \dots, \eta_m$ and closed with respect to the Lie bracket operation $[\ell_1, \ell_2] = \ell_1\ell_2 - \ell_2\ell_1$. This Lie algebra plays a central role in the analysis of series generated by systems due to its relation with coefficients (2.4), which is used below.

2.3. Realizability conditions

Now we return to the realizability problem. Suppose that the series (2.6) is given; does there exist a system of the form (2.1), (2.2) satisfying equalities (2.4)? We recall well-known realizability conditions [10], [9].

First, we notice that the series (2.6) generates the linear map $c : \mathcal{F} \to \mathbb{R}$ defined on basis elements as

$$c(\eta_I) = c_I, \quad I \in M.$$

It is convenient to introduce the unitary algebra $\mathcal{F}^e = \mathcal{F} + \mathbb{R}$ so that $1 = \eta_\varnothing$ (the empty word). We expand the linear map c to $\mathcal{F}^e$ assuming $c(1) = 0$. Below we use the notation $M_0 = M \cup \{0\}$.

Now, for any $\ell \in \mathcal{L}$, let us consider the series

$$F_c(\ell) = \sum_{I \in M_0} c(\eta_I \ell)\eta_I. \tag{2.7}$$

The number $\rho_L(S) = \dim\{F_c(\ell) : \ell \in \mathcal{L}\}$ is called *the Lie rank* of the series (2.6).

Theorem 2.1 ([10], [9]). *The series* (2.6) *satisfying condition* (2.5) *is realizable if and only if $\rho_L(S) < \infty$. If this is the case, then $n = \rho_L(S)$ equals the minimal possible dimension of the system* (2.1) *satisfying equalities* (2.4)*. Moreover, this system is defined uniquely up to a change of variables.*

Below we refer to the system mentioned in Theorem 2.1 as *the minimal realization of the series S*.

2.4. Minimal realization and its homogeneous approximation

The minimal realization is of the form (2.1), where $n = \rho_L(S)$; its trajectory $x(t; u)$ starting at the origin can be expressed as a series of iterated integrals with *vector* coefficients

$$x(t; u) = \sum_{I \in M} \widetilde{c}_I \eta_I(t, u), \quad \widetilde{c}_{i_1 \ldots i_k} = X_{i_k} \ldots X_{i_1} E(0) \in \mathbb{R}^n,$$

where $E(x) = x$. Below we denote by $\widetilde{S}$ the corresponding series,

$$\widetilde{S} = \sum_{I \in M} \widetilde{c}_I \eta_I.$$

Also, denote by $\widetilde{c}$ the linear map $\widetilde{c} : \mathcal{F} \to \mathbb{R}^n$ defined on basis elements as $\widetilde{c}(\eta_I) = \widetilde{c}_I$. It satisfies the Rashevsky-Chow condition

$$\dim \widetilde{c}(\mathcal{L}) = n.$$

This means that the minimal realization is locally controllable in a neighborhood of the origin of $\mathbb{R}^n$.

It may be useful to introduce another nonlinear system that is close to the system (2.1) but has a simpler structure; in particular, with the series having a finite number of terms. The most known choice is a homogeneous approximation [4], [20], [7], [3].

We introduce the homogeneous approximation of the minimal realization in algebraic terms, which allows us to avoid finding explicitly the coefficients of the series $\widetilde{S}$ [2]. Below we use the notation

$$M_k = \{I \in M : I = (i_1, \dots, i_k)\}, \quad k \geq 1,$$

i.e., the set M_k contains multi-indices of length k. First, we recall that the algebra $\mathcal{F}$ admits a grading structure

$$\mathcal{F} = \sum_{k=1}^{\infty} \mathcal{F}^k, \quad \mathcal{F}^k = \mathrm{Lin}\{\eta_I : I \in M_k\}, \ k \geq 1,$$

which is inherited by the Lie algebra $\mathcal{L}$,

$$\mathcal{L} = \sum_{k=1}^{\infty} \mathcal{L}^k, \quad \mathcal{L}^k = \mathcal{F}^k \cap \mathcal{L}, \ k \geq 1.$$

Below we say that $a \in \mathcal{F}^k$ is homogeneous of order k and write $\mathrm{ord}(a) = k$.

Let us consider the following subspaces

$$\begin{aligned} \mathcal{P}^1 &= \{\ell \in \mathcal{L}^1 : c(a\ell) = 0 \text{ for any } a \in \mathcal{F}^e\}, \\ \mathcal{P}^k &= \{\ell \in \mathcal{L}^k : \text{there exists } \ell' \in \mathcal{L}^1 + \cdots + \mathcal{L}^{k-1} \text{ such that} \\ &\qquad c(a(\ell - \ell')) = 0 \text{ for any } a \in \mathcal{F}^e\}, \quad k \geq 2. \end{aligned} \tag{2.8}$$

One can show that the subspace

$$\mathcal{L}_S = \sum_{k=1}^{\infty} \mathcal{P}^k \tag{2.9}$$

is a graded Lie subalgebra of $\mathcal{L}$ of codimension n [2]. We call it *a core Lie subalgebra of the series S*.

Since $\mathrm{codim}(\mathcal{L}_S) = n$, there exist n elements $\ell_1, \dots, \ell_n \in \mathcal{L}$ such that

$$\mathrm{Lin}\{\ell_1, \dots, \ell_n\} \dotplus \mathcal{L}_S = \mathcal{L},$$

where $\dotplus$ means a direct sum. Without loss of generality we can assume that $\ell_1, \dots, \ell_n$ are homogeneous and $\mathrm{ord}(\ell_i) \leq \mathrm{ord}(\ell_j)$ for $1 \leq i < j \leq n$.

Now, let us choose a homogeneous basis of $\mathcal{L}_S$; denote it by $\{\ell_i\}_{i=n+1}^{\infty}$; then $\{\ell_i\}_{i=1}^{\infty}$ is a basis of $\mathcal{L}$. Due to the Poincaré-Birkhoff-Witt Theorem [16], the set

$$\{\ell_{i_1}^{q_1} \cdots \ell_{i_k}^{q_k} : k \geq 1, \ 1 \leq i_1 < \cdots < i_k, \ q_1, \dots, q_k \geq 1\} \tag{2.10}$$

is a basis of $\mathcal{F}$. Here $\ell^q = \ell \cdots \ell$ (q times).

We define the inner product $\langle \cdot, \cdot \rangle$ in $\mathcal{F}$ assuming that the basis $\{\eta_I : I \in M\}$ is orthonormal. Denote by $\{d_{i_1 \dots i_k}^{q_1 \dots q_k} : k \geq 1,\ 1 \leq i_1 < \dots < i_k,\ q_1, \dots, q_k \geq 1\}$ the dual basis to the basis (2.10), i.e.,

$$\langle \ell_{i_1}^{q_1} \cdots \ell_{i_k}^{q_k}, d_{j_1 \dots j_r}^{p_1 \dots p_r} \rangle = \begin{cases} 1 \text{ if } k = r, i_t = j_t, q_t = p_t \text{ for } t = 1, \dots, k, \\ 0 \text{ otherwise.} \end{cases}$$

In order to express elements of the dual basis in a more explicit form, we introduce the shuffle product $\sqcup\!\sqcup$ in $\mathcal{F}^e$ [6] defined recursively on basis elements as

$$\eta_{i_1 I_1} \sqcup\!\sqcup \eta_{i_2 I_2} = \eta_{i_1}(\eta_{I_1} \sqcup\!\sqcup \eta_{i_2 I_2}) + \eta_{i_2}(\eta_{i_1 I_1} \sqcup\!\sqcup \eta_{I_2}),$$

where $1 \sqcup\!\sqcup a = a \sqcup\!\sqcup 1 = a$ for any $a \in \mathcal{F}^e$. Then the elements of the dual basis satisfy the following equalities [15]

$$d_{i_1 \dots i_k}^{q_1 \dots q_k} = \frac{1}{q_1! \cdots q_k!} d_{i_1}^{\sqcup\!\sqcup q_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_{i_k}^{\sqcup\!\sqcup q_k}, \tag{2.11}$$

where the notation $d_i = d_i^1$ is used. Here $d^{\sqcup\!\sqcup q} = d \sqcup\!\sqcup \cdots \sqcup\!\sqcup d$ (q times).

It can be shown that there exist such coordinates in the minimal realization in which the series $\widetilde{S}$ has the following "triangular" form

$$\widetilde{S}_j = d_j + \sum_{k \geq w_j + 1} \sum_{I \in M_k} \widetilde{c}_I \eta_I, \quad j = 1, \dots, n, \tag{2.12}$$

where $w_j = \operatorname{ord}(d_j) = \operatorname{ord}(\ell_j)$, $j = 1, \dots, n$. This representation justifies the following definition: a homogeneous approximation of the system (2.1) is a system whose series by some change of variables is reduced to the form

$$\widehat{\widetilde{S}}_j = d_j,\ j = 1, \dots, n. \tag{2.13}$$

Thus, each component of the series $\widehat{\widetilde{S}}$ is homogeneous and approximates the corresponding component of the series $\widetilde{S}$ in the sense of grading in $\mathcal{F}$.

We mention the following properties of core Lie subalgebras.

Theorem 2.2 ([8], [2]). *(i) Any graded Lie subalgebra of a finite codimension $n > 0$ is a core Lie subalgebra of some one-dimensional series of the form* (2.6)*.*

(ii) Minimal realizations of two series of the form (2.6) *have the same homogeneous approximation if and only if the core Lie subalgebras of these series coincide.*

Along with the core Lie subalgebra $\mathcal{L}_S$, we consider the (graded) left ideal generated by $\mathcal{L}_S$ of the form

$$\mathcal{J}_S = \operatorname{Lin}\{a\ell : a \in \mathcal{F}^e, \ell \in \mathcal{L}_S\}. \tag{2.14}$$

One can show [17], [18] that the set

$$\{d_1^{\sqcup\!\sqcup q_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_n^{\sqcup\!\sqcup q_n} : q_1, \ldots, q_n \geq 0,\ q_1 + \cdots + q_n \geq 1\} \tag{2.15}$$

is a basis of the subspace $\mathcal{J}_S^\perp$.

2.5. Approximation in the sense of time optimality

Let us consider two nonlinear systems, namely, the initial system (2.1) and its homogeneous approximation, i.e., the system

$$\dot{x} = \sum_{i=1}^{m} \widehat{X}_i(x) u_i, \tag{2.16}$$

whose trajectory starting at the origin corresponds to the series (2.13). The latter system approximates the former one not only in the algebraic sense mentioned above, but also in the sense of time optimality [18]. More specifically, let us consider systems (2.1) and (2.16) and assume that the trajectories of these systems correspond to the series (2.12) and (2.13) respectively. For an arbitrary point $\widetilde{s}$ from a neighborhood of the origin, we consider the time-optimal control problems for these two systems,

$$\dot{x} = \sum_{i=1}^{m} X_i(x) u_i,\ x(0) = 0, \quad x(\theta) = \widetilde{s}, \quad u \in B^\theta,\ \theta \to \min, \tag{2.17}$$

$$\dot{x} = \sum_{i=1}^{m} \widehat{X}_i(x) u_i,\ x(0) = 0, \quad x(\theta) = \widetilde{s}, \quad u \in B^\theta,\ \theta \to \min, \tag{2.18}$$

where the set B^θ of admissible controls is defined by (2.3).

Theorem 2.3 ([18]). *Suppose that the problem* (2.18) *has a unique solution; denote by $\widehat{\theta}_{\widetilde{s}}^*$ and $\widehat{u}_{\widetilde{s}}^*(t)$ the optimal time and the optimal control. For the problem* (2.17)*, let us denote by $\theta_{\widetilde{s}}^*$ the optimal time and by $U_{\widetilde{s}}^*$ the set of optimal controls. Then*

$$\frac{\theta_{\widetilde{s}}^*}{\widehat{\theta}_{\widetilde{s}}^*} \to 1 \quad \textit{as} \quad \widetilde{s} \to 0, \tag{2.19}$$

$$\frac{1}{\theta} \int_0^\theta |\widehat{u}_{\widetilde{s}i}^*(t) - u_{\widetilde{s}i}^*(t)|\, dt \to 0,\ i = 1, \ldots, m, \quad \textit{as} \quad \widetilde{s} \to 0 \tag{2.20}$$

for any $u_{\widetilde{s}}^*(t) \in U_{\widetilde{s}}^*$, where $\theta = \min\{\widehat{\theta}_{\widetilde{s}}^*, \theta_{\widetilde{s}}^*\}$. This means that, after some change of variables, the optimal time and the optimal control of the homogeneous problem (2.18) *approximates the optimal time and optimal controls of the problem* (2.17).

Actually, the uniqueness requirement for the optimal control of the problem (2.18) can be weakened [18].

In the next section we consider control systems with one-dimensional output. We show that another approximation is suitable in this situation.

3. Homogeneous approximation of a one-dimensional series of iterated integrals

3.1. Homogeneous series and its minimal realization

We start with considering a *homogeneous* one-dimensional series of the form

$$S = \sum_{I \in M_r} c_I \eta_I. \tag{3.1}$$

In other words, S contains only terms of order r, i.e., $S \in \mathcal{F}^r$ (we assume that at least one of its coefficients is nonzero). For consistency, we call S a "series", although the sum in the right-hand side of (3.1) is finite.

In what follows, we adopt the following definition.

Definition 3.1. We say that a linear subspace $\mathcal{J}' \subset \mathcal{F}$ is a *graded Lie generated left ideal* if there exists a graded Lie subalgebra $\mathcal{L}' \subset \mathcal{L}$ such that

$$\mathcal{J}' = \mathrm{Lin}\{a\ell : a \in \mathcal{F}^e, \ell \in \mathcal{L}'\}.$$

In this case we say that $\mathcal{J}'$ *is generated by* $\mathcal{L}'$.

For example, the left ideal $\mathcal{J}_S$ is a graded Lie generated left ideal; it is generated by the core Lie subalgebra $\mathcal{L}_S$, which follows from its definition (2.14).

Now, let us consider the set of all graded Lie generated left ideals that are orthogonal to S,

$$D = \{\mathcal{J} : \mathcal{J} \text{ is a graded Lie generated left ideal and } \mathcal{J} \subset S^\perp\}. \tag{3.2}$$

The following lemma gives an alternative way to find the core Lie subalgebra for a homogeneous series.

Lemma 3.1. *Let $S \in \mathcal{F}^r$ be nonzero. Let $\mathcal{L}_S$ be a core Lie subalgebra of S and $\mathcal{J}_S$ be a left ideal generated by $\mathcal{L}_S$. Then $\mathcal{J}_S$ is the maximal (in the sense of inclusion) left ideal from the set* (3.2).

Proof. Obviously, the maximal left ideal, which contains all other left ideals from D, exists and is unique. Denote it by $\mathcal{J}_{\max}$; let $\mathcal{L}_{\max}$ be the graded Lie subalgebra that generates $\mathcal{J}_{\max}$.

First, let us consider the minimal realization of the series S. As was explained in the previous section, the core Lie subalgebra $\mathcal{L}_S$ of the series S has the form (2.8), (2.9). However, the map c equals zero on any element of order other than r. Let us show that $c(a\ell) = 0$ for any $\ell \in \mathcal{L}_S$ and $a \in \mathcal{F}^e$. If not so, then $c(a\ell) \neq 0$ for some $\ell \in \mathcal{L}_S$ and some $a \in \mathcal{F}^e$. Then $a\ell \in \mathcal{F}^r$. Without loss of generality we can assume that a is homogeneous, then $a \in \mathcal{F}^k$ and $\ell \in \mathcal{L}^{r-k}$, where $0 \leq k \leq r-1$. Due to (2.8), there exists $\ell' \in \mathcal{L}^1 + \cdots + \mathcal{L}^{r-k-1}$ such that $c(a\ell) = c(a\ell') \neq 0$. Hence, $a\ell' \in \mathcal{F}^r$. On the other hand, $a\ell' \in \mathcal{L}^{k+1} + \cdots + \mathcal{L}^{r-1}$, which gives a contradiction.

Thus, $c(a\ell) = 0$ for any $\ell \in \mathcal{L}_S$ and $a \in \mathcal{F}^e$. This means that $c(\mathcal{J}_S) = 0$ or, what is the same, $S \perp \mathcal{J}_S$. Hence, S belongs to $\mathcal{J}_S^\perp$ and therefore $\mathcal{J}_S \subset S^\perp$, which implies $\mathcal{J}_S \subset \mathcal{J}_{\max}$ and therefore $\mathcal{L}_S \subset \mathcal{L}_{\max}$.

Now, let us consider $\mathcal{L}_{\max}$. Since $S \in \mathcal{F}^r$, it is orthogonal to any subspace $\mathcal{F}^k$ with $k \geq r+1$; therefore, $\mathcal{L}^k \subset \mathcal{L}_{\max}$ for any $k \geq r+1$. Hence, $\mathcal{L}_{\max}$ is of finite codimension.

Now we repeat the construction described above. Denote $n' = \operatorname{codim}(\mathcal{L}_{\max})$. We choose homogeneous elements $\ell'_1, \ldots, \ell'_{n'} \in \mathcal{L}$ such that $\mathcal{L} = \operatorname{Lin}\{\ell'_1, \ldots, \ell'_{n'}\} \dot{+} \mathcal{L}_{\max}$, choose a homogeneous basis $\{\ell'_i\}_{i=n'+1}^\infty$ of $\mathcal{L}_{\max}$, and construct the Poincaré-Birkhoff-Witt basis analogous to (2.10) and the dual basis $d'^{q_1 \cdots q_k}_{i_1 \ldots i_k}$ analogous to (2.11) as is explained in the previous section. Then re-expand the series (3.1) in the dual basis.

By construction, $S \in \mathcal{J}_{\max}^\perp$. Analogously to the subspace $J_S^\perp$, see (2.15), shuffle monomials of $d'_1, \ldots, d'_{n'}$ form a basis of $\mathcal{J}_{\max}^\perp$,

$$\mathcal{J}_{\max}^\perp = \operatorname{Lin}\{d'^{\sqcup\!\sqcup q_1}_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup d'^{\sqcup\!\sqcup q_{n'}}_{n'} : q_1, \ldots, q_{n'} \geq 0, q_1 + \cdots + q_{n'} \geq 1\}.$$

Therefore, S is a shuffle polynomial of $d'_1, \ldots, d'_{n'}$,

$$S = \sum_{q_1 w'_1 + \cdots + q_{n'} w'_{n'} = r} \alpha_{q_1 \ldots q_{n'}} d'^{\sqcup\!\sqcup q_1}_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup d'^{\sqcup\!\sqcup q_{n'}}_{n'},$$

where $w'_j = \operatorname{ord}(d'_j)$, $j = 1, \ldots, n'$. This implies that S has a realization of dimension n'. Namely, let us consider the n'-dimensional series S' with components

$$S'_k = d'_k, \quad k = 1, \ldots, n'.$$

Obviously, it corresponds to the linear map $c' : \mathcal{F} \to \mathbb{R}^{n'}$ defined on the Poincaré-Birkhoff-Witt basis (2.10) as follows:

$$\begin{aligned} &c'(\ell'^{q_1}_{i_1} \cdots \ell'^{q_k}_{i_k}) = 0 \ \text{ if } \ q_1 + \cdots + q_k \geq 2, \\ &c'(\ell'_i) = 0 \ \text{ if } \ i \geq n'+1, \\ &c'(\ell'_i) = e_i \ \text{ if } \ 1 \leq i \leq n', \end{aligned}$$

where e_i is a unit vector with 1 on the i-th place. One can show that the series S' satisfies the realizability conditions [9], i.e., there exists a control system of the form (2.1) whose trajectory is expressed as

$$x_k(\theta; u) = d'_k(\theta, u), \quad k = 1, \ldots, n'.$$

Moreover, this system can be efficiently constructed [18], [19]. Therefore, the initial series S corresponds to the polynomial output

$$y = h(x) = \sum_{q_1 w'_1 + \cdots + q_{n'} w'_{n'} = r} \alpha_{q_1 \ldots q_{n'}} x_1^{q_1} \cdots x_{n'}^{q_{n'}}$$

for this system.

Thus, the series S can be realized as an output of some n'-dimensional system, where $n' = \operatorname{codim}(\mathcal{L}_{\max})$. Therefore, $n' \geq n$, i.e., $\operatorname{codim}(\mathcal{L}_{\max}) \geq \operatorname{codim}(\mathcal{L}_S)$, which implies $\mathcal{L}_{\max} \subset \mathcal{L}_S$. On the other hand, as is shown above, $\mathcal{L}_S \subset \mathcal{L}_{\max}$. Hence, $\mathcal{L}_S = \mathcal{L}_{\max}$ and therefore $\mathcal{J}_S = \mathcal{J}_{\max}$. □

3.2. Homogeneous approximation of a series

Let us consider a series of elements of $\mathcal{F}$

$$S = \sum_{I \in M} c_I \eta_I, \tag{3.3}$$

where c_I are scalar coefficients. Below we assume that S is realizable, that is, its Lie rank is finite. By the homogeneous approximation of S we mean the sum of terms of the minimal order. More specifically, we adopt the following definition.

Definition 3.2. For the series (3.3), let r be the minimal order of terms of S, that is,

$$r = \min\{k : c_I \neq 0 \text{ for some } I \in M_k\}.$$

We say that

$$\widehat{S} = \sum_{I \in M_r} c_I \eta_I$$

is *a homogeneous approximation of the series* (3.3).

Since $\widehat{S}$ contains a finite number of terms, it is realizable. Theorem 3.1 below describes the relation between the Lie ranks and core Lie subalgebras of series S and $\widehat{S}$.

Theorem 3.1. *Let $\widehat{S}$ be a homogeneous approximation of S. Then $\rho_L(\widehat{S}) \leq \rho_L(S)$ and $\mathcal{L}_S \subset \mathcal{L}_{\widehat{S}}$.*

Proof. Let $\widehat{S} \in \mathcal{F}^r$. First, let us show that the left ideal $\mathcal{J}_S$, which obviously is graded and Lie generated, is orthogonal to $\widehat{S}$.

Denote by $\widetilde{S}$ the series of the minimal realization of S, which has the form (2.12) with $n = \rho_L(S) = \operatorname{codim}(\mathcal{L}_S)$, i.e.,

$$\widetilde{S}_k = d_k + R_k, \ k = 1, \dots, n,$$

where $R_k \in \sum_{i \geq w_k+1} \mathcal{F}^i$, $w_k = \operatorname{ord}(d_k)$. Then S has the form

$$S = \sum_{q_1+\dots+q_n \geq 1} \alpha_{q_1 \dots q_n} (d_1 + R_1)^{\sqcup\!\sqcup q_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup (d_n + R_n)^{\sqcup\!\sqcup q_n},$$

where $\alpha_{q_1\dots q_n} \in \mathbb{R}$. Since $\widehat{S}$ contains elements of the minimal order from S, we obviously get that $\widehat{S}$ is a shuffle polynomial of $d_1, \dots, d_n$,

$$\widehat{S} = \sum_{q_1 w_1 + \dots + q_n w_n = r} \alpha_{q_1\dots q_n} d_1^{\sqcup\!\sqcup q_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_n^{\sqcup\!\sqcup q_n}.$$

However, shuffle monomials (2.15) of $d_1, \dots, d_n$ form a basis of $\mathcal{J}_S^\perp$, therefore, $\widehat{S} \subset \mathcal{J}_S^\perp$, which implies $\mathcal{J}_S \subset \widehat{S}^\perp$.

Let $\mathcal{J}_{\max}$ be the maximal (in the sense of inclusion) graded Lie generated left ideal orthogonal to $\widehat{S}$ and let $\mathcal{L}_{\max}$ be the Lie subalgebra that generates $\mathcal{J}_{\max}$. Then $\mathcal{J}_S \subset \mathcal{J}_{\max}$ and therefore $\mathcal{L}_S \subset \mathcal{L}_{\max}$.

On the other hand, as shown in the previous subsection, $\mathcal{L}_{\max}$ is a core Lie subalgebra of $\widehat{S}$, that is, $\mathcal{L}_{\widehat{S}} = \mathcal{L}_{\max}$ and $\rho_L(\widehat{S}) = \operatorname{codim}(\mathcal{L}_{\max})$. Hence, $\mathcal{L}_S \subset \mathcal{L}_{\widehat{S}}$ and therefore $\rho_L(\widehat{S}) \le \rho_L(S)$. □

Example 3.1. Let us consider the series

$$S = \eta_1 + \eta_{21} + \eta_{221} + \eta_{2221} + \cdots = \sum_{k=0}^{\infty} \eta_2^k \eta_1,$$

which describes the output of the following one-dimensional system

$$\dot{x} = u_1 + x\,u_2, \quad y = h(x) = x.$$

Its homogeneous approximation is $\widehat{S} = \eta_1$; it corresponds to the homogeneous one-dimensional system

$$\dot{x} = u_1, \quad y = h(x) = x.$$

Thus, here $\mathcal{L}_S = \mathcal{L}_{\widehat{S}} = \operatorname{Lin}\{\eta_2\} + \sum_{k=2}^{\infty} \mathcal{L}^k$ and $\rho_L(S) = \rho_L(\widehat{S}) = 1$.

Note that the minimal realization of any truncated series $S' = \sum_{k=0}^{p-1} \eta_2^k \eta_1$ for $p \ge 2$ is of dimension p. For example, it can be chosen as

$$\dot{x}_1 = u_1, \ \dot{x}_2 = x_1 u_2, \ \cdots, \dot{x}_p = \tfrac{1}{(p-1)!} x_1^{p-1} u_2$$

with the output $y = h(x) = x_1 + \cdots + x_p$. Thus, here we have $\rho_L(S') > \rho_L(S)$.

Example 3.1 shows that Theorem 3.1 cannot be generalized to an arbitrary truncation of one-dimensional series.

3.3. Classification theorem

Theorem 3.1 describes the relation between the core Lie subalgebras of the series and of its homogeneous approximation. The following theorem shows that any pair of nested graded Lie subalgebras of finite nonzero codimension are the core Lie subalgebras of some series and its homogeneous approximation.

Theorem 3.2. *Let $\mathcal{L}_1 \subset \mathcal{L}_2 \subset \mathcal{L}$ be two graded Lie subalgebras such that* $0 < \operatorname{codim}(\mathcal{L}_2) \leq \operatorname{codim}(\mathcal{L}_1) < \infty$. *Then there exists a one-dimensional series S of the form* (2.6) *such that $\mathcal{L}_1$ is its core Lie subalgebra and $\mathcal{L}_2$ is a core Lie subalgebra of its homogeneous approximation $\widehat{S}$, i.e., $\mathcal{L}_S = \mathcal{L}_1$ and $\mathcal{L}_{\widehat{S}} = \mathcal{L}_2$.*

Proof. If $\mathcal{L}_1 = \mathcal{L}_2$, then we can take a homogeneous series. Namely, let $n = \operatorname{codim}(\mathcal{L}_1)$. Then choose n homogeneous elements $\ell_1, \dots, \ell_n$ such that

$$\operatorname{Lin}\{\ell_1, \dots, \ell_n\} \dotplus \mathcal{L}_1 = \mathcal{L}. \tag{3.4}$$

Then choose a homogeneous basis $\{\ell_i\}_{i=n+1}^{\infty}$ of $\mathcal{L}_1$ and construct the Poincaré-Birkhoff-Witt basis (2.10) and the dual basis (2.11) as explained above. Let us consider the series

$$S = d_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup\, d_n.$$

As was shown in [2, proof of Theorem 4], $\mathcal{L}_1$ is the core Lie subalgebra of S.

Suppose now that $\mathcal{L}_1 \neq \mathcal{L}_2$. Let us denote $n = \operatorname{codim}(\mathcal{L}_1)$ and $q = \operatorname{codim}(\mathcal{L}_2)$, then $0 < q < n < \infty$. In this case we also choose n homogeneous elements satisfying (3.4). However now, without loss of generality, we assume that $n - q$ of these elements belong to $\mathcal{L}_2$. Namely, we choose elements $\ell_1, \dots, \ell_n$ so that

$$\operatorname{Lin}\{\ell_1, \dots, \ell_q\} \dotplus \mathcal{L}_2 = \mathcal{L}$$

and

$$\ell_{q+1}, \dots, \ell_n \in \mathcal{L}_2.$$

Then we choose a homogeneous basis $\{\ell_i\}_{i=n+1}^{\infty}$ of $\mathcal{L}_1$ and construct the bases (2.10) and (2.11). Let us consider the series

$$S = d_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup\, d_q + d_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup\, d_n.$$

This series is not homogeneous; its homogeneous approximation equals

$$\widehat{S} = d_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup\, d_q.$$

The core Lie subalgebra of $\widehat{S}$ equals $\mathcal{L}_2$, i.e., $\mathcal{L}_{\widehat{S}} = \mathcal{L}_2$, which directly follows from [2, proof of Theorem 4]. Let us prove that the core Lie subalgebra of S equals $\mathcal{L}_1$.

We follow arguments from [2]. Namely, let us denote $p = \operatorname{ord}(\ell_1 \cdots \ell_n)$ and notice that $\operatorname{ord}(\ell_1 \cdots \ell_q) < p$. Therefore, the map $c : \mathcal{F} \to \mathbb{R}$ corresponding to S is such that on elements of the Poincaré-Birkhoff-Witt basis (2.10) from the subspace $\mathcal{F}^p$

$$c(\ell_{i_1} \cdots \ell_{i_k}) = \begin{cases} 1 & \text{if } (i_1, \dots, i_k) = (1, \dots, n), \\ 0 & \text{otherwise.} \end{cases}$$

Repeating the proof of Lemma 1 from [2] we get that for any tuple $(j_1, \dots, j_k)$ such that $\operatorname{ord}(\ell_{j_1} \cdots \ell_{j_k}) \geq p$

$$c(\ell_{j_1} \cdots \ell_{j_k}) = \begin{cases} 1 & \text{if } (j_1, \dots, j_k) \text{ is a permutation of } \{1, \dots, n\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Now we consider series (2.7), which are used to find the Lie rank as in Theorem 2.1. We can re-expand them in the dual basis (2.11),

$$F_c(\ell) = \sum_{j_1<\cdots<j_k} \frac{1}{q_1!\cdots q_k!} c(\ell_{j_1}^{q_1}\cdots\ell_{j_k}^{q_k}\ell) d_{j_1}^{\sqcup\!\sqcup q_1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_{j_k}^{\sqcup\!\sqcup q_k}.$$

Using this representation, we prove that the series $F_c(\ell_1), \ldots, F_c(\ell_n)$ are linearly independent. Actually, using the property (3.5), we prove analogously to [2] that for $1 \le k \le j \le n$

$$c(\ell_1\cdots\ell_{k-1}\ell_{k+1}\cdots\ell_n\ell_j) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j > k. \end{cases}$$

Hence, the $n \times n$ matrix whose elements are the coefficients of $d_1 \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_{k-1} \sqcup\!\sqcup d_{k+1} \sqcup\!\sqcup \cdots \sqcup\!\sqcup d_n$ in the series $F_c(\ell_j)$ is nonsingular (it has units on the diagonal and zeros above the diagonal). This implies that $\rho_L(S) = \dim\{F_c(\ell) : \ell \in \mathcal{L}\} \ge n$.

On the other hand, the series S obviously has the n-dimensional realization with the series

$$\widetilde{S}_k = d_k, \quad k = 1, \ldots, n, \tag{3.6}$$

and the output $y = h(x) = x_1\cdots x_q + x_1\cdots x_n$. This means that $\rho_L(S) \le n$. Thus, $\rho_L(S) = n = \operatorname{codim}(\mathcal{L}_1)$ and therefore the realization (3.6) is minimal. Since its core Lie subalgebra obviously equals $\mathcal{L}_1$, we get that $\mathcal{L}_S = \mathcal{L}_1$. □

4. Time-optimal problem for one-dimensional series

4.1. Time-optimal problem for homogeneous series

Let us consider a homogeneous series

$$\widehat{S} = \sum_{I\in M_r} c_I \eta_I$$

(which can be a homogeneous approximation of some other series S). Let (2.16) be its minimal realization and let $y = \widehat{h}(x)$ be the corresponding output. We consider the following time-optimal problem

$$\dot{x} = \sum_{i=1}^{m} \widehat{X}_i(x)u_i, \; x(0) = 0, \quad \widehat{h}(x(\theta)) = s, \quad u \in B^\theta, \; \theta \to \min, \tag{4.1}$$

where B^θ is of the form (2.3) and $s \in \mathbb{R}$ is a given nonzero number. Thus, the problem is to find a control steering the system from the origin to the surface $\widehat{h}(x) = s$ in the minimal possible time. If there exists a control and a time moment T satisfying the equality $\widehat{h}(x(T; u)) = s$, then the problem (4.1) has a solution due to Filippov's Theorem [5]. In this case we denote the optimal time by $\widehat{\theta}_s^*$ and an optimal control by $\widehat{u}_s^*(t)$.

The time-optimal problem (4.1) can be rewritten in the form

$$\sum_{I\in M_r} c_I\eta_I(\theta,u) = s, \quad u\in B^\theta,\ \theta\to\min. \tag{4.2}$$

Since the series is homogeneous, we can simplify the problem. Namely, we notice that

$$\eta_I(\theta,u) = \int_0^\theta\int_0^{\tau_1}\cdots\int_0^{\tau_{k-1}} u_{i_1}(\tau_1)\cdots u_{i_k}(\tau_k)d\tau_k\cdots d\tau_2 d\tau_1 =$$

$$= \theta^k\int_0^1\int_0^{\tau_1}\cdots\int_0^{\tau_{k-1}} u_{i_1}(\tau_1\theta)\cdots u_{i_k}(\tau_k\theta)d\tau_k\cdots d\tau_2 d\tau_1 = \theta^k\eta_I(1,\bar{u}),$$

where $\bar{u}(t) = u(t\theta)$, $t\in[0,1]$. When $u(t)$ runs through the set (2.3), $\bar{u}(t)$ runs through the set B^1 (of the form (2.3) with $\theta = 1$). Thus, the problem (4.2) is reduced to

$$\theta^r\sum_{I\in M_r} c_I\eta_I(1,u) = s, \quad u\in B^1,\ \theta\to\min.$$

In order to solve this problem, it is sufficient to solve the optimization problem for the nonlinear homogeneous functional on the unit ball of the space $L_\infty([0,1];\mathbb{R}^m)$. Namely, for $s>0$, we consider the problem

$$\sum_{I\in M_r} c_I\eta_I(1,u)\to\max, \quad u\in B^1.$$

If the maximum value equals $\mu>0$, then $\widehat{\theta}_s^* = (\frac{s}{\mu})^{\frac{1}{r}}$ is the optimal time in the problem (4.1). For any function $\bar{u}(t)$ on which the maximum value is achieved, $\widehat{u}_s^*(t) = \bar{u}(t/\widehat{\theta}_s^*)$ is an optimal control in (4.1). Analogously, for $s<0$, we consider the problem

$$\sum_{I\in M_r} c_I\eta_I(1,u)\to\min, \quad u\in B^1.$$

If the minimum value equals $\mu<0$, then $\widehat{\theta}_s^* = (\frac{s}{-\mu})^{\frac{1}{r}}$ is the optimal time in (4.1).

Thus, the following lemma holds.

Lemma 4.1. *Let us fix the number $\bar{s}>0$ (resp., $\bar{s}<0$) and suppose that the following time-optimal problem has a solution*

$$\dot{x} = \sum_{i=1}^m \widehat{X}_i(x)u_i,\ x(0)=0,\ \widehat{h}(x(\theta)) = \bar{s}, \quad u\in B^\theta,\ \theta\to\min.$$

Let $\widehat{\theta}_{\bar{s}}^*$ be the optimal time and $\widehat{u}_{\bar{s}}^*(t)$ be an optimal control. Then for any $s>0$ (resp., $s<0$) the problem (4.1) has a solution: the optimal time equals

$$\widehat{\theta}_s^* = \left(\frac{s}{\bar{s}}\right)^{\frac{1}{r}}\widehat{\theta}_{\bar{s}}^*$$

and the function

$$\widehat{u}^*_s(t) = \widehat{u}^*_{\bar{s}}\left(t(\bar{s}/s)^{\frac{1}{r}}\right),\ t \in [0, \widehat{\theta}^*_s],$$

is an optimal control.

Example 4.1. Consider the following system, which describes the Grushin plane [3]

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= x_1 u_2 \end{aligned} \tag{4.3}$$

with the output $y = x_1 x_2$. This system and the output are homogeneous, the corresponding series has the form

$$\widehat{S} = \eta_1 ⧢ \eta_{21} = \eta_{121} + 2\eta_{211}.$$

As mentioned above, it is sufficient to consider one particular s, for example, $s = 1$. (The case $s = -1$ can be considered analogously.) The corresponding optimization problem is to maximize the functional $\eta_{121}(1, u) + 2\eta_{211}(1, u)$ on the set B^1 of controls satisfying the constraint $u_1^2(t) + u_2^2(t) \le 1$, $t \in [0, 1]$.

The solution of the time-optimal problem for this system is well known [1]; we briefly explain how it can be obtained by applying the Pontryagin Maximum Principle. First, we write the Hamilton-Pontryagin function and the dual system

$$H = \psi_1 u_1 + \psi_2 x_1 u_2, \quad \dot{\psi}_1 = -\psi_2 u_2,\ \dot{\psi}_2 = 0.$$

Hence, $\psi_2(t)$ is a constant; denote it by c. Following the Pontryagin Maximum Principle, we find the maximum value of H on the set of controls satisfying the inequality $u_1^2 + u_2^2 \le 1$. To do this, we write the Lagrange function $L = \psi_1 u_1 + c x_1 u_2 + \lambda(u_1^2 + u_2^2 - 1)$. Equating its partial derivatives on u_1 and u_2 to zero, we get two equations

$$\psi_1 + 2\lambda u_1 = 0, \quad c x_1 + 2\lambda u_2 = 0.$$

Hence, $\psi_1 u_2 = c x_1 u_1$ identically, which implies $\psi_1 \dot{\psi}_1 + c^2 x_1 \dot{x}_1 = 0$, and therefore, $\psi_1^2 + c^2 x_1^2 = const$. Since $\psi_1 = -2\lambda u_1$ and $c x_1 = -2\lambda u_2$, we get $4\lambda^2(u_1^2 + u_2^2) = const$. However, it is known [18] that actually the optimal control satisfies the equality $u_1^2 + u_2^2 \equiv 1$, hence, $|\lambda| = const$. Since H has a maximum value on the optimal control, λ is non-positive, hence, $\lambda = const$. Moreover, if $\lambda = 0$, then $\psi_1 = 0$; hence, $c \neq 0$ and $x_1 = 0$. This means that $\lambda \neq 0$ for any optimal trajectory that ends at the curve $x_1 x_2 = 1$. Thus,

$$u_1(t) = -\tfrac{1}{2\lambda}\psi_1(t), \quad u_2(t) = -\tfrac{c}{2\lambda}x_1(t),$$

where λ is constant. Therefore, optimal controls $u_1(t)$ and $u_2(t)$ are differentiable and satisfy the differential equations

$$\dot{u}_1(t) = -\xi u_2(t), \quad \dot{u}_2(t) = \xi u_1(t), \qquad \text{where } \xi = -\tfrac{c}{2\lambda}.$$

This gives $u_1(t)=\cos(\xi t+\varphi)$ and $u_2(t)=\sin(\xi t+\varphi)$, where, taking into account the initial condition $u_2(0)=\xi x_1(0)=0$, we get $\varphi=\pi k$, $k\in\mathbb{Z}$. Actually, one can show that the points on the optimal trajectory where $x_1(\tau)=0$ are conjugate, hence, for the optimal trajectory, only $t\in(0,\frac{\pi}{|\xi|})$ is possible. Thus, optimal controls can be

$$u_1(t)=\alpha\cos(\xi t),\quad u_2(t)=\alpha\sin(\xi t),$$

where $\alpha=\pm1$. Substituting such controls, we get the optimal trajectories

$$x_1(t)=\tfrac{\alpha}{\xi}\sin(\xi t),\quad x_2(t)=\tfrac{t}{2\xi}-\tfrac{1}{4\xi^2}\sin(2\xi t).$$

Now let us apply the transversality condition that means that at the end time moment the vector $(\psi_1(T),\psi_2(T))$ is proportional to a normal vector to the curve $x_1x_2=1$ at the point $(x_1(T),x_2(T))$. This obviously leads to the equality

$$\frac{\alpha\cos(\xi T)}{\xi}=\frac{\frac{T}{2\xi}-\frac{1}{4\xi^2}\sin(2\xi T)}{\frac{\alpha}{\xi}\sin(\xi T)},$$

which gives $\sin(2\xi T)=\frac{2\xi T}{3}$. Denote the solution of the equation $\sin z=\frac{z}{3}$, which belongs to the interval $[0,\pi]$, by $\bar z$; it equals $\bar z\approx2.27886$. Then $2\xi T=\pm\bar z$ (depending on the sign of ξ).

On the other hand, the optimal time $T=\widehat\theta_1^*$ equals the time moment when the trajectory reaches the curve $x_1x_2=1$, hence,

$$\tfrac{\alpha}{\xi}\sin(\xi T)\left(\tfrac{T}{2\xi}-\tfrac{1}{4\xi^2}\sin(2\xi T)\right)=1.$$

Combining two last equations, we obtain the explicit expressions for the optimal time and the optimal controls

$$\widehat\theta_1^*=\left(\tfrac{3\bar z^2}{4\sin(\bar z/2)}\right)^{1/3}\approx1.62458,\ \widehat u_1^*(t)=(\widehat u_{11}^*(t),\widehat u_{12}^*(t))=(\alpha\cos(\xi t),\alpha\sin(\xi t)),$$

where $\alpha=\pm1$ and $\xi=\alpha\bar z/(2\widehat\theta_1^*)\approx0.70137\alpha$. Thus, there exist two optimal controls corresponding to $\alpha=+1$ and $\alpha=-1$; they steer the system from the origin to the points $(\frac{\sin(\xi T)}{\xi},\frac{\xi}{\sin(\xi T)})\approx(1.2952,0.7721)$ and $(-\frac{\sin(\xi T)}{\xi},-\frac{\xi}{\sin(\xi T)})\approx(-1.2952,-0.7721)$ on the curve $x_1x_2=1$.

For the final condition $x_1x_2=-1$ the result is analogous: the optimal time is the same whereas the optimal controls are $\widehat u_1^*(t)=(\alpha\cos(\xi t),-\alpha\sin(\xi t))$, $\alpha=\pm1$.

For any final condition $x_1x_2=s\neq0$ the optimal time and optimal controls can be found as described in Lemma 4.1.

4.2. Approximation in the sense of time optimality

Now we consider an arbitrary realizable series S. Suppose that its minimal realization is n-dimensional and has the form (2.1) with the output (2.2). Let us consider the time-optimal problem

$$\dot{x} = \sum_{i=1}^{m} X_i(x)u_i, \ x(0) = 0, \quad h(x(\theta)) = s, \quad u \in B^\theta, \ \theta \to \min, \tag{4.4}$$

where $s \neq 0$ is a fixed number. If an optimal control exists, we denote the optimal time by θ_s^* and the set of optimal controls by U_s^*.

Also, we consider the homogeneous approximation $\widehat{S}$ of the series S and the corresponding time-optimal problem (4.1) for its minimal realization (2.16). If the problem (4.1) has a solution, we denote by $\widehat{\theta}_s^*$ the optimal time and by $\widehat{U}_s^*$ the set of optimal controls.

Theorem 4.1. *Let us consider the time-optimal problems* (4.4) *and* (4.1) *for the minimal realizations of the series S and of its homogeneous approximation $\widehat{S}$ respectively. Suppose the problem* (4.1) *has a solution for $s = 1$. Then there exists $\varepsilon > 0$ such that for any $s \in (0, \varepsilon)$ the problem* (4.4) *also has a solution and*

$$\frac{\theta_s^*}{\widehat{\theta}_s^*} \to 1 \ \textit{as} \ \ s \to +0. \tag{4.5}$$

For $s < 0$ the analogous result holds.

Proof. We follow the idea of [13] developed in [18]. First, let us fix $\delta \in (0,1)$ and for any $s > 0$ consider the set $Q_s = [s(1-\delta), s(1+\delta)]$. Obviously, $0 \notin Q_s$. Also, let us introduce the notation

$$R_k(\theta, u) = \sum_{I \in M_k} c_I \eta_I(\theta, u), \ k \geq r+1, \quad R(\theta, u) = \sum_{k=r+1}^{\infty} R_k(\theta, u),$$

then

$$S(\theta, u) = \widehat{S}(\theta, u) + R(\theta, u). \tag{4.6}$$

Since the system (2.1), (2.2) is analytic, there exist $C_1, C_2, C > 0$ such that the series $S(\theta, u)$ converges for any $\theta \in (0, 1/C)$ and any $u \in B^\theta$ and

$$|R_k(\theta, u)| \leq C_1 (C\theta)^k, \ k \geq r+1, \quad |R(\theta, u)| \leq C_2 (C\theta)^{r+1}.$$

Let us introduce the operator

$$G_s(x) = s - R(\widehat{\theta}_x^*, \widehat{u}_x^*)$$

defined on x for which the series $S(\widehat{\theta}_x^*, \widehat{u}_x^*)$ converges. Let us show that the operator G_s is defined on Q_s and maps Q_s to itself.

If $x \in Q_s$, then $0 < s(1-\delta) \le x \le s(1+\delta) < 2s$. Therefore, due to Lemma 4.1, $\widehat{\theta}^*_x < (2s)^{1/r}\widehat{\theta}^*_1$. Therefore, the series $S(\widehat{\theta}^*_x, \widehat{u}^*_x)$ converges for all $x \in Q_s$ if $0 < s < 1/(2(C\widehat{\theta}^*_1)^r) = \varepsilon_1$. Moreover,

$$|R(\widehat{\theta}^*_x, \widehat{u}^*_x)| \le C_2(C\widehat{\theta}^*_x)^{r+1} < (2s)^{\frac{r+1}{r}} C_2(C\widehat{\theta}^*_1)^{r+1} = s(C_2 s^{\frac{1}{r}} 2^{\frac{r+1}{r}} (C\widehat{\theta}^*_1)^{r+1}).$$

Thus, $|R(\widehat{\theta}^*_x, \widehat{u}^*_x)| \le s\delta$ if $C_2 s^{\frac{1}{r}} 2^{\frac{r+1}{r}} (C\widehat{\theta}^*_1)^{r+1} < \delta$. This inequality holds if $0 < s < \delta^r/(2^{r+1}C_2^r(C\widehat{\theta}^*_1)^{(r+1)r}) = \varepsilon_2$. Let us choose $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$; then for any $s \in (0, \varepsilon)$ the operator G_s maps the set Q_s to itself.

Now let us prove that $G_s(x)$ is continuous. Let x_q be a sequence from Q_s and $x_q \to x \in Q_s$ as $q \to \infty$. Then

$$R_k(\widehat{\theta}^*_{x_q}, \widehat{u}^*_{x_q}) - R_k(\widehat{\theta}^*_x, \widehat{u}^*_x) = \Big(\Big(\frac{x_q}{x}\Big)^{\frac{k}{r}} - 1\Big) R_k(\widehat{\theta}^*_x, \widehat{u}^*_x) \tag{4.7}$$

since $\eta_I(\widehat{\theta}^*_{x_q}, \widehat{u}^*_{x_q}) = (\widehat{\theta}^*_{x_q}/\widehat{\theta}^*_x)^k \eta_I(\widehat{\theta}^*_x, \widehat{u}^*_x)$ for any $I \in M_k$ and $\widehat{\theta}^*_{x_q}/\widehat{\theta}^*_x = (x_q/x)^{\frac{1}{r}}$ due to Lemma 4.1. Moreover,

$$|R_k(\widehat{\theta}^*_x, \widehat{u}^*_x)| \le C_1((1+\delta)/2)^{\frac{k}{r}}, \quad |R_k(\widehat{\theta}^*_{x_q}, \widehat{u}^*_{x_q})| \le C_1((1+\delta)/2)^{\frac{k}{r}}$$

since $0 < x, x_q < s(1+\delta)$ and $(2s)^{\frac{1}{r}} C\widehat{\theta}^*_1 < 1$. Thus, for any $\varepsilon' > 0$ we can choose $N > r$ such that for any q

$$\sum_{k=N+1}^{\infty} |R_k(\widehat{\theta}^*_{x_q}, \widehat{u}^*_{x_q})| < \varepsilon'/4, \qquad \sum_{k=N+1}^{\infty} |R_k(\widehat{\theta}^*_x, \widehat{u}^*_x)| < \varepsilon'/4.$$

Then, taking into account (4.7), we choose q_0 such that for any $q > q_0$

$$|R_k(\widehat{\theta}^*_{x_q}, \widehat{u}^*_{x_q}) - R_k(\widehat{\theta}^*_x, \widehat{u}^*_x)| < \varepsilon'/(2(N-r)), \ k = r+1, \ldots, N.$$

Thus, for any $\varepsilon' > 0$ we can choose q_0 such that $|G_s(x) - G_s(x_q)| < \varepsilon'$ for all $q > q_0$, which means that G_s is continuous.

Thus, the continuous function maps the closed interval Q_s to itself, therefore, it has a fixed point in Q_s. Let us denote it by s^1, i.e., $G_s(s^1) = s^1$, which implies $s = s^1 + R(\widehat{\theta}^*_{s^1}, \widehat{u}^*_{s^1})$. However, $s^1 = \widehat{S}(\widehat{\theta}^*_{s^1}, \widehat{u}^*_{s^1})$. Thus, (4.6) implies $s = S(\widehat{\theta}^*_{s^1}, \widehat{u}^*_{s^1})$. This means that the control $\widehat{u}^*_{s^1}(t)$ steers the system (2.1) to the surface $h(x) = s$ in the time $\widehat{\theta}^*_{s^1}$. Therefore, the time-optimal problem (4.4) has a solution and

$$\theta^*_s \le \widehat{\theta}^*_{s^1}. \tag{4.8}$$

Now, let us consider the point $s^0 = s - R(\theta^*_s, u^*_s)$. The series $R(\theta^*_s, u^*_s)$ converges due to (4.8). Moreover, arguing as above we get that $|R(\theta^*_s, u^*_s)| \le C_2(C\theta^*_s)^{r+1} \le C_2(C\widehat{\theta}^*_{s^1})^{r+1} < s\delta$, which implies $s^0 \in Q_s$. Since $s = S(\theta^*_s, u^*_s) = \widehat{S}(\theta^*_s, u^*_s) + R(\theta^*_s, u^*_s)$, we get $s^0 = \widehat{S}(\theta^*_s, u^*_s)$, which implies $\widehat{\theta}^*_{s^0} \le \theta^*_s$. Thus, we get the two-sided estimate

$$\widehat{\theta}^*_{s^0} \le \theta^*_s \le \widehat{\theta}^*_{s^1},$$

which holds for any $s \in (0, \varepsilon)$. Then

$$\frac{\widehat{\theta}^*_{s^0}}{\widehat{\theta}^*_s} \leq \frac{\theta^*_s}{\widehat{\theta}^*_s} \leq \frac{\widehat{\theta}^*_{s^1}}{\widehat{\theta}^*_s}. \tag{4.9}$$

Recall that $\widehat{\theta}^*_{s^0}/\widehat{\theta}^*_s = (s^0/s)^{\frac{1}{r}}$ and $\widehat{\theta}^*_{s^1}/\widehat{\theta}^*_s = (s^1/s)^{\frac{1}{r}}$. Repeating the arguments given above we obtain

$$\frac{|s^1 - s|}{s} = \frac{|R(\widehat{\theta}^*_{s^1}, \widehat{u}^*_{s^1})|}{s} \leq \frac{(2s)^{\frac{r+1}{r}} C_2 (C\widehat{\theta}^*_1)^{r+1}}{s} = 2^{\frac{r+1}{r}} C_2 (C\widehat{\theta}^*_1)^{r+1} s^{\frac{1}{r}},$$

hence, $|s^1 - s|/s \to 0$ as $s \to +0$. This implies that $s^1/s \to 1$ and therefore $\widehat{\theta}^*_{s^1}/\widehat{\theta}^*_s = (s^1/s)^{\frac{1}{r}} \to 1$ as $s \to +0$. Analogously we get $\widehat{\theta}^*_{s^0}/\widehat{\theta}^*_s = (s^0/s)^{\frac{1}{r}} \to 1$ as $s \to +0$. Hence, (4.9) implies (4.5). □

Theorem 4.1 means that the optimal times for the series and its homogeneous approximation are equivalent in a neighborhood of the origin. This result partially generalizes the approximation theorem for control systems [18, Theorem 7.17], see (2.19). In that case, an approximation property holds also for optimal controls, see (2.20). For one-dimensional series a direct generalization is impossible since in typical situations an optimal control for the homogeneous approximation is not unique, which is demonstrated by Example 4.1.

However, we can prove the following property.

Lemma 4.2. *For any sequence $s_q \to +0$, let us consider a sequence of optimal controls $u^*_{s_q}(t) \in U^*_{s_q}$. Then there exists a subsequence s_{q_k} and a vector function $v(t) \in B^1$ such that*

$$\int_0^1 \left| u^*_{s_{q_k} i}(t\theta^*_{s_{q_k}}) - v_i(t) \right| dt \to 0 \quad as \quad k \to \infty, \quad i = 1, \ldots, m, \tag{4.10}$$

and, moreover, $v(t/\widehat{\theta}^*_1) \in \widehat{U}^*_1$.

Proof. Let us consider the sequence $v_q(t) = u^*_{s_q}(t\theta^*_{s_q})$ as a sequence in the Hilbert space $L_2([0,1]; \mathbb{R}^m)$. Since $v_q(t)$ belongs to the unit ball of the Hilbert space, it has a weakly convergent subsequence $v_{q_k}(t)$. Denote the weak limit by $v(t)$; one can show that $v(t) \in B^1$. On the other hand, any $\eta_I(1, u)$ is weakly continuous, hence, $\widehat{S}(1, v_{q_k}) \to \widehat{S}(1, v)$ as $k \to \infty$ [18].

We have $s_q = \widehat{S}(\theta^*_{s_q}, u^*_{s_q}) + R(\theta^*_{s_q}, u^*_{s_q})$. However, $\widehat{S}(\theta^*_{s_q}, u^*_{s_q})/(\theta^*_{s_q})^r = \widehat{S}(1, v_q)$ and $|R(\theta^*_{s_q}, u^*_{s_q})|/(\theta^*_{s_q})^r \leq C_2 C^{r+1} \theta^*_{s_q} \to 0$ as $s_q \to 0$. Besides, $s_q/(\theta^*_{s_q})^r = s_q/(\widehat{\theta}^*_{s_q})^r \cdot (\widehat{\theta}^*_{s_q}/\theta^*_{s_q})^r \to 1/(\widehat{\theta}^*_1)^r$ due to Lemma 4.1 and Theorem 4.1. Hence, taking the limit for the both sides of the following equality

$$\frac{s_{q_k}}{(\theta^*_{s_{q_k}})^r} = \frac{\widehat{S}(\theta^*_{s_{q_k}}, u^*_{s_{q_k}}) + R(\theta^*_{s_{q_k}}, u^*_{s_{q_k}})}{(\theta^*_{s_{q_k}})^r}$$

we get $1/(\widehat{\theta}_1^*)^r = \widehat{S}(1,v)$, which implies $1 = \widehat{S}(\widehat{\theta}_1^*, \widetilde{v})$, where $\widetilde{v}(t) = v(t/\widehat{\theta}_1^*)$, $t \in [0, \widehat{\theta}_1^*]$. This means that $\widetilde{v}(t) = v(t/\widehat{\theta}_1^*) \in \widehat{U}_1^*$.

Since $u^*_{s_{q_k}}(t\theta^*_{s_{q_k}})$ and $v(t)$ are time-optimal controls, they satisfy the equalities $\sum_{i=1}^m u^{*2}_{s_{q_k}i}(t) = 1$ and $\sum_{i=1}^m v_i^2(t) = 1$ a.e. [18]. Therefore, they belong to the boundary of the unit ball in the Hilbert space $L_2([0,1];\mathbb{R}^m)$. Hence, the weakly convergence of $u^*_{s_{q_k}}(t\theta^*_{s_{q_k}})$ to $v(t)$ implies the strong convergence, which in turn implies (4.10) with $v(t/\widehat{\theta}_1^*) \in \widehat{U}_1^*$. We notice that this means that $v(t) \in \widehat{U}_{\bar{s}}^*$ for $\bar{s} = 1/(\widehat{\theta}_1^*)^r$. □

The proof of Lemma 4.2 shows that any weak partial limit of the sequence $u^*_{s_q}(t\theta^*_{s_q})$ tends to some optimal control $v(t)$ of the problem (4.1). If the optimal control is unique, then the sequence $u^*_{s_q}(t\theta^*_{s_q})$ tends to this control. However, as was mentioned above, typically an optimal control is not unique. However, we can specify the result of Lemma 4.2 for one important case.

Theorem 4.2. *Let us consider the time-optimal problems* (4.4) *and* (4.1) *for the minimal realization of the series S and for its homogeneous approximation $\widehat{S}$ respectively. Suppose the problem* (4.1) *has a solution for $s = 1$ and, moreover, the set of optimal controls $\widehat{U}_1^*$ is finite. Then for any sequence $s_q \to +0$ and any sequence of optimal controls $u^*_{s_q}(t) \in U^*_{s_q}$ there exists a sequence $\widehat{u}^*_{s_q}(t) \in \widehat{U}^*_{s_q}$ such that*

$$\int_0^1 \left| u^*_{s_q i}(t\theta^*_{s_q}) - \widehat{u}^*_{s_q i}(t\widehat{\theta}^*_{s_q}) \right| dt \to 0 \quad \text{as} \quad s_q \to +0, \quad i = 1, \dots, m. \tag{4.11}$$

For $s_q \to -0$ the analogous result holds.

Proof. By our assumption, the set $\widehat{U}_1^*$ contains a finite number of elements; denote them by $w_j(t/\widehat{\theta}_1^*)$, $t \in [0, \theta_1^*]$, $j = 1, \dots, N$. Then the set $\widehat{U}_{\bar{s}}^* \subset B^1$ for $\bar{s} = 1/(\widehat{\theta}_1^*)^r$ consists of the elements $w_1(t), \dots w_N(t)$, $t \in [0,1]$.

Let us consider all controls as elements of the Hilbert space $L_2([0,1];\mathbb{R}^m)$; denote by $\|\cdot\|$ the norm in this space, i.e., $\|u\| = (\int_0^1 \sum_{i=1}^m u_i^2(t)dt)^{1/2}$. Obviously, there exists $\varepsilon > 0$ such that $\|w_j - w_k\| > \varepsilon$ for any $j \neq k$.

On the other hand, as was shown in the proof of Lemma 4.2, any weak partial limit of the sequence $u^*_{s_q}(t\theta^*_{s_q})$ is a strong partial limit and belongs to the set $\widehat{U}_{\bar{s}}^*$, i.e., coincides with one of the controls $w_1(t), \dots w_N(t)$. Therefore, there exists q_0 such that for any $q > q_0$ there exists a unique $j = j(q)$ such that $\|u^*_{s_q}(t\theta^*_{s_q}) - w_{j(q)}(t)\| < \varepsilon/2$. We notice that $w_{j(q)}(t/\widehat{\theta}^*_{s_q}) \in \widehat{U}^*_{s_q}$, therefore, we can use the notation $w_{j(q)}(t/\widehat{\theta}^*_{s_q}) = \widehat{u}^*_{s_q}(t)$, that is, $w_{j(q)}(t) = \widehat{u}^*_{s_q}(t\widehat{\theta}^*_{s_q})$. Thus, for any s_q (with $q \geq q_0$) we have chosen the unique control $\widehat{u}^*_{s_q}(t) \in \widehat{U}^*_{s_q}$. Obviously, for any $\varepsilon' \in (0, \varepsilon/2)$ there exists $q_0' \geq q_0$ such that $\|u^*_{s_q}(t\theta^*_{s_q}) - \widehat{u}^*_{s_q}(t\widehat{\theta}^*_{s_q})\| < \varepsilon'$ for any $q > q_0'$, which means that

$$\|u^*_{s_q}(t\theta^*_{s_q}) - \widehat{u}^*_{s_q}(t\widehat{\theta}^*_{s_q})\| \to 0 \text{ as } q \to \infty. \tag{4.12}$$

Finally, (4.12) obviously implies (4.11). □

Example 4.2. Let us again consider the system (4.3) but define the output as $y = h(x) = x_1 + x_2$. The corresponding series $S = \eta_1 + \eta_{21}$ is not homogeneous. Its homogeneous approximation is $\widehat{S} = \eta_1$ and its minimal realization is one-dimensional, namely, $\dot{x}_1 = u_1$, $y = \widehat{h}(x) = x_1$. If $s > 0$, then the optimal control obviously equals $\widehat{u}^*_s(t) = (1, 0)$, $t \in [0, \widehat{\theta}^*_s]$, where $\widehat{\theta}^*_s = s$.

For the system (4.3), arguing similarly to Example 4.1 and applying the transversality conditions we get two equations

$$\tfrac{\alpha}{\xi}\cos(\xi T) = 1, \quad \tfrac{\alpha}{\xi}\sin(\xi T) + \tfrac{T}{2\xi} - \tfrac{1}{4\xi^2}\sin(2\xi T) = s.$$

Hence, ξT equals the root of the equation $F(z) = s$, where $F(z) = \frac{\sin z}{2\cos z} + \frac{z}{2\cos^2 z}$. The function $F(z)$ strictly increases for $z \in [0, \frac{\pi}{2})$, $F(z) \to +\infty$ as $z \to \frac{\pi}{2}$, and $F'(0) = 1$. Hence, the equation $F(z) = s$ has a unique solution $\bar{z}(s) > 0$ for any $s > 0$ and $\bar{z}'(0) = 1$. The optimal time equals $\theta^*_s = T = \bar{z}(s)/\cos(\bar{z}(s))$; using L'Hôpital's rule one can show that $\lim_{s\to+0} \frac{\theta^*_s}{s} = 1$. This means that $\lim_{s\to+0} \frac{\theta^*_s}{\widehat{\theta}^*_s} = 1$, which illustrates Theorem 4.1. For the optimal control, we get $u^*_s(t\theta^*_s) = (\cos(\bar{z}(s)t), \sin(\bar{z}(s)t))$, $t \in [0, 1]$. Therefore,

$$\int_0^1 \left| u^*_{s\,1}(t\theta^*_s) - \widehat{u}^*_{s\,1}(t\widehat{\theta}^*_s) \right| dt = \int_0^1 |\cos(\bar{z}(s)t) - 1|\, dt \le 1 - \cos(\bar{z}(s)) \to 0,$$
$$\int_0^1 \left| u^*_{s\,2}(t\theta^*_s) - \widehat{u}^*_{s\,2}(t\widehat{\theta}^*_s) \right| dt = \int_0^1 |\sin(\bar{z}(s)t)|\, dt \le \sin(\bar{z}(s)) \to 0$$

as $s \to +0$ since $\bar{z}(s) \to +0$, which illustrates Theorem 4.2.

References

1. A. Agrachev, D. Barilari, U. Boscain, *A comprehensive introduction to sub-Riemannian geometry. Cambridge University Press*, 2020.
2. D. M. Andreieva, S. Yu. Ignatovich, *Homogeneous approximation for minimal realizations of series of iterated integrals*, *Visnyk of V. N. Karazin Kharkiv National University. Ser. Math., Applied Math. and Mech.*, **96** (2022), 23–39.
3. A. Bellaïche, *The tangent space in sub-Riemannian geometry*, in Progress in Mathematics, A. Bellaïche and J.J. Risler, eds., 144, *Birkhäuser Basel*, 1996, 1–78.
4. P. E. Crouch, *Solvable approximations to control systems*, *SIAM J. Control Optimiz.*, **22** (1984), 40–54.
5. A. F. Filippov, *On certain questions in the theory of optimal control*, *J. SIAM Control Ser. A*, **1** (1962), 76–84.
6. M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, *Bull. Soc. Math. France*, **109** (1981), P. 3–40.
7. H. Hermes, *Nilpotent and high-order approximations of vector field systems*, *SIAM Rev.*, **33** (1991), 238–264.
8. S. Yu. Ignatovich, *Realizable growth vectors of affine control systems*, *J. Dyn. Control Syst.*, **15** (2009), 557–585.
9. A. Isidori, *Nonlinear control systems. Springer-Verlag, London*, 1995.
10. B. Jakubczyk, *Existence and uniqueness of realizations of nonlinear systems*, *SIAM J. Control and Optimiz.*, **18** (1980), 455–471.

11. M. Kawski, *Combinatorial algebra in controllability and optimal control*, in Algebra and Applications 2: Combinatorial Algebra and Hopf Algebras, A. Makhlouf, ed., Chapter 5, 2021, 221–286.

12. M. Kawski, H. J. Sussmann, *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in Operators, Systems and Linear Algebra. European Consortium for Mathematics in Industry, U. Helmke, D. Prätzel-Wolters, E. Zerz, eds., *Teubner*, 1997, 111–128.

13. V. I. Korobov, G. M. Sklyar, *The Markov moment problem on the smallest possible interval*, *Sov. Math. Dokl.*, **40** (1990), 334–337.

14. S. M. LaValle, *Planning algorithms*, *Cambridge Univ. Press*, 2006.

15. G. Melançon, C. Reutenauer, *Lyndon words, free algebras and shuffles*, *Canad. J. Math.*, **41** (1989), 577–591.

16. C. Reutenauer, *Free Lie algebras*, *Clarendon Press, Oxford*, 1993.

17. G. M. Sklyar, S. Yu. Ignatovich, *Approximation of time-optimal control problems via nonlinear power moment min-problems*, *SIAM J. Control Optimiz.*, **42** (2003), 1325–1346.

18. G. M. Sklyar, S. Yu. Ignatovich, *Free algebras and noncommutative power series in the analysis of nonlinear control systems: an application to approximation problems*, *Dissertationes Math. (Rozprawy Mat.)*, **504** (2014), 1–88.

19. G. Sklyar, P. Barkhayev, S. Ignatovich, V. Rusakov, *Implementation of the algorithm for constructing homogeneous approximations of nonlinear control systems*, *Mathematics of Control, Signals, and Systems*, **34** (2022), 883–907.

20. G. Stefani, *Polynomial approximations to control systems and local controllability*, in 1985 24th IEEE Conference on Decision and Control, 1985, 33–38.

Received 10.11.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 24–49, DOI 10.15421/142309

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

SYSTEMS OF SINGULAR DIFFERENTIAL EQUATIONS AS THE BASIS FOR NEURAL NETWORK MODELING OF CHAOTIC PROCESSES

Vasiliy Ye. Belozyorov[*], Oleksandr A. Inkin[†]

Abstract. Currently, systems of neural ordinary differential equations (ODEs) have become widespread for modeling various dynamic processes. However, in forecasting tasks, priority remains with the classical neural network approach to building a model. This is due to the fact that by choosing the neural network architecture, a more accurate approximation of the trajectories of a dynamic system can be achieved. It is known that the accuracy of the mentioned approximation significantly depends on the settings of the neural network parameters and their initial values. In this regard, the main idea of the article is that the initial values of the neural network parameters are taken to be equal to the parameters of the neural ODE system obtained by modeling the same process, which will then be simulated using a neural network. Subsequently, the singular ODE system was used to adjust the parameters of the LSTM (Long Short Term Memory) neural network. The results obtained were used to model the process of epilepsy.

Key words: time series, system of differential equations, compact region of attraction, neural network.

2010 Mathematics Subject Classification: 34A34, 34D20, 37D45, 93A30.

Communicated by Prof. V. Kapustyan

1. Introduction

Let

$$x_0 = x(t_0), x_1 = x(t_1), ..., x_N = x(t_N) \tag{1.1}$$

be a finite sequence (time series) of numerical values of some scalar dynamical variable $x(t)$ measured with the constant time step Δt in the moments $t_i = t_0 + i\Delta t$; $x_i = x(t_i)$; $i = 0, 1, ..., N$ (thus, $\Delta t = t_N/N$) [5, 6, 9, 11, 12, 16, 21].

The choice of equations for a model that describes the dynamics of certain processes is a difficult task. Experiments show that the most logical approach to constructing models that describe the dynamics of the passage of electrical

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

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

signals through certain objects is based on the use of well-known physical laws (for example Ohm, Maxwell, Joule-Lenz, the law of conservation of energy), in which the interaction between measured quantities is described with using quadratic functions.

In addition, in rapidly oscillating processes there is a sharp change in the sign of the derivative. It is this characteristic that most often determines chaotic processes. Therefore, we believe that a sufficiently informative model of chaos can be described by differential equations, on the right sides of which there are rational functions with quadratic functions in the numerator and periodic functions that take sufficiently small non-zero values in the denominator. Such ODE systems are called singular [9].

In order to construct the mentioned system of differential equations using a known time series (1.1), it is necessary to know its dimension. The last characteristic (dimension n of the embedding space) and the optimal time delay τ (at which time t must be shifted to obtain a new variable $y(t_i) = x(t_i + \tau)$) can be determined using recurrent qualitative analysis (RQA) methods [21]. (Note that the number τ must be such that $t_i + \tau \in \{t_0, t_1, ..., t_N\}; i \in \{0, ..., N - (n-1) \cdot \tau\}$.)

Having parameters n and τ, we can assume that to model a process described by time series (1.1), a certain system of differential equations has already built. (In what follows, we will assume that a recurrent neural network (RNN), which is a discrete analogue of the mentioned ODE system, was also constructed [4, 6, 11, 16, 21].)

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 [17, 18]. 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 [3, 7, 8].)

The final sections of the article are devoted to the development of an algorithm for determining the parameters of ODE systems for a known time series. The essence of this algorithm is that it uses a special structure of neural ODEs (antisymmetric neural ODEs), with which it is possible to generate a limit cycle [6, 10, 13]. After this, by selecting weight coefficients, we obtain such bifurcations of the indicated cycle that lead to the modeling of a real chaotic process. Subsequently, the found weighting coefficients are used as initial data for adjusting the parameters of the LSTM neural network [1].

2. Mathematical preliminaries

By $\mathbf{x} = (x_1, ..., x_n)^T$ it denotes an arbitrary vector of real space $\mathbb{R}^n$. Consider the real system of ordinary autonomous differential equations

$$\begin{cases} \dot{x}_1(t) = \dfrac{f_1(x_1(t), ..., x_n(t))}{1 - \vartheta \cdot u_1(x_1(t), ..., x_n(t))}, \\ \ldots\ldots\ldots\ldots\ldots\ldots , \\ \dot{x}_n(t) = \dfrac{f_n(x_1(t), ..., x_n(t))}{1 - \vartheta \cdot u_n(x_1(t), ..., x_n(t))} \end{cases} \tag{2.1}$$

of order n with the vector of initial data $\mathbf{x}^T(0) = (x_{10}, ..., x_{n0})$. Here $f_i(x_1, ..., x_n)$, $u_i(x_1, ..., x_n)$; $i = 1, ..., n$, are continuous functions of their arguments, and for functions $u_i(x_1, ..., x_n)$ the condition

$$\Omega = \max_{1 \le i \le n} \sup_{\|\mathbf{x}\| \to \infty} (|u_i(x_1, ..., x_n)|) < \infty$$

is satisfied. In addition, ϑ is a real parameter such that $0 \le |\vartheta| < 1/\Omega$.

Definition 2.1. System (2.1) will be called singular.

Let $A = (a_{ij}), B_1, ..., B_n \in \mathbb{R}^{n \times n}$ be real matrices. In addition, let the matrices $B_1 = (b_{ij}^{(1)}), ..., B_n = (b_{ij}^{(n)})$ be symmetrical; $i, j = 1, ..., n$. Let us consider one special case of system (2.1):

$$\begin{cases} \dot{x}_1(t) = \dfrac{\sum\limits_{j=1}^{n} a_{1j} x_j(t) + \mathbf{x}^T(t) B_1 \mathbf{x}(t)}{1 - \vartheta \cdot u_1(x_1(t), ..., x_n(t))}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots , \\ \dot{x}_n(t) = \dfrac{\sum\limits_{j=1}^{n} a_{nj} x_j(t) + \mathbf{x}^T(t) B_n \mathbf{x}(t)}{1 - \vartheta \cdot u_n(x_1, ..., x_n(t))}. \end{cases} \tag{2.2}$$

Below we recall some of the results obtained in [7,9].

Consider the system of ordinary autonomous quadratic differential equations

$$\begin{cases} \dot{x}_1(t) = \sum\limits_{j=1}^{n} a_{1j} x_j(t) + \mathbf{x}^T(t) B_1 \mathbf{x}(t), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots , \\ \dot{x}_n(t) = \sum\limits_{j=1}^{n} a_{nj} x_j(t) + \mathbf{x}^T(t) B_n \mathbf{x}(t). \end{cases} \tag{2.3}$$

Assume that the region of attraction for the solutions of system (2.3) is a ball

$$\mathbb{B} \equiv (x_1 + \gamma_1)^2 + ... + (x_n + \gamma_n)^2 - R^2 \le 0$$

of radius R with center at point $(-\gamma_1, ..., -\gamma_n)^T$.

Let also the elements of matrices $B_1, ..., B_n$ satisfy the following three groups of restrictions:

C_n^1 one-term restrictions

$$b_{ii}^{(i)} x_i^3 \equiv 0; i = 1, ..., n; \tag{2.4}$$

$2C_n^2$ two-term restrictions

$$b_{jj}^{(i)} x_i x_j^2 + b_{ij}^{(j)} x_i x_j^2 \equiv 0; i \neq j; i, j = 1, ..., n; \tag{2.5}$$

C_n^3 three-term restrictions

$$b_{jk}^{(i)} x_i x_j x_k + b_{ik}^{(j)} x_i x_j x_k + b_{ij}^{(k)} x_i x_j x_k \equiv 0; i \neq j \neq k; i, j, k = 1, ..., n. \tag{2.6}$$

As shown in [5], for small values of n system (2.3), taking into account restrictions (2.4), (2.5), and (2.6), has the following form:

$n = 3$

$$\begin{cases} \dot{x}(t) = a_{11}x + \cdots + a_{13}z + b_{12}xy + b_{13}xz + b_{22}y^2 + b_{23}yz + b_{33}z^2, \\ \dot{y}(t) = a_{21}x + \cdots + a_{23}z - b_{12}x^2 - b_{22}xy + c_{13}xz + c_{23}yz + c_{33}z^2, \\ \dot{z}(t) = a_{31}x + \cdots + a_{33}z - b_{13}x^2 - (b_{23} + c_{13})xy - b_{33}xz - c_{23}y^2 - c_{33}yz; \end{cases} \tag{2.7}$$

$n = 4$

$$\begin{cases} \dot{x}(t) = a_{11}x + \cdots + a_{14}u + b_{12}xy + b_{13}xz + b_{14}xu + b_{22}y^2 \\ \qquad +b_{23}yz + b_{24}yu + b_{33}z^2 + b_{34}zu + b_{44}u^2, \\ \dot{y}(t) = a_{21}x + \cdots + a_{24}u - b_{12}x^2 - b_{22}xy + c_{13}xz + c_{14}xu \\ \qquad +c_{23}yz + c_{24}yu + c_{33}z^2 + c_{34}zu + c_{44}u^2, \\ \dot{z}(t) = a_{31}x + \cdots + a_{34}u - b_{13}x^2 - (b_{23} + c_{13})xy - b_{33}xz \\ \qquad +d_{14}xu - c_{23}y^2 - c_{33}yz + d_{24}yu + d_{34}zu + d_{44}u^2, \\ \dot{u}(t) = a_{41}x + \cdots + a_{44}u - b_{14}x^2 - (b_{24} + c_{14})xy - (b_{34} + d_{14})xz \\ \qquad -b_{44}xu - c_{24}y^2 - (c_{34} + d_{24})yz - c_{44}yu - d_{34}z^2 - d_{44}zu. \end{cases} \tag{2.8}$$

Note that equations (2.7) – (2.8) are presented in this detailed form solely for the convenience of users. In the case of arbitrary n, the system that satisfies the conditions (2.4) – (2.6) looks like this:

$$\dot{\mathbf{x}}(t) = (A + B(\mathbf{x}) - B^T(\mathbf{x})) \cdot \mathbf{x}. \tag{2.9}$$

Here

$$B(\mathbf{x}) = \begin{pmatrix} 0 & b_{12}^1 x_1 + b_{22}^1 x_2 & b_{13}^1 x_1 + b_{23}^1 x_2 + b_{33}^1 x_3 & ... & \sum\limits_{i=1}^{n} b_{in}^1 x_i \\ 0 & 0 & b_{13}^2 x_1 + b_{23}^2 x_2 + b_{33}^2 x_3 & ... & \sum\limits_{i=1}^{n} b_{in}^2 x_i \\ 0 & 0 & 0 & ... & \sum\limits_{i=1}^{n} b_{in}^3 x_i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \sum\limits_{i=1}^{n} b_{in}^{n-1} x_i \\ 0 & 0 & 0 & ... & 0 \end{pmatrix},$$

$b_{ij}^k \in \mathbb{R}; i, j, k \in \{1, ..., n\}$. (It is clear that $\mathbf{x}^T \cdot (B(\mathbf{x}) - B^T(\mathbf{x})) \cdot \mathbf{x} \equiv 0$.)

The method for finding the radius R of sphere $\mathbb{B}$ and its center $(-\gamma_1, ..., -\gamma_n)^T$ is presented in [5].

Below we will use the following well-known result:

Theorem 2.1. *(***LaSalle's Theorem** *[14]). Let $\mathbb{H} \subset \mathbb{R}^n$ be a compact set that is positively invariant with respect to (2.2). Let $V : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function such that $\dot{V}(\mathbf{x}) \leq 0$ (or $\dot{V}(\mathbf{x}) \geq 0$) in $\mathbb{H}$. Let $\mathbb{E}$ be the set of all points in $\mathbb{H}$ where $\dot{V}(\mathbf{x}) = 0$. Let $\mathbb{M}$ be the largest invariant set in $\mathbb{E}$. Then every solution starting in $\mathbb{H}$ approaches $\mathbb{M}$ as $t \to +\infty$.*

Let $s_1, ..., s_n$ be unknown real constants. Let us construct from matrices A and $B_1, ..., B_n$ of system (2.2) the following matrix:

$$F(s_1, ..., s_n) := (A^T + A)/2 + s_1 B_1 + ... + s_n B_n.$$

Introduce also the following function $V : \mathbb{R}^n \to \mathbb{R}$:

$$V(x_1, ..., x_n) = \frac{1}{2} \sum_{i=1}^{n} \int (1 - \vartheta \cdot u_i(x_1, ..., x_n))(x_i + s_i)\, dx_i,$$

where $\forall i\ (1 - \vartheta \cdot u_i(x_1, ..., x_n)) > 0$. (The indefinite integral symbol is used here.)

It is obvious that

$$\dot{V}_t = \frac{1}{2} \sum_{i=1}^{n} (1 - \vartheta \cdot u_i(x_1, ..., x_n))(x_i + s_i)\dot{x}_i$$

$$= \mathbf{x}^T (\frac{1}{2}(A + A^T) + \sum_{i=1}^{n} s_i B_i)\mathbf{x} + L(\mathbf{x}), \quad (2.10)$$

(Here the derivative $\dot{V}_t$ is defined by virtue of the equations (2.2); $L(\mathbf{x})$ is a cubic function of variables $x_1, ..., x_n$ without quadratic terms.)

Let R be a positive constant. We define the set $\mathbb{B}_R \subset \mathbb{R}^n$ as follows:

$$\mathbb{B}_R := \{(x_1, ..., x_n) \in \mathbb{R}^n | V(x_1, ..., x_n) - R^2 \leq 0\}. \quad (2.11)$$

Introduce the following sets:

$$\mathbb{D}_- := \{(x_1, ..., x_n) \in \mathbb{R}^n | \dot{V}_t(x_1, ..., x_n) \leq 0\}, \quad (2.12)$$

$$\mathbb{D}_+ := \{(x_1, ..., x_n) \in \mathbb{R}^n | \dot{V}_t(x_1, ..., x_n) \geq 0\}, \quad (2.13)$$

and

$$\mathbb{L} := \{(x_1, ..., x_n) \in \mathbb{R}^n | \dot{V}_t(x_1, ..., x_n) = 0\}. \quad (2.14)$$

Theorem 2.2. *Let's assume that for system (2.2) the following conditions:*

1) the matrices B_i satisfy the restriction $\mathbf{x}^T(B(\mathbf{x}) - B^T(\mathbf{x}))\mathbf{x} \equiv 0$;

2) there are real constants $s_1^*, ..., s_n^*$ such that the matrix $F(s_1^*, ..., s_n^*)$ is negative definite, are satisfied.

Then there exists the compact region of attraction $\mathbb{H} = \mathbb{D}_+ \neq \emptyset$ for trajectories of system (2.2).

Proof. Condition 1) guarantees that the function $\dot{V}_t(x_1, ..., x_n)$ contains only linear and quadratic terms and does not contain cubic terms.

It remains only to clarify condition 2). So, let there exist numbers $s_1^*, ..., s_n^*$ such that the matrix $F(s_1^*, ..., s_n^*)$ is negative definite.

Proof of condition 2) split into two parts.

2a) The function $V(x_1, ..., x_n)$ for $s_i = 0$ positive definite and in this case $\lim\limits_{\|\mathbf{x}\| \to \infty} V(x_1, ..., x_n) = \infty$. Thus, by virtue of the construction of the function $V(x_1, ..., x_n)$, the sets $\mathbb{B}_R$ and $\mathbb{L}$ are compact. Therefore, we can choose R such that $\mathbb{B}_R \cap \mathbb{D}_- \neq \emptyset$ and $\mathbb{L} \subset \mathbb{B}_R$. Then, in the region $\mathbb{B}_R \cap \mathbb{D}_-$, we can assert that $V(x_1(t), ..., x_n(t))$ is a decreasing function of t. Since $V(x_1(t), ..., x_n(t))$ is continuous on the compact set $\mathbb{B}_R$, it is bounded from below on $\mathbb{B}_R$. Therefore, $V(x_1(t), ..., x_n(t))$ has a finite limit as $t \to \infty$. Then, according to Theorem (2.1), we can assume that $\mathbb{H} = \mathbb{B}_R \cap \mathbb{D}_-$ and $\mathbb{H}$ is a the compact region of attraction for trajectories of system (2.2).

2b) Now we choose the radius R so large that the set $\mathbb{B}_R \cap \mathbb{D}_+ = \mathbb{D}_+ \neq \emptyset$. (Note that, by virtue of 2), the set $\mathbb{D}_+$ is compact. Therefore, we have $\mathbb{L} \subset \mathbb{B}_R$.) Then, in the domain $\mathbb{B}_R \cap \mathbb{D}_+$ the function $V(x_1(t), ..., x_n(t))$ is an increasing function of t. Since the function $V(x_1(t), ..., x_n(t))$ is continuous on the compact set $\mathbb{D}_+$, it is bounded from above on $\mathbb{D}_+$ and has a finite limit as $t \to \infty$.

Thus, from items 2a) and 2b) it follows that, regardless of the starting point $\mathbf{x}^T(0) \in \mathbb{R}^n$, the trajectory $V(x_1(t)), ..., x_n(t))$ will be attracted to the boundary $\dot{V}_t(x_1, ..., x_n) = 0$ (this is $\mathbb{L}$) of the compact set $\mathbb{D}_+$. This means that there exists an attractor belonging to the region $\mathbb{D}_+$. (An equilibrium point can act as such attractor.) □

2.1. Model design

This article is a continuation of work [9]. Therefore, the motives leading to certain results are based on the assumptions introduced in article [9].

The main object of study in this work will be the electroencephalograms (EEGs) of the brain of patients suffering from epilepsy. Now, we will consider EEGs obtained for healthy and sick patients (see Fig.2.1). (The main features of the processes Fig.2.1 are described in [20].)

Naturally, when modeling the real process of epilepsy, it is impossible to take into account all these features. However, we will try to at least establish the trend accompanying such processes.

The latest considerations allow us to use system (2.2) for modeling the processes represented on encephalograms, in which $cos(...)$ is used as functions $u_i(...)$; $i = 1, ..., n$. The final appearance of this system is as follows:

$$\left\{ \begin{array}{l} \dot{x}_1(t) = \dfrac{\displaystyle\sum_{j=1}^{n} a_{1j}x_j(t) + \mathbf{x}^T(t)B_1\mathbf{x}(t)}{1 - \vartheta \cdot \cos(x_1(t))}, \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; , \\ \dot{x}_n(t) = \dfrac{\displaystyle\sum_{j=1}^{n} a_{nj}x_j(t) + \mathbf{x}^T(t)B_n\mathbf{x}(t)}{1 - \vartheta \cdot \cos(x_n(t))}, \end{array} \right. \tag{2.15}$$

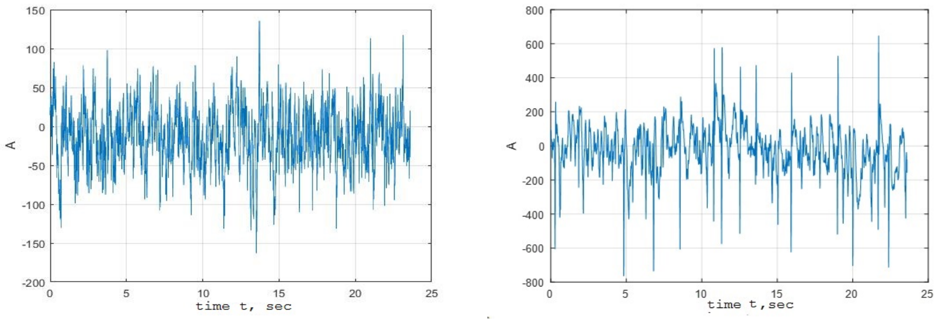
where ϑ is a real parameter such that $0 \leq |\vartheta| < 1/\Omega = 1/1 = 1$. (Note that the simplest case of system (2.15), in which $B_1 = ... = B_n = 0$ was investigated in [9].)

In what follows, instead of the system (2.15), we will sometimes consider the system

$$\left\{ \begin{array}{l} \dot{x}_1(t) = \dfrac{a_{10} + \displaystyle\sum_{j=1}^{n} a_{1j}x_j(t) + \mathbf{x}^T(t)B_1\mathbf{x}(t)}{1 - \vartheta \cdot \cos(x_1(t))}, \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; , \\ \dot{x}_n(t) = \dfrac{a_{n0} + \displaystyle\sum_{j=1}^{n} a_{nj}x_j(t) + \mathbf{x}^T(t)B_n\mathbf{x}(t)}{1 - \vartheta \cdot \cos(x_n(t))}, \end{array} \right. \tag{2.16}$$

where $a_{10}, ..., a_{n0} \in \mathbb{R}$.

We assume that $x_1 = \phi_1, ..., x_n = \phi_n$ is a real solution to the system of



(a1) (a2)

Fig. 2.1. The electroencephalogram taken from a certain point in the cerebral cortex: (a1) a healthy patient, (a2) a patient with an epileptic disease (see [20]).

algebraic equations

$$a_{10}+\sum_{j=1}^{n}a_{1j}x_j+\mathbf{x}^TB_1\mathbf{x}=0,...,a_{n0}+\sum_{j=1}^{n}a_{nj}x_j+\mathbf{x}^TB_n\mathbf{x}=0.$$

Let us introduce new variables $y_1,...,y_n$ into system (2.16) using the formulas: $x_1=y_1+\phi_1,...,x_n=y_n+\phi_n$. Then we have

$$\begin{cases} \dot{y}_1(t)=\dfrac{\displaystyle\sum_{j=1}^{n}c_{1j}y_j(t)+\mathbf{y}^T(t)B_1\mathbf{y}(t)}{1-\vartheta\cdot\cos(y_1(t)+\phi_1)}, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ , \\ \dot{y}_n(t)=\dfrac{\displaystyle\sum_{j=1}^{n}c_{nj}y_j(t)+\mathbf{y}^T(t)B_n\mathbf{y}(t)}{1-\vartheta\cdot\cos(y_n(t)+\phi_n)}, \end{cases} \tag{2.17}$$

where $c_{ij}\in\mathbb{R}; i,j=1,...,n$. Then the statements of Theorems 2.1 and 2.2 can be applied to system (2.17) (and therefore (2.16)).

3. Two-stage neural network modeling procedure

It is known that one of the most common methods for adjusting the weighting coefficients of a neural network is the steepest descent method. This method is a type of gradient descent, which means it descends the error surface, continuously adjusting the weights towards the minimum. The error surface of a complex network is highly rugged and consists of hills, valleys, folds and ravines in high-dimensional space. A network may fall into a local minimum (shallow valley) when there is a much deeper minimum nearby. At the local minimum point, all directions lead upward, and the network is unable to escape from it. The main difficulty in training neural networks is precisely the methods for exiting local minima: each time you exit a local minimum, the next local minimum is again searched until it is no longer possible to find a way out of it.

In this regard, the following two-stage modeling method suggests finding an initial point in the space of weighting coefficients (parameters) at which the error function would be as close as possible to the local minimum point. Subsequently, the found point is used as a starting point for adjusting the parameters of some recurrent neural network (the LSTM neural network) using the back propagation method.

3.1. First stage

At this stage, we will try to solve the problem of parametric identification of system (2.16).

Let us write the equations of system (2.16) in the following form

$$\dot{x}_i(t) = \frac{a_{i0} + a_{i1}x_i + \cdots + a_{in}x_n + \mathbf{x}^T B_i \mathbf{x}}{1 - \vartheta \cdot \cos(x_i)} = \phi_i(x_1, ..., x_n); i = 1, ..., n. \tag{3.1}$$

Now we rewrite the equations of system (3.1) as follows

$$\dot{x}_i(t) = a_{i0} + a_{i1}x_i + \cdots + a_{in}x_n + \mathbf{x}^T B_i \mathbf{x} + \dot{x}_i \vartheta \cdot \cos(x_i) = \psi_i(x_1, ..., x_n); i = 1, ..., n. \tag{3.2}$$

From the point of view of the theory of differential equations, systems (3.1) and (3.2) describe the same dynamics. However, from the point of view of approximation theory (determining the coefficients $a_{i0}, ..., a_{in}, B_i, \vartheta$ from the known values of the functions $x_i(t), i = 1, ..., n$), these are different problems for systems (3.1) and (3.2).

Indeed, in case of system ((3.1)) it is necessary to minimize by $a_{10}, ..., B_n, \vartheta$ the loss function $\sum_{i=1}^{n} |\dot{x}_i - \phi_i(x_1, ..., x_n, a_{10}, ..., B_n, \vartheta)|$, and in case of system (3.2) it is necessary to minimize by $a_{10}, ..., B_n, \vartheta$ the loss function $\sum_{i=1}^{n} |\dot{x}_i - \psi_i(x_1, ..., x_n, a_{10}, ..., B_n, \vartheta)|$, where the equations (3.1) are rational and the equations (3.2) are linear.

It is clear that in the case of system (3.2), the approximation problem will be simpler than in the case of system (3.1). That is why we chose system (3.2) for solving the approximation problem. (It should be remembered that the approximation results for system (3.2) may be worse than for system (3.1).)

In the study of dynamic processes, as a rule, only a few variables describing the process are available for direct measurement. The remaining variables (the so-called hidden variables) are inaccessible to observation. This raises the problem of reconstructing these unobserved variables from known observable variables. The first step towards solving this problem is to establish the minimum number of all variables (measured and hidden) on which the dynamic process depends.

In article [9] it was shown that for time series obtained using EEG, the dimension of the embedding space is $n = 5$. This means that in addition to the measured variable, it is also necessary to recover 4 hidden variables. Therefore, in the following we will demonstrate the modeling procedure only for a $5D$ system.

In addition, we will assume that the non-diagonal elements of matrix $A = \{a_{ij}\}; i, j = 1, ..., n; i \neq j$, form an antisymmetric matrix, and the elements of matrices $B_1, ..., B_n$ satisfy the condition $\mathbf{x}^T (B(\mathbf{x}) - B^T(\mathbf{x}))\mathbf{x} \equiv 0$ (see Theorem 2.2). (Transferring the algorithm to an arbitrary n is not difficult.)

3.2. Algorithm for quadratic model

In order for system (3.1) to be stable, we introduce into it the diffusion parameter $\mu > 0$ [10, 13]. Then, we have

$$\dot{\mathbf{x}}(t) = T \cdot (\mathbf{a}_0 + (A - \mu I) \cdot \mathbf{x} + K(\mathbf{x}) \cdot \mathbf{x}), \tag{3.3}$$

where $\mathbf{x} = (x_1, ..., x_n)^T$,

$$K(\mathbf{x}) = \begin{pmatrix} k_{11}(\mathbf{x}) & b_{12}x_2 & b_{13}x_3 & \cdots & b_{1n}x_n \\ b_{21}x_1 & k_{22}(\mathbf{x}) & b_{23}x_3 & \cdots & b_{2n}x_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1}x_1 & b_{n2}x_2 & b_{n3}x_3 & \cdots & k_{nn}(\mathbf{x}) \end{pmatrix},$$

$$k_{ii}(\mathbf{x}) = -\sum_{j=1, j\neq i}^{n} (b_{ij}x_j); i = 1, ..., n.$$

This system is a complicated version of system

$$\dot{\mathbf{x}}(t) = \sigma[T \cdot ((\mathbf{W} - \mu I)\mathbf{x}(t) + \mathbf{V}\mathbf{u}(t) + \mathbf{b})], \tag{3.4}$$

where $\mathbf{x}$ is the hidden state, $\mathbf{u}$ is the input, and σ is the activation function. (Thus, we can assume that system (3.3) is system (3.4) closed by nonlinear state feedback $\mathbf{u} = K(\mathbf{x})\mathbf{x}$, where V is the identity matrix.)

In the case $n = 5$, system (3.3) takes the following form:

$$\begin{cases} \dot{x}(t) = \dfrac{a_{10} + (a_{11} - \mu)x + a_{12}y + a_{13}z + a_{14}u + a_{15}v}{1 - \vartheta \cdot \cos(x)} \\ \qquad + \dfrac{b_{22}y^2 + b_{33}z^2 + b_{44}u^2 + b_{55}v^2 - c_{11}yx - d_{11}zx - e_{11}ux - f_{11}vx}{1 - \vartheta \cdot \cos(x)}, \\ \dot{y}(t) = \dfrac{a_{20} - a_{12}x + (a_{22} - \mu)y + a_{23}z + a_{24}u + a_{25}v}{1 - \vartheta \cdot \cos(y)} \\ \qquad + \dfrac{c_{11}x^2 + c_{33}z^2 + c_{44}u^2 + c_{55}v^2 - b_{22}xy - d_{22}zy - e_{22}uy - f_{22}vy}{1 - \vartheta \cdot \cos(y)}, \\ \dot{z}(t) = \dfrac{a_{30} - a_{13}x - a_{23}y + (a_{33} - \mu)z + a_{34}u + a_{35}v}{1 - \vartheta \cdot \cos(z)} \\ \qquad + \dfrac{d_{11}x^2 + d_{22}y^2 + d_{44}u^2 + d_{55}v^2 - b_{33}xz - c_{33}yz - e_{33}uz - f_{33}vz}{1 - \vartheta \cdot \cos(z)}, \\ \dot{u}(t) = \dfrac{a_{40} - a_{14}x - a_{24}y - a_{34}z + (a_{44} - \mu)u + a_{45}v}{1 - \vartheta \cdot \cos(u)} \\ \qquad + \dfrac{e_{11}x^2 + e_{22}y^2 + e_{33}z^2 + e_{55}v^2 - b_{44}xu - c_{44}yu - d_{44}zu - f_{44}vu}{1 - \vartheta \cdot \cos(u)}, \\ \dot{v}(t) = \dfrac{a_{50} - a_{15}x - a_{25}y - a_{35}z - a_{45}u + (a_{55} - \mu)v}{1 - \vartheta \cdot \cos(v)} \\ \qquad + \dfrac{f_{11}x^2 + f_{22}y^2 + f_{33}z^2 + f_{44}u^2 - b_{55}xv - c_{55}yv - d_{55}zv - e_{55}uv}{1 - \vartheta \cdot \cos(v)}, \end{cases} \tag{3.5}$$

where the parameter $\vartheta (0 \leq |\vartheta| < 1)$ is assigned. (In total in system (3.5) we have $5 + 15 + 20 = 40$ unknown parameters.)

To find the coefficients of system (3.5), the following algorithm is proposed.

1. Fix parameters ϑ and μ that exclude the appearance of singularities in the iterative process. Let, for example, be $\vartheta = 0.95, \nu = 0.1$. Additionally, we choose the diffusion parameter $\mu = 0.00$.

2. Based on the known time series $\mathbf{x}(t) = \{x_0, x_1, ..., x_N\}$, determine the dimension of the embedding space n and the delay time τ.

3. Based on the known n (here $n = 5$) and τ (here $\tau = 1$) , construct five time series

$$\mathbf{x}(t) = (x_0, x_1, x_2, ..., x_L)^T, \mathbf{y}(t) = \mathbf{x}(t+\tau) = (y_0, y_1, y_2, ..., y_L)^T$$
$$\mathbf{z}(t) = \mathbf{x}(t+2\tau) = (z_0, z_1, z_2, ..., z_L)^T, \mathbf{u}(t) = \mathbf{x}(t+3\tau) = (u_0, u_1, u_2, ..., u_L)^T$$
$$\mathbf{v}(t) = \mathbf{x}(t+4\tau) = (v_0, v_1, v_2, ..., v_L)^T$$

that are given on the same time interval $T_L \leq t_0 + (n-1)\tau \leq T$ in equally spaced $L \leq N$ nodes: $0, \Delta t, ..., k\Delta t, ..., L\Delta t = T_L \leq T$. Thus, $\Delta t = T_L/L$.

4. Fix a learning selections

$$x_0, x_1, ..., x_L; y_0, y_1, ..., y_L; z_0, z_1, ..., z_L; u_0, u_1, ..., u_L; v_0, v_1, ..., v_L,$$

where $L \geq 40$.

5. Construct the columns of numerical derivatives D_x, D_y, D_z, D_u, D_v, where

$$D_x = \begin{pmatrix} D_{x1} \\ \vdots \\ D_{x_L} \end{pmatrix} = \frac{1}{\Delta t} \begin{pmatrix} x_1 - x_0 \\ \vdots \\ x_L - x_{L-1} \end{pmatrix} \in \mathbb{R}^L,$$

$$..., D_v = \begin{pmatrix} D_{v1} \\ \vdots \\ D_{v_L} \end{pmatrix} = \frac{1}{\Delta t} \begin{pmatrix} v_1 - v_0 \\ \vdots \\ v_L - v_{L-1} \end{pmatrix} \in \mathbb{R}^L,$$

$$\mathbf{D} = \begin{pmatrix} D_x \\ \vdots \\ D_v \end{pmatrix} \in \mathbb{R}^{5L}.$$

6. Calculate the disturbances introduced by the diffusion parameter

$$R_x = \begin{pmatrix} \dfrac{x_0}{1 - \vartheta \cdot \cos(x_0)} \\ \vdots \\ \dfrac{x_{L-1}}{1 - \vartheta \cdot \cos(x_{L-1})} \end{pmatrix} \in \mathbb{R}^L, ..., R_v = \begin{pmatrix} \dfrac{v_0}{1 - \vartheta \cdot \cos(v_0)} \\ \vdots \\ \dfrac{v_{L-1}}{1 - \vartheta \cdot \cos(v_{L-1})} \end{pmatrix} \in \mathbb{R}^L,$$

$$\mathbf{R} = \begin{pmatrix} R_x \\ \vdots \\ R_v \end{pmatrix} \in \mathbb{R}^{5L}.$$

7. Introduce the designations:

$$\mathbf{0} = (0, ..., 0)^T, \mathbf{1} = (1, ..., 1)^T \in \mathbb{R}^L, E_L \in \mathbb{R}^{L\times L} \text{ is the identity matrix,}$$

$$\mathbf{x} = (x_0, ..., x_{L-1})^T \in \mathbb{R}^L, ..., \mathbf{v} = (v_0, ..., v_{L-1})^T \in \mathbb{R}^L;$$

$$\mathbf{x} \odot \mathbf{x} = (x_0^2, ..., x_{L-1}^2)^T, \mathbf{y} \odot \mathbf{y} = (y_0^2, ..., y_{L-1}^2)^T, ...,$$

$$\mathbf{u} \odot \mathbf{u} = (u_0^2, ..., u_{L-1}^2)^T, \mathbf{v} \odot \mathbf{v} = (v_0^2, ..., v_{L-1}^2)^T,$$

$$\mathbf{x} \odot \mathbf{y} = (x_0 y_0, ..., x_{L-1} y_{L-1})^T, ..., \mathbf{x} \odot \mathbf{v} = (x_0 v_0, ..., x_{L-1} v_{L-1})^T, ...$$

$$\mathbf{v} \odot \mathbf{x} = (v_0 x_0, ..., v_{L-1} x_{L-1})^T, ..., \mathbf{v} \odot \mathbf{u} = (v_0 u_0, ..., v_{L-1} u_{L-1})^T;$$

$$\cos(\mathbf{x}) = \operatorname{diag}(\cos(x_0), ..., \cos(x_{L-1})),$$

$$..., \cos(\mathbf{v}) = \operatorname{diag}(\cos(v_0), ..., \cos(v_{L-1})),$$

$$T_1 = \operatorname{diag}(E_L - \vartheta \cdot \cos(\mathbf{x}))^{-1}, ..., T_5 = \operatorname{diag}(E_L - \vartheta \cdot \cos(\mathbf{v}))^{-1} \in \mathbb{R}^{L\times L},$$

$$T = \begin{pmatrix} T_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & T_5 \end{pmatrix} \in \mathbb{R}^{5L\times 5L}.$$

$$J_1 = \begin{pmatrix} \mathbf{1} & \mathbf{x} & \mathbf{y} & \mathbf{z} & \mathbf{u} & \mathbf{v} \\ \mathbf{0} & \mathbf{0} & -\mathbf{x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{x} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{x} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{x} \end{pmatrix} \in \mathbb{R}^{5L\times 6},$$

$$J_2 = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{y} & \mathbf{z} & \mathbf{u} & \mathbf{v} \\ \mathbf{0} & \mathbf{0} & -\mathbf{y} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{y} \end{pmatrix} \in \mathbb{R}^{5L\times 5},$$

$$J_3 = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{z} & \mathbf{u} & \mathbf{v} \\ \mathbf{0} & \mathbf{0} & -\mathbf{z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{z} \end{pmatrix} \in \mathbb{R}^{5L\times 4},$$

$$J_4 = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{u} & \mathbf{v} \\ \mathbf{0} & \mathbf{0} & -\mathbf{u} \end{pmatrix} \in \mathbb{R}^{5L\times 3}, J_5 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{v} \end{pmatrix} \in \mathbb{R}^{5L\times 2},$$

$$J_6 = \begin{pmatrix} \mathbf{y}\odot\mathbf{y} & \mathbf{z}\odot\mathbf{z} & \mathbf{u}\odot\mathbf{u} & \mathbf{v}\odot\mathbf{v} \\ -\mathbf{x}\odot\mathbf{y} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{x}\odot\mathbf{z} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{x}\odot\mathbf{u} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{x}\odot\mathbf{v} \end{pmatrix} \in \mathbb{R}^{5L\times 4}.$$

$$J_7 = \begin{pmatrix} -\mathbf{y}\odot\mathbf{x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}\odot\mathbf{x} & \mathbf{z}\odot\mathbf{z} & \mathbf{u}\odot\mathbf{u} & \mathbf{v}\odot\mathbf{v} \\ \mathbf{0} & -\mathbf{y}\odot\mathbf{z} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{y}\odot\mathbf{u} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{y}\odot\mathbf{v} \end{pmatrix} \in \mathbb{R}^{5L\times 4}.$$

$$J_8 = \begin{pmatrix} -\mathbf{z}\odot\mathbf{x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{z}\odot\mathbf{y} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}\odot\mathbf{x} & \mathbf{y}\odot\mathbf{y} & \mathbf{u}\odot\mathbf{u} & \mathbf{v}\odot\mathbf{v} \\ \mathbf{0} & \mathbf{0} & -\mathbf{z}\odot\mathbf{u} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{z}\odot\mathbf{v} \end{pmatrix} \in \mathbb{R}^{5L\times 4}.$$

$$J_9 = \begin{pmatrix} -\mathbf{u}\odot\mathbf{x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{u}\odot\mathbf{y} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{u}\odot\mathbf{z} & \mathbf{0} \\ \mathbf{x}\odot\mathbf{x} & \mathbf{y}\odot\mathbf{y} & \mathbf{z}\odot\mathbf{z} & \mathbf{v}\odot\mathbf{v} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{u}\odot\mathbf{v} \end{pmatrix} \in \mathbb{R}^{5L\times 4}.$$

$$J_{10} = \begin{pmatrix} -\mathbf{v}\odot\mathbf{x} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{v}\odot\mathbf{y} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{v}\odot\mathbf{z} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{v}\odot\mathbf{u} \\ \mathbf{x}\odot\mathbf{x} & \mathbf{y}\odot\mathbf{y} & \mathbf{z}\odot\mathbf{z} & \mathbf{u}\odot\mathbf{u} \end{pmatrix} \in \mathbb{R}^{5L\times 4}.$$

8. Construct Jacobi matrix:

$$\mathbf{H} = T\cdot(J_1, J_6, J_2, J_7, J_3, J_8, J_4, J_9, J_5, J_{10}) \in \mathbb{R}^{5L\times 40}.$$

9. Using the least squares method (see [12]), compute the vector: :

$$\mathbf{P} := (\mathbf{H}^T\cdot\mathbf{H} + \nu I)^{-1}\cdot\mathbf{H}^T\cdot(\mathbf{D} + \mu\mathbf{R}) =$$

$$(a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, b_{22}, b_{33}, b_{44}, b_{55}, a_{20}, a_{22}, a_{23}, a_{24}, a_{25},$$

$$c_{11}, c_{33}, c_{44}, c_{55}, a_{30}, a_{33}, a_{34}, a_{35}, d_{11}, d_{22}, d_{44}, d_{55},$$

$$a_{40}, a_{44}, a_{45}, e_{11}, e_{22}, e_{33}, e_{55}, a_{50}, a_{55}, f_{11}, f_{22}, f_{33}, f_{44})^T \in \mathbb{R}^{40}.$$

(Here $I \in \mathbb{R}^{40\times 40}$ is the identity matrix.)

10. Solve system (3.5) the weight coefficients of which are the coordinates of vector $\mathbf{P}$. If the solution of system (3.5) is unstable, then increase the parameter μ (for example, $\mu = 0.01$) and go to step 9, and repeat the algorithm. (After a few iterations, the solution to system (3.5) will become bounded.)

    If the use of diffusion parameter μ is undesirable ($\mu = 0$), then in system (3.5) should be assigned $a_{10} = ... = a_{50} = a_{11} = ... = a_{55} = 0$. In this case, the vector $\mathbf{P} \in \mathbb{R}^{30}$ and for any of its coordinates the system (3.5) has a bounded solution. (In step 9 of the algorithm, we have $\mathbf{P} \in \mathbb{R}^{30}$ and $\mu = 0$.)

3.3. Algorithm for linear model

In this case, items 1 – 6 are the same as in the quadratic model.
In item 7 matrices $J_6, J_7, J_8, J_9, J_{10}$ are not calculated.
Items 8, 9, and 10 are rewritten as follows:
8. Construct Jacobi matrix:

$$\mathbf{H} = T \cdot (J_1, J_2, J_3, J_4, J_5) \in \mathbb{R}^{5L \times 20}.$$

9. Compute the vector:

$$\mathbf{P} := (\mathbf{H}^T \cdot \mathbf{H} + \nu I)^{-1} \cdot \mathbf{H}^T \cdot (\mathbf{D} + \mu \mathbf{R}) =$$

$$(a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15}, a_{20}, a_{22}, a_{23}, a_{24}, a_{25}, a_{30}, a_{33}, a_{34}, a_{35},$$

$$a_{40}, a_{44}, a_{45}, a_{50}, a_{55})^T \in \mathbb{R}^{20}.$$

(Here $I \in \mathbb{R}^{20 \times 20}$ is the identity matrix.)

10. Solve system (3.5) the weight coefficients of which are the coordinates of vector $\mathbf{P}$ and all twenty coefficients $b_{22}, ..., e_{55}$ at nonlinear terms are equal to zero. In the future, the actions of item 10 of the quadratic model algorithm are repeated. (If $\mu = 0$, then in the linear model we should put $a_{10} = ... = a_{50} = a_{11} = ... = a_{55} = 0$ and $\mathbf{P} \in \mathbb{R}^{10}$. In addition, in the Jacobian matrix $\mathbf{H}$ block J_5 is absent and each of blocks $J_1 - J_4$ has 2 columns less. In this case, we have $\mathbf{H} \in \mathbb{R}^{5L \times 10}$.)

The final step is to estimate the parameters of vector $\mathbf{P}$, which will be used for further calculations or predictions based on the input time series. This algorithm allows you to adjust model parameters based on input data and improve their suitability for analysis or prediction.

3.4. Second stage: using the LSTM method

The function $\sigma(x)$ (hyperbolic tangent) satisfies the inequality $\forall x \in \mathbb{R}\ 0 \leq |\sigma(x)| < 1$. Therefore, from the boundedness of solutions of system (3.3) with initial conditions $\mathbf{x}_0$ it follows the boundedness of solutions of system (3.4) with initial conditions $\sigma(\mathbf{x}_0)$ [14]. Consequently, the weight matrices of system (3.3) can be taken as the initial weight matrices W and V for system (3.4).

In order to use the LSTM method it is necessary to insert under the sign σ in the system (3.4) equations (3.5) with known coefficients $a_{10}, ..., f_{44}$ (for the quadratic model) or $a_{10}, ..., a_{55}$ (for the linear model) as initial data.

On the basis of the parameter vector $\mathbf{P}$, the antisymmetric matrix W and the rectangular matrix V are formed. They represent the connections between the input and hidden layers of neurons and are key components of the LSTM structure.

The obtained matrices W and V are transformed into weight matrices of the LSTM model (input weight) taking into account the architectural features of the LSTM and their dimensions. On the basis of the weight matrices, the architecture of the LSTM neural network is created in Matlab, including the definition of the number of layers, the number of neurons in each layer, activation functions and other parameters that determine the behavior of the network.

We have

$$W = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ -a_{12} & a_{22} & a_{23} & a_{24} & a_{25} \\ -a_{13} & -a_{23} & a_{33} & a_{34} & a_{35} \\ -a_{14} & -a_{24} & -a_{34} & a_{44} & a_{45} \\ -a_{15} & -a_{25} & -a_{35} & -a_{45} & a_{55} \end{pmatrix} \in \mathbb{R}^{5\times 5}, \mathbf{b} = \begin{pmatrix} a_{10} \\ a_{20} \\ a_{30} \\ a_{40} \\ a_{50} \end{pmatrix} \in \mathbb{R}^{5}$$

$$\mathbf{u} = (x^2, y^2, z^2, u^2, v^2, xy, xz, xu, xv, yz, yu, yv, zu, zv, uv)^T \in \mathbb{R}^{15}$$

$$V = \left(\begin{array}{ccccc} 0 & b_{22} & b_{33} & b_{44} & b_{55} \\ c_{11} & 0 & c_{33} & c_{44} & c_{55} \\ d_{11} & d_{22} & 0 & d_{44} & d_{55} \\ e_{11} & e_{22} & e_{33} & 0 & e_{55} \\ f_{11} & f_{22} & f_{33} & f_{44} & 0 \end{array}\right| \rightarrow$$

$$\left|\begin{array}{cccccccccc} -c_{11} & -d_{11} & -e_{11} & -f_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ -b_{22} & 0 & 0 & 0 & -d_{22} & -e_{22} & -f_{22} & 0 & 0 & 0 \\ 0 & -b_{33} & 0 & 0 & -c_{33} & 0 & 0 & -e_{33} & -f_{33} & 0 \\ 0 & 0 & -b_{44} & 0 & 0 & -c_{44} & 0 & -d_{44} & 0 & -f_{44} \\ 0 & 0 & 0 & -b_{55} & 0 & 0 & -c_{55} & 0 & -d_{55} & -e_{55} \end{array}\right) \in \mathbb{R}^{5\times 15}.$$

(For linear model $V = 0$!)

After initializing the weight matrices and building the LSTM network, you can start training the network on the input data or use it for various tasks such as time series prediction or data analysis.

Further, in this study, an algorithm for determining and forming weighting coefficients, as well as a neural network for achieving forecasting goals, was developed and implemented. The development was carried out in the MATLAB 2020a environment. Below is a block diagram (see Fig.3.1), which shows the main process of the developed algorithm and neural network.

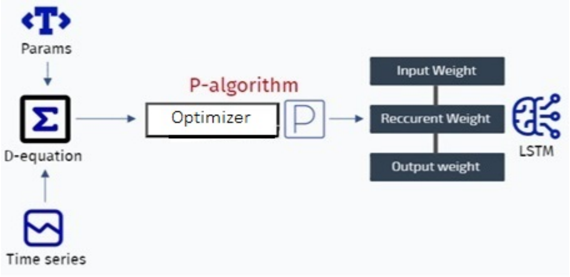


Fig. 3.1. Block diagram of the general EEG data processing algorithm (see ([15])

In particular, attention was focused on the development of a neural network using the Long Short Time Memory (LSTM) layer. LSTM is a powerful tool for processing serial data, and it shows the most accurate forecasting results after proper training of input parameters because it has the ability to consider and analyze long-term dependencies in serial data. The LSTM layer is able to store and use information from previous time steps, allowing the neural network to effectively model and predict complex sequences.

LSTMs are a type of recurrent neural networks (RNNs) designed to model sequential data. This architecture was specifically designed to solve the gradient vanishing problem that often occurs in conventional RNNs. The main characteristics of LSTMs are the ability to store and use information from previous time steps, supervised forgetting, and the assignment of weights to control the flow of information.

LSTM consists of the following main components:

1. Cell state (Cell State) is the main memory of LSTM. It allows a neural network to store and transfer information over many time steps. This memory is controlled by interface weights that determine which information should be forgotten or retained.

2. Input layer (Input Gate) - this input decides what information should be updated in the cellular state. It is activated by a weighted multiplication of the input data and the previous state.

3. Output layer (Output Gate) - determines what information should be output from the cellular state. It is also governed by weighting factors and the internal state of the model.

4. Forgetting layer (Forget Gate) - allows LSTM to decide what information should be forgotten from the cell state based on the current input data and the previous state.

5. Internal weights (Internal Weights) - internal weight coefficients that allow the model to interact and calculate the new state of the cell based on the input data and the previous state.

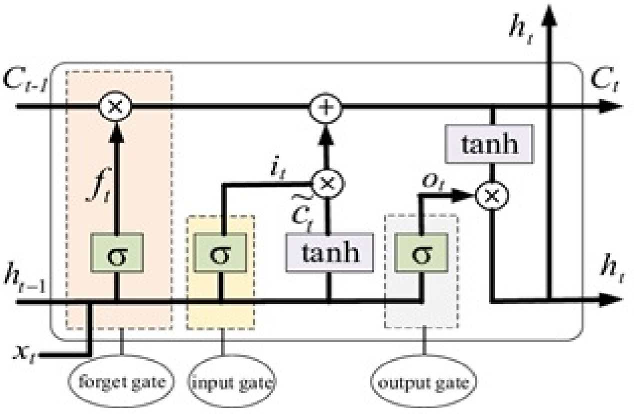


Fig. 3.2. Block diagram of the LSTM layer (see [1, 2])

The main idea behind LSTM is that it can effectively handle and model long-term dependencies in sequential data due to its ability to control the flow of internal cell information. This architecture has found wide application in areas where it is important to model complex sequences, such as language analysis, machine learning, and many other areas (see Fig.3.2).

4. Real applications and numerical analysis of the obtained results

The above algorithm generally describes the steps involved in processing an input time series and is designed to help researchers and practitioners analyze and model complex systems using time series data. In particular, the algorithm provides step-by-step instructions for finding coefficients in a special system, and the system itself is a set of differential equations that can be used to model a wide range of physical, biological, and social phenomena.

Solving the corresponding systems allows you to find the values of unknown parameters that accurately describe the dynamics of the modeled system. Once the coefficients are found, they can be used to predict the behavior of the system over time. For example, the coefficients found in our study can be set by the input weights of the neural network for predicting EEG behavior. Similarly, if the system is a model of a biological process such as the spread of a disease, unraveling the system can help predict the future number of infected individuals given the current state of the population. In general, solving a system allows you to gain insight into the underlying dynamics of complex systems and make predictions about their behavior, which can be useful in many industries.

The initial stages of the described algorithm determine significantly influential parameters of the system state, such as singularities, errors, nesting dimensions, and delays. Based on the known time series $x_0, x_1, ..., x_N$, the dimension of the embedding space and the delay time are determined. This can be done using the delay method, which involves constructing a set of time-delayed copies of the original time series and using them to reconstruct the underlying attractor that defines any chaotic system [17, 18].

In general, in the work, a user interface was implemented using MATLAB2020a tools, which provides a functional opportunity to process the input time step with three algorithms of a similar nature (as described above), but with different system parameters and, accordingly, their essential structured difference. Below is a comparative result of the work of each algorithm, which ended with the original parameter matrix and the solution of the simulated system.

4.1. Modeling epilepsy based on EEG data

Consider the implementation of the above algorithm on real data, taking into account the primary processing of the signal from the encephalograph cap by the primary noise filter [19]. For this, software tools were used for automatic data initialization in the system of multiple space, which means a time series is formed from each electrode of the EEG cap. The input series is divided into a series of trajectories with a predetermined displacement ($\tau = 10$), which was indicated above (Fig.4.1,4.2).

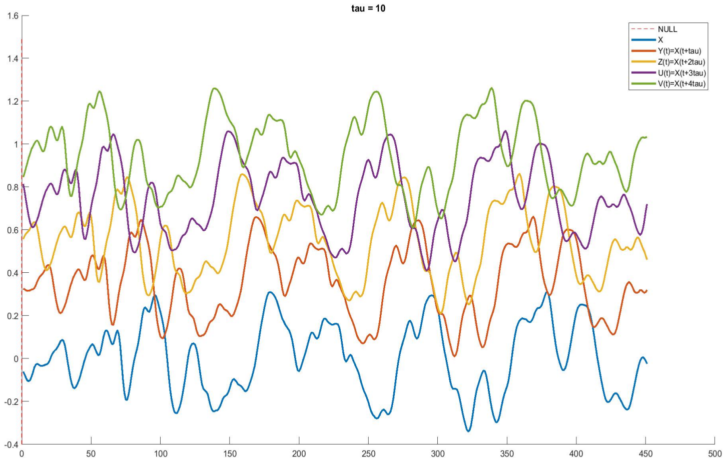


Fig. 4.1. EEG time series of a sick patient with shift $\tau = 10$

We will present the reconstruction (Fig.4.3,4.4) of both sequences and run the algorithm for their processing to obtain the parameters of the system that models the specified sequences with the defined control of the process of propagation of trajectories.

The methodology, which is developed on the basis of MATLAB2020a tools, was tested on two patients with pre-processing of the data to determine the weights of the neural network through the EEG behavior modeling algorithm and the singularities that were added to this process. Note that the developed

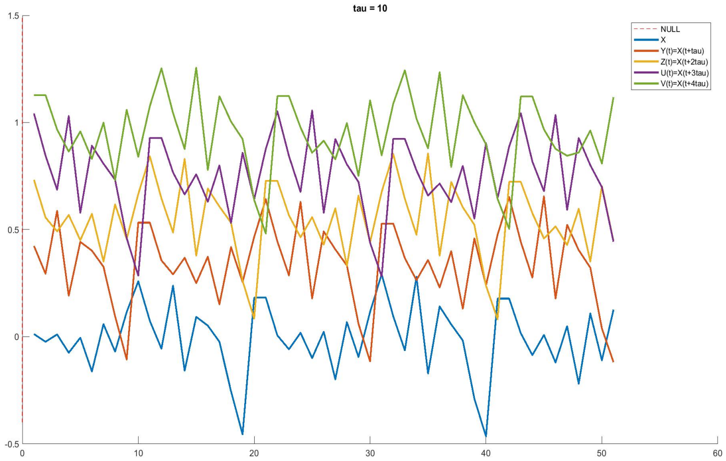


Fig. 4.2. EEG time series of a healthy patient with shift $\tau = 10$

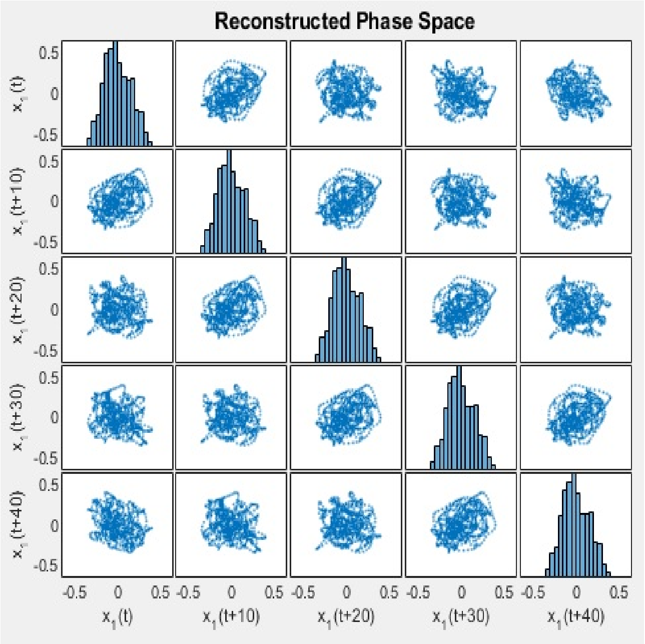


Fig. 4.3. Phase space of a sick patient

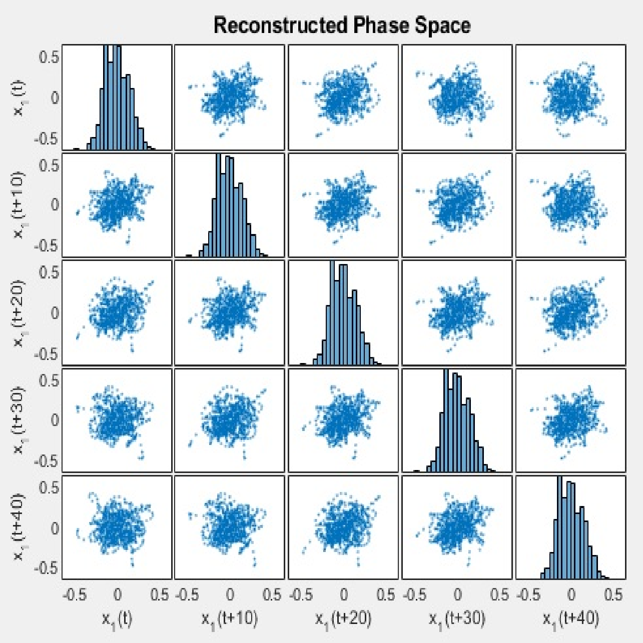
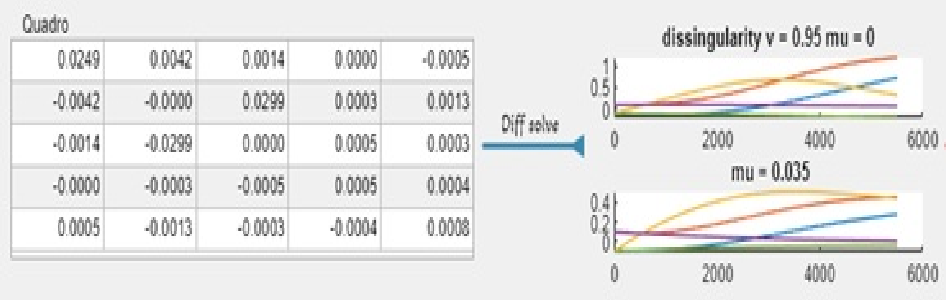


Fig. 4.4. Phase space of a healthy patient

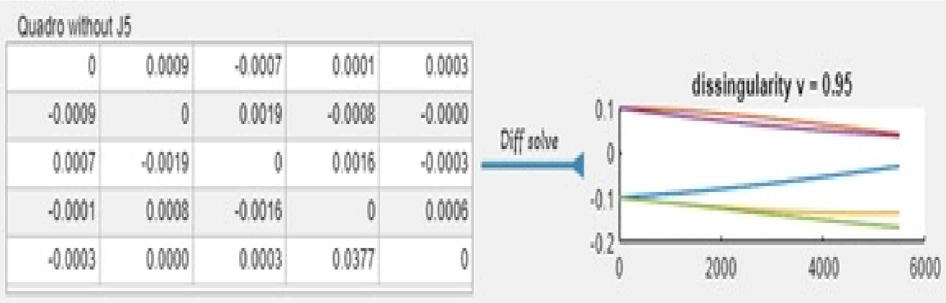
algorithms have a characteristic difference in the presence of quadratic elements and the diffusion parameter. Therefore, three different cases were analyzed: a quadratic algorithm, a quadratic algorithm with a diffusion parameter, and a linear algorithm.

A neural network with defined weights is evaluated using the mean square error (MSE) method. The results of forecasting in relation to real data and the forecast of unknown values for the future 30 steps are displayed graphically (Fig.4.5 -4.7). This allows us to evaluate the effectiveness and accuracy of the model in forecasting based on the training data provided for training the LSTM neural network.


Quadro

0.0249	0.0042	0.0014	0.0000	-0.0005
-0.0042	-0.0000	0.0299	0.0003	0.0013
-0.0014	-0.0299	0.0000	0.0005	0.0003
-0.0000	-0.0003	-0.0005	0.0005	0.0004
0.0005	-0.0013	-0.0003	-0.0004	0.0008

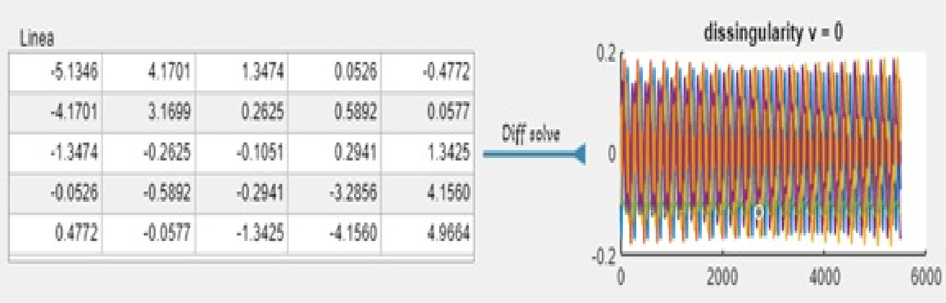
Fig. 4.5. The weight matrix of the quadratic algorithm with the diffusion parameter ($\mu = 0, \mu = 0.035$) of a sick patient with the result of data modeling; $MSE = 0.0041193$



Quadro without J5

0	0.0009	-0.0007	0.0001	0.0003
-0.0009	0	0.0019	-0.0008	-0.0000
0.0007	-0.0019	0	0.0016	-0.0003
-0.0001	0.0008	-0.0016	0	0.0006
-0.0003	0.0000	0.0003	0.0377	0

Fig. 4.6. The weight matrix of the quadratic algorithm of the sick patient with the result of data modeling; $MSE = 0.00511702$



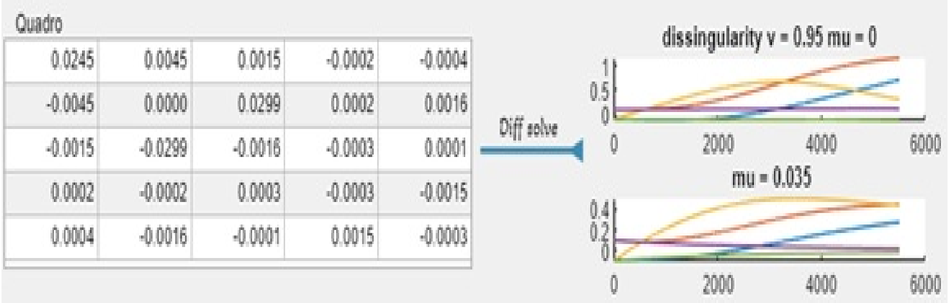
Linea

-5.1346	4.1701	1.3474	0.0526	-0.4772
-4.1701	3.1699	0.2625	0.5892	0.0577
-1.3474	-0.2625	-0.1051	0.2941	1.3425
-0.0526	-0.5892	-0.2941	-3.2856	4.1560
0.4772	-0.0577	-1.3425	-4.1560	4.9664

Fig. 4.7. The weight matrix of the patient-to-patient linear algorithm with the result of data modeling; $MSE = 0.0030853$

We note the appearance of nonlinear effects in modeling by the quadratic algorithm with diffusion, which provides the possibility of further adjustment of the system to obtain a more similar solution. At the same time, we note the effective operation of the linear algorithm, which repeats the input trajectory, but with falling into a periodic process, which distinguishes the simulation result from the real scenario. According to the values of neural network training errors, we observe the smallest deviations precisely in the linear algorithm, and the worst is the quadratic algorithm without the diffusion parameter.

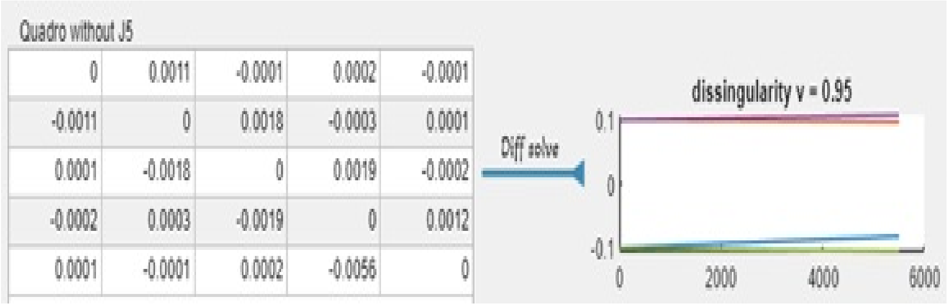
Let's try to repeat these actions based on the data of a healthy patient (Fig.4.8 - 4.10).



Quadro

0.0245	0.0045	0.0015	-0.0002	-0.0004
-0.0045	0.0000	0.0299	0.0002	0.0016
-0.0015	-0.0299	-0.0016	-0.0003	0.0001
0.0002	-0.0002	0.0003	-0.0003	-0.0015
0.0004	-0.0016	-0.0001	0.0015	-0.0003

Fig. 4.8. The weight matrix of the quadratic algorithm with the diffusion parameter ($\mu = 0, \mu = 0.035$) for a healthy patient with the result of data modeling; $MSE = 0.017608$



Quadro without J5

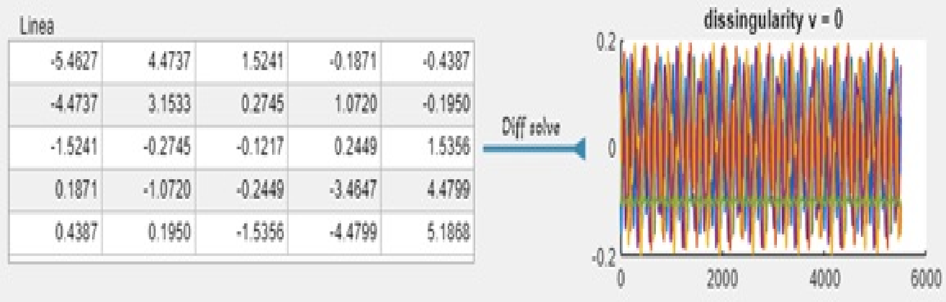
0	0.0011	-0.0001	0.0002	-0.0001
-0.0011	0	0.0018	-0.0003	0.0001
0.0001	-0.0018	0	0.0019	-0.0002
-0.0002	0.0003	-0.0019	0	0.0012
0.0001	-0.0001	0.0002	-0.0056	0

Fig. 4.9. The weight matrix of the quadratic algorithm for a healthy patient with the result of data modeling; $MSE = 0.017553$

These results demonstrate a significant difference in simulation results compared to sick patients, with an order of magnitude higher error. Such results demonstrate differences in input amplitudes and embedding dimensions that directly affect the training outcome and allow classification of healthy patients with more chaotic behavior of brain signals.

The next step is to present the results of the neural network and compare the prediction results.

In all the following figures, the blue line is the line obtained from the results of



Linea

-5.4627	4.4737	1.5241	-0.1871	-0.4387
-4.4737	3.1533	0.2745	1.0720	-0.1950
-1.5241	-0.2745	-0.1217	0.2449	1.5356
0.1871	-1.0720	-0.2449	-3.4647	4.4799
0.4387	0.1950	-1.5356	-4.4799	5.1868

Fig. 4.10. The weight matrix of the linear algorithm for a healthy patient with the result of data modeling; $MSE = 0.018206$

EEG measurements, the red line is obtained as a result of adjusting the weighting coefficients of the neural network model, and the yellow line is the prediction line.

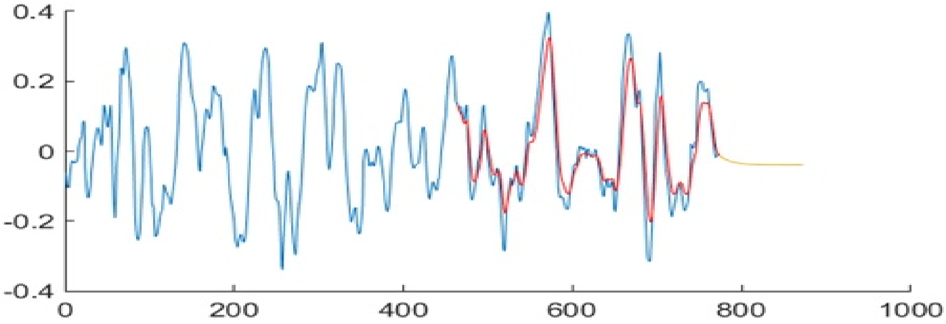


Fig. 4.11. The result of LSTM prediction of a patient's neural network with weighting coefficients of the quadratic algorithm and a diffusion parameter

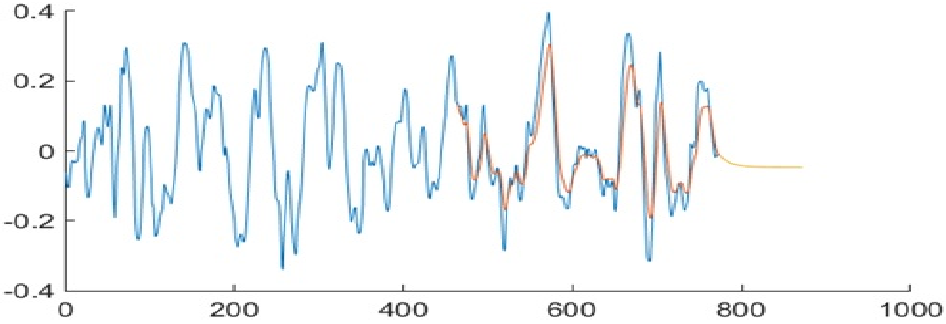


Fig. 4.12. The result of LSTM prediction of a patient's neural network with weighting coefficients of the quadratic algorithm

These results allow us to draw conclusions about the sufficiently effective result of neural network training using pre-processing algorithms and to determine a more optimal approach to the classification of EEG data. However, further use of

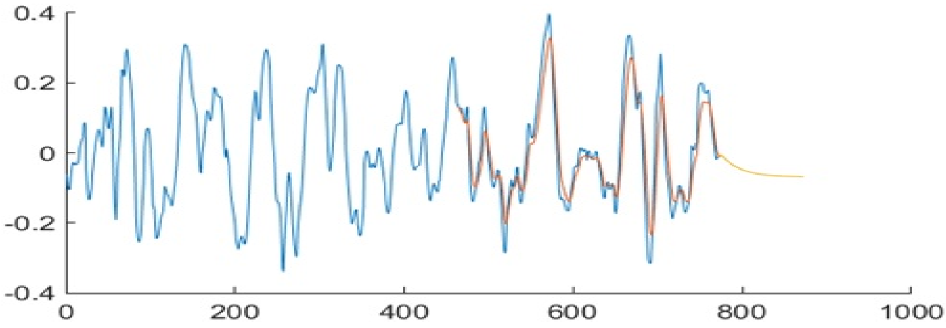


Fig. 4.13. The result of LSTM prediction of the patient's neural network with the weighting coefficients of the linear algorithm

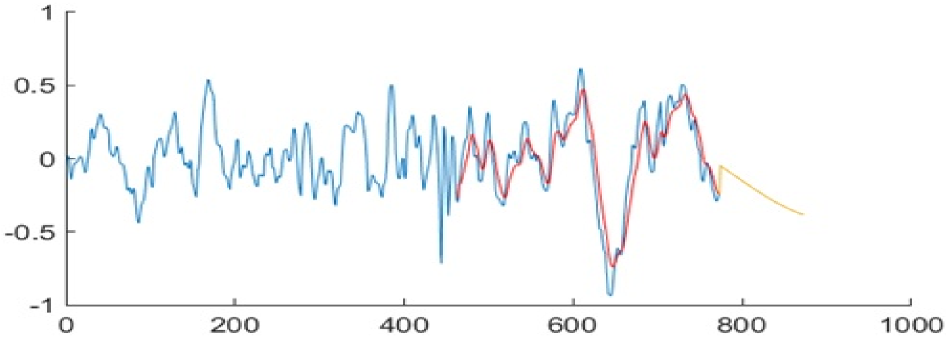


Fig. 4.14. Prediction result of LSTM neural network of a healthy patient with quadratic algorithm weights and diffusion parameter

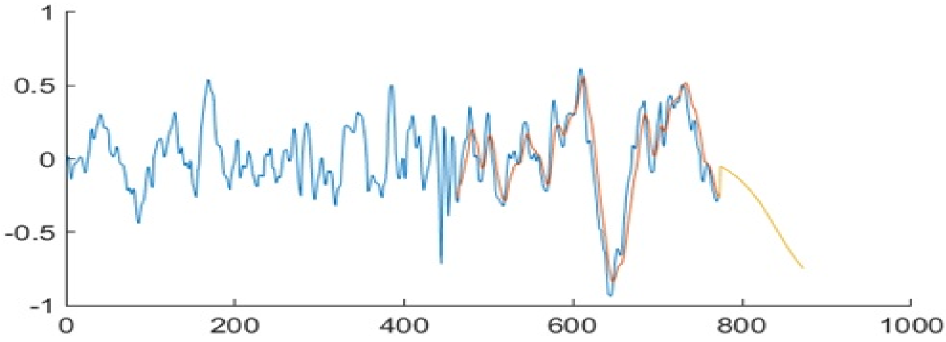


Fig. 4.15. LSTM neural network prediction result for a healthy patient with quadratic algorithm weights

the neural network to predict future values is possible for very short time intervals (about 10 steps), after which either a steady-state mode of the system is observed (in other words, there is no signal), or a mode of constant monotonicity, which excludes the occurrence of further chaos (see Fig.4.11 - 4.16).

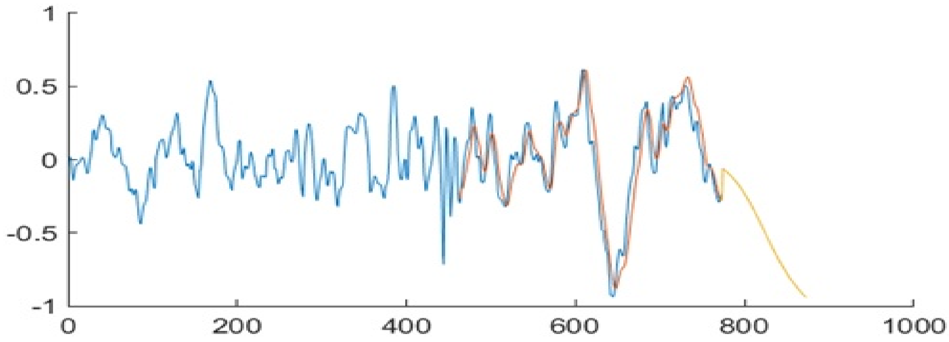


Fig. 4.16. LSTM neural network prediction result for a healthy patient with linear algorithm weights

References

1. U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, J. S. Suri, *Automated EEG analysis of epilepsy: A review*, Knowledge-Based Systems, **45**, (2013), 147–165.
2. B. Ay, O. Yildirim, M. Talo, U. B. Baloglu, G. Aydin, S. D. Puthankattil, U. R. Acharya, *Automated depression detection using deep representation and sequence learning with EEG signals*, Journal of Medical Systems, **43**(7), (2019), 1–24.
3. V. Ye. Belozyorov, *Universal approach to the problem of emergence of chaos in autonomous dynamical systems*, Nonlinear Dynamics, **95**(1), (2019), 579–595.
4. 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.
5. V. Ye. Belozyorov, D. V. Dantsev, Y. V. Koshel, *On the equivalence of real dynamic process and its models*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **29**(2)(2021), 76–91.
6. V. Ye. Belozyorov, D. V. Dantsev, *Modeling of chaotic processes by means of antisymmetric neural ODEs*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **30**(1)(2022), 1–41.
7. V. Ye. Belozyorov, S. A. Volkova, *Odd and even functions in the design problem of new chaotic attractors*, International Journal of Bifurcation and Chaos, **32**, (2022), Article ID 2250218, 26 pages
8. V. Ye. Belozyorov, S. A. Volkova, *Discrete processes and chaos in systems of ordinary differential equations*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **30**(2)(2022), 62–78.
9. V. Ye. Belozyorov, S. A. Volkova, V. G. Zaytsev, *Singular differential equations and their applications for modeling strongly oscillating processes*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **31**(1)(2023), 22–52.
10. 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.
11. R. T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, *Neural ordinary differential equations*, *arXiv preprint arXiv:1806.07366v5[cs.LG]*, (2019), 1–18.

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

13. E. Haber, L. Ruthotto, *Stable Architectures for Deep Neural Networks*, arXiv *preprint arXiv: 1705.03341v1[cs.LG]*, (2019), 1–23.

14. H. K. Khalil, *Nonlinear Systems – 2nd Edition*, Prentice Hall, New-Jersy, 1996.

15. J. S. Kumar, P. Bhuvaneswari, *Analysis of electroencephalography (EEG) signals and its categorization*, A Study Procedia Engineering, **38**, (2012), 2525–2536.

16. Q. Li, *Dynamical Systems and Machine Learning*, Summer School, Peking University, 2020.

17. N. A. Magnitskii, *Universal theory of dynamical chaos in nonlinear dissipative systems of differential equations*, Commun. Nonlinear Sci. Numer. Simul., **13**, (2008), 416–433.

18. 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.

19. National Center for Biotechnology Information, *National Library of Medicine at the U.S. National Institutes of Health (NIH)*, USA, https://www.ncbi.nlm.nih.gov.

20. Toolboxes for Complex Systems, *Potsdam Institute for Climate Impact Research (PIK)*, Germany, http://tocsy.pik-potsdam.de.

21. C. L. Webber, N. Marwan (eds), *Reccurence Quantification Analysis. Theory and Best Practice*, Springer, NY, Dordrecht, Heidelberg, London, 2015.

Received 27.09.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 50–66, DOI 10.15421/142311

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

UNIVARIATE TIME SERIES ANALYSIS WITH HYPER NEURAL ODE

Yevhen V. Koshel[*], Vasiliy Ye. Belozyorov[†]

Abstract. Neural ordinary differential equations (NODE) are ordinary differential equations whose right-hand side is determined by a neural network. Hyper NODE (hNODE) is a special type of neural network architecture, which is aimed at creating such NODE system that regulates its own parameters based on known input data. The article uses a new approach to the study of one-dimensional time series, the basis of which is the hNODE system. This system takes into account the relationship between the input data and its latent representation in the network and uses an explicit parametrization when controlling the latent flow. The proposed model is tested on artificial time series of data. The influence of some activation functions (besides sigmoid and hyperbolic tangent) on the quality of the forecast is also considered.

Key words: hypernetwork, neural ODE, time series analysis, power activation function.

2010 Mathematics Subject Classification: 34A34, 34D20, 37D45, 93A30.

Communicated by Prof. P. Kogut

1. Introduction

Constructing and fitting models that can reliably predict time series data has been a subject of thorough research for many decades. The problem of time series forecasting is important in many fields, from economics and finance to meteorology and biology. The future time series values predicted by a model can be used for increasing the planning horizon or decreasing the incident response time, both of which can be invaluable for business and research.

There are two broad approaches to constructing a model — stochastic and deterministic [2]. The stochastic approach aims to identify the underlying statistical patterns in the time series data and use them to estimate its future values. The main downside of this approach is that it requires the researchers to make a priori assumptions about the statistical distribution of the data. The deterministic approach aims to create a model that doesn't contain random variables in its definition. Deterministic models come in many forms, but the most popular one is an

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

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

artificial neural network (ANN). ANNs are universal function approximators that can be used to detect complex patterns in the data and perform either pattern recognition [1], time series forecasting [14], sequence transformation [15], new data generation [5], or other tasks. Even though the ANNs are data-driven and self-adaptive, the researchers are still required to select the proper architecture and size for the model to achieve optimal performance. Pattern recognition capabilities of ANNs were substantially improved by introduction of Convolutional neural networks [12], accurate time series forecasting as well as sequence-to-sequence conversion can be achieved by recurrent, long-short term memory, time-lagged, or seasonal neural networks [9], etc.

In this article, the neural ODE (NODE) architecture is of particular interest. NODE is a new family of residual neural network models that, instead of specifying a discrete sequence of hidden layers, parametrizes the derivative of the hidden state using a neural network [8]. The original paper proposes the NODE as an alternative to the residual neural networks that perform a series of composable transformations to their hidden state:

$$\mathbf{h}_{t+1} = \mathbf{h}_t + f(\mathbf{h}_t, \Theta), \tag{1.1}$$

where $t \in [0..T]$ and $\mathbf{h} \in \mathbb{R}^D$. By viewing the equation (1.1) as an Euler discretization of a continuous transformation, the authors transition from the discrete to a continuous representation of the original sequence:

$$\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}, t, \Theta). \tag{1.2}$$

Thus, a residual neural network of infinite depth is achieved, the forward pass of which is calculated as follows:

$$F(x_0, T) = x_0 + \int_0^T f(x, \Theta, t)\, dt, \tag{1.3}$$

where x_0 is the initial point and Θ is a set of parameters.

This approach, however, is not limited to residual networks. The efficient way of backpropagating the errors via reverse-mode automatic differentiation developed by the authors of this model allows construction and training of any kind of neural network that contains a NODE as its component.

One of the disadvantages of the base NODE model is that it is only able to learn one continuous flow F. In other words, the set of parameters Θ is static and cannot react to the new inputs or change with time. This is fairly limiting, especially when building models for time series data that can change their qualitative characteristics over time. Another disadvantage is that the behavior of model (1.3) is largely determined by the dimensionality of the space of inputs because the model structure limits F to only be a homeomorphism of the input space onto itself which is very limiting in the context of univariate time series analysis.

To address these issues and expand the capabilities of NODE-based models, different approaches were proposed. Neural ODE Process [11] model aims to achieve the data-dependence of Θ by adopting a stochastic approach and maintaining an adaptive data-dependent distribution over the underlying ODE. The core idea of this approach is to transition to the latent representation of the modeled data using an encoder network, obtain the set of parameters from the provided context, evolve the initial point in the latent space, and finally decode each point in the resulting trajectory back to the input space using a decoder.

The neurally-controlled ODE [6] (N-CODE) model adopts a deterministic approach by introducing a coupled system for determining the set of parameters at each point in time during integration. It effectively merges the input data with the inferred set of parameters into a single system and evolves it, greatly increasing the dimensionality and expressiveness of the hidden representation of the data.

Both Neural ODE Process and N-CODE approaches are examples of hypernetworks because they infer their set of parameters at runtime based on the input data. However, the Neural ODE Process' way of representing the parameters and the input data is more flexible because they are not being coupled into a single space. This provides the designers of the model with the freedom to both define the dimensionality of the latent space to increase the range of possible behaviors of the model and constrain the values of the parameters to reduce the possibility of unwanted behaviors arising in the system.

The proposed hNODE model is effectively a simplified and deterministic Neural ODE Process with N-CODE-inspired approach to control.

2. hNODE definition

Consider a series of pairs of real values $X = \{(t_0, x_0), (t_1, x_1), \dots, (t_n, x_n)\}$, $t_{i+1} - t_i = \Delta t$, $X_i = (t_i, x_i)$ that represent the univariate time series data augmented with timestamps at which the data was recorded. The goal of time series analysis is to extract meaningful information from X and enable forecasting of its future values. To accomplish this, a model F that maps the set X onto itself is to be constructed:

$$X^* = F(X, \Theta). \tag{2.1}$$

The model F, governed by the set of parameters Θ, must interpret the sequence X and infer the future values X^*. But if the underlying process that produces the data changes, the model F becomes useless because it is unable to adapt its parameters to continue producing reliable outputs. Consider a simple linear model:

$$F(x, A, b) = Ax + b, \tag{2.2}$$

which takes a number of observations from X and produces a prediction X^*. The parameters of the model are static and are optimized to fit the training data. Control functions $A'(x)$ and $b'(x)$ with parameters θ_A and θ_b, can be introduced to adjust the parameters depending on the input data:

$$F^*(x, A, b, \theta_A, \theta_b) = (A + A'(x))x + (b + b'(x)) = \mathcal{A}(x)x + \mathcal{B}(x). \tag{2.3}$$

The expressions $\mathcal{A}$ and $\mathcal{B}$ are matrix and vector functions respectively that represent the rules for inferring parameter values for equation (2.2) based on the input values and the set of hyperparameters that define behaviors of functions $A'(x)$ and $b'(x)$. If we combine the hyperparameters into a set $\Theta = \{\theta_A, \theta_b\}$, and denote $\mathcal{H}(X, \Theta)$ as a map from the Cartesian product of the set of hyperparameters Θ and the input values X into the set $\Theta^* = \{A^*, b^*\}$ of adjusted parameters for (2.2), the equation (2.3) becomes

$$F^*(x, A, b, \Theta) = F(x, \mathcal{H}(x, \Theta)) \tag{2.4}$$

If instead of the equation (2.2) one were to use the model (1.3), the dimensionality of the inputs X might become a problem so to increase or decrease it, one may use an extra pair of vector functions — an encoder-decoder couple:

$$\begin{aligned} Y &= \mathcal{E}(X, \theta_{\mathcal{E}}) \\ X &= \mathcal{D}(Y, \theta_{\mathcal{D}}). \end{aligned} \tag{2.5}$$

Definition 2.1. A model $\mathcal{E}$ that disentangles the input data and maps the time series data into the latent space is called an *encoder*. A model $\mathcal{D}$ that interprets the points in the latent space and maps them back onto the time series data space called a *decoder*. A pair of models $\mathcal{E}$ and $\mathcal{D}$ such that $X \equiv \mathcal{D} \circ \mathcal{E}(X)$ is called an *encoder-decoder couple.*

So the model (2.4) becomes

$$F^*(x, A, b, \Theta) = \mathcal{D} \circ F(\mathcal{E}(x, \Theta), \mathcal{H}(x, \Theta)). \tag{2.6}$$

Finally, the hNODE model is defined as (2.6) where function F is the NODE model (1.3):

$$hNODE(X, \Theta) = \mathcal{D} \circ G(\mathcal{E}(X, \Theta), \mathcal{H}(X, \Theta)), \tag{2.7}$$

where $\mathcal{E} : X \times \Theta \to L$ is an encoder that maps the time series data X into the latent space L, $\mathcal{D} : L \to X$ is a decoder that interprets the points from L and maps them into X, $G : L \times \Theta^* \to L$ is a system of ordinary differential equations that evolves the state in the latent space, $\mathcal{H} : X \times \Theta \to \Theta^*$ is a function that produces control rule for G. A visual representation of model (2.7) is provided on Fig. 2.1

3. Activation functions in neural network modeling of time series

It is known that in neural network modeling three types of activation functions are most often used: sigmoid, hyperbolic tangent and rectified linear unit (ReLU). The increased attention to these functions is explained by a number of reasons,

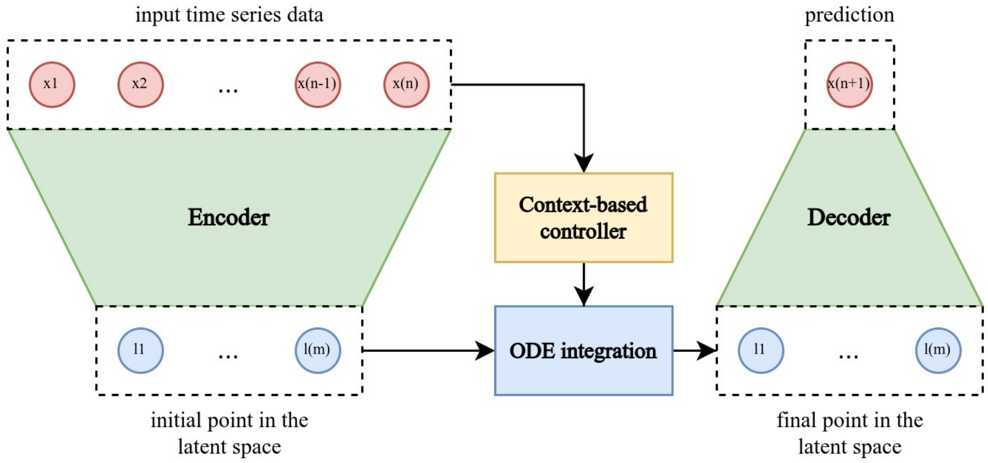


Fig. 2.1. Top-level representation of the model for predicting univariate time series data.

the main one being the stability that they provide to the neural network models. In this regard, in modeling problems it seems interesting to use other activation functions. Naturally, one of the requirements for such functions must be the stability of the resulting neural network models. Further, from the point of view of stability, we will consider power-law activation functions.

Definition 3.1. [4] 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 $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}$ ($\mathbf{x}_1 \neq \mathbf{x}_2$), there exists a function $f \in \mathbb{F}$ such that $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$.

Let $f(w) \in \mathbb{F}$ be the function of one real variable w such that $f(0) = 0$ and

$$\text{either conditions for } f(w) : \begin{cases} \text{if } w < 0 \text{ then } f(w) = -\psi(-w) < 0, \\ \text{if } w > 0 \text{ then } f(w) = \phi(w) > 0 \end{cases} \tag{3.1}$$

$$\text{or conditions for } f(w) : \begin{cases} \text{if } w < 0 \text{ then } f(w) = \psi(-w) > 0, \\ \text{if } w > 0 \text{ then } f(w) = \phi(w) > 0 \end{cases} \tag{3.2}$$

are fulfilled. (Here $\phi(w), \psi(w)$ are differentiable functions of one variable w.)

Definition 3.2. [4] Representation (3.1), ((3.2)) is called an odd (even) activation function.

For example, the ReLU-like function can be represented in the form:

$$f(w) = \ln \frac{a \cdot \exp(b \cdot w)}{1 + (a-1) \cdot \exp(c \cdot w)}; a > 1, b \geq c > 0.$$

Consider the following system

$$\begin{cases} \dot{x}_1(t) &= f_1(a_{11}x_1 + \ldots + a_{1n}x_n + b_1) \\ \dot{x}_2(t) &= f_2(a_{21}x_1 + \ldots + a_{2n}x_n + b_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \dot{x}_n(t) &= f_n(a_{n1}x_1 + \ldots + a_{nn}x_n + b_n), \end{cases} \tag{3.3}$$

where a_{ij}, b_i are known real constants; $i, j = 1 \ldots n$.

Along with system (3.3), we will also consider the system

$$\begin{cases} \dot{x}_1(t) &= a_{11}f_1(x_1) + \ldots + a_{1n}f_n(x_n) \\ \dot{x}_2(t) &= a_{21}f_1(x_1) + \ldots + a_{2n}f_n(x_n) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \dot{x}_n(t) &= a_{n1}f_1(x_1) + \ldots + a_{nn}f_n(x_n). \end{cases} \tag{3.4}$$

Note that if $\det A \neq 0$, then with the help of the change of variables $\mathbf{x} \to A\mathbf{x} + \mathbf{b}$ it is always possible to pass from system (3.3) to system (3.4) and vice versa; $A \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$. Therefore, in what follows, we restrict ourselves to the study of system (3.4).

We introduce the notation

$$\mathbf{f}(\mathbf{x}) = (f_1(x_1), \ldots, f_n(x_n))^T. \tag{3.5}$$

Taking into account (3.5), we introduce the following function

$$V(x_1, \ldots, x_n) = \int f_1(s_1)\, ds_1 + \ldots + \int f_n(s_n)\, ds_n,$$

where the integral is understood in the sense of an indefinite integral.

The integral of the function $\mathbf{f}(\mathbf{x})$ is defined by formulas:

1. If $f_i(s_i)$ is even then

$$\int f_i(s_i)\, ds_i = \begin{cases} \displaystyle\int_0^{x_i} \phi_i(s_i)\, ds_i, \text{ if } x_i \geq 0 \\ \displaystyle\int_{x_i}^{0} \psi_i(s_i)\, ds_i, \text{ if } x_i < 0 \end{cases}.$$

2. If $f_i(s_i)$ is odd then

$$\int f_i(s_i)\, ds_i = \begin{cases} \displaystyle\int_0^{x_i} \phi_i(s_i)\, ds_i, \text{ if } x_i \geq 0 \\ \displaystyle -\int_{x_i}^{0} \psi_i(s_i)\, ds_i, \text{ if } x_i < 0 \end{cases}.$$

Theorem 3.1. *[3] Assume that in system* (3.3) *the matrix* $A = \{a_{ij}\}$ *is anti-symmetric and invertible:* $A^T + A = 0$ *and* $\det(A) \neq 0$. *Let all components of the vector function* $\mathbf{f}(\mathbf{x})$ *be odd activation functions. Then any solution* $\mathbf{x}(t, \mathbf{x}_0)$ *of system* (3.3) *is bounded and is either periodic or chaotic.*

Proof. It is known [4] that the integral of the odd activation function $f_i(x_i), i = 1, \dots, n$, is the even activation function. Therefore, we have $V(x_1, \dots, x_n) \geq 0$.

Compute the total derivative with respect to t of the function $V(x_1, \dots, x_n)$:

$$\begin{aligned} \dot{V}_t(x_1, \dots, x_n) &= 0.5\left[\dot{x}_1(t)\tfrac{\partial V(x_1,\dots,x_n)}{\partial x_1} + \dots + \dot{x}_n(t)\tfrac{\partial V(x_1,\dots,x_n)}{\partial x_n}\right] \\ &\quad +0.5\left[\tfrac{\partial V(x_1,\dots,x_n)}{\partial x_1}\dot{x}_1(t) + \dots + \tfrac{\partial V(x_1,\dots,x_n)}{\partial x_n}\dot{x}_1(t)\right] \\ &= 0.5\mathbf{f}^T(\mathbf{x})(A^T + A)\mathbf{f}(\mathbf{x}) = 0. \end{aligned} \tag{3.6}$$

Further, by virtue of inequalities (3.1), (3.2) and the fact that $V(x_1, \dots, x_n) \geq 0$, we have

$$\lim_{x_i \to \infty} \int_{-x_i}^{x_i} f_i(s_i)\, ds_i = \lim_{x_i \to \infty} \int_{-x_i}^{0} f_i(s_i) ds_i + \lim_{x_i \to \infty} \int_{0}^{x_i} f_i(s_i)\, ds_i \geq 0; i = 1, \dots, n.$$

This means that $\lim_{\|\mathbf{x}\| \to \infty} V(x_1, \dots, x_n) \geq 0$. In addition, from (3.6) it follows that for a sufficiently large value $\|\mathbf{x}_0\|$, we have $V(x_1(t), \dots, x_n(t)) = \mathit{const} = V(x_{10}, \dots, x_{n0}) = V(\mathbf{x}_0) > 0$. This implies that the set $\mathbb{S} = \{V(x_1(t), \dots, x_n(t)) - V(x_{10}, \dots, x_{n0}) = 0\}$ is compact. Therefore, any trajectory $\mathbf{x}(t, \mathbf{x}_0)$ of system (3.3) (or (3.4)) is bounded and is either periodic (if $n \geq 2$) or chaotic (if $n \geq 3$). □

Theorem 3.2. *[3] Assume that in system* (3.3) *the matrix* $A + A^T$ *is non-negative definite. Let also all components of the vector function* $\mathbf{f}(\mathbf{x})$ *be odd activation functions. Then any solution* $\mathbf{x}(t, \mathbf{x}_0)$ *of system* (3.3) *is stable.*

Proof. It's clear that $V(0, \dots, 0) = 0$. Then, under the conditions of Theorem 3.2, equality (3.6) must be replaced by inequality $\dot{V}_t(x_1, \dots, x_n) \leq 0$. Now it remains to apply Lyapunov's theorem [2] on the stability of solutions of a system of ordinary differential equations to system (3.4). □

3.1. Generalization of the concept of power activation function

Introduce the following power functions [4]:

$$g(u) = \begin{cases} -(-u)^\beta & \text{if}(u < 0 \text{ and } \beta > 0);\ 0 \text{ if}(u < 0 \text{ and } \beta = 0) \\ u^\alpha & \text{if}(u \geq 0 \text{ and } \alpha > 0);\ 0 \text{ if}(u \geq 0 \text{ and } \alpha = 0) \end{cases} \tag{3.7}$$

or

$$g(u) = \begin{cases} (-u)^\beta & \text{if}(u < 0 \text{ and } \beta > 0);\ 0 \text{ if}(u < 0 \text{ and } \beta = 0) \\ u^\alpha & \text{if}(u \geq 0 \text{ and } \alpha > 0);\ 0 \text{ if}(u \geq 0 \text{ and } \alpha = 0). \end{cases} \tag{3.8}$$

It is clear that representation (3.7) ((3.8)) is an odd (even) activation function.

Formulas (3.7) and (3.8), which introduce power activation functions, have two drawbacks:

1. If $0 < \alpha \leq 1$ or $0 < \beta \leq 1$, then the functions (3.7) and (3.8) are non-differentiable;
2. Functions (3.7) and (3.8) do not take into account the shift of the argument.

In this connection, we introduce the following function (see Fig.3.1):

$$w(u,\alpha,\beta,b,c) = \mathit{piecewise}\Big[u+\frac{b}{c} < -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{1}{\beta}\Big(-\Big(u+\frac{b}{c}\Big)\Big)^{\beta},$$
$$u+\frac{b}{c} \leq c^{\frac{1}{\alpha-1}}, c\cdot\Big(u+\frac{b}{c}\Big), \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{1}{\alpha}\Big(u+\frac{b}{c}\Big)^{\alpha}\Big]. \tag{3.9}$$

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 (3.9) $b = 0$. Then we will have

$$w(u,\alpha,\beta,c) = \mathit{piecewise}\Big[u < -c^{\frac{1}{\beta-1}}, -\frac{\beta-1}{\beta}c^{\frac{\beta}{\beta-1}} - \frac{(-u)^{\beta}}{\beta},$$
$$u \leq c^{\frac{1}{\alpha-1}}, cu, \frac{\alpha-1}{\alpha}c^{\frac{\alpha}{\alpha-1}} + \frac{u^{\alpha}}{\alpha}\Big]. \tag{3.10}$$

Formula (3.10) can be obtained from formula (3.9) by introducing a new variable $z := u + b/c$, which in (3.10)) 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:

$$\dot{w}_u(u,\alpha,\beta,c) = \mathit{piecewise}\Big[u < -c^{\frac{1}{\beta-1}}, (-u)^{\beta-1}, u \leq c^{\frac{1}{\alpha-1}}, c, u^{\alpha-1}\Big] \tag{3.11}$$

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

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

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

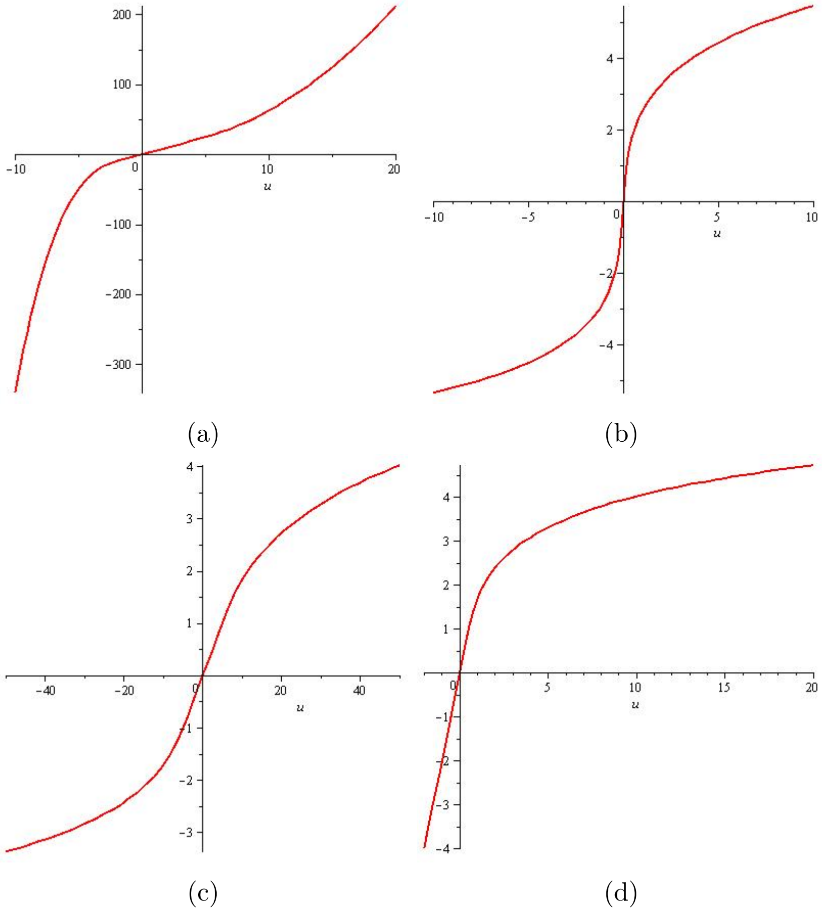
If $\lim\alpha \to 1$, then

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

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

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 (3.10) is transformed into formula (3.9) if we put in (3.10) $u := u + b/c$. Thus, we have $w(u + b/c,\alpha,\beta,c) \equiv w(u,\alpha,\beta,b,c)$.



(a)

(b)

(c)

(d)

Fig. 3.1. The activation differentiable power function (3.10) for different values of the parameters α, β, and c: (a) $c = 5, \alpha = 2, \beta = 3$; (b) $c = 2, \alpha = 0.01, \beta = 2$; (c) $c = 7, \alpha = 0.3, \beta = 0.1$; (d) $c = 0.2, \alpha = 0.1, \beta = 0.01$.

Finally, if we put $c = 0$ in formula (3.10), then we obtain (with insignificant additions) function (3.7):

$$w(u, \alpha, \beta) = \mathit{piecewise}\Big[u < 0, -\frac{(-u)^\beta}{\beta}, \frac{u^\alpha}{\alpha}\Big], \tag{3.12}$$

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

Note that the functions (3.9) and (3.10) 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 (3.12) 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 (3.9) and (3.10) are by a generalization of the power odd activation function (3.7) (or(3.12)). This generalization is that function (3.10) (unlike function (3.7)) is differentiable. Therefore, it becomes possible to use these functions in the gradient methods of search algorithms.

3.2. Example

Consider the example of a generalized cubic root function that is differentiable on the entire real line. To do this we will use formulas (3.10) and (3.11) at $\alpha = \beta = 1/3$; $c = 1$. Then we have

$$w(u) = \mathit{piecewise}\Big[u < -1, 2 - 3(-u)^{1/3}, u \leq 1, u, -2 + 3u^{1/3}\Big]$$

and

$$\dot{w}_u(u) = \mathit{piecewise}\Big[u < -1, (-u)^{-2/3}, u \leq 1, 1, u^{-2/3}\Big].$$

Thus, the function $w(u)$ is differentiable over the entire interval $(-\infty, \infty)$. The derivative of this function $\dot{w}_u$ is continuous (but not differentiable!) and bounded also over the entire interval $(-\infty, \infty)$: $\dot{w}_u(u) \in (0, 1]$ (see Fig. 3.2).

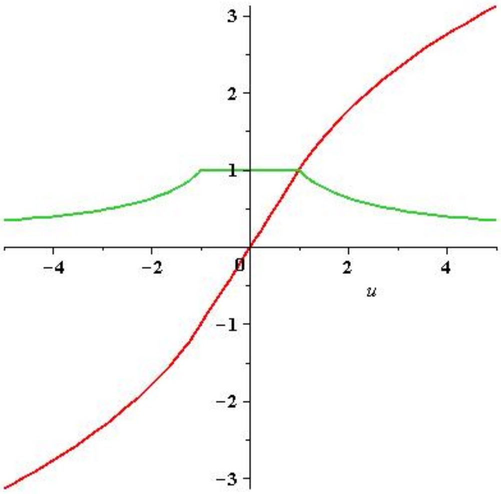


Fig. 3.2. Differentiable power activation function and its derivative: $w(u)$ (red), $\dot{w}_u(u)$(green)

4. Time series prediction

4.1. Direct prediction without explicit parametrization

As mentioned earlier, the main flaw of the basic NODE block is that it restricts the dimensionality of the model by requiring the produced trajectories to be in the same space as the modelled series. Since the NODE block is an autonomous system, in case of modeling the univariate time series data the model must be one-dimensional. This restricts the model's behavior to only 3 basic types depending on its Lyapunov exponent and makes it impossible to model any periodic or quasi-periodic processes. One way to avoid this issue is to turn the univariate time series into multivariate data by embedding it with estimated minimum embedding dimension d and the lag times $\{\tau_1, \tau_2, \ldots, \tau_{d-1}\}$ by using any of the available delay embedding procedures [10, 13]. This operation will provide the basic context for the model to work with by producing unique combinations of points $X^* \in \mathbb{R}^d$, $X^* = \left\{x_n, x_{n-\tau_1}, \ldots, x_{n-\tau_{d-1}}\right\}$ that can be used to create autoregressive models of type $x_{n+1} = f(X^*, \Theta)$. However, if a regular NODE is used in such a model, the prediction will have dimension d. This can be partially solved by simply discarding all the dimensions except one. The model of this type will impose unnecessary restrictions on the underlying ODE which will essentially be required to fit every coordinate of its trajectories to the same set of data which simply was time-lagged. This is not how most high-dimensional ODEs normally behave. To resolve this issue, the ODE can be allowed to function in its own latent space L that does not impose any such restrictions. The initial values in this latent space can be produced by a more complicated encoder $\mathcal{E}$ that delay-embeds the time series data and applies extra transformations in an attempt to decorrelate the dimensions of X^*.

The encoding step removes the hard coupling of the ODE and the time series data. However, the trajectories produced by the ODE in its latent space L will have to be mapped back to the original one-dimensional time series space. This is handled by the decoder $\mathcal{D}$.

The encoder and decoder make the ODE completely decoupled from the original time series which allows the researchers to freely select its structure and dimensionality. For the purposes of modeling univariate time series data which often exhibits quasi-periodic behavior, it is reasonable to select the model that can capture such behaviors. One such model is based on the AntisymmetricRNN [7] which is given by the equation:

$$\mathbf{h}_n = \mathbf{h}_{n-1} + \epsilon\sigma((\mathbf{W_h} - \mathbf{W_h}^T - \gamma\mathbf{I})\mathbf{h}_{n-1} + \mathbf{V}_h x_n + \mathbf{b_h}), \tag{4.1}$$

where $\mathbf{h}$ is the hidden state, $\mathbf{x}$ is the input, and σ is the activation function.

The idea of AntisymmetricRNN is to structurally enforce the periodicity of the model's trajectories by making sure that the eigenvalues of its Jacobian have either zero or slightly negative real parts. The model (4.1) is designed to prevent the hidden state $\mathbf{h}$ from growing or diminishing rapidly as it is carried from one data point to another. The hNODE model doesn't have hidden state, and it aims to model processes that may exhibit drifting behavior which may require the model's Jacobian eigenvalues to have positive values. With these considerations, the latent ODE G in (2.7) then becomes:

$$G(l, \mathbf{W}, \mathbf{D}, \mathbf{b}) = l_0 + \int_0^T \sigma((\mathbf{W} - \mathbf{W}^T + \mathbf{D})l + \mathbf{b})\, dt, \tag{4.2}$$

where the parameters $\mathbf{W}$, $\mathbf{D}$, and $\mathbf{b}$ are produced by the hypermodel $\mathcal{H}(X, \Theta)$.

W is matrix with unrestricted values that is converted into its antisymmetric form by subtracting a transposed version of it from itself, **D** is a diagonal matrix, and **b** is the bias vector. Pÿhe vector $l \in L$ is a point in the latent space L.

Considering the possibility of drift or other qualitative changes that may occur in the time series, the model (4.2) has to be equipped to react to such changes. And this is where the function $\mathcal{H}$ for generating control weights comes in. It produces a new set of parameters for (4.2) each time it is evaluated, allowing it change behavior based on where the current point is in the latent space. The activation function σ used in the examples below was selected as (3.10) with parameters $\alpha = 0.5$, $\beta = 0.3$, and $c = 5$; the plot of the activation function and its derivative is provided on Fig. 4.1.

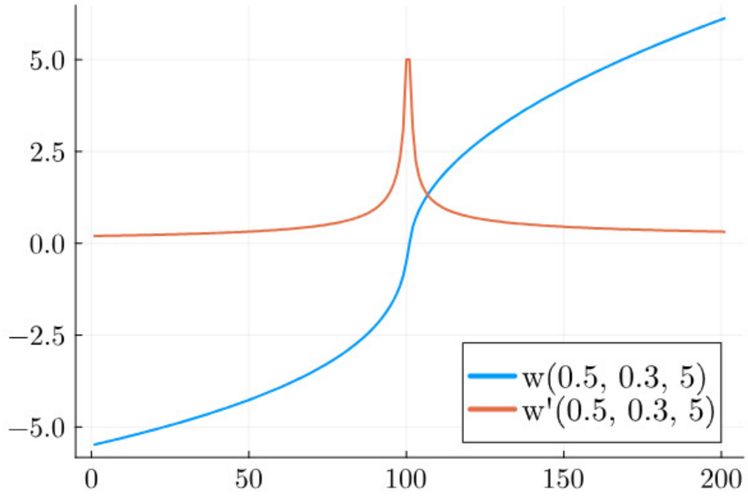
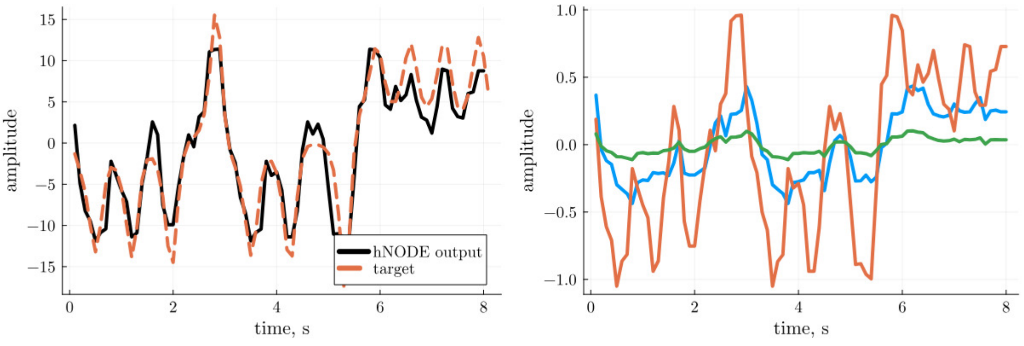


Fig. 4.1. Continuously differentiable activation function and its non-differentiable first derivative.

The final architecture of the hNODE with all the pieces assembled together is given on the Fig. 2.1.

A model with the architecture described above was used to learn and predict the values of the first coordinate of the Lorenz system's ($\dot{x} = \sigma(y - x); \dot{y} = x(\rho - z) - y; \dot{z} = xy - \beta z$) chaotic attractor. The attractor was generated with the parameters $\sigma = 10.0$, $\rho = 28$, $\beta = 8/3$ after which the y and z coordinates were discarded. The remaining x coordinate was delay-embedded with dimension $d = 3$ and delay $\tau = 1$. The resulting dataset split into chunks of size $d \cdot \tau + M$ ($M \geq 1$) and shaped as follows: $\{X_m, Y_m\}$, $m \in [(d-1) \cdot \tau, n - M]$, $X_m = \left\{x_m, x_{m-\tau}, \ldots, x_{m-(d-1)\cdot\tau}\right\}$, $Y_m = \{x_{m+1}, x_{m+2}, \ldots, x_{m+M}\}$. In other words, the first $d\tau$ points from a trajectory are picked as input context and the following M points are the expected output of the model.

The output of the model was produced recursively — the input data was used to predict the new point in the series after which the point was added to the



(a) (b)

Fig. 4.2. hNODE generates Lorenz-like output (a) from the latent trajectories of the underlying NODE (b).

input data and the first point in the input was removed. This process was applied iteratively until the desired time series prediction was obtained.

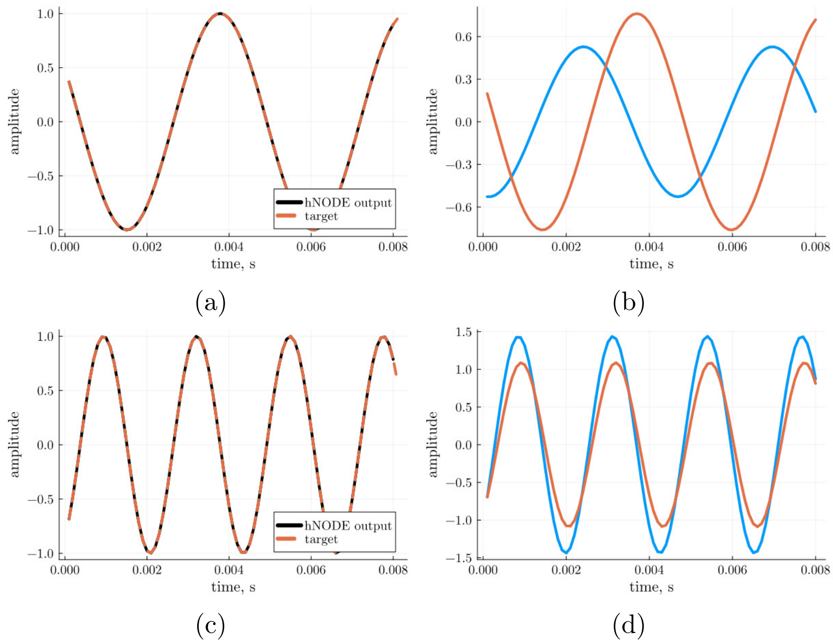
4.2. Prediction with explicit parametrization with parameter injection

Since the time series data is completely divorced from the latent space of the nested NODE, it is possible to augment the time series context with extra data to guide the model to specific behaviors. For example, the hNODE model can be trained to generate a sine wave with a specific frequency that is passed to the $\mathcal{H}$ function along with the current context.

As in the previous section, the model (4.2) was used for representing the nested NODE. The time series generated as a sine wave of different frequencies sampled at 8 kHz was delay-embedded with dimension $d = 2$ and delay $\tau = 1$ and the resulting dataset was processed similarly to the Lorenz attractor one. Each input data point X_m was augmented with the frequency F_0 of the waveform it was selected from: $X_m = \{x_m, x_{m-\tau}, F_0\}$. But unlike the previous example, the model's output was produced by obtaining the full trajectory in the latent space and decoding each point all at once which demonstrates that the two approaches are both valid and yield satisfactory results. The model's output for different injected frequencies are provided on the Fig. 4.3.

5. Activation functions performance comparison

As discussed in previous sections, the selection of an activation function is an important consideration when building a model. An incorrectly chosen activation function can slow down the learning rate or even make the model unfit for the problem. To demonstrate this, we selected a basic example similar to the one in the previous section — a sine wave that the model has to approximate without any extra parameters being injected into it. We compare the most widespread activation functions, ReLU and hyperbolic tangent, against the cubic



(a) (b)

(c) (d)

Fig. 4.3. hNODE generates sine waves for injected frequency parameter, the model's output is shown on (a, c) and the latent trajectories of the underlying NODE are shown on (b, d).

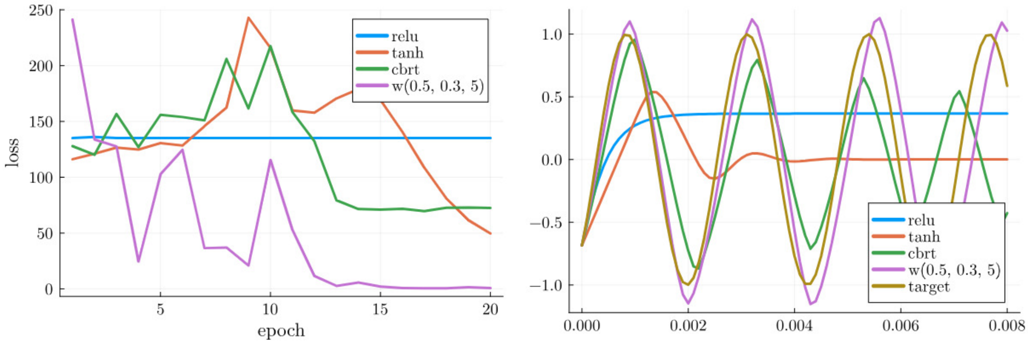
root ($cbrt(x) = \sqrt[3]{x}$) and the function (3.10) with parameters $\alpha = 0.5$, $\beta = 0.3$, and $c = 5$ to see how they perform when used inside the NODE nested in hNODE.

Exploiting the fact that the hNODE model consists of several standalone models, we train the encoder and decoder parts of it separately before proceeding to train the nested NODE in the latent space. Each NODE in the comparison was trained on two batches of data where the data points differ in length. The first batch contains pieces of latent trajectory of the length 5 and the second one 50. The rate of learning ε on the first batch is 0.01 and the second one is 0.001. Each model is trained on both batches back-to-back for 10 epochs. The results of the training are provided on Fig. 5.1.

As can be seen on the provided figures, the function (3.10) achieved the best accuracy and converged faster than any other activation function. The second-closest was the hyperbolic tangent which accelerated convergence towards the end of the training. The cubic root training plateaued at around the same point as the piecewise power function but with much worse loss values. The ReLU turned out to be the worst and turned out to be unfit for this particular modeling problem.

6. Conclusion

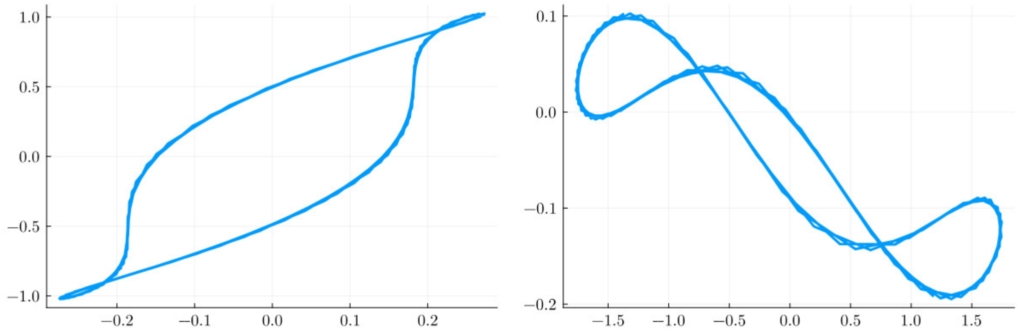
While the hNODE is versatile and can be adapted to a variety of applications, there are still some caveats that are mostly related to the nature of the underlying NODE model.



(a) (b)

Fig. 5.1. Loss over epochs for different activation functions (a) and trajectories predicted when using those functions (b).

The first caveat is that the latent space trajectories produced by the encoder from the input data cannot intersect as this would violate the theorem about existence and uniqueness of the ODE solution that the model is trying to approximate. An example of malformed latent space trajectories are given on Fig. 6.1.



(a) (b)

Fig. 6.1. Intersecting trajectories in the latent space generated for the sine data from the previous section.

The second caveat is the integration step used for obtaining numerical solutions for the nested NODE. It may seem reasonable to use the time intervals between the point in the input data as integration step. For example, one might use the inverse of the sampling frequency of the data but such inverse might be too small for the model to handle properly — even simple examples like the sine wave sampled at 8 kHz give unreasonably small integration step of 0.000125. This pitfall is more subtle than the previous one, but it can slow down the learning rate considerably. The example of using different integration steps to learn the example from the previous section with the piecewise power function is given on the Fig. 6.2.

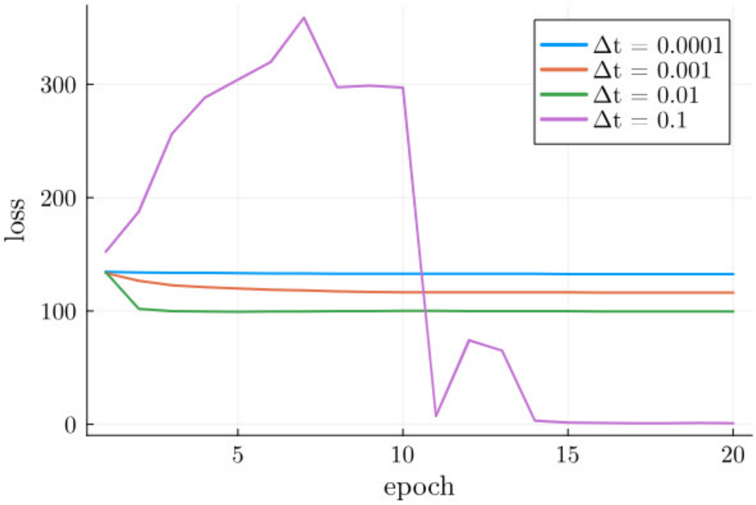


Fig. 6.2. Integration step Δt influencing the rate of learning.

In conclusion, we demonstrated the viability of the new approach to using NODE model by utilizing the latent space mappings using dedicated decoder and encoder models, and learning the dynamics of the modeled process in the latent space instead of the space of the process itself. It was also demonstrated that for the purposes of modeling time series in the latent space it can be beneficial to use continuously differentiable activation functions that do not approach constants when their argument approaches infinity.

References

1. O. I. Abiodun et al, *Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition*, *IEEE Access*, **7** (2019), 158820-158846.
2. R. Adhikari and R. K. Agrawal, *An Introductory Study on Time Series Modeling and Forecasting*, *CoRR*, abs/1302.6613, 2013.
3. V. Ye. Belozyorov and D. V. Dantsev, *Modeling of chaotic processes by means of antisymmetric neural ODEs*, *Journal of Optimization, Differential Equations and Their Applications (JODEA)*, **30**(1)(2022), 1–41.
4. V. Ye. Belozyorov and D. V. Dantsev, *Stability of neural ordinary differential equations with power nonlinearities*, *Journal of Optimization, Differential Equations, and Their Applications*, **28** (2) (2020), 21–46.
5. M. Castelli and L. Manzoni, *Generative models in artificial intelligence and their applications*, *Applied Sciences*, **12** (9) (2022), 4127.
6. M. Chalvidal, M. Ricci, R. VanRullen and T. Serre, *Neural Optimal Control for Representation Learning*, *CoRR*, abs/2006.09545 (2020).
7. B. Chang, M. Chen, E. Haber and E. H. Chi, *AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks*, *New Journal of Physics*, 1902.09689 (2019).
8. R. T. Q. Chen, Y. Rubanova, J. Bettencourt and D. K. Duvenaud, *Neural Ordinary Differential Equations*, *Advances in Neural Information Processing Systems*, **31** (2018).

9. P. Hendikawati et al, *A survey of time series forecasting from stochastic method to soft computing*, Journal of Physics: Conference Series, **1613** (1) (2020).

10. K. H. Kraemer, G. Datseris, J. Kurths, I. Z. Kiss, J. L. Ocampo-Espindola and N. Marwan, *A unified and automated approach to attractor reconstruction*, New Journal of Physics, **23** (3) (2021), 033017.

11. A. Norcliffe, C. Bodnar, B. Day, J. Moss and P. Liò, *Neural ODE Processes*, *CoRR*, abs/2103.12413 (2018).

12. K. O'Shea and R. Nash, *An Introduction to Convolutional Neural Networks*, *CoRR*, abs/1511.08458 (2015).

13. F. Takens, *Detecting strange attractors in turbulence*, Dynamical Systems and Turbulence, Warwick, 1981, 366–381.

14. G. Zhang, B. E. Patuwo and M. Y. Hu, *Forecasting with artificial neural networks:: The state of the art*, International Journal of Forecasting, **14** (1) (1998), 35-62.

15. J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang and L.-R. Dai, *Sequence-to-Sequence Acoustic Modeling for Voice Conversion*, *IEEE/ACM* Transactions on *Audio, Speech, and Language Processing*, **27** (3) (2019), 631-644.

Received 30.08.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 67–88, DOI 10.15421/142310

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

NONLINEAR PARABOLIC VARIATIONAL INEQUALITIES WITH VARIABLE TIME-DELAY IN TIME UNBOUNDED DOMAINS

Mykola M. Bokalo[*], Olga V. Ilnytska[†], Taras M. Bokalo[‡]

Communicated by Prof. O. Kupenko

Abstract. We research the well-posedness of the problem without initial condition for nonlinear parabolic variational inequalities with variable time-delay. To justify our results, we impose some assumptions on the solution behavior and growth of the data-in as time variable tends to $-\infty$. Also, we obtain estimates for weak solutions of this problem.

Key words: evolution variational inequality, evolution subdifferential inclusion, time delay, Fourier problem, unbounded domain.

2010 Mathematics Subject Classification: 26D10, 49J40, 47J20, 47J22.

1. Introduction

In this paper we consider the problem without initial condition or, in other words, Fourier problem for evolution variational inequalities (inclusions) with a time-depended delay. Let us introduce an example of the problem being studied here.

Let $p \geq 2$, Ω be a bounded domain in $\mathbb{R}^n$ ($n \in \mathbb{N}$), $\partial\Omega$ be the boundary of Ω. We put $Q := \Omega \times (-\infty, 0]$, $\Sigma := \partial\Omega \times (-\infty, 0]$, $\Omega_t := \Omega \times \{t\}$ $\forall\, t \in \mathbb{R}$. Let $L^p(\Omega)$ and $L^p(Q)$ be the standard Lebesgue spaces. Denote by $W^{1,p}(\Omega) := \{v \in L^p(\Omega) \mid v_{x_i} \in L^p(\Omega),\ i = \overline{1,n}\}$ the standard Sobolev space with the norm $||v||_{W^{1,p}(\Omega)} := \Big(\int_\Omega [|\nabla v|^p + |v|^p]\Big)^{1/p}$, where $\nabla v := (v_{x_1}, \ldots, v_{x_n})$.

Let K be a convex closed set in $W^{1,p}(\Omega)$ which contains 0. Let us consider the problem of finding a function $u \in L^p(Q)$ such that $u_{x_i} \in L^p(Q)$, $i = \overline{1,n}$, $u_t \in L^2(Q)$, and, for a.e. $t \in (-\infty, 0]$, $u(\cdot, t) \in K$ and

$$\int_{\Omega_t} \Big[u_t(v-u) + |\nabla u|^{p-2}\nabla u \nabla(v-u) + |u|^{p-2}u(v-u) + \widehat{b}u(v-u) \\ + \widehat{c}(v-u)\int_{t-\tau(t)}^{t} u(x,s)\,ds\Big]\,dx \geq \int_{\Omega_t} f(v-u)\,dx \quad \forall\, v \in K, \tag{1.1}$$

[*] Ivan Franko National University of Lviv, Department of Mathematical Statistics and Differential Equations, 1, Universytetska St., Lviv, 79000, Ukraine, mm.bokalo@gmail.com

[†] Ivan Franko National University of Lviv, Department of Computational Mathematics, 1, Universytetska St., Lviv, 79000, Ukraine, ol.ilnytska@gmail.com

[‡] Ivan Franko National University of Lviv, Department of Mathematical Statistics and Differential Equations, 1, Universytetska St., Lviv, 79000, Ukraine, tbokalo@gmail.com

where $\widehat{b}, \widehat{c}$ are positive constants, $f \in L^2(Q)$, $\tau \in C((-\infty,0])$, $\tau(t) \geq 0$ $\forall t \in (-\infty,0]$, $\tau^+ := \sup_{t\in(-\infty,0]} \tau(t) < \infty$.

As it will be shown below, this problem, which we will call Problem (1.1), has a unique solution, when $\widehat{b} - \widehat{c}\tau^+ > 0$.

Note that Problem (1.1) can be written in more abstract way. Indeed, after an appropriate identification of functions and functionals, we have continuous and dense imbedding

$$W^{1,p}(\Omega) \subset L^2(\Omega) \subset (W^{1,p}(\Omega))',$$

where $(W^{1,p}(\Omega))'$ is dual to $W^{1,p}(\Omega)$ space. Clearly, for any $h \in L^2(\Omega)$ and $v \in W^{1,p}(\Omega)$ we have $\langle h, v\rangle = (h,v)$, where $\langle\cdot,\cdot\rangle$ is the notation for scalar product on the dual pair $\big[(W^{1,p}(\Omega))', W^{1,p}(\Omega)\big]$, and $(\cdot,\cdot)$ is the scalar product in $L^2(\Omega)$. Thus, we will use the notation $(\cdot,\cdot)$ instead of $\langle\cdot,\cdot\rangle$.

Now, we denote $S := (-\infty,0]$, $V := W^{1,p}(\Omega)$, $H := L^2(\Omega)$ and define an operator $A : V \to V'$ as follows

$$(A(v), w) = \int_\Omega \big[|\nabla v|^{p-2}\nabla v \nabla w + |v|^{p-2}vw + \widehat{b}vw\big]\,dx, \quad v, w \in V.$$

Then, Problem (1.1) becomes equivalent to the next problem: to find a function $u \in L^p(S;V)$ such that $u' \in L^2(S;H)$, and, for a.e. $t \in S$, $u(t) \in K$ and

$$(u'(t) + A(u(t)) + \widehat{c}\int_{t-\tau(t)}^{t} u(s)\,ds, v - u(t)) \geq (f(t), v - u(t)) \quad \forall\, v \in K. \tag{1.2}$$

Here $f \in L^2(S;H)$, τ is as above.

Note that variational inequality (1.2) can be written as a subdifferential inclusion. For this purpose, we put $I_K(v) := 0$ if $v \in K$, and $I_K(v) := +\infty$ if $v \in V \setminus K$, and also

$$\Phi(v) := \int_\Omega \big[p^{-1}|\nabla v|^p + p^{-1}|v|^p + 2^{-1}\widehat{b}|v|^2\big]\,dx + I_K(v), \quad v \in V.$$

It is easy to verify that the functional $\Phi : V \to \mathbb{R}_\infty := (-\infty; +\infty]$ is proper, convex and semi-lower-continuous. By the known results (see, e.g., [22, p. 83]), it follows that the problem of finding a solution of variational inequality (1.2) can be written as the following subdifferential inclusion: to find a function $u \in L^p(S;V)$ such that $u' \in L^2(S;H)$ and, for a.e. $t \in S$, $u(t) \in D(\partial\Phi)$ and

$$u'(t) + \partial\Phi(u(t)) + \widehat{c}\int_{t-\tau(t)}^{t} u(s)\,ds \ni f(t) \quad \text{in} \quad H, \tag{1.3}$$

where $\partial\Phi : V \to 2^{V'}$ is a subdifferential of Φ, $D(\partial\Phi)$ is a domain of $\partial\Phi$ ($\partial\Phi$ and $D(\partial\Phi)$ will be defined of later).

The aim of this paper is to investigate problems for inclusions of type (1.3).

Let us mention that initial-value problems for evolution inclusions with constant delay were studied in [19], [25], [26] and others. Many results on such problems were obtained by using the semi-group theory. Refer to [25] for more comments and citations. In [19], [26] the fixed point theorems were used.

Problem without initial conditions for evolution equations arise in modeling different nonstationary processes in nature, that started long time ago and initial conditions do not affect on them in the actual time moment. Thus, we can assume that the initial time is $-\infty$, while 0 is the final time, and initial conditions can be replaced with the behaviour of the solution as time variable turns to $-\infty$. Such problems appear in modeling in many fields of science such as ecology, economics, physics, cybernetics, etc. The research of the problem without initial conditions for the evolution equations and variational inequalities (without delay) were conducted in the monographs [15], [17], [22], and the papers [3], [4], [5], [7], [10], [14], [16], [18], [23] and others. In particular, R.E. Showalter in the paper [21] proved the existence of unique solution $u \in \mathrm{e}^{2\omega\cdot}H^1(S;H)$, where H is a Hilbert space, of the problem without initial condition

$$u'(t) + \mu u(t) + A\big(u(t)\big) \ni f(t), \quad t \in S,$$

for $\omega + \mu > 0$ and $f \in \mathrm{e}^{2\omega\cdot}H^1(S;H)$ in case when $A : H \to 2^H$ is a maximal monotone operator such that $0 \in A(0)$. Moreover, if $A = \partial\varphi$, where $\varphi : H \to (-\infty, +\infty]$ is proper, convex and lower-semi-continuous functional such that $\varphi(0) = 0 = \inf\{\varphi(v) \mid v \in H\}$, then this problem has a unique solution for each $\mu > 0$, $f \in L^2(S;H)$ and $\omega = 0$.

Note that the uniqueness of the solutions of problem without initial conditions for linear parabolic equations and variational inequalities is possible only under some restrictions on the behavior of solutions when time variable tends to $-\infty$. For the first time it was strictly justified by A.N. Tikhonov [24] in the case of heat equation. However, as it was shown by M.M. Bokalo [3], problem without initial conditions for some nonlinear parabolic equations has a unique solution in the class of functions without behavior restriction as time variable tends to $-\infty$. Similar results were also obtained for evolutionary variational inequalities in the paper [4].

Previously, problems without initial conditions of evolution equations with constant delay were studied in [6], [11], and with variable delay, as far as we know, only in [13]. Let us note that problems without initial conditions for variational inequalities or inclusions with delay have not been considered in the literature, which serves as one of the motivations for the study of such problems.

The outline of this paper is as follows. In Section 2, we give notations, definitions of function spaces and auxiliary results. In Section 3, we formulate the problem and main result. In Section 4, we prove the main result.

2. Preliminaries

We set, as above, $S := (-\infty, 0]$. Let V be a reflexive and separable Banach space with norm $\|\cdot\|$, and H be a separable Hilbert spaces with the scalar product

$(\cdot,\cdot)$ and norm $|\cdot|$. Suppose that $V \subset H$ with dense, continuous and compact injection.

Let V' and H' be the dual spaces to V and H, respectively. We suppose (after appropriate identification of functionals), that the space H' is a subspace of V'. By the Riesz-Fréchet representation theorem, identifying the spaces H and H', we obtain the dense and continuous embeddings

$$V \subset H \subset V'. \tag{2.1}$$

Note that in this case $\langle g, v\rangle_V = (g, v)$ for every $v \in V, g \in H$, where $\langle\cdot,\cdot\rangle_V$ is the scalar product for the dual pair $\big[V', V\big]$. Thus, further we will be using notation $(\cdot,\cdot)$ instead of $\langle\cdot,\cdot\rangle_V$.

We introduce some spaces of functions and distributions. Let X be an arbitrary Banach space with the norm $\|\cdot\|_X$. By $C(S;X)$ we mean the linear space of continuous functions defined on S with values in X. We say that $w_m \longrightarrow_{m\to\infty} w$ in $C(S;X)$ if for each $t_1, t_2 \in S,\ t_1 < t_2,$ the sequence of the restrictions of the functions $\{w_m\}_{m=1}^{\infty}$ to segment $[t_1, t_2]$ converges in $C([t_1, t_2]; X)$ to the restriction of w to the same segment.

Let $q \in [1, \infty]$, q' be dual to q, i.e., $1/q + 1/q' = 1$. Denote by $L^q_{\text{loc}}(S;X)$ the linear space of measurable functions defined on S with values in X, whose restrictions to any segment $[t_1, t_2] \subset S$ belong to the space $L^q(t_1, t_2; X)$. We say that a sequence $\{w_m\}$ is bounded (respectively, strongly, weakly or $*$-weakly convergent to w) in $L^q_{\text{loc}}(S;X)$, if for each $t_1, t_2 \in S,\ t_1 < t_2,$ the sequence of restrictions of $\{w_m\}$ to the segment $[t_1, t_2]$ is bounded (respectively, strongly, weakly or $*$-weakly convergent to the restriction of w to segment $[t_1, t_2]$) in $L^q(t_1, t_2; X)$.

By $D'(-\infty, 0; V')$ we mean the space of continuous linear functionals on $D(-\infty, 0)$ with values in V'_w. Hereafter $D(-\infty, 0)$ is a space of test functions, that is, the space of infinitely differentiable on $(-\infty, 0)$ functions with compact supports, equipped with the corresponding topology, and V'_w is the linear space V' equipped with weak topology. It is easy to see (using (2.1)), that spaces $L^q_{\text{loc}}(S;V)$, $L^2_{\text{loc}}(S;H)$, $L^{q'}_{\text{loc}}(S;V')$ can be identified with the corresponding subspaces of $D'(-\infty, 0; V')$. In particular, this allows us to talk about derivatives w' of functions w from $L^q_{\text{loc}}(S;V)$ or $L^2_{\text{loc}}(S;H)$ in the sense of distributions $D'(-\infty, 0; V')$ and belonging of such derivatives to $L^{q'}_{\text{loc}}(S;V')$ or $L^2_{\text{loc}}(S;H)$.

Let us define the spaces

$$H^1(S;H) := \{w \in L^2(S;H) \,|\, w' \in L^2(S;H)\},$$

$$W^1_{q,\text{loc}}(S;V) := \{w \in L^q_{\text{loc}}(S;V) \,|\, w' \in L^{q'}_{\text{loc}}(S;V')\}, \quad q > 1.$$

From known results (see., for example, [12, p. 177-179]) it follows that $H^1(S;H) \subset C(S;H)$ and $W^1_{q,\text{loc}}(S;V) \subset C(S;H)$. Moreover, for every w from $H^1(S;H)$ or $W^1_{q,\text{loc}}(S;V)$ the function $t \to |w(t)|^2$ is absolutely continuous on any segment of the interval S and the following equality holds

$$\frac{d}{dt}|w(t)|^2 = 2(w'(t), w(t)) \quad \text{for a.e.} \quad t \in S. \tag{2.2}$$

Remark 2.1. For $w \in L^2(S;H)$, we have

$$\lim_{\sigma\to-\infty}\int_{\sigma-1}^{\sigma}|w(t)|^2dt=0.$$

If $w \in L^2(S;H)\cap C(S;H)$ then there exists a sequence $\{t_k\}_{k=0}^{\infty}\subset S$ such that $t_k\to-\infty$ for $k\to+\infty$ and

$$\lim_{k\to+\infty}|w(t_k)|^2=0.$$

In this paper we use the following well-known facts.

Proposition 2.1 (Cauchy-Schwarz-Bunjakovsky inequality; see, for example, [12, p. 158])**.** Let $t_1,t_2\in\mathbb{R}$, $t_1<t_2$, and X is a Hilbert space with the scalar product $(\cdot,\cdot)_X$. Then, if $v\in L^2(t_1,t_2;X)$ and $w\in L^2(t_1,t_2;X)$, we have $(w(\cdot),v(\cdot))_X\in L^1\big(t_1,t_2\big)$ and

$$\int_{t_1}^{t_2}(w(t),v(t))_X\,dt\leqslant\|w\|_{L^2(t_1,t_2;X)}\|v\|_{L^2(t_1,t_2;X)}.$$

Proposition 2.2 ([27, p. 173,179])**.** Let Y be a Banach space with the norm $\|\cdot\|_Y$, and $\{v_k\}_{k=1}^{\infty}$ be a sequence of elements of Y, which is weakly or $*$-weakly convergent to v in Y. Then, $\underline{\lim}_{k\to\infty}\|v_k\|_Y\geqslant\|v\|_Y$.

Proposition 2.3 (Aubin theorem [1], [2, p. 393])**.** Let $q>1, r>1$, $t_1,t_2\in\mathbb{R}$, $t_1<t_2$, and B_0,B_1,B_2 are Banach spaces such that $B_0\subset^c B_1\subset B_2$ (here $\subset^c$ means compact embedding and $\subset$ means continuous embedding). Then

$$\{w\in L^q(t_1,t_2;B_0)\mid w'\in L^r(t_1,t_2;B_2)\}\overset{c}{\subset}\big(L^q(t_1,t_2;B_1)\cap C([t_1,t_2];B_2)\big).\quad(2.3)$$

Note that we understand the embedding (2.3) as follows: if a sequence $\{w_m\}$ is bounded in the space $L^q(t_1,t_2;B_0)$ and the sequence $\{w_m'\}_{m\in\mathbb{N}}$ is bounded in the space $L^r(t_1,t_2;B_2)$, then there exist a function $w\in L^q(t_1,t_2;B_1)\cap C([t_1,t_2];B_2)$ and a subsequence $\{w_{m_j}\}$ of the sequence $\{w_m\}$ such that $w_{m_j}\longrightarrow_{j\to\infty}w$ in $C([t_1,t_2];B_2)$ and strongly in $L^q(t_1,t_2;B_1)$.

Lemma 2.1. *If a sequence $\{w_m\}$ is bounded in the space $L^q_{\mathrm{loc}}(S;V), q>1$, and the sequence $\{w_m'\}$ is bounded in the space $L^2_{\mathrm{loc}}(S;H)$, then there exist a function $w\in L^q_{\mathrm{loc}}(S;V)$, $w'\in L^2_{\mathrm{loc}}(S;H)$, and a subsequence $\{w_{m_j}\}$ of the sequence $\{w_m\}$ such that $w_{m_j}\longrightarrow_{j\to\infty}w$ in $C(S;H)$ and weakly in $L^q_{\mathrm{loc}}(S;V)$, and, $w_{m_j}'\longrightarrow_{j\to\infty}w'$ weakly in $L^2_{\mathrm{loc}}(S;H)$.*

Proof. The Proposition 2.3 for $r=2$, $B_0=V$, $B_1=B_2=H$ and the reflexiveness of V and H yields, for every $t_1,t_2\in S$, $t_1<t_2$, from the sequence of restrictions of the elements $\{w_m\}$ to the segment $[t_1,t_2]$ one can choose a subsequence which is convergent in $C([t_1,t_2];H)$ and weakly in $L^q(t_1,t_2;V)$, and the

sequence of derivatives of the elements of this subsequence is weakly convergent in $L^2(t_1,t_2;H)$. For each $k\in\mathbb{N}$, we choose a subsequence $\{w_{m_{k,j}}\}_{j=1}^{\infty}$ of the given sequence, which is convergent in $C([-k,0];H)$ and weakly in $L^q(-k,0;V)$ to some function $\widehat{w}_k\in C([-k,0];H)\cap L^q(-k,0;V)$, and the sequence $\{w'_{m_{k,j}}\}_{j=1}^{\infty}$ is weakly convergent to the derivative $\widehat{w}'_k$ in $L^2(-k,0;H)$. Making this choice we ensure that the sequence $\{w_{m_{k+1,j}}\}_{j=1}^{\infty}$ was a subsequence of the sequence $\{w_{m_{k,j}}\}_{j=1}^{\infty}$. Now, according to the diagonal process, we select the desired subsequence as $\{w_{m_{j,j}}\}_{j=1}^{\infty}$, and we define the function w as follows: for each $k\in\mathbb{N}$ we take $w(t):=\widehat{w}_k(t)$ for $t\in(-k,-k+1]$. □

In the sequel the Cauchy inequality of the following form will be used

$$ab\le\varepsilon a^2+(4\varepsilon)^{-1}b^2\quad\forall a,b\in\mathbb{R},\,\forall\varepsilon>0.\tag{2.4}$$

3. Setting of the problem and main result

Let $\Phi:V\to(-\infty,+\infty]$ be a proper functional, i.e., $\mathrm{dom}(\Phi):=\{v\in V:\Phi(v)<+\infty\}\neq\emptyset$, which satisfies such conditions:

$(\mathcal{A}_1)$ $\quad\Phi\big(\alpha v+(1-\alpha)w\big)\leqslant\alpha\Phi(v)+(1-\alpha)\Phi(w)\quad\forall\,v,w\in V,\ \forall\,\alpha\in[0,1],$

i.e., the functional Φ is *convex*,

$(\mathcal{A}_2)$ $\quad v_k\longrightarrow_{k\to\infty}v$ in $V\quad\Longrightarrow\quad\underline{\inf}_{k\to\infty}\Phi(v_k)\ge\Phi(v),$

i.e., the functional Φ is *lower semicontinuous.*

Recall that the *subdifferential* of functional Φ is a mapping $\partial\Phi:V\to 2^{V'}$, defined as follows

$$\partial\Phi(v):=\{v^*\in V'\mid\Phi(w)\geqslant\Phi(v)+(v^*,w-v)\quad\forall\,w\in V\},$$

for any $v\in V$, and the *domain* of the subdifferential $\partial\Phi$ is the set $D(\partial\Phi):=\{v\in V\,|\,\partial\Phi(v)\neq\emptyset\}$. We identify the subdifferential $\partial\Phi$ with its graph, assuming that $[v,v^*]\in\partial\Phi$ if and only if $v^*\in\partial\Phi(v)$, i.e., $\partial\Phi=\{[v,v^*]\mid v\in D(\partial\Phi),\ v^*\in\partial\Phi(v))\}$. R. Rockafellar (see [20, Theorem A]) proves that the subdifferential $\partial\Phi$ is a *maximal monotone operator*, that is,

$$(v_1^*-v_2^*,v_1-v_2)\geqslant 0\quad\forall\,[v_1,v_1^*],\ [v_2,v_2^*]\in\partial\Phi,$$

and for every element $[v_1,v_1^*]\in V\times V'$ we have the implication

$$(v_1^*-v_2^*,v_1-v_2)\geqslant 0\quad\forall\,[v_2,v_2^*]\in\partial\Phi\implies[v_1,v_1^*]\in\partial\Phi.$$

Additionally, assume that the following conditions hold:

($\mathcal{A}_3$) there exist constants $p \geq 2,\ K_1 > 0$ such that

$$\Phi(v) \geqslant K_1 \|v\|^p \quad \forall\ v \in \mathrm{dom}(\Phi);$$

moreover, $\Phi(0) = 0$;

($\mathcal{A}_4$) there exists a constant $K_2 > 0$ such that

$$(v_1^* - v_2^*, v_1 - v_2) \geqslant K_2 |v_1 - v_2|^2 \quad \forall\ [v_1, v_1^*],\ [v_2, v_2^*] \in \partial\Phi.$$

Remark 3.1. Condition $\Phi(0) = 0$ (see ($\mathcal{A}_3$)) implies that $\Phi(v) \geq \Phi(0) + (0, v - 0)$ $\forall v \in V$, hence, $[0, 0] \in \partial\Phi$. From this and condition ($\mathcal{A}_4$) we have

$$(v^*, v) \geq K_2 |v|^2 \quad \forall\, [v, v^*] \in \partial\Phi. \tag{3.1}$$

Let $\tau : S \to \mathbb{R}$ be a function such that

($\mathcal{T}$) $\tau \in C(S)$, $\tau(t) \geq 0$ for all $t \in S$, $\tau^+ := \sup_{t \in S} \tau(t) < \infty$.

Let $c : \Pi_\tau \times H \to H$, where $\Pi_\tau := \{(t, s)\,|\, t \leq 0,\ t - \tau(t) \leq s \leq t\}$ and τ satisfies condition ($\mathcal{T}$), be a function which satisfies the condition:

($\mathcal{C}$) for any $v \in H$ the mapping $c(\cdot, \cdot, v) : \Pi_\tau \to H$ is measurable, and there exists a constant $L \geq 0$ such that following inequality holds

$$|c(t, s, v_1) - c(t, s, v_2)| \leq L |v_1 - v_2|$$

for a.e. $(t, s) \in \Pi_\tau$ and for all $v_1, v_2 \in H$; in addition, $c(t, s, 0) = 0$ for a.e. $(t, s) \in \Pi_\tau$.

Remark 3.2. From the condition ($\mathcal{C}$), it follows that, for a.e. $(t, s) \in \Pi_\tau$ and for every $v \in H$, the following estimate is valid:

$$|c(t, s, v)| \leq L|v|. \tag{3.2}$$

Remark 3.3. Conditions ($\mathcal{T}$), ($\mathcal{C}$) and remark 3.2 yield, for any function $w \in L^2(S; H)$ the function $t \mapsto \int_{t-\tau(t)}^t c(t, s, w(s))\, ds : S \to H$ belongs to $L^2(S; H)$.

Indeed, by (3.2), assuming that $w(t) = 0$ for all $t > 0$, using Cauchy-Schwarz-Bunjakovsky inequality and changing the order of integration, we have

$$\begin{aligned}
\int_\sigma^0 \Big| \int_{t-\tau(t)}^t c(t, s, w(s))\, ds \Big|^2 dt &\leq L^2 \tau^+ \int_\sigma^0 \int_{t-\tau^+}^t |w(s)|^2\, ds dt \\
\leq L^2 \tau^+ \int_{\sigma-\tau^+}^0 |w(s)|^2\, ds \int_s^{s+\tau^+} dt &= (L\tau^+)^2 \int_{\sigma-\tau^+}^0 |w(s)|^2\, ds \\
&\leq (L\tau^+)^2 ||w||^2_{L^2(S;H)}
\end{aligned} \tag{3.3}$$

for each $\sigma \in S$. Thus, the function $t \mapsto \int_{t-\tau(t)}^t c(t, s, w(s))\, ds$ belongs to $L^2(S; H)$.

Let us consider the evolutionary variational inequality

$$u'(t) + \partial\Phi\big(u(t)\big) + \int_{t-\tau(t)}^{t} c(t,s,u(s))\,ds \ni f(t), \quad t \in S, \tag{3.4}$$

where $f : S \to V'$ is a given measurable function, and $u : S \to V$ is an unknown function.

Definition 3.1. Let conditions $(\mathcal{A}_1)$ – $(\mathcal{A}_3)$, $(\mathcal{T})$, $(\mathcal{C})$ hold, and $f \in L^{p'}_{\mathrm{loc}}(S;V')$. The *solution* of variational inequality (3.4) is a function $u : S \to V$ that satisfies the following conditions:

1) $u \in W^1_{p,\mathrm{loc}}(S;V)$;

2) $u(t) \in D(\partial\Phi)$ for a.e. $t \in S$;

3) there exists a function $g \in L^{p'}_{\mathrm{loc}}(S;V')$ such that, for a.e. $t \in S$, $g(t) \in \partial\Phi\big(u(t)\big)$ and

$$u'(t) + g(t) + \int_{t-\tau(t)}^{t} c(t,s,u(s))\,ds = f(t) \quad \text{in} \;\; V'.$$

We consider the problem of finding a solution u of variational inequality (3.4) for given Φ, c, τ and f such that

$$\int_S |u(t)|^2\,dt < +\infty, \quad \text{that is} \;\; u \in L^2(S;H). \tag{3.5}$$

This problem is called the *problem* $\mathbf{P}(\Phi,\tau,c,f)$, and the function u is called its *solution.*

Theorem 3.1. *Let conditions* $(\mathcal{A}_1)$ – $(\mathcal{A}_4)$, $(\mathcal{T})$, $(\mathcal{C})$ *hold,* $f \in L^2(S;H)$, *and*

$$K_2 - L\tau^+ > 0. \tag{3.6}$$

Then the problem $\boldsymbol{P}(\Phi,\tau,c,f)$ *has a unique solution, it belongs to the space* $L^\infty(S;V) \cap L^p(S;V) \cap H^1(S;H)$ *and satisfies the estimate*

$$\operatorname*{ess\,sup}_{t\in S} \|u(t))\|^p + \int_S \big(\|u(t)\|^p + |u(t)|^2 + |u'(t)|^2\big)\,dt$$

$$+ \int_S \Phi(u(t))dt \leqslant C_1 \int_S |f(t)|^2\,dt, \tag{3.7}$$

where C_1 *is a positive constant depending on* $K_1, K_2, L,$ *and* τ^+ *only.*

Remark 3.4. The problem $\mathbf{P}(\Phi, \tau, c, f)$ can be replaced by the following one. Let K be a convex and closed set in V, $A : V \to V'$ be a monotone, bounded and semicontinuous operator such that $(A(v), v) \geq \widetilde{K}_1 \|v\|^p \quad \forall v \in V$, where $p \geq 2, \widetilde{K}_1 = \text{const} > 0$. The problem is to find a function $u \in W^1_{p,\text{loc}}(S; V) \cap L^2(S; H)$ such that for a.e. $t \in S$, $u(t) \in K$ and

$$(u'(t) + A(u(t)) + \int_{t-\tau(t)}^{t} c(t, s, u(s))\, ds, v - u(t)) \geq (f(t), v - u(t)) \quad \forall\, v \in K.$$

4. Proof of the main result

We divide the proof of Theorem 3.1 into seven steps.

Step 1 (*uniqueness of solution*). Assume the contrary. Let u_1, u_2 be two solutions of the problem $\mathbf{P}(\Phi, \tau, c, f)$. Then for every $i \in \{1, 2\}$ there exists function $g_i \in L^{p'}_{\text{loc}}(S; V')$ such that, for a.e. $t \in S$, we have $g_i(t) \in \partial\Phi\big(u_i(t)\big)$ and

$$u_i'(t) + g_i(t) + \int_{t-\tau(t)}^{t} c(t, s, u_i(s))\, ds = f(t) \quad \text{in } V', \quad i = 1, 2. \tag{4.1}$$

We put $w := u_1 - u_2$. From equalities (4.1), for a.e. $t \in S$, we obtain

$$w'(t) + g_1(t) - g_2(t) + \int_{t-\tau(t)}^{t} \big(c(t, s, u_1(s)) - c(t, s, u_2(s))\big) ds = 0 \;\text{ in }\; V'. \tag{4.2}$$

Let $t_1, t_2 \in S$ be arbitrary numbers such that $t_1 < t_2$. Multiplying equality (4.2) by $w(t)$ and integrating from t_1 to t_2, we have

$$\int_{t_1}^{t_2} (w'(t), w(t))\, dt + \int_{t_1}^{t_2} \big(g_1(t) - g_2(t), u_1(t) - u_2(t)\big)\, dt \\ + \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} (c(t, s, u_1(s)) - c(t, s, u_2(s)))\, ds, w(t)\Big)\, dt = 0. \tag{4.3}$$

Consider the third term from left-hand side of equality (4.3). By condition $(\mathcal{T})$, $(\mathcal{C})$, the Fubini Theorem and the Cauchy-Schwarz-Bunjakovsky inequality, we get

$$\Big| \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} \big(c(t, s, u_1(s)) - c(t, s, u_2(s))\big)\, ds, w(t)\Big) dt \Big| \\ \leq \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} \big|c(t, s, u_1(s)) - c(t, s, u_2(s))\big| ds\Big) |w(t)|\, dt \\ \leq L \int_{t_1}^{t_2} \Big(\int_{t-\tau^+}^{t} |w(s)|\, ds\Big) |w(t)|\, dt \\ \leq L\sqrt{\tau^+} \Big(\int_{t_1}^{t_2} |w(t)|^2\, dt\Big)^{1/2} \Big(\int_{t_1}^{t_2} \Big(\int_{t-\tau^+}^{t} |w(s)|^2\, ds\Big)\, dt\Big)^{1/2}. \tag{4.4}$$

Changing the order of integration, we have

$$\int_{t_1}^{t_2}\Big(\int_{t-\tau^+}^{t}|w(s)|^2ds\Big)\,dt \le \int_{t_1-\tau^+}^{t_2}|w(s)|^2\,ds\int_{s}^{s+\tau^+}\,dt \\ = \tau^+\Big(\int_{t_1}^{t_2}|w(s)|^2\,ds+\int_{t_1-\tau^+}^{t_1}|w(s)|^2\,ds\Big). \quad (4.5)$$

Substituting in (4.4) the last term from relations chain (4.5) instead of the first one, and using inequalities: $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, $\sqrt{a}\sqrt{b}\le\varepsilon a+(4\varepsilon)^{-1}b$, $a\ge 0$, $b\ge 0$, $\varepsilon>0$, we obtain

$$\Big|\int_{t_1}^{t_2}\Big(\int_{t-\tau(t)}^{t}\big(c(t,s,u_1(s))-c(t,s,u_2(s))\big)\,ds,w(t)\Big)\,dt\Big| \\ \le L\tau^+\Big((1+\varepsilon)\int_{t_1}^{t_2}|w(t)|^2\,dt+(4\varepsilon)^{-1}\int_{t_1-\tau^+}^{t_1}|w(t)|^2\,dt\Big), \quad (4.6)$$

where $\varepsilon>0$ is an arbitrary.

By equality (2.2), inequality (4.6), condition $(\mathcal{A}_4)$ and the fact that $g_i(t)\in\partial\Phi(u_i(t))$ for a.e. $t\in S$, $i=1,2$, from (4.3) for a.e. $t\in S$, we obtain

$$\frac{1}{2}\int_{t_1}^{t_2}\big(|w(t)|^2\big)'\,dt \\ +\big(K_2-(1+\varepsilon)L\tau^+\big)\int_{t_1}^{t_2}|w(t)|^2\,dt-(4\varepsilon)^{-1}L\tau^+\int_{t_1-\tau^+}^{t_1}|w(t)|^2dt\le 0. \quad (4.7)$$

Using the integration-by-parts formula, we have

$$|w(t)|^2\Big|_{t_1}^{t_2}+2\big(K_2-(1+\varepsilon)L\tau^+\big)\int_{t_1}^{t_2}|w(t)|^2\,dt \\ \le(2\varepsilon)^{-1}L\tau^+\int_{t_1-\tau^+}^{t_1}|w(t)|^2\,dt. \quad (4.8)$$

By inequality (3.6) and taking $\varepsilon>0$ such that $K_2-(1+\varepsilon)L\tau^+\ge 0$, from (4.8) we obtain

$$|w(t_2)|^2\le|w(t_1)|^2+C_3\int_{t_1-\tau^+}^{t_1}|w(t)|^2dt, \quad (4.9)$$

where $C_3>0$ is a constant independent of t_1,t_2.

Let us fix an arbitrary $t_2\in S$. Since $w\in L^2(S;H)\cap C(S;H)$, according to Remark 2.1 there exists a sequence $\{t_{1,k}\}_{k=1}^{\infty}\subset S$ such that $t_{1,k}<t_2$ for all $k\in\mathbb{N}$, $t_{1,k}\longrightarrow_{k\to+\infty}-\infty$, and

$$|w(t_{1,k})|^2+C_3\int_{t_{1,k}-\tau^+}^{t_{1,k}}|w(t)|^2\,dt\underset{k\to+\infty}{\to}0.$$

Taking $t_{1,k}$ ($k \in \mathbb{N}$) instead of t_1 in (4.9) and passing to the limit as $k \to +\infty$ we obtain $|w(x,t_2)|^2 = 0$. Since $t_2 \in S$ is an arbitrary number, we have $w(t) = 0$ for a.e. $t \in S$, this contradicts our assumption. Therefore, a solution of the problem $\mathbf{P}(\Phi,\tau,c,f)$ is unique.

Step 2 (*auxiliary statements*). We define the functional $\Phi_H : H \to (-\infty,+\infty]$ by the rule: $\Phi_H(v) := \Phi(v)$, if $v \in V$, and $\Phi_H(v) := +\infty$ otherwise. Note that conditions $(\mathcal{A}_1)$, $(\mathcal{A}_2)$, Lemma IV.5.2 and Proposition IV.5.2 of the monograph [22] imply that Φ_H is a proper, convex, and lower-semi-continuous functional on H, $\mathrm{dom}(\Phi_H) = \mathrm{dom}(\Phi) \subset V$ and $\partial\Phi_H = \partial\Phi \cap (V \times H)$, where $\partial\Phi_H : H \to 2^H$ is the subdifferential of the functional Φ_H. Moreover, condition $(\mathcal{A}_3)$ yields $0 \in \partial\Phi_H(0)$ (see Remark 3.1).

The following statements will be used in the sequel.

Proposition 4.1 ([22, Lemma IV.4.3]). Let $-\infty < a < b < +\infty$, $w \in H^1(a,b;H)$, and $g \in L^2(a,b;H)$ such that $w(t) \in D(\partial\Phi_H)$ and $g(t) \in \partial\Phi_H\big(w(t)\big)$ for a.e. $t \in (a,b)$. Then the function $\Phi_H\big(w(\cdot)\big)$ is absolutely continuous on the interval $[a,b]$ and for any function $h : [a,b] \to H$ such that $h(t) \in \partial\Phi_H\big(w(t)\big)$ for a.e. $t \in [a,b]$ the following equality holds

$$\frac{d}{dt}\Phi_H\big(w(t)\big) = (h(t), w'(t)) \quad \text{for a.e. } t \in [a,b].$$

Proposition 4.2 ([8, Proposition 3.12], [22, Proposition IV.5.2]). Let $T > 0$, $\widetilde{f} \in L^2(0,T;H)$ and $w_0 \in \mathrm{dom}(\Phi)$. Then there exists a unique function $w \in H^1(0,T;H)$ such that $w(0) = w_0$ and, for a.e. $t \in (0,T]$, we have $w(t) \in D(\partial\Phi_H)$ and

$$w'(t) + \partial\Phi_H\big(w(t)\big) \ni \widetilde{f}(t) \quad \text{in } H. \tag{4.10}$$

i.e., there exists a function $\widetilde{g} \in L^2(0,T;H)$ such that, for a.e. $t \in (0,T]$, we have $\widetilde{g}(t) \in \partial\Phi_H(w(t))$ and

$$w'(t) + \widetilde{g}(t) = \widetilde{f}(t) \quad \text{in } H. \tag{4.11}$$

Lemma 4.1. *Let $t_0 < 0$, $\widetilde{f} \in L^2(t_0,0;H)$, and $w_0 \in C([\tau_0,t_0];H)$, $w_0(t) \in \mathrm{dom}(\Phi)$ for every $t \in [\tau_0,t_0]$, where $\tau_0 := \min_{t\in[t_0,0]}(t-\tau(t))$ (if $\tau_0 = t_0$ then $[\tau_0,t_0] = \{t_0\}$). Then there exists a unique function $w \in C([\tau_0,0];H) \cap H^1(t_0,0;H)$ such that $w(t) = w_0(t)$ for every $t \in [\tau_0,t_0]$, and, for a.e. $t \in (t_0,0]$, we have $w(t) \in D(\partial\Phi_H)$ and*

$$w'(t) + \partial\Phi_H\big(w(t)\big) + \int_{t-\tau(t)}^{t} c(t,s,w(s))\,ds \ni \widetilde{f}(t) \quad \text{in } H, \tag{4.12}$$

that is, there exists function $\widetilde{g} \in L^2(t_0,0;H)$ such that, for a.e. $t \in (t_0,0]$ we have $\widetilde{g}(t) \in \partial\Phi_H(w(t))$ and

$$w'(t) + \widetilde{g}(t) + \int_{t-\tau(t)}^{t} c(t,s,w(s))\,ds = \widetilde{f}(t) \quad \text{in } H. \tag{4.13}$$

Proof of Lemma 4.1. Let $M := \{w \in C([\tau_0, 0]; H) \mid w(t) = w_0(t)\ \forall t \in [\tau_0, t_0]\}$ be a set with the metric

$$\rho(w_1, w_2) = \max_{t \in [t_0, 0]} \big[e^{-\alpha(t-t_0)}|w_1(t) - w_2(t)|\big], \quad w_1, w_2 \in M,$$

where $\alpha > 0$ is an arbitrary fixed number. It is obvious that the metric space (M, ρ) is complete. Now let us consider an operator $A : M \to M$ defined as follows: for any given function $\widetilde{w} \in M$, it defines a function $\widehat{w} \in M \cap H^1(t_0, 0; H)$ such that, for a.e. $t \in [t_0, 0]$, we have $\widehat{w}(t) \in D(\partial\Phi_H)$ and

$$\widehat{w}\,'(t) + \partial\Phi_H(\widehat{w}(t)) \ni \widetilde{f}(t) - \int_{t-\tau(t)}^{t} c(t, s, \widetilde{w}(s))\, ds \quad \text{in} \quad H. \tag{4.14}$$

Clearly, variational inequality (4.14) coincides with variational inequality (4.10) after replacing $[0, T]$ by $[t_0, 0]$, $\widetilde{f}(t)$ by $\widetilde{f}(t) - \int_{t-\tau(t)}^{t} c(t, s, \widetilde{w}(s))\, ds$ and the condition $w(0) = w_0$ by the condition $\widehat{w}(0) = w_0(t_0)$. Thus, using Proposition 4.2, we get that operator A is well-defined. Let us show that the operator A is a contraction for some $\alpha > 0$. Indeed, let $\widetilde{w}_1, \widetilde{w}_2$ be arbitrary functions from M and $\widehat{w}_1 := A\widetilde{w}_1$, $\widehat{w}_2 = A\widetilde{w}_2$. According to (4.14) (see (4.11)) there exist functions $\widehat{g}_1$, $\widehat{g}_2$ from $L^2(t_0, 0; H)$ such that, for each $k \in \{1, 2\}$ and a.e. $t \in (t_0, 0]$, we have $\widehat{g}_k(t) \in \partial\Phi_H(\widehat{w}_k(t))$ and

$$\widehat{w}_k'(t) + \widehat{g}_k(t) = \widetilde{f}(t) - \int_{t-\tau(t)}^{t} c(t, s, \widetilde{w}_k(s))\, ds \quad \text{in} \quad H, \tag{4.15}$$

while $\widehat{w}_k(t) = w_0(t)$ for a.e. $t \in [\tau_0, t_0]$.

Subtracting identity (4.15) for $k = 2$ from identity (4.15) for $k = 1$, and, for a.e. $t \in (t_0, 0]$, multiplying the obtained identity by $\widehat{w}_1(t) - \widehat{w}_2(t)$, we get

$$\begin{aligned} &\big((\widehat{w}_1(t) - \widehat{w}_2(t))', \widehat{w}_1(t) - \widehat{w}_2(t)\big) + (\widehat{g}_1(t) - \widehat{g}_2(t), \widehat{w}_1(t) - \widehat{w}_2(t)) \\ &\qquad = -\Big(\int_{t-\tau(t)}^{t} \big(c(t, s, \widetilde{w}_1(s)) - c(t, s, \widetilde{w}_2(s))\big)\, ds, \widehat{w}_1(t) - \widehat{w}_2(t)\Big), \end{aligned} \tag{4.16}$$

$$\widehat{w}_1(t) - \widehat{w}_2(t) = 0 \quad \text{for a.e.} \quad t \in [\tau_0, t_0]. \tag{4.17}$$

We integrate equality (4.16) by t from t_0 to $\sigma \in [t_0, 0]$, taking into account that for a.e. $t \in (t_0, 0]$ we have

$$\big((\widehat{w}_1(t) - \widehat{w}_2(t))', \widehat{w}_1(t) - \widehat{w}_2(t)\big) = \frac{1}{2}\frac{d}{dt}|\widehat{w}_1(t) - \widehat{w}_2(t)|^2.$$

As a result, we get the equality

$$\begin{aligned} &\frac{1}{2}|\widehat{w}_1(\sigma) - \widehat{w}_2(\sigma)|^2 + \int_{t_0}^{\sigma} (\widehat{g}_1(t) - \widehat{g}_2(t), \widehat{w}_1(t) - \widehat{w}_2(t))\, dt \\ &\qquad = -\int_{t_0}^{\sigma} \Big(\int_{t-\tau(t)}^{t} \big(c(t, s, \widetilde{w}_1(s)) - c(t, s, \widetilde{w}_2(s))\big)\, ds, \widehat{w}_1(t) - \widehat{w}_2(t)\Big) dt. \end{aligned} \tag{4.18}$$

By condition $(\mathcal{A}_4)$, for a.e. $t \in (t_0, 0]$, we have the inequality

$$(\widehat{g}_1(t) - \widehat{g}_2(t), \widehat{w}_1(t) - \widehat{w}_2(t)) \geqslant K_2 |\widehat{w}_1(t) - \widehat{w}_2(t))|^2. \tag{4.19}$$

Taking into account conditions $(\mathcal{T})$, $(\mathcal{C})$ and the Cauchy inequality (2.4), for a.e. $t \in (t_0, 0]$, we obtain

$$\begin{aligned}
&\Big|\Big(\int_{t-\tau(t)}^{t} \big(c(t,s,\widetilde{w}_1(s)) - c(t,s,\widetilde{w}_2(s))\big)\,ds, \widehat{w}_1(t) - \widehat{w}_2(t)\Big)\Big| \\
&\le \Big(\int_{t-\tau(t)}^{t} \big|c(t,s,\widetilde{w}_1(s)) - c(t,s,\widetilde{w}_2(s))\big|\,ds\Big)\,|\widehat{w}_1(t) - \widehat{w}_2(t)| \\
&\le L\Big(\int_{t-\tau^+}^{t} |\widetilde{w}_1(s) - \widetilde{w}_2(s)|\,ds\Big)|\widehat{w}_1(t) - \widehat{w}_2(t)| \\
&\le \varepsilon|\widehat{w}_1(t) - \widehat{w}_2(t)|^2 + (4\varepsilon)^{-1}L^2\Big(\int_{t-\tau^+}^{t} |\widetilde{w}_1(s) - \widetilde{w}_2(s)|\,ds\Big)^2 \\
&\le \varepsilon|\widehat{w}_1(t) - \widehat{w}_2(t)|^2 + (4\varepsilon)^{-1}L^2\tau^+ \int_{t-\tau^+}^{t} |\widetilde{w}_1(s) - \widetilde{w}_2(s)|^2\,ds,
\end{aligned} \tag{4.20}$$

where $\varepsilon > 0$ is an arbitrary number, $\widetilde{w}_1(s) - \widetilde{w}_2(s) := 0\ \forall s \le \tau_0$.

From (4.18), according to (4.19) and (4.20), we have

$$\begin{aligned}
|\widehat{w}_1(\sigma) - \widehat{w}_2(\sigma)|^2 &+ 2(K_2 - \varepsilon)\int_{t_0}^{\sigma} |\widehat{w}_1(t) - \widehat{w}_2(t)|^2\,dt \\
&\le (2\varepsilon)^{-1}L^2\tau^+ \int_{t_0}^{\sigma}\Big(\int_{t-\tau^+}^{t} |\widetilde{w}_1(s) - \widetilde{w}_2(s)|^2\,ds\Big)\,dt.
\end{aligned} \tag{4.21}$$

Let us consider the right-hand side of the inequality (4.21). Using the assumption that $\widetilde{w}_1(s) - \widetilde{w}_2(s) = 0$ for $s \le t_0$ and $s \ge 0$, we obtain

$$\int_{t_0}^{\sigma}\Big(\int_{t-\tau^+}^{t} |\widetilde{w}_1(s) - \widetilde{w}_2(s)|^2\,ds\Big)\,dt \le t_0 \int_{t_0}^{\sigma} |\widetilde{w}_1(t) - \widetilde{w}_2(t)|^2\,dt. \tag{4.22}$$

From (4.21) and (4.22), choosing $\varepsilon = K_2$, we get

$$|\widehat{w}_1(\sigma) - \widehat{w}_2(\sigma)|^2 \leqslant C_2 \int_{t_0}^{\sigma} |\widetilde{w}_1(t) - \widetilde{w}_2(t)|^2\,dt, \quad \sigma \in (t_0, 0], \tag{4.23}$$

where $C_2 > 0$ is a constant depending on L, K_2, τ^+, and t_0 only.

Multiplying (4.23) by $e^{-2\alpha(\sigma-t_0)}$, we obtain

$$\begin{aligned}
&e^{-2\alpha(\sigma-t_0)}|\widehat{w}_1(\sigma)-\widehat{w}_2(\sigma)|^2\\
&\qquad\leqslant C_2e^{-2\alpha(\sigma-t_0)}\int_{t_0}^{\sigma} e^{2\alpha(t-t_0)}e^{-2\alpha(t-t_0)}|\widetilde{w}_1(t)-\widetilde{w}_2(t)|^2\,dt\\
&\qquad\leqslant C_2e^{-2\alpha(\sigma-t_0)}\max_{t\in[t_0,0]}\big[e^{-\alpha(t-t_0)}|\widetilde{w}_1(t)-\widetilde{w}_2(t)|\big]^2\int_{t_0}^{\sigma} e^{2\alpha(t-t_0)}\,dt\\
&\qquad=\frac{C_2}{2\alpha}\big(1-e^{-2\alpha(\sigma-t_0)}\big)\big[\rho(\widetilde{w}_1,\widetilde{w}_2)\big]^2\\
&\qquad\leqslant\frac{C_2}{2\alpha}\big[\rho(\widetilde{w}_1,\widetilde{w}_2)\big]^2,\quad \sigma\in[t_0,0]. \qquad (4.24)
\end{aligned}$$

From (4.24) it follows

$$\rho(\widehat{w}_1,\widehat{w}_2)\leqslant\sqrt{C_2/(2\alpha)}\rho(\widetilde{w}_1,\widetilde{w}_2).$$

This, choosing $\alpha>0$ such that $C_2/(2\alpha)<1$, yields, operator A is a contraction. Hence, we may apply the Banach fixed-point theorem (in other words, the contraction mapping principle; see, for example, [9, Theorem 5.7]) and deduce that there exists a unique function $w\in M$ such that $Aw=w$, i.e., we have proved our proposition. □

Step 3 (*solution approximations*). We construct a sequence of functions which, in some sense, approximate the solution of the problem $\mathbf{P}(\Phi,\tau,c,f)$.

Let $\{\varkappa_k\}_{k=1}^{\infty}$ be a monotonically decreasing sequence of numbers from S such that $\varkappa_1<0$ and $\lim_{k\to\infty}\varkappa_k=-\infty$. Denote $\widehat{f}_k(t):=f(t)$ for $t\in[\varkappa_k,0]$, $\tau_k:=\min_{t\in[\varkappa_k,0]}(t-\tau(t))$, $k\in\mathbb{N}$.

For each $k\in\mathbb{N}$ consider the problem of finding a function $\widehat{u}_k\in C([\tau_k,0];H)\cap H^1(\varkappa_k,0;H)$ such that, for a.e. $t\in(\varkappa_k,0]$, we have $\widehat{u}_k(t)\in D(\partial\Phi_H)$ and

$$\widehat{u}_k'(t)+\partial\Phi_H\big(\widehat{u}_k(t)\big)+\int_{t-\tau(t)}^{t}c(t,s,\widehat{u}_k(s))\,ds\ni\widehat{f}_k(t)\quad\text{in } H, \qquad (4.25)$$

and

$$\widehat{u}_k(t)=0,\quad t\in[\tau_k,\varkappa_k]. \qquad (4.26)$$

Inclusion (4.25) means that there exists a function $\widehat{g}_k\in L^2(\varkappa_k,0;H)$ such that, for a.e. $t\in(\varkappa_k,0]$, we have $\widehat{g}_k(t)\in\partial\Phi_H(\widehat{u}_k(t))$ and

$$\widehat{u}_k'(t)+\widehat{g}_k(t)+\int_{t-\tau(t)}^{t}c(t,s,\widehat{u}_k(s))\,ds=\widehat{f}_k(t)\quad\text{in}\ \ H. \qquad (4.27)$$

Lemma 4.1 implies the existence and uniqueness of solution of the problem (4.25), (4.26). Since $D(\partial\Phi_H)\subset\operatorname{dom}(\Phi_H)=\operatorname{dom}(\Phi)$ then $\widehat{u}_k(t)\in V$ for a.e. $t\in[\varkappa_k,0]$. According to the definition of the subdifferential of a functional and the fact that $\widehat{g}_k(t)\in\partial\Phi_H(\widehat{u}(t))$ for a.e. $t\in(\varkappa_k,0]$, we have

$$\Phi_H(0)\geq\Phi_H(\widehat{u}_k(t))+(\widehat{g}_k(t),0-\widehat{u}_k(t))\quad\text{for a.e.}\ \ t\in(\varkappa_k,0].$$

This and condition $(\mathcal{A}_3)$ yield that for a.e. $t \in (\varkappa_k, 0]$ we have

$$(\widehat{g}_k(t), \widehat{u}_k(t)) \geq \Phi(\widehat{u}_k(t)) \geq K_1 \|\widehat{u}_k(t)\|^p. \tag{4.28}$$

Since the left side of this chain of inequalities belongs to $L^1(S_k)$ then $\widehat{u}_k$ belongs to $L^p(\varkappa_k, 0; V)$.

For each $k \in \mathbb{N}$ we extend functions $\widehat{f}_k, \widehat{u}_k$ and $\widehat{g}_k$ by zero for the entire interval S, and denote these extensions by f_k, u_k and g_k, respectively. From the above it follows that for each $k \in \mathbb{N}$ the function u_k belongs to $L^p(S; V)$, its derivative u_k' belongs to $L^2(S; H)$ and for a.e. $t \in S$ the inclusion $g_k(t) \in \partial\Phi_H\big(u_k(t)\big)$ and the following equality (see (4.27)) hold

$$u_k'(t) + g_k(t) + \int_{t-\tau(t)}^{t} c(t, s, u_k(s))\, ds = f_k(t) \quad \text{in} \quad H. \tag{4.29}$$

In order to show the convergence $\{u_k\}_{k=1}^{+\infty}$ to the solution of the problem $\mathbf{P}(\Phi, \tau, c, f)$ we need some estimates of the functions u_k, $k \in \mathbb{N}$.

Step 4 (*first order estimates of solution approximations*).

Let t_1, $t_2 \in S$ be arbitrary numbers such that $t_1 < t_2$. Multiplying identity (4.29) for a.e. $t \in S$ by $u_k(t)$ and integrating by t from t_1 to t_2, we obtain

$$\int_{t_1}^{t_2} (u_k'(t), u_k(t))\, dt + \int_{t_1}^{t_2} (g_k(t), u_k(t))\, dt \\ + \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} c(t, s, u_k(s))\, ds, u_k(t) \Big)\, dt = \int_{t_1}^{t_2} (f_k(t), u_k(t))\, dt. \tag{4.30}$$

Equality (2.2) yield

$$\int_{t_1}^{t_2} (u_k'(t), u_k(t)) dt = \int_{t_1}^{t_2} \frac{d}{dt} |u_k(t)|^2 dt = \frac{1}{2}(|u_k(t_2)|^2 - |u_k(t_1)|^2). \tag{4.31}$$

From Remark 3.1, it follows

$$(g_k(t), u_k(t)) \geq K_2 |u_k(t)|^2 \quad \text{for a.e.} \quad t \in S. \tag{4.32}$$

By inequalities (4.28) and (4.32), for a.e. $t \in S$, we have

$$(g_k(t), u_k(t)) \geq \delta (g_k(t), u_k(t)) + (1-\delta)(g_k(t), u_k(t)) \\ \geq \delta K_2 |u_k(t)|^2 + \frac{1}{2}(1-\delta) K_1 \|u_k(t)\|^p + \frac{1}{2}(1-\delta)\Phi(u_k(t)), \tag{4.33}$$

where $\delta \in (0,1)$ is an arbitrary number.

Let us estimate the second term of the left-hand side of equality (4.30) by using (4.33), in this way

$$\int_{t_1}^{t_2} (g_k(t), u_k(t))\, dt \\ \geq \frac{1}{2} \int_{t_1}^{t_2} \Big((1-\delta)\Phi\big(u_k(t)\big) + (1-\delta) K_1 \|u_k(t)\|^p + 2\delta K_2 |u_k(t)|^2 \Big)\, dt. \tag{4.34}$$

We estimate the third term on the left-hand side of equality (4.30) by using the Cauchy-Schwartz-Bunjakovsky inequality and (3.2). As the result, we obtain

$$\begin{aligned}\Big|\int_{t_1}^{t_2}\Big(\int_{t-\tau(t)}^{t} c(t,s,u_k(s))\,ds, u_k(t)\Big)\,dt\Big| &\\ \le \int_{t_1}^{t_2}\Big(\int_{t-\tau(t)}^{t}\big|c(t,s,u_k(s))\big|\,ds\Big)|u_k(t)|\,dt &\\ \le L\int_{t_1}^{t_2}\Big(\int_{t-\tau^+}^{t}|u_k(s)|\,ds\Big)\,|u_k(t)|\,dt &\\ \le L\sqrt{\tau^+}\Big(\int_{t_1}^{t_2}|u_k(t)|^2\,dt\Big)^{1/2}\Big(\int_{t_1}^{t_2}\Big(\int_{t-\tau^+}^{t}|u_k(s)|^2\,ds\Big)\,dt\Big)^{1/2}. &\qquad (4.35)\end{aligned}$$

Now, let us estimate the last item on the inequality chain above. Changing the order of integration, we have

$$\begin{aligned}\int_{t_1}^{t_2}\Big(\int_{t-\tau^+}^{t}|u_k(s)|^2 ds\Big)\,dt &\\ \le \int_{t_1-\tau^+}^{t_2}|u_k(s)|^2\,ds\int_{s}^{s+\tau^+}dt = \tau^+\int_{t_1-\tau^+}^{t_2}|u_k(t)|^2\,dt. &\qquad (4.36)\end{aligned}$$

From (4.35), (4.36) with $t_1 < \varkappa_k$, and definition of u_k, it follows

$$\Big|\int_{t_1}^{t_2}\Big(\int_{t-\tau(t)}^{t} c(t,s,u_k(s))\,ds, u_k(t)\Big)dt\Big| \le L\tau^+\int_{t_1}^{t_2}|u_k(t)|^2 dt. \qquad (4.37)$$

Now we estimate the first term of the right-hand side of equality (4.30) by using the Cauchy-Schwartz-Bunjakovsky and Cauchy inequalities (2.4). As the result, we obtain

$$\int_{t_1}^{t_2}(f_k(t),u_k(t))\,dt \leqslant \varepsilon\int_{t_1}^{t_2}|u_k(t)|^2\,dt + (4\varepsilon)^{-1}\int_{t_1}^{t_2}|f_k(t)|^2\,dt, \qquad (4.38)$$

where $\varepsilon > 0$ is an arbitrary number.

From (4.30), taking into account (4.31), (4.34), (4.37) and (4.38), for any $t_1, t_2 \in S$ such that $t_1 < \min\{\varkappa_k, t_2\}$, we obtain

$$\begin{aligned}|u_k(t_2)|^2 + (1-\delta)\int_{t_1}^{t_2}\Phi\big(u_k(t)\big)dt + (1-\delta)K_1\int_{t_1}^{t_2}\|u_k(t)\|^p\,dt &\\ + 2[\delta K_2 - L\tau^+ - \varepsilon]\int_{t_1}^{t_2}|u_k(t)|^2\,dt \le (2\varepsilon)^{-1}\int_{t_1}^{t_2}|f_k(t)|^2\,dt.&\end{aligned}$$

First we choose $\delta \in (0,1)$ such that $\delta K_2 - L\tau^+ > 0$ (see (3.6)). Then take $\varepsilon = (\delta K_2 - L\tau^+)/2 > 0$. As a result, we obtain

$$|u_k(t_2)|^2 + \int_{t_1}^{t_2}\Phi\big(u_k(t)\big)\,dt + \int_{t_1}^{t_2}\big[\|u_k(t)\|^p + |u_k(t)|^2\big]\,dt \le C_4\int_{t_1}^{t_2}|f_k(t)|^2\,dt,$$

where C_4 is a positive constant depended on K_1, K_2, L, and τ^+ only.

Since $t_2 \in S$ is an arbitrary, by the definition of f_k, we have

$$\sup_{t\in S} |u_k(t)|^2 + \int_S \Phi\big(u_k(t)\big)\,dt + \int_S \big[\|u_k(t)\|^p + |u_k(t)|^2\big]\,dt \leqslant C_5 \int_S |f(t)|^2\,dt, \quad (4.39)$$

where $C_5 > 0$ is a positive constant depended on K_1, K_2, L, and τ^+ only.

From (4.39) it follows

$$\{u_k\}_{k=1}^{+\infty} \text{ is bounded in } L^\infty(S;H) \cap L^p(S;V) \cap L^2(S;H). \quad (4.40)$$

Step 5 (*second order estimates of solution approximations*). Now we shell estimate the functions u_k', $k \in \mathbb{N}$. Let t_1, t_2 be arbitrary numbers such that t_1, $t_2 \in S$, $t_1 < t_2$. For almost every $t \in [t_1, t_2]$ we multiply equality (4.29) by the function $u_k'(t)$ (recall that $u_k' \in L^2(S;H)$) and integrate the resulting equality from t_1 to t_2. Then we obtain

$$\int_{t_1}^{t_2} |u_k'(t)|^2\,dt + \int_{t_1}^{t_2} (g_k(t), u_k'(t))\,dt$$
$$= \int_{t_1}^{t_2} (f_k(t), u_k'(t))\,dt - \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} c(t,s,u_k(s))\,ds, u_k'(t)\Big)\,dt. \quad (4.41)$$

Since $g_k \in L^2(S;H)$ and $g_k(t) \in \partial\Phi(u_k(t))$ for a.e. $t \in S$, Proposition 4.1 implies that the function $\Phi_H\big(u_k(\cdot)\big)$ is absolutely continuous on $[t_1, t_2]$ and

$$\frac{d}{dt}\Phi_H\big(u_k(t)\big) = (g_k(t), u_k'(t)) \quad \text{for a.e. } t \in (t_1, t_2). \quad (4.42)$$

By (4.42) we can rewrite the second term on the left-hand side of the equality (4.41) as follows

$$\int_{t_1}^{t_2} (g_k(t), u_k'(t))\,dt = \int_{t_1}^{t_2} \frac{d}{dt}\Phi_H\big(u_k(t)\big)\,dt$$
$$= \Phi_H\big(u_k(t)\big)\Big|_{t_1}^{t_2} = \Phi\big(u_k(t)\big)\Big|_{t_1}^{t_2}. \quad (4.43)$$

By the Cauchy inequality (2.4), changing the order of integration (see 4.36) and (3.2), we have

$$\Big|\int_{t_1}^{t_2} (f_k(t), u_k'(t))\,dt\Big| \le \int_{t_1}^{t_2} |f_k(t)||u_k'(t)|\,dt$$
$$\le \int_{t_1}^{t_2} |f_k(t)|^2\,dt + \frac{1}{4}\int_{t_1}^{t_2} |u_k'(t)|^2\,dt. \quad (4.44)$$

$$
\begin{aligned}
\Big| \int_{t_1}^{t_2} & \Big(\int_{t-\tau(t)}^{t} c(t,s,u_k(s))\, ds, u_k'(t) \Big)\, dt \Big| \\
&\le \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} \big| c(t,s,u_k(s)) \big|\, ds \Big) |u_k'(t)|\, dt \\
&\le L \int_{t_1}^{t_2} \Big(\int_{t-\tau(t)}^{t} |u_k(s)|\, ds \Big) |u_k'(t)|\, dt \\
&\le L^2\tau^+ \int_{t_1}^{t_2} \Big(\int_{t-\tau^+}^{t} |u_k(s)|^2\, ds \Big)\, dt + \frac{1}{4}\int_{t_1}^{t_2} |u_k'(t)|^2\, dt \\
&\le (L\tau^+)^2 \int_{t_1-\tau^+}^{t_2} |u_k(t)|^2 dt + \frac{1}{4}\int_{t_1}^{t_2} |u_k'(t)|^2\, dt, \qquad (4.45)
\end{aligned}
$$

By (4.43), (4.45), (4.44), from (4.41) we get

$$
\frac{1}{2}\int_{t_1}^{t_2} |u_k'(t)|^2\, dt + \Phi_H\big(u_k(t)\big)\Big|_{t_1}^{t_2} \le (L\tau^+)^2 \int_{t_1-\tau^+}^{t_2} |u_k(t)|^2 dt + \int_{t_1}^{t_2} |f_k(t)|^2\, dt. \qquad (4.46)
$$

By the definitions of u_k and f_k we pass to the limit in (4.46) when $t_1 \to -\infty$. Taking into account condition $\Phi(0) = 0$ (see $(\mathcal{A}_3)$) and estimate (4.39), from (4.46), taking $t_2 = \sigma \in S$, we have

$$
\Phi\big(u_k(\sigma)\big) + \int_{-\infty}^{\sigma} |u_k'(t)|^2\, dt \leqslant C_6 \int_{-\infty}^{\sigma} |f(t)|^2\, dt, \qquad (4.47)
$$

where $C_6 > 0$ is a positive constant depended on K_1, K_2, L, and τ^+ only.

According to the definition of the functional Φ_H, condition $(\mathcal{A}_3)$ (recall that $u_k(t) \in V$ for a.e. $t \in S$) from (4.47), we have

$$
\sup_{\sigma\in S} ||u_k(\sigma)||^p + \int_S |u_k'(t)|^2\, dt \leqslant C_7 \int_S |f(t)|^2\, dt, \qquad (4.48)
$$

where C_7 is a positive constant depended on K_1, K_2, L, and τ^+ only.

Estimate (4.48) implies, that

$$
\text{the sequence } \big\{u_k\big\}_{k=1}^{+\infty} \text{ is bounded in } L^\infty(S;V), \qquad (4.49)
$$

$$
\text{the sequence } \big\{u_k'\big\}_{k=1}^{+\infty} \text{ is bounded in } L^2(S;H). \qquad (4.50)
$$

Let us show that

$$
\text{the sequence } \{g_k\}_{k=1}^{+\infty} \text{ is bounded in } L^2(S;H). \qquad (4.51)
$$

Indeed, from (3.3) for $w = u_k$, using (4.39), we have

$$\int_S \Big| \int_{t-\tau(t)}^t c(t,s,u_k(s))\,ds \Big|^2 dt \le (L\tau^+)^2 \int_S |u_k(t)|^2\,dt \le C_8 \int_S |f(t)|^2\,dt, \quad (4.52)$$

where C_8 is a positive constant dependent on K_1, K_2, L, and τ^+ only.

From (4.29), (4.48), (4.52), and the definitions of u_k, f_k we obtain (4.51) *Step 6* (*passing to the limit*). From (4.40), (4.49)–(4.51) and Lemma 2.1 we have that there exist functions $u \in L^\infty(S;V) \cap L^p(S;V) \cap H^1(S;H)$, $g \in L^2(S;H)$ and a subsequence of the sequence $\{u_k, g_k\}_{k=1}^{+\infty}$ (still denoted by $\{u_k, g_k\}_{k=1}^{+\infty}$) such that

$$\begin{gathered} u_k \longrightarrow_{k\to\infty} u \quad *\text{-weakly in } L^\infty(S;V), \\ \text{weakly in } L^p(S;V), \text{ weakly in } H^1(S;H), \end{gathered} \quad (4.53)$$

$$u_k \underset{k\to\infty}{\longrightarrow} u \quad \text{in } C(S;H), \quad (4.54)$$

$$g_k \underset{k\to\infty}{\longrightarrow} g \quad \text{weakly in } L^2(S;H). \quad (4.55)$$

Using condition $(\mathcal{T})$, $(\mathcal{C})$, (4.54), the Cauchy-Schwarz-Bunjakovsky inequality and changing the order of integration (see (4.36)), for any $t_1, t_2 \in S$, $t_1 < t_2$, we obtain

$$\begin{aligned} \int_{t_1}^{t_2} \Big| \int_{t-\tau(t)}^t & c(t,s,u_k(s))\,ds - \int_{t-\tau(t)}^t c(t,s,u(s))\,ds \Big|^2 dt \\ &\le L^2\tau^+ \int_{t_1}^{t_2} \Big(\int_{t-\tau^+}^t |u_k(s) - u(s)|^2 ds \Big)\,dt \\ &\le (L\tau^+)^2 \int_{t_1-\tau^+}^{t_2} |u_k(t) - u(t)|^2\,dt \underset{k\to\infty}{\longrightarrow} 0. \quad (4.56) \end{aligned}$$

Thus, we have

$$\int_{t-\tau(t)}^t c(t,s,u_k(s))\,ds \underset{k\to\infty}{\longrightarrow} \int_{t-\tau(t)}^t c(t,s,u(s))\,ds \quad \text{strongly in } L^2_{\text{loc}}(S;H). \quad (4.57)$$

Let $v \in H, \varphi \in D(-\infty, 0)$ be an arbitrary. For a.e. $t \in S$ we multiply equality (4.29) by v, and then we multiply the obtained equality by φ and integrate in t on S. As a result, we obtain the equality

$$\begin{aligned} \int_S (u_k'(t), v\varphi(t))\,dt + \int_S (g_k(t), v\varphi(t))\,dt + \int_S \Big(\int_{t-\tau(t)}^t c(t,s,u_k(s))\,ds, v\varphi(t) \Big)\,dt \\ = \int_S (f_k(t), v\varphi(t))\,dt, \quad k \in \mathbb{N}. \quad (4.58) \end{aligned}$$

We pass to the limit in (4.58) as $k \to \infty$, taking into account (4.53), (4.55), (4.57) and convergence of $\{f_k\}_{k=1}^\infty$ to f in $L^2_{\text{loc}}(S;H)$. As a result, since $v \in H, \varphi \in D(-\infty, 0)$ are arbitrary, for a.e. $t \in S$ we obtain the equality

$$u'(t) + g(t) + \int_{t-\tau(t)}^t c(t,s,u(s))\,ds = f(t) \quad \text{in } H.$$

Step 7 (*completion of the proof*). In order to complete the proof of the theorem it remains only to show that $u(t) \in D(\partial\Phi)$ and $g(t) \in \partial\Phi\big(u(t)\big)$ for a.e. $t \in S$.

Let $k \in \mathbb{N}$ be an arbitrary number. Since $u_k(t) \in D(\partial\Phi)$ and $g_k(t) \in \partial\Phi_H\big(u_k(t)\big)$ for every $t \in S \setminus \widetilde{S}_k$, where $\widetilde{S}_k \subset S$ is a set of measure zero, applying the monotonicity of the subdifferential $\partial\Phi_H$, we obtain that for every $t \in S \setminus \widetilde{S}_k$ the following equality holds

$$(g_k(t) - v^*, u_k(t) - v) \geqslant 0 \quad \forall\, [v, v^*] \in \partial\Phi_H. \tag{4.59}$$

Let $\sigma \in S$, $h > 0$ be arbitrary numbers. We integrate (4.59) on $[\sigma - h, \sigma]$:

$$\int_{\sigma-h}^{\sigma} (g_k(t) - v^*, u_k(t) - v)\, dt \geqslant 0 \quad \forall\, [v, v^*] \in \partial\Phi_H. \tag{4.60}$$

Now according to (4.54) and (4.55) we pass to the limit in (4.60) as $k \to \infty$. As a result we obtain

$$\int_{\sigma-h}^{\sigma} (g(t) - v^*, u(t) - v)\, dt \geq 0 \quad \forall\, [v, v^*] \in \partial\Phi_H. \tag{4.61}$$

The monograph [27, Theorem 2, p. 192] and (4.61) imply that for every $[v, v^*] \in \partial\Phi_H$ there exists a set $R_{[v,v_*]} \subset S$ of measure zero such that

$$\begin{gathered} 0 \leq \lim_{h\to+0} \frac{1}{h} \int_{\sigma-h}^{\sigma} \big(g(t) - v^*, u(t) - v\big)\, dt = \big(g(\sigma) - v^*, u(\sigma) - v\big) \\ \forall\, \sigma \in S \setminus R_{[v,v_*]}. \end{gathered} \tag{4.62}$$

Let us show that there exists a set $R \subset S$ of measure zero such that for all $\sigma \in S \setminus R$ the following inequality holds

$$\big(g(\sigma) - v^*, u(\sigma) - v\big) \geq 0 \quad \forall\, [v, v^*] \in \partial\Phi_H. \tag{4.63}$$

Since V and H are separable spaces, there exists a countable set $F \subset \partial\Phi_H \subset V \times H$ which is dense in $\partial\Phi_H$. Let us denote $R := \cup_{[v,v^*]\in F} R_{[v,v_*]}$. Since the set F is countable, and any countable union of sets of measure zero is a set of measure zero, R is a set of measure zero. Therefore, by (4.62) for any $\sigma \in S \setminus R$ inequality $\big(g(\sigma) - v^*, u(\sigma) - v\big) \geq 0$ holds for every $[v, v^*] \in F$. Let $[\widehat{v}, \widehat{v}^*]$ be an arbitrary element from $\partial\Phi_H$. Then from the density F in $\partial\Phi_H$ we have the existence of a sequence $\{[v_l, v_l^*]\}_{l=1}^{\infty}$ such that $v_l \to \widehat{v}$ in V, $v_l^* \to \widehat{v}^*$ in H and for all $\sigma \in S \setminus R$

$$(g(\sigma) - v_l^*, u(\sigma) - v_l) \geqslant 0 \quad \forall l \in \mathbb{N}. \tag{4.64}$$

Thus, passing to the limit in this equality as $l \to \infty$, we get (4.63). Therefore, for a.e. $t \in S$ we have

$$(g(t) - v^*, u(t) - v) \geqslant 0 \quad \forall\, [v, v^*] \in \partial\Phi_H.$$

From this, according to maximal monotonicity of $\partial\Phi_H$, we obtain that $[u(t), g(t)] \in \partial\Phi_H$ for a.e. $t \in S$.

Estimate (3.1) of the solution of the problem $\mathbf{P}(\Phi, \tau, c, f)$ follows directly from (4.39) and (4.48), (4.53), (4.54), and Proposition 2.2, Fatou's Lemma and the fact that Φ_H is lower semicontinuous in H.

References

1. J.-P. Aubin, *Un theoreme de compacite*, Comptes rendus hebdomadaires des seances de l'academie des sciences, **256** (24) (1963), 5042–5044.
2. F. Bernis, *Existence results for doubly nonlinear higher order parabolic equations on unbounded domains*, Mathematische Annalen, **279** (1988), 373–394.
3. M. Bokalo, *Problem without initial conditions for some classes of nonlinear parabolic equations*, Journal of Soviet Mathematics, **51** (1990), 2291–2322.
4. M. Bokalo, *Well-posedness of problems without initial conditions for nonlinear parabolic variational inequalities*, Nonlinear boundary problem, **8** (1998), 58–63.
5. M. Bokalo, A. Lorenzi, *Linear evolution first-order problems without initial conditions*, Milan Journal of Mathematics, **77** (2009), 437–494.
6. M. Bokalo, *On a Fourier problem for coupled evolution system of equations with integral time delays*, Visnyk of the Lviv University. Series Mechanics and Mathematics, **60** (2002), 32–49.
7. M. Bokalo, I. Skira, *Fourier problem for weakly nonlinear evolution inclusions with functionals*, Journal of optimization, differential equations and their applications (JODEA), **27** (1) (2019), 1–20.
8. H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Publishing Comp., Amsterdam, London, 1973.
9. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, Dordrecht, Heidelberg, London, 2011.
10. O. Buhrii, *Some parabolic variational inequalities without initial conditions*, Visnyk of the Lviv University. Series Mechanics and Mathematics, **49** (1998), 113–121.
11. V. Dmytriv, *On a Fourier problem for coupled evolution system of equations with time delays*, Matematychni Studii, **16** (2001), 141–156.
12. H. Gayevskyy, K. Greger, K. Zaharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.
13. O. Ilnytska, *Fourier problem for nonlinear parabolic equations with time-dependent delay*, Visnyk of the Lviv University. Series Mechanics and Mathematics, **82** (2017), 137–150.
14. S. Ivasishen, *Parabolic boundary-value problems without initial conditions*, Ukrainian Mathematical Journal, **34** (5) (1982), 439–443.
15. J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Paris: Dunod, 1969.
16. O. Oleinik, G. Iosifjan, *Analog of Saint-Venant's principle and uniqueness of solutions of the boundary problems in unbounded domain for parabolic equations*, Uspekhi Mat. Nauk, **31** (6) (1976), 142–166.
17. A. Pankov, *Bounded and almost periodic solutions of nonlinear operator differential equations*, Kluwer, Dordrecht, 1990.
18. P. Pukach, *On problem without initial conditions for some nonlinear degenarated parabolic system*, Ukrainian Mathematical Journal, **46** (4) (1994), 484–487.
19. H.-H. Rho, J.-M. Jeong, *Regularity for nonlinear evolution variational inequalities with delay terms*, Journal of Inequalities and Applications, **2014:387** (2014), 1–14.

20. R. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, *Pacific Journal of Mathematics*, **33** (1) (1970), 209–216.

21. R. Showalter, *Singular nonlinear evolution equations*, The Rocky Mountain Journal of Mathematics, **10** (3) (1980), 499–507.

22. R. Showalter, *Monotone operators in Banach space and nonlinear partial differential equations*, **49**, *American Mathematical Society*, Providence, 1997.

23. A. Tikhonov, A. Samarskii, *Equations of mathematical physics*, Moscow: *Nauka*, 1972.

24. A. Tychonoff, *Théorèmes d'unicité pour l'équation de la chaleur*, *Mat. Sb.*, **42** (2) (1953), 199–216.

25. I. Vrabie, *Global solutions for nonlinear delay evolution inclusions with nonlocal initial conditions*, Set-Valued and Variational Analysis, **20** (3) (2012), 477–497.

26. R. Wang, Q. Xiang, P. Zhu, *Existence and approximate controllability for systems governed by fractional delay evolution inclusions*, *Optimization*, **63** (8) (2014), 1191–1204.

27. K. Yoshida, *Functional Analysis*, *Springer-Verlag*, Berlin, Heidelberg, 1995.

Received 12.06.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 89–114, DOI 10.15421/1423012

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

HOW CAN WE MANAGE REPAIRING A BROKEN FINITE VIBRATING STRING? FORMULATIONS OF THE PROBLEM

Vladimir L. Borsch[*], Peter I. Kogut[†]

Communicated by Prof. A. Plotnikov

Abstract. Three approaches to solve a well-known IBVP posed for vibrating composite string with piece-wise constant properties have been applied. The main issue of the stufy is the number of the matching conditions to be imposed for the solution to the IBVP to be obtained.

Key words: separation of variables, the Laplace transform, eigenvalues and eigenfunctions, the energy equation, the transmission and matching conditions.

2010 Mathematics Subject Classification: 35L05, 35L80.

1. Introduction and the problem formulation

The current study was inspired by a series of our previous publications [1–5] dealing with an IBVP for the 1D degenerate wave equation. Interpreting the IBVP as that of a vibrating 'string', driven by: *a*) initial disturbances of the shape and velocity of 'the string', and *b*) known external controls imposed at both ends of 'the string', is convenient to discuss the IBVP formulation and approaches to solve it. Since the degeneracy was related to the coefficient function of the wave equation vanishing in an interior point of the spatial segment, being interpreted as the undisturbed position of 'the string', the point of degeneracy, in its turn, can be interpreted as a hinge, where the known transmission conditions (continuity of 'the string' and the transverse component of the longitudinal tension, usually referred to as the flux) hold.

We studied different approaches to solve the IBVP for the 1D degenerate wave equation based on separation of variables (SV). Some of them imply that we pose and solve separately associated IBVPs for both regular parts of 'the string' and then match the obtained solutions at the degeneracy point using the transmission conditions. Physically this means that we: *a*) 'cut' (or, 'break') 'the string' at the degeneracy point, *b*) observe vibrations of both parts of 'the string', and then *c*) match both parts again at any instant to restore 'the entire string' (or, in other

[*] Dept. of Math Analysis and Optimization, Faculty of Mech & Math, Oles Honchar Dnipro National University, 72, Gagarin av., Dnipro, 49010, Ukraine, bvl@dsu.dp.ua

[†] Dept. of Math Analysis and Optimization, Faculty of Mech & Math, Oles Honchar Dnipro National University, 72, Gagarin av., Dnipro, 49010, Ukraine, p.kogut@i.ua

words, 'to repair' it) by forcing some corrections to positions and inclinations of both regular parts for the transmission conditions to hold. We found out that these conditions, generally speaking, are not sufficient for matching (or, for 'repairing' the broken 'degenerate string'), therefore some other conditions are necessary to be applied. To clarify the situation, we have decided to transfer our approaches to the entire regular string composed of two parts with different elastic properties. Since the way to solve the IBVP of the vibrating composite string is known, we expect that will have ideal opportunities for studying our previous approaches and comparing them with those known for the composite string.

Thus, we currently deal with the following well-known initial boundary value problem (referred to below as the *original* IBVP, or the IBVPO for short) for the 1D homogeneous wave-equation in closed space-time rectangle $[0,T] \times [0,l]$

$$\begin{cases} Q\,[W(t,x)] = 0\,, & (t,x) \in G_0\,, \\ \left.\begin{aligned} \dfrac{\partial W(0,x)}{\partial t} &= \overset{**}{W}(x) \\ W(0,x) &= \overset{*}{W}(x) \end{aligned}\right\}, & x \in [0,l]\,, \\ \left.\begin{aligned} W(t,0) &= \chi_1(t) \\ W(t,\,l) &= \chi_2(t) \end{aligned}\right\}, & t \in [0,T]\,, \end{cases} \tag{1.1}$$

where the second order differential operator

$$Q\,[W(t,x)] = \frac{\partial^2 W(t,x)}{\partial t^2} - \frac{\partial}{\partial x}\left(c(x)\,\frac{\partial W(t,x)}{\partial x}\right) \tag{1.2}$$

is defined in $G_0 = G_1 \bigcup G_2$, $G_1 = (0,T] \times (0,x_0)$, $G_2 = (0,T] \times (x_0,l)$, $0 < x_0 < l$, for functions $W(t,x) \in \mathscr{C}^{(2,2)}(G_0)$; the coefficient function is piece-wise constant

$$c(x) = \begin{cases} c_1 > 0\,, & x \in [0,x_0)\,, \\ c_2 > 0\,, & x \in (x_0,\,l]\,, \end{cases} \tag{1.3}$$

being defined non-uniquely at the jump discontinuity: $c_1 \leqslant c(x_0) \leqslant c_2$; the given *a*) control $\chi_1(t), \chi_2(t) \in \mathscr{C}^1[0,T] \bigcap \mathscr{C}^2(0,T]$ and *b*) initial $\overset{*}{W}(x), \overset{**}{W}(x) \in \mathscr{C}^1[0,l]$ functions obey the compatibility conditions

$$\begin{cases} \chi_1(0) = \overset{*}{W}(0)\,, & \overset{*}{W}(l) = \chi_2(0)\,, \\ \chi_1'(0) = \overset{**}{W}(0)\,, & \overset{**}{W}(l) = \chi_2'(0)\,. \end{cases} \tag{1.4}$$

Along the dividing segment $[0,T] \times \{x_0\}$ of subrectangles G_1, G_2 the required solution $W(t,x)$ to the IBVPO obeys the transmission conditions

$$\begin{cases} W(t,x_0-0) = W(t,x_0+0)\,, \\ \mathcal{F}\,(t,x_0-0) = \mathcal{F}\,(t,x_0+0)\,, \end{cases} \quad t \in [0,T]\,, \tag{1.5}$$

the second condition being expressed in terms of the flux

$$\mathcal{F}(t,x) = c(x)\,\frac{\partial W(t,x)}{\partial x}\,\cdot \tag{1.6}$$

The only solution $W(t,x)$ to the IBVPO is known can be interpreted as the distributed over segment $[0,l]$ displacements of a string, subject to known external controls $\chi_1(t)$ and $\chi_2(t)$, imposed at both ends of the string as the Dirichlet boundary conditions, and having the initial distributed displacements $\overset{*}{W}(x)$ and velocities $\overset{**}{W}(x)$. Therefore, the first transmission condition (1.5) expresses the continuity of the string, whereas the second one expresses the continuity of the transverse component of the tension produced in the string.

The original IBVP, posed in $\bar{G}_0 = [0,T]\times[0,l]$, is known can be solved by applying SV [6,8], but our concern related to the original IBVP is based on the following underlying ideas: *a*) reformulating the original IBVP into two associated IBVPs in the above closed space-time subrectangles $\bar{G}_1$, $\bar{G}_2$; *b*) adding conditions of smooth matching the solutions to the associated IBVPs along the dividing segment; *c*) solving such a *composite* IBVP.

For the sake of convenience (and to establish a relation with [1–5] dealing with the IBVP similar to (1.1) – (1.5) for a 1D degenerate wave equation), we introduce the following transformation of the independent variables

$$\begin{cases} t = \dfrac{l}{\sqrt{c_0}}\,\underline{t}\,, & \underline{t}\in[0,\underline{T}]\,,\\[2ex] x = l_1\,\underline{x} + l_1, & \underline{x}\in K_1\,,\\[2ex] x = l_2\,\underline{x} + l_1\,, & \underline{x}\in K_2\,, \end{cases} \qquad \begin{cases} \underline{t} = \dfrac{\sqrt{c_0}}{l}\,t\,, & t\in[0,T]\,,\\[2ex] \underline{x} = \dfrac{x-l_1}{l_1}\,, & x\in[0,\,l_1]\,,\\[2ex] \underline{x} = \dfrac{x-l_1}{l_2}\,, & x\in[l_1,\,l]\,, \end{cases} \tag{1.7}$$

where $K_1 := [-1,0]$, $K_2 := [0,+1]$, $c_0 = \frac{1}{2}\,(c_1+c_2)$, $l_1 = x_0$, $l_2 = l - l_1$; then replace the given in (1.1) functions with the following ones

$$\begin{cases} \overset{**}{u}(\underline{x}) = \dfrac{\sqrt{c_0}}{l}\,\overset{**}{W}(x)\big|_{x\to\underline{x}}\,,\\[1ex] \overset{*}{u}(\underline{x}) = \phantom{\dfrac{\sqrt{c_0}}{l}}\,\overset{*}{W}(x)\big|_{x\to\underline{x}}\,, \end{cases} \qquad \begin{cases} h_1(\underline{t}) = \chi_1(t)\big|_{t\to\underline{t}}\,,\\[1ex] h_2(\underline{t}) = \chi_2(t)\big|_{t\to\underline{t}}\,. \end{cases}$$

Finally, to simplify notation, we drop the bars under the new independent variables and present the IBVPO in the following form

$$\begin{cases} S\,[u(t,x)] = 0\,, & (t,x)\in D_0\,,\\[1ex] \left.\begin{aligned} \dfrac{\partial u(0,x)}{\partial t} &= \overset{**}{u}(x)\\ u(0,x) &= \overset{*}{u}(x) \end{aligned}\right\}, & x\in K_0\,,\\[3ex] \left.\begin{aligned} u\,(t,-1) &= h_1(t)\\ u\,(t,+1) &= h_2(t) \end{aligned}\right\}, & t\in[0,T]\,, \end{cases} \tag{1.8}$$

where: *a*) $K_0 := [-1,+1]$; *b*) the second order differential operator

$$S\,[u(t,x)] = \frac{\partial^2 u(t,x)}{\partial t^2} - \frac{\partial}{\partial x}\left(a(x)\,\frac{\partial u(t,x)}{\partial x}\right) \tag{1.9}$$

is defined in $D_0 = D_1 \bigcup D_2$, $D_1 = (0,T] \times I_1$, $D_2 = (0,T] \times I_2$, $I_1 := (-1,0)$, and $I_2 := (0,+1)$; *c*) the piece-wise constant coefficient function is as follows

$$a(x) = \begin{cases} a_1 = \dfrac{c_1}{c_0}\left(\dfrac{l}{l_1}\right)^2 > 0\,, & x \in J_1\,, \\[2ex] a_2 = \dfrac{c_2}{c_0}\left(\dfrac{l}{l_2}\right)^2 > 0\,, & x \in J_2\,, \end{cases} \tag{1.10}$$

$J_1 := [-1,0)$, $J_2 := (0,+1]$; *d*) the given 1) control $h_1(t), h_2(t)$ and 2) initial $\overset{*}{u}(x)$, $\overset{**}{u}(x)$ functions obey the compatibility conditions

$$\begin{cases} h_1(0) = \overset{*}{u}(-1)\,, & \overset{*}{u}(+1) = h_2(0)\,, \\[1ex] h_1'(0) = \overset{**}{u}(-1)\,, & \overset{**}{u}(+1) = h_2'(0)\,; \end{cases} \tag{1.11}$$

e) the transmission conditions read

$$\begin{cases} u(t,0-0) = u(t,0+0)\,, \\[1ex] f(t,0-0) = f(t,0+0)\,, \end{cases} \qquad t \in [0,T]\,, \tag{1.12}$$

where the flux is defined similarly to that of (1.6)

$$f(t,x) = a(x)\,\frac{\partial u(t,x)}{\partial x}\,\cdot \tag{1.13}$$

The above formulation of the IBVPO below is referred to as the *derived* IBVP, or shortly as the IBVPD. We believe, that confusing the independent variables (t,x), being dimensional in the IBVPO and non-dimensional in the IBVPD, is impossible, due to the context they will be used in.

Following the underlying idea of the current study, we formulate two IBVPs, respectively in the closures $\bar{D}_1$ and $\bar{D}_2$ of subrectangles D_1 and D_2, associated with the derived IBVPD (1.8) – (1.12). The left associated IBVP$_1$ yields to

$$\begin{cases} S\,[u_1(t,x)] = 0\,, & (t,x) \in D_1\,, \\[1ex] \left.\begin{aligned} \frac{\partial u_1(0,x)}{\partial t} &= \overset{**}{u}(x) \\ u_1(0,x) &= \overset{*}{u}(x) \end{aligned}\right\}, & x \in K_1\,, \\[3ex] \left.\begin{aligned} u_1(t,-1) &= h_1(t) \\ |u_1(t,\ \ 0)| &< \infty \end{aligned}\right\}, & t \in [0,T]\,, \end{cases} \tag{1.14}$$

supplemented with the following compatibility conditions

$$\begin{cases} h_1(0) = \overset{*}{u}(-1)\,, & u_1(0,0) \ = \overset{*}{u}(0)\,; \\ h_1'(0) = \overset{**}{u}(-1)\,, & \dfrac{\partial u_1(0,0)}{\partial t} = \overset{**}{u}(0)\,; \end{cases} \tag{1.15}$$

whereas the right associated IBVP$_2$ yields to

$$\begin{cases} S\,[u_2(t,x)] = 0\,, & (t,x) \in D_2\,, \\ \left.\begin{aligned} \dfrac{\partial u_2(0,x)}{\partial t} &= \overset{**}{u}(x) \\ u_2(0,x) \ &= \overset{*}{u}(x) \end{aligned}\right\}, & x \in K_2\,, \\ \left.\begin{aligned} |u_2(t,\ \ 0)| &< \infty \\ u_2(t,+1) \ &= h_2(t) \end{aligned}\right\}, & t \in [0,T]\,, \end{cases} \tag{1.16}$$

supplemented with the following compatibility conditions

$$\begin{cases} \overset{*}{u}(0) = \ \ u_2(0,0)\ , & \overset{*}{u}(+1) = h_2(0)\,; \\ \overset{**}{u}(0) = \dfrac{\partial u_2(0,0)}{\partial t}\,, & \overset{**}{u}(+1) = h_2'(0)\,. \end{cases} \tag{1.17}$$

For the composite solution

$$u(t,x) = \begin{cases} u_1(t,x)\,, & (t,x) \in \bar{D}_1\,, \\ u_2(t,x)\,, & (t,x) \in \bar{D}_2\,, \end{cases} \tag{1.18}$$

to be one-valued, continuous and to have the continuous flux, the following matching conditions, inheriting the transmission conditions (1.12), are imposed along the dividing segment

$$\left\{\begin{aligned} u_1(t,0) &= u_2(t,0)\,, \\ f_1(t,0) &= f_2(t,0)\,, \end{aligned}\right. \qquad t \in [0,T]\,, \tag{1.19}$$

where both fluxes $f_1(t,x), f_2(t,x)$ are defined similarly to that of (1.13).

To answer the question of how to match the solutions to the above associated IBPVs (1.14), (1.16) we will take the following steps.

In Sect. 2 we refer to the energy equation and recall some its properties used further. In Sect. 3 we consider in some detail a classical approach [6], based on SV, to solve the IBVPO (1.8). The solution obtained will be used further not only as the master solution, but also to demonstrate continuous differentiability of the flux at the midpoint $x_0 = 0$, where the coefficient function $a(x)$ is discontinuous.

In Sect. 4 we apply the Laplace transformation (LT) to the associated IBVPs, obtain the solutions to the images of the IBVPs followed by matching the obtained solutions. We consider this approach to be classical as well, since applying the integral transformation of the associated IBVPs precedes matching their solutions. The key issue we are interested is how many conditions we need to match the solutions, whereas the inversion of the image solutions will be postponed to the next publication on the subject.

In Sect. 5 we apply SV to the associated IBVPs to obtain their solutions, then match the obtained solutions followed by applying LT to the matching conditions. Again we are interested in the number of conditions necessary for the matching and show that the matching conditions (1.19) should be supplemented with two more conditions, one of which is local, similarly to (1.19), and the other is nonlocal and presented by the energy equation.

In Sect. 6 we briefly summarize our observations concerning the matching procedures and the number of the matching conditions.

2. The energy equation

In case of treating the string as a whole, being obeyed the IBVPD (1.8), (1.12), (1.11), the energy equation can be obtained following a well known procedure: 1) multiplying the wave equation by the local velocity of the transverse motion, and 2) integrating the product in x over the whole segment K_0

$$\int_{K_0} \frac{\partial u}{\partial t}\left[\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\left(a\,\frac{\partial u}{\partial x}\right)\right] \mathrm{d}x = 0\,. \tag{2.1}$$

Integration is then performed by parts for the first

$$\int_{K_0} \frac{\partial u}{\partial t}\,\frac{\partial^2 u}{\partial t^2}\,\mathrm{d}x = \int_{K_0} \frac{\partial}{\partial t}\left[\frac{1}{2}\left(\frac{\partial u}{\partial t}\right)^2\right] \mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\int_{K_0}\left[\frac{1}{2}\left(\frac{\partial u}{\partial t}\right)^2\right]\mathrm{d}x}_{\Omega(t)}$$

and second

$$\int_{K_0} \frac{\partial u}{\partial t}\,\frac{\partial}{\partial x}\left(a\,\frac{\partial u}{\partial x}\right)\,\mathrm{d}x = \underbrace{\left[\frac{\partial u}{\partial t}\left(a\,\frac{\partial u}{\partial x}\right)\right]\Bigg|_{x=-1}^{x=+1}}_{\mathrm{A}(t)} - \frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\int_{K_0}\left[\frac{a}{2}\left(\frac{\partial u}{\partial x}\right)^2\right]\mathrm{d}x}_{\Pi(t)}$$

terms in (2.1) respectively. The resulting equation for the total energy rate reads

$$\Theta'(t) = \mathrm{A}(t)\,, \qquad t \in [0,T]\,, \tag{2.2}$$

where $\Theta(t)$ is the total energy, composed of the kinetic $\Omega(t)$ and potential $\Pi(t)$ one, and $\mathrm{A}(t)$ is the power of the external forces acting on the ends of the 'string' as the known controls $h_1(t)$, $h_2(t)$. Integrating the above equation in t yields to the required total energy equation

$$\Theta(t) = \Theta(0) + \int_0^t \mathrm{A}(\tau)\,\mathrm{d}\tau\,, \qquad t \in [0,T]\,, \tag{2.3}$$

where the integral over $[0,t]$ represents the work done by the external forces.

In case of treating the string as consisting of two parts, being obeyed the left and the right associated IBVPs of Sect. 1, (1.14), (1.15) and (1.16), (1.17) respectively, one finds the following *a*) total energy rate equations

$$\Theta_j'(t) = \mathrm{A}_j(t)\,, \qquad t \in [0,T]\,, \tag{2.4}$$

and *b*) total energy equations

$$\Theta_j(t) = \Theta_j(0) + \int_0^t \mathrm{A}_j(\tau)\,\mathrm{d}\tau\,, \qquad t \in [0,T]\,, \tag{2.5}$$

where $j = 1,2$ (for the left and right parts of the string respectively), and the powers of the external forces acting on the ends of both parts of the string read

$$\mathrm{A}_1(t) = \left[\frac{\partial u_1}{\partial t}\left(a_1\frac{\partial u_1}{\partial x}\right)\right]\Bigg|_{x=-1}^{x=\ \ 0}\,, \quad \mathrm{A}_2(t) = \left[\frac{\partial u_2}{\partial t}\left(a_2\frac{\partial u_2}{\partial x}\right)\right]\Bigg|_{x=\ \ 0}^{x=+1}\,, \quad t \in [0,T]\,.$$

Both equations (2.2) and (2.3) can be readily derived from (2.4) and (2.5) respectively, since: *a*) the total energy $\Theta(t)$ is additive; *b*) the following equality

$$\left[\frac{\partial u_1}{\partial t}\left(a_1\frac{\partial u_1}{\partial x}\right)\right]\Bigg|_{x=0} = \left[\frac{\partial u_2}{\partial t}\left(a_2\frac{\partial u_2}{\partial x}\right)\right]\Bigg|_{x=0}\,, \qquad t \in [0,T]\,,$$

holds, due to the matching conditions (1.19).

3. A classical approach to solve the IBVPD based on SV

3.1. Preliminaries to SV

A classical approach to solve the IBVPD (1.8) – (1.12) is based on: *a*) reducing the former to an auxiliary IBVP with the homogeneous boundary conditions and *b*) invoking the anzatz

$$\underset{\smile}{v}(t,x) = \underset{\smile}{Q}(t)\,\underset{\smile}{X}(x) \tag{3.1}$$

for finding particular solutions to the auxiliary IBVP by applying SV as follows

$$\frac{\underset{\smile}{Q}''(t)}{\underset{\smile}{Q}(t)} = \frac{\underset{\smile}{F}'(x)}{\underset{\smile}{X}(x)} = -\underset{\smile}{\lambda} = \mathtt{const}\,, \qquad \underset{\smile}{\lambda} > \mathtt{0}\,, \tag{3.2}$$

where $\underset{\smile}{F}(x) = a(x)\,\underset{\smile}{X}'(x)$ is 'the flux' of $\underset{\smile}{X}(x)$. The key issue in invoking (3.1) is to build the complete set of functions $\underset{\smile}{X}(x)$ that satisfy (3.2) and the proper boundary and transmission conditions. According to [6] (a collection of problems, supplementing the textbook [8]; refer, for example, to problems 164 – 166 on p. 37, problem 57 on p. 128, etc.), building the required set is based on

Proposition 3.1. Let the following composite BVP be given

$$\begin{cases} \underset{\smile}{F}'(x) + \underset{\smile}{\lambda}\underset{\smile}{X}(x) = 0\,, & x \in I_1 \bigcup I_2\,, \\ \underset{\smile}{X}(-1) = \underset{\smile}{X}(+1) = 0\,, & \end{cases} \tag{3.3}$$

supplemented with the transmission conditions

$$\begin{cases} \underset{\smile}{X}(0-0) = \underset{\smile}{X}(0+0)\,, \\ \underset{\smile}{F}(0-0) = \underset{\smile}{F}(0+0)\,, \end{cases} \tag{3.4}$$

then: *a*) the complete countable set of the continuous and piece-wise smooth eigenfunctions of the problem is determined as follows

$$\underset{\smile}{X}_\mu(x) = \begin{cases} \left[\sin\left(\dfrac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_1}}\right)\right]^{-1} \sin\left(\dfrac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_1}}\,(1+x)\right), & x \in J_1\,, \\ \left[\sin\left(\dfrac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_2}}\right)\right]^{-1} \sin\left(\dfrac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_2}}\,(1-x)\right), & x \in J_2\,, \end{cases} \tag{3.5}$$

whereas the eigenvalues $\{\underset{\smile}{\lambda}_\mu\}_{\mu=1}^{\infty} \equiv \{\underset{\smile}{\alpha}_\mu^2\}_{\mu=1}^{\infty}$ associated with (3.5) are determined as the set of the squared roots of the following transcendental equation wrt $\underset{\smile}{\alpha}_\mu$

$$\sqrt{a_1}\,\cot\!\left(\frac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_1}}\right) + \sqrt{a_2}\,\cot\!\left(\frac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_2}}\right) = 0\,; \tag{3.6}$$

b) the eigenfunctions (3.5) are orthogonal in $\mathscr{L}_{2,0} \coloneqq \mathscr{L}_2\,(K_0)$, that is

$$\left(\underset{\smile}{X}_\mu\,, \underset{\smile}{X}_\gamma\right)_0 = \int_{K_0} \underset{\smile}{X}_\mu(x)\,\underset{\smile}{X}_\gamma(x)\,\mathrm{d}x = \|\underset{\smile}{X}_\mu\|_0^2\,\delta_{\mu,\gamma}\,, \tag{3.7}$$

$(p,q)_0$ is the inner product of two elements $p(x), q(x) \in \mathscr{L}_{2,0}$, $\|r\|_0$ is the norm of an element $r(x) \in \mathscr{L}_{2,0}$, $\mu, \gamma \in \mathbb{N}$.

We note, in addition to the properties specified in Prop. 3.1, that the eigenfunctions $\underset{\smile}{X}_\mu(x)$ (3.5) have one more remarkable and easily verified property of continuous differentiability of their fluxes $\underset{\smile}{F}_\mu(x) = a(x)\,\underset{\smile}{X}'_\mu(x)$ over the segment K_0, including the midpoint: $\underset{\smile}{F}'_\mu(x)(0-0) = \underset{\smile}{F}'_\mu(0+0)$, where the piece-wise constant coefficient function $a(x)$ is discontinuous. We will apply this property in Sect. 5 to solve the IBVPD using a non-classical approach.

To clarify building $\underset{\smile}{X}_\mu(x)$ (3.5), we consider 3 cases, indicated in Tbl. 1 (the origin of the cases is explained by taking $c_1 = 1$, $c_2 = 4$, $l = 2$ and $x_0 = 0.25$, 0.50, 1.00 in (1.3) and then applying (1.10)). The infinite series of the roots $\underset{\smile}{\alpha}_\mu$ of (3.6) for the cases are formed by periodic shifts of the 8-tuples, 4-tuples, and 2-tuples (the octuples, quadruples, and couples), respectively, placed in Tbl. 1. Some eigenfunctions for the cases are shown in Figs. 3.1, 3.2, 3.3.

Table 1. Roots of transcendental equation (3.6) (cases 1 – 3)

No	a_1	a_2	$\underset{\smile}{\alpha}_\mu$
1	25.6	2.0897959183...	4.11705742307002802074... 7.36681251991832973695... 9.95011848102124236679... 13.80472706515471812453... 17.98595538539280582565... 21.84056396952628158339... 24.42386993062919421323... 27.67362502747749592944...
2	6.4	2.8(4)	3.30228352255807050061... 6.16702110866456655154... 9.72832011660919542354... 12.59305770271569147447...
3	1.6	6.4	2.90995823927664691729... 5.03771237336023407025...

In case 4, with the coefficients $a_1 = a_2 = 4$ (originated from $c_1 = c_2 = 1$, $l = 2$, and $x_0 = 1$ in (1.3)), the roots of (3.6) and the eigenfunctions (3.5) are reduced to

$$\frac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_0}} = \mu\pi - \frac{\pi}{2} \equiv \hat{\alpha}_\mu \,, \qquad \underset{\smile}{X}_\mu(x) = \cos\left(\frac{\underset{\smile}{\alpha}_\mu}{\sqrt{a_0}}\, x\right) = \cos\frac{(2\mu - 1)\,\pi x}{2} \cdot \tag{3.8}$$

On the other hand, independent of Prop. 3.1 considering identically constant coefficient function (1.10) (case 5), that is $a(x) \equiv a_0$, is based on

Proposition 3.2. Let the following BVP be given

$$\begin{cases} a_0 X''(x) + \lambda X(x) = 0\,, & x \in I_0\,, \\ X(-1) = X(+1) = 0\,, \end{cases} \tag{3.9}$$

where $I_0 := (-1, +1)$, then: *a*) the countable sets of the eigenvalues and the eigenfunctions of the problem of two kinds are determined as follows (see Fig. 3.3)

$$\lambda_{1,\mu} \equiv \sigma_{1,\mu}^2 = a_0\,(\mu\pi)^2, \qquad \lambda_{2,\mu} \equiv \sigma_{2,\mu}^2 = a_0\Big(\mu\pi - \frac{\pi}{2}\Big)^2, \tag{3.10}$$

$$\begin{cases} X_{1,\mu}(x) = \sin\left(\dfrac{\sigma_{1,\mu}}{\sqrt{a_0}}\, x\right) = \sin\,(\mu\pi x)\,, \\ X_{2,\mu}(x) = \cos\left(\dfrac{\sigma_{2,\mu}}{\sqrt{a_0}}\, x\right) = \cos\dfrac{(2\mu - 1)\,\pi x}{2}\,; \end{cases} \tag{3.11}$$

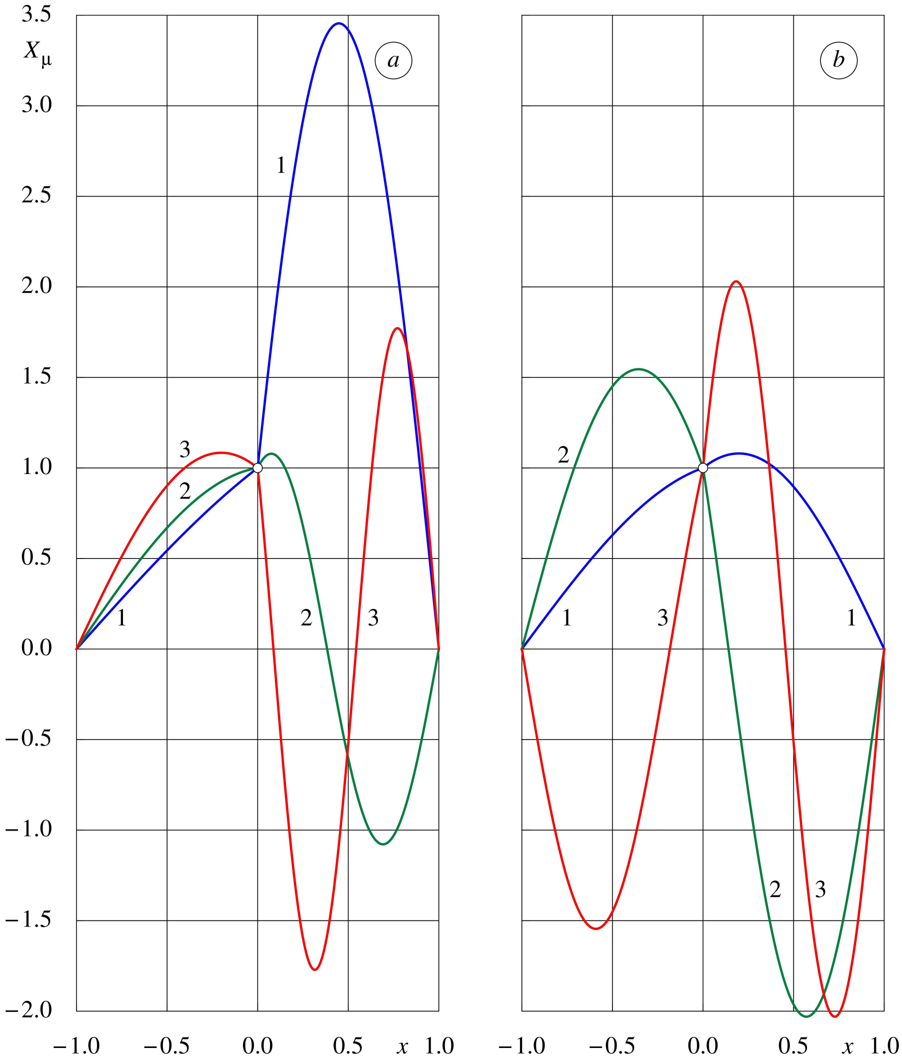


Fig. 3.1. Eigenfunctions (3.5): case 1 (a), case 2 (b); $\mu = 1, 2, 3$

b) eigenfunctions (3.11): 1) of each kind are orthogonal in $\mathscr{L}_{2,0}$; 2) of both kinds are biorthogonal in $\mathscr{L}_{2,0}$, that is

$$\left(X_{k,\mu}\,, X_{k,\gamma}\right)_0 = \|X_{k,\mu}\|_0^2\, \delta_{\mu,\gamma} = \delta_{\mu,\gamma}\,, \qquad \left(X_{1,\mu}\,, X_{2,\gamma}\right)_0 = 0\,. \tag{3.12}$$

We note that the eigenfunctions (3.8) are evidently to be the same as those (3.11) of the second kind (cf. Figs. 3.2, b and 3.3, b).

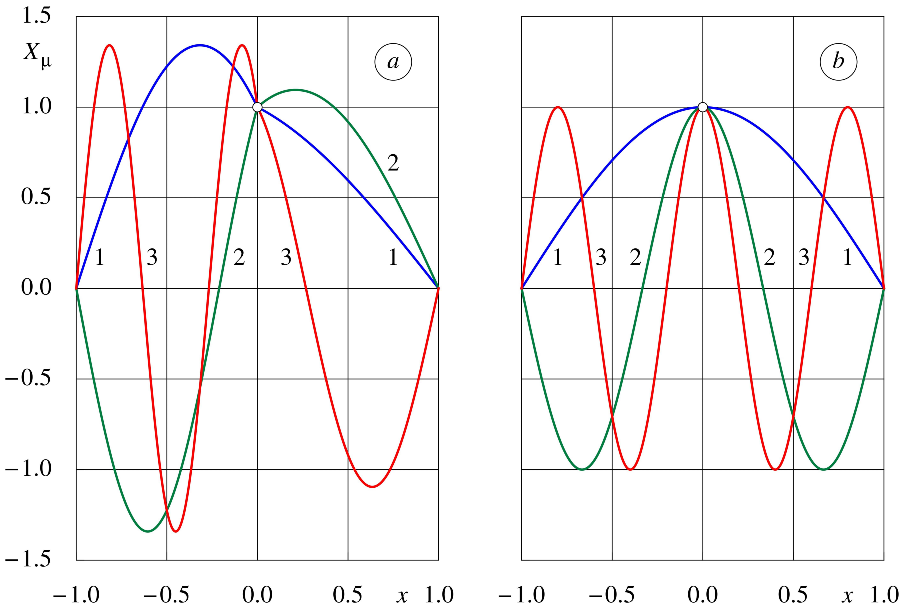


Fig. 3.2. Eigenfunctions (3.5): case 3 (a), and eigenfunctions (3.8): case 4 (b); $\mu = 1, 2, 3$

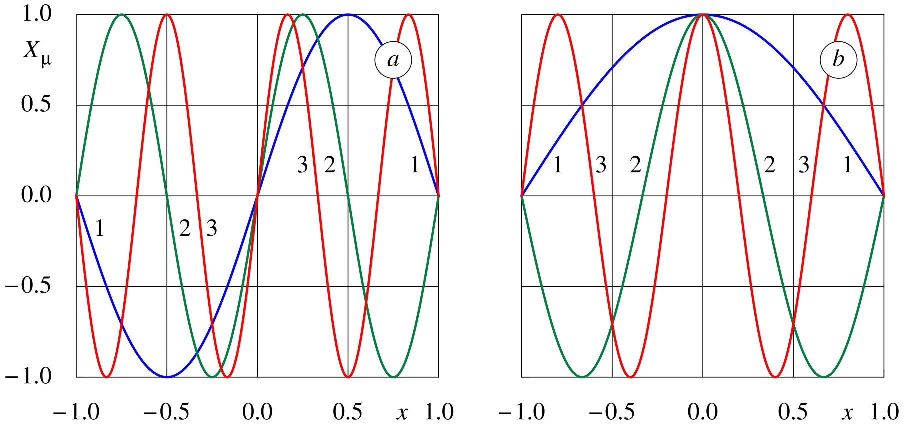


Fig. 3.3. Eigenfunctions (3.11), case 5: of the first (a) and second (b) kinds; $\mu = 1, 2, 3, 4$

3.2. Applying SV to the IBVPD

The ansatz for the solution to the derived IBVPD (1.8) – (1.12) read

$$\underset{\smile}{u}(t,x) = \underset{\smile}{v}(t,x) + \underset{\smile}{w}(t,x)\,, \tag{3.13}$$

where: a) function $\underset{\smile}{v}(t,x)$ is required; b) function $\underset{\smile}{w}(t,x)$ is given as follows

$$\underset{\smile}{w}(t,x) = \underset{\smile}{\phi}_1(x)\,h_1(t) + \underset{\smile}{\phi}_2(x)\,h_2(t)\,; \tag{3.14}$$

c) blending functions $\underset{\smile}{\phi}_1(x), \underset{\smile}{\phi}_2(x)$ satisfy the following conditions

$$\begin{cases} \underset{\smile}{\phi}_1(-1) = 1\,; \qquad \underset{\smile}{\phi}_1(x) \equiv 0\,, \quad x \in K_1\,; \qquad \underset{\smile}{\phi}{}_1''(x) \in \mathscr{C}\,(K_0)\,; \\ \underset{\smile}{\phi}_2(+1) = 1\,; \qquad \underset{\smile}{\phi}_2(x) \equiv 0\,, \quad x \in K_2\,; \qquad \underset{\smile}{\phi}{}_2''(x) \in \mathscr{C}\,(K_0)\,; \end{cases} \tag{3.15}$$

whereas their fluxes $\underset{\smile}{\varphi}_j(x) = a(x)\,\underset{\smile}{\phi}{}_j'(x)$ vanish on the dividing segment: $\underset{\smile}{\varphi}_j(0) = 0$, to meet the transmission conditions (1.12).

Combining (3.13) – (3.15) with the IBVPD formulation (1.8), (1.11) yields to: a) the initial conditions for the required function $\underset{\smile}{v}(t,x)$

$$\begin{cases} \dfrac{\partial \underset{\smile}{v}(0,x)}{\partial t} = \dfrac{\partial \underset{\smile}{u}(0,x)}{\partial t} - \dfrac{\partial \underset{\smile}{w}(0,x)}{\partial t} = \overset{**}{u}(x) - \dfrac{\partial \underset{\smile}{w}(0,x)}{\partial t} \equiv \overset{**}{\underset{\smile}{v}}(x)\,, \\ \underset{\smile}{v}(0,x) \;=\; \underset{\smile}{u}(0,x) \;-\; \underset{\smile}{w}(0,x) \;= \overset{*}{u}(x) \;-\; \underset{\smile}{w}(0,x) \;\equiv \overset{*}{\underset{\smile}{v}}(x)\,, \end{cases} \tag{3.16}$$

and b) reformulation of the IBVPD into the following auxiliary IBVPA wrt $\underset{\smile}{v}(t,x)$

$$\begin{cases} S\,[\underset{\smile}{v}(t,x)] = \underset{\smile}{g}(t,x)\,, & (t,x) \in D_0\,, \\ \left.\begin{aligned} \dfrac{\partial \underset{\smile}{v}(0,x)}{\partial t} &= \overset{**}{\underset{\smile}{v}}(x) \\ \underset{\smile}{v}(0,x) &= \overset{*}{\underset{\smile}{v}}(x) \end{aligned}\right\}, & x \in K_0\,, \\ \left.\begin{aligned} \underset{\smile}{v}(t,-1) &= 0 \\ \underset{\smile}{v}(t,+1) &= 0 \end{aligned}\right\}, & t \in [0,T]\,, \end{cases} \tag{3.17}$$

where $\underset{\smile}{g}(t,x) = -S\,[\underset{\smile}{w}(t,x)]$ is the right hand side of the above non-homogeneous wave equation, and the compatibility conditions hold: $\overset{*}{\underset{\smile}{v}}(-1) = \overset{*}{\underset{\smile}{v}}(+1) = 0\,$, $\overset{**}{\underset{\smile}{v}}(-1) = \overset{**}{\underset{\smile}{v}}(+1) = 0\,$.

Then we: a) expand $\underset{\smile}{g}(t,x)$ and $\overset{*}{\underset{\smile}{v}}(x)$, $\overset{**}{\underset{\smile}{v}}(x)$ into the series wrt $\underset{\smile}{X}_\mu(x)$ (3.5)

$$\underset{\smile}{g}(t,x) = \sum_{\mu=1}^{\infty} \underset{\smile}{g}_\mu(t)\,\underset{\smile}{X}_\mu(x)\,, \qquad \overset{*}{\underset{\smile}{v}}(x) = \sum_{\mu=1}^{\infty} \overset{*}{\underset{\smile}{v}}_\mu\,\underset{\smile}{X}_\mu(x)\,, \quad \overset{**}{\underset{\smile}{v}}(x) = \sum_{\mu=1}^{\infty} \overset{*}{\underset{\smile}{v}}_\mu\,\underset{\smile}{X}_\mu(x)\,,$$

where the coefficients are determined directly by integration: $\|\underset{\smile}{X}_\mu\|_0^2\; y_\mu = \big(y, \underset{\smile}{X}_\mu\big)_0$, for y being $\underset{\smile}{g}(t,x)$, $\overset{*}{\underset{\smile}{v}}(x)$, and $\overset{**}{\underset{\smile}{v}}(x)$, respectively; b) account for Prop. 3.1 and apply ansatz (3.1) for the required solution to the IBVPA (3.17) as follows

$$\underset{\smile}{v}(t,x) = \sum_{\mu=1}^{\infty} \underset{\smile}{Q}_\mu(t)\,\underset{\smile}{X}_\mu(x)\,; \tag{3.18}$$

c) obtain the following Cauchy problems wrt the coefficient functions of (3.18)

$$\begin{cases} \underset{\smile}{Q}{}_\mu''(t) + \underset{\smile}{\alpha}{}_\mu^2\,\underset{\smile}{Q}_\mu(t) = \underset{\smile}{g}_\mu(t)\,, \qquad t \in (0,T]\,, \\ \left.\begin{aligned} \underset{\smile}{Q}{}_\mu'(0) &= \overset{**}{\underset{\smile}{v}}_\mu \\ \underset{\smile}{Q}_\mu(0) &= \overset{*}{\underset{\smile}{v}}_\mu \end{aligned}\right\}, \end{cases} \qquad \mu \in \mathbb{N}\,. \tag{3.19}$$

The solutions to the above Cauchy problems

$$\underset{\smile}{Q}_\mu(t) = \underset{\smile}{\overset{*}{v}}_\mu \cos\big(\underset{\smile}{\alpha}_\mu t\big) + \underset{\smile}{\alpha}_\mu^{-1}\, \underset{\smile}{\overset{**}{v}}_\mu \sin\big(\underset{\smile}{\alpha}_\mu t\big) + \underset{\smile}{\alpha}_\mu^{-1} \int_0^t \underset{\smile}{g}_\mu(\theta) \sin\big[\underset{\smile}{\alpha}_\mu(t-\theta)\big]\,\mathrm{d}\theta$$

can be presented shortly as

$$\underset{\smile}{Q}_\mu(t) = \underset{\smile}{\overset{*}{v}}_\mu \cos\big(\underset{\smile}{\alpha}_\mu t\big) + \underset{\smile}{\alpha}_\mu^{-1}\, \underset{\smile}{\overset{**}{v}}_\mu \sin\big(\underset{\smile}{\alpha}_\mu t\big) + \underset{\smile}{\alpha}_\mu^{-1}\, \underset{\smile}{g}_\mu(t) * \sin\big(\underset{\smile}{\alpha}_\mu t\big)\,, \tag{3.20}$$

by invoking the notion of convolution.

Finally, the solution to the IBVPD (1.8), (1.11) reads

$$\underset{\smile}{u}(t,x) = \sum_{\mu=1}^{\infty} \underset{\smile}{Q}_\mu(t)\, \underset{\smile}{X}_\mu(x) + \underset{\smile}{\phi}_1(x)\, h_1(t) + \underset{\smile}{\phi}_2(x)\, h_2(t)\,. \tag{3.21}$$

We note, that the flux of the above solution

$$\underset{\smile}{f}(t,x) = a(x)\,\frac{\partial \underset{\smile}{u}(t,x)}{\partial x} = \sum_{\mu=1}^{\infty} \underset{\smile}{Q}_\mu(t)\, \underset{\smile}{F}_\mu(x) + \underset{\smile}{\varphi}_1(x)\, h_1(t) + \underset{\smile}{\varphi}_2(x)\, h_2(t) \tag{3.22}$$

is continuous and continuously differentiable wrt x over the segment K_0, due to continuity and continuous differentiabilty of the fluxes $\underset{\smile}{F}_\mu(x)$ (refer to p. 96).

4. A classical approach to solve the IBVPD based on LT

4.1. Preliminaries to LT

For a function $p(t)$, $t \in [0,\infty)$, its Laplace transform [7] is defined as follows

$$P(\tau) = \mathfrak{L}\,[p(t)] := \int_0^\infty p(t)\,\mathrm{e}^{-\tau t}\,\mathrm{d}t\,, \qquad \tau = \xi + i\eta \in \mathbb{C}\,, \tag{4.1}$$

provided the original function $p(t)$ satisfies the known sufficient conditions for the image function $P(\tau)$ to exist.

Applying the Laplace transformation for solving second order partial differential equations and integro-differential equations of convolution type is based on:

a) the rule for the images of the first and second derivatives of the function $p(t)$

$$\mathfrak{L}\big[p'(t)\big] = P(\tau)\,\tau - p(0)\,, \quad \mathfrak{L}\big[p''(t)\big] = P(\tau)\,\tau^2 - p(0)\,\tau - p'(0)\,;$$

b) the convolution theorem

$$\mathfrak{L}\,[p(t) * q(t)] = \mathfrak{L}\,[p(t)] \cdot \mathfrak{L}\,[q(t)] = P(\tau) \cdot Q(\tau)\,.$$

If the image function $P(\tau)$ for an original function $p(t)$ is known, then applying the inverse Laplace transformation [7], sometimes referred to as the Bromwich integral, yields to the required original function as follows

$$p(t) = \mathfrak{L}^{-1}\,[P(\tau)] = \frac{1}{2\pi i}\int_{\xi^*-i\infty}^{\xi^*+i\infty} P(\tau)\,\mathrm{e}^{+t\tau}\,\mathrm{d}\tau\,, \tag{4.2}$$

where $\Re\,\tau = \xi^*$ is a vertical straight line on the whole τ-plane lying to the right of all the singularities of the image $P(\tau)$.

4.2. Applying LT to the BVP$_1$ and BVP$_2$

Applying (4.1) to the associated IBVP$_j$ (1.14), (1.16) yields to their images being the following BVP$_j$

$$\begin{cases} \dfrac{\mathrm{d}F_1(\tau,x)}{\mathrm{d}x} - \tau^2 U_1(\tau,x) = -\tau\,\overset{*}{u}(x) - \overset{**}{u}(x)\,, \qquad x \in I_1\,, \\ U_1(\tau,-1) = H_1(\tau)\,, \end{cases} \tag{4.3}$$

$$\begin{cases} \dfrac{\mathrm{d}F_2(\tau,x)}{\mathrm{d}x} - \tau^2 U_2(\tau,x) = -\tau\,\overset{*}{u}(x) - \overset{**}{u}(x)\,, \qquad x \in I_2\,, \\ U_2(\tau,+1) = H_2(\tau)\,, \end{cases} \tag{4.4}$$

wrt the images $U_j(\tau,x)$ of the solutions $u_j(t,x)$ to the IBVP$_j$, where

$$F_j(\tau,x) = a_j\,\frac{\mathrm{d}U_j(\tau,x)}{\mathrm{d}x} \tag{4.5}$$

are the fluxes of the images $U_j(\tau,x)$ (or, it is the same, the images of the fluxes $f_j(t,x)$ of the solutions $u_j(t,x)$). Both BVP$_j$ are supplemented with the images of the matching conditions (1.19)

$$\begin{cases} U_1(\tau,0) = U_2(\tau,0)\,, \\ F_1(\tau,0) = F_2(\tau,0)\,. \end{cases} \tag{4.6}$$

4.3. Finding the image functions $U_j(\tau,x)$

The equations of the BVP$_j$ (4.3), (4.4) are linear non-homogeneous ordinary differential equations of the second order, and their 2-parameter solutions (usually referred to as the general solutions) can be readily written as

$$\begin{cases} U_1(\tau,x) = A_1(\tau,x)\,\mathsf{e}^{-\frac{\tau x}{s_1}} + B_1(\tau,x)\,\mathsf{e}^{+\frac{\tau x}{s_1}}\,, \\ U_2(\tau,x) = A_2(\tau,x)\,\mathsf{e}^{-\frac{\tau x}{s_2}} + B_2(\tau,x)\,\mathsf{e}^{+\frac{\tau x}{s_2}}\,, \end{cases} \tag{4.7}$$

where: *a*) the 1-parameter coefficient functions $A_j(\tau,x)$, $B_j(\tau,x)$ read

$$\begin{cases} A_1(\tau,x) = A_{1,0}(\tau) + \dfrac{\Lambda_1^+(\tau;-1,x)}{s_1\tau}\,, \\ B_1(\tau,x) = B_{1,0}(\tau) - \dfrac{\Lambda_1^-(\tau;-1,x)}{s_1\tau}\,, \end{cases} \tag{4.8}$$

$$\begin{cases} A_2(\tau, x) = A_{2,0}(\tau) - \dfrac{\Lambda_2^+(\tau; x, +1)}{s_2\tau}\,, \\[2ex] B_2(\tau, x) = B_{2,0}(\tau) + \dfrac{\Lambda_2^-(\tau; x, +1)}{s_2\tau}\,; \end{cases} \tag{4.9}$$

b) the auxiliary functions involved in (4.8), (4.9) are as follows

$$\Lambda_j^{\mp}(\tau; a, b) = \frac{1}{2}\int_a^b [\tau\,\overset{*}{u}(\xi) - \overset{**}{u}(\xi)]\ \mathrm{e}^{\mp\frac{\tau\xi}{s_j}}\,\mathrm{d}\xi\,; \tag{4.10}$$

c) the parameter functions $A_{j,0}(\tau)$, $B_{j,0}(\tau)$ are undetermined; and d) $s_j^2 = a_j$.

To determine four functions $A_{j,0}(\tau)$, $B_{j,0}(\tau)$ we need four conditions, they are known to be: a) two boundary conditions of the BVP$_1$ (4.3) and the BVP$_2$ (4.4); b) two matching conditions (4.6).

Due to the known images (4.7), we have for their boundary values at $|x| = 1$ the following expressions

$$U_1(\tau, -1) = \left[A_{1,0}(\tau) + \underbrace{\frac{\Lambda_1^+(\tau; -1, -1)}{s_1\tau}}_{0}\right]\mathrm{e}^{+\frac{\tau}{s_1}} + \left[B_{1,0}(\tau) - \underbrace{\frac{\Lambda_1^-(\tau; -1, -1)}{s_1\tau}}_{0}\right]\mathrm{e}^{-\frac{\tau}{s_1}},$$

$$U_2(\tau, +1) = \left[A_{2,0}(\tau) - \underbrace{\frac{\Lambda_2^+(\tau; +1, +1)}{s_2\tau}}_{0}\right]\mathrm{e}^{-\frac{\tau}{s_2}} + \left[B_{2,0}(\tau) + \underbrace{\frac{\Lambda_2^-(\tau; +1, +1)}{s_2\tau}}_{0}\right]\mathrm{e}^{+\frac{\tau}{s_2}},$$

and after simplifications the above expressions reduce to

$$\begin{cases} U_1(\tau, -1) = A_{1,0}(\tau)\ \mathrm{e}^{+\frac{\tau}{s_1}} + B_{1,0}(\tau)\ \mathrm{e}^{-\frac{\tau}{s_1}}\,, \\[2ex] U_2(\tau, +1) = A_{2,0}(\tau)\ \mathrm{e}^{-\frac{\tau}{s_2}} + B_{2,0}(\tau)\ \mathrm{e}^{+\frac{\tau}{s_2}}\,. \end{cases} \tag{4.11}$$

Doing in the same way, we obtain: a) the values of the images $U_j(\tau, x)$ and b) their fluxes $F_j(\tau, x)$ at the midpoint $x_0 = 0$

$$\begin{cases} U_1(\tau, 0) = \left[A_{1,0}(\tau) + \dfrac{\Lambda_1^+(\tau; -1, 0)}{s_1\tau}\right] + \left[B_{1,0}(\tau) - \dfrac{\Lambda_1^-(\tau; -1, 0)}{s_1\tau}\right], \\[2ex] U_2(\tau, 0) = \left[A_{2,0}(\tau) - \dfrac{\Lambda_2^+(\tau; 0, +1)}{s_2\tau}\right] + \left[B_{2,0}(\tau) + \dfrac{\Lambda_2^-(\tau; 0, +1)}{s_2\tau}\right], \end{cases} \tag{4.12}$$

$$\begin{cases} F_1(\tau, 0) = -s_1\tau\, A_{1,0}(\tau) - \Lambda_1^+(\tau; -1, 0) + s_1\tau\, B_{1,0}(\tau) - \Lambda_1^-(\tau; -1, 0)\,, \\[2ex] F_2(\tau, 0) = -s_2\tau\, A_{2,0}(\tau) + \Lambda_2^+(\tau; 0, +1) + s_2\tau\, B_{2,0}(\tau) + \Lambda_2^-(\tau; 0, +1)\,. \end{cases} \tag{4.13}$$

Finally, substituting: a) the values (4.11) into the boundary conditions of the BVP$_1$ (4.3) and the BVP$_2$ (4.4) respectively; then b) the values (4.12), (4.13) into the first and the second matching conditions (4.6) respectively, yields to the following linear algebraic system wrt the required parameter functions

$$\begin{pmatrix} \mathsf{e}^{+\frac{\tau}{s_1}} & \mathsf{e}^{-\frac{\tau}{s_1}} & 0 & 0 \\ 0 & 0 & \mathsf{e}^{-\frac{\tau}{s_2}} & \mathsf{e}^{+\frac{\tau}{s_2}} \\ +s_1 s_2\,\tau & +s_1 s_2 \tau & -s_1 s_2\,\tau & -s_1 s_2\,\tau \\ +s_1\,\tau & -s_1 \tau & -s_2 \tau & +s_2 \tau \end{pmatrix} \begin{pmatrix} A_{1,0}(\tau) \\ B_{1,0}(\tau) \\ A_{2,0}(\tau) \\ B_{2,0}(\tau) \end{pmatrix} = \begin{pmatrix} H_1(\tau) \\ H_2(\tau) \\ R_1(\tau) \\ R_2(\tau) \end{pmatrix}, \tag{4.14}$$

where the rhs of the two last equations read

$$\begin{cases} R_1(\tau) = -\, s_2 \Lambda_1^+(\tau;-1,0) + s_2 \Lambda_1^-(\tau;-1,0) \\ \qquad\qquad -\, s_1 \Lambda_2^+(\tau;0,+1) + s_1 \Lambda_2^-(\tau;0,+1)\,, \\ R_2(\tau) = -\quad \Lambda_1^+(\tau;-1,0) - \quad \Lambda_1^-(\tau;-1,0) \\ \qquad\qquad -\quad \Lambda_2^+(\tau;0,+1) - \quad \Lambda_2^-(\tau;0,+1)\,. \end{cases} \tag{4.15}$$

We present the expression for the determinant of the system (4.14), (4.15)

$$\Delta_0(\tau) = s_1 s_2\,\tau \Big[(s_1 - s_2) \left(\mathsf{e}^{-\theta_1 \tau} - \mathsf{e}^{+\theta_1 \tau} \right) + (s_1 + s_2) \left(\mathsf{e}^{-\theta_2 \tau} - \mathsf{e}^{+\theta_2 \tau} \right) \Big], \tag{4.16}$$

$$\theta_1 = \frac{(s_1 - s_2)\,\tau}{s_1 s_2}\,, \qquad \theta_2 = \frac{(s_1 + s_2)\,\tau}{s_1 s_2}\,,$$

since this expression is important both for: *a*) solving the system and *b*) choosing the method of inversion of the images $U_j(\tau, x)$ of the solutions $u_j(\tau, x)$ to the associated IBVP$_j$ (1.14), (1.16).

5. A non-classical approach to solve the IBVPD based on SV+LT

5.1. Preliminaries to applying SV

In case of identically constant coefficient function $a(x) \equiv a_0$ an alternative approach to SV (compared to that of (3.2))

$$\frac{1}{a_0} \frac{O''(t)}{O(t)} = \frac{X''(x)}{X(x)} = -\lambda\,, \qquad \lambda > 0\,, \tag{5.1}$$

yields to the following BVP wrt $X(x)$

$$\begin{cases} X''(x) + \lambda X(x) = 0\,, \qquad x \in I_0\,, \\ X(\mp 1) = 0\,. \end{cases} \tag{5.2}$$

not involving a_0 (cf. the BVP (3.9) and its eigenvalues (3.10) and eigenfunctions (3.11)). The above BVP is known to have the countable sets of the eigenvalues and eigenfunctions of two kinds

$$\begin{cases} \lambda_{1,\mu} \equiv \alpha_\mu^2 = (\mu\pi)^2\,, & Y_\mu(x) = \sin\big(\alpha_\mu x\big)\,, \\ \lambda_{2,\mu} \equiv \omega_\mu^2 = \Big(\mu\pi - \dfrac{\pi}{2}\Big)^2, & Z_\mu(x) = \cos\big(\omega_\mu x\big)\,. \end{cases} \tag{5.3}$$

The eigenfunctions (5.3): *a*) of each kind are orthogonal in $\mathscr{L}_{2,0}$; *b*) of both kinds are biorthogonal in $\mathscr{L}_{2,0}$. The eigenfunctions $Y_\mu(x)$ of the first kind (see Fig. 3.3, *a*) are responsible for the string inclination to the x-axis (the undisturbed string position) at the midpoint $x_0 = 0$ of the segment K_0, whereas the eigenfunctions $Z_\mu(x)$ of the second kind (see Fig. 3.3, *b*) are responsible for the string standoff from the x-axis at the midpoint.

For solving the IBVPD, using a non-classical approach, we cut the segment K_0 down the midpoint $x_0 = 0$ and obtain functions of two kinds: *a*) $Y_{1,\mu}(x) = Y_\mu(x)$, $Z_{1,\mu}(x) = Z_\mu(x)$ for the left subsegment K_1 and *b*) $Y_{2,\mu}(x) = Y_\mu(x)$, $Z_{2,\mu}(x) = Z_\mu(x)$ for the right subsegment K_2. The functions $\big\{Y_{j,\mu}(x)\big\}_{\mu=1}^\infty$, $\big\{Z_{j,\mu}(x)\big\}_{\mu=1}^\infty$ of each kind are orthogonal in $\mathscr{L}_{2,j} = \mathscr{L}_2(K_j)$, but they are not biorthogonal, that is

$$\big(Y_{j,\mu}, Y_{j,\gamma}\big)_j = \big(Z_{j,\mu}, Z_{j,\gamma}\big)_j = \frac{1}{2}\,\delta_{\mu,\gamma}\,, \qquad \big(Y_{j,\mu}, Z_{j,\gamma}\big)_j = \mp\frac{\alpha_\mu}{\alpha_\mu^2 - \omega_\gamma^2} \neq 0\,, \tag{5.4}$$

where

$$(p_j, q_j)_j = \int_{K_j} p_j(x)\, q_j(x)\; \mathrm{d}x = \mp\int_0^{\mp 1} p_j(x)\, q_j(x)\; \mathrm{d}x \tag{5.5}$$

is the inner product of two elements $p_j(x), q_j(x) \in \mathscr{L}_{2,j}$, whereas $(r_j, r_j)_j = \|r_j\|_j^2$ is the norm of an element $r_j(x) \in \mathscr{L}_{2,j}$.

Due to completeness of $\big\{Y_{j,\mu}(x)\big\}_{\mu=1}^\infty$ and $\big\{Z_{j,\mu}(x)\big\}_{\mu=1}^\infty$ (separetely) on K_j, any function $r_j(x)$, smooth on K_j and vanishing at the end points of K_j, can be expanded in series of $\big\{Y_{j,\mu}(x)\big\}_{\mu=1}^\infty$ or $\big\{Z_{j,\mu}(x)\big\}_{\mu=1}^\infty$, uniformly convergent on K_j. But, for smooth functions $r_j(x)$ not vanishing at the midpoint $x_0 = 0$ (this is the case as for: *a*) the solutions $u_j(t,x)$ to the associated IBVP$_j$ of Sect. 1, as *b*) the solutions $v_j(t,x)$ to the auxiliary IBVPA$_j$ (5.11), (5.12), introduced below) expansions in series of $\big\{Y_{j,\mu}(x)\big\}_{\mu=1}^\infty$ are not valid. Therefore, choosing $\big\{Z_{j,\mu}(x)\big\}_{\mu=1}^\infty$ for representing on K_j smooth functions $r_j(x)$, we nevertheless account that 'the fluxes' $\Psi_{j,\mu}(x) = a_j\, Z'_{j,\mu}(x)$ vanish at the midpoint identically on μ and meet the second condition (1.19) only in a trivial way. To avoid this, we will introduce 'manually' correction terms $\phi_{j+4}(x)\, k_j(t)$ in the series solutions $u_j(t,x)$ for the string to have a non-vanishing inclination at the midpoint.

5.2. Applying SV to the IBVP$_1$ and the IBVP$_2$

The ansatze for the solutions to the associated IBVP$_j$ of Sect. 1 read

$$u_j(t,x) = v_j(t,x) + w_j(t,x)\,, \tag{5.6}$$

where: *a*) the functions $v_j(t,x)$ are required; *b*) the functions $w_j(t,x)$ are given as follows

$$w_j(t,x) = \phi_j(x)\,h_j(t) + \phi_{j+2}(x)\,h_{j+2}(t) + \phi_{j+4}(x)\,k_j(t)\,, \tag{5.7}$$

c) the blending functions $\phi_{j+\nu}(x)$, $\nu\in\{0,2,4\}$, satisfy the following boundary

$$\begin{cases} \phi_1(-1) = 1\,, & \phi_1(0) = 0\,, \\ \phi_3(-1) = 0\,, & \phi_3(0) = 1\,, \\ \phi_5(-1) = 0\,, & \phi_5(0) = 0\,, \end{cases} \qquad \begin{cases} \phi_2(0) = 0\,, & \phi_2(+1) = 1\,, \\ \phi_4(0) = 1\,, & \phi_4(+1) = 0\,, \\ \phi_6(0) = 0\,, & \phi_6(+1) = 0\,, \end{cases} \tag{5.8}$$

and regularity

$$\begin{cases} \psi_{1+\nu}(x) \equiv \varphi'_{1+\nu}(x) \equiv \big(a_1\,\phi'_{1+\nu}(x)\big)' \equiv a_1\,\phi''_{1+\nu}(x) \in \mathscr{C}[-1,0]\,, \\ \psi_{2+\nu}(x) \equiv \varphi'_{2+\nu}(x) \equiv \big(a_2\,\phi'_{2+\nu}(x)\big)' \equiv a_2\,\phi''_{2+\nu}(x) \in \mathscr{C}[0,+1]\,, \end{cases} \tag{5.9}$$

conditions, whereas their fluxes $\varphi_{j+\nu}(x) = a_j\,\phi'_{j+\nu}(x)$ obey the following conditions at the midpoint: $\varphi_j(0) = 0$, $\varphi_{j+2}(0) = 0$, $\varphi_{j+4}(0) = a_j$, $\psi_{j+4}(0) = 0$, and for convenience we denote $b_{j+2} := \psi_{j+2}(0)$; *d*) $h_{j+2}(t)$ and $k_j(t)$ are the required corrections to the standoff and inclination of the left ($j=1$) and right ($j=2$) parts of the string on the dividing segment to meet the matching conditions (1.19).

Assuming that $h_{j+2}(0)=\overset{*}{u}(0)$, $h'_{j+2}(0)=\overset{**}{u}(0)$ and combining (5.6) – (5.8) yield to: *a*) the initial conditions for $v_j(t,x)$

$$\begin{cases} v_j(0,x) = u_j(0,x) - w_j(0,x) \equiv \overset{*}{v}_j(x)\,, \\ \dfrac{\partial v_j(0,x)}{\partial t} = \dfrac{\partial u_j(0,x)}{\partial t} - \dfrac{\partial w_j(0,x)}{\partial t} \equiv \overset{**}{v}_j(x)\,; \end{cases} \tag{5.10}$$

b) reformulation of the IBVP$_1$ (1.14), (1.15) into the auxiliary IBVPA$_1$ wrt $v_1(t,x)$

$$\begin{cases} S\,[v_1(t,x)] = g_1(t,x)\,, & (t,x) \in D_1\,, \\ \left.\begin{aligned} \dfrac{\partial v_1(0,x)}{\partial t} &= \overset{**}{v}(x) \\ v_1(0,x) &= \overset{*}{v}(x) \end{aligned}\right\}, & x \in [-1,0]\,, \\ \left.\begin{aligned} v_1(t,-1) &= 0 \\ |v_1(t,\ \ 0)| &< \infty \end{aligned}\right\}, & t \in [0,T]\,; \end{cases} \tag{5.11}$$

c) reformulation of the IBVP$_2$ (1.16), (1.17) into the auxiliary IBVPA$_2$ wrt $v_2(t,x)$

$$\left\{\begin{array}{ll} S\left[v_2(t,x)\right] = g_2(t,x)\,, & (t,x)\in D_2\,,\\ \left.\begin{array}{l}\dfrac{\partial v_2(0,x)}{\partial t} = \overset{**}{v}(x)\\ v_2(0,x) = \overset{*}{v}(x)\end{array}\right\}, & x\in[0,+1]\,,\\ \left.\begin{array}{l}|v_2(t,\ \ 0)| < \infty\\ v_2(t,+1) = 0\end{array}\right\}, & t\in[0,T]\,;\end{array}\right. \tag{5.12}$$

where

$$\begin{aligned} g_j(t,x) = -S\,w_j(t,x) = &-\phi_j(x)\,h_j''(t) - \phi_{j+2}(x)\,h_{j+2}''(t) - \phi_{j+4}(x)\,k_j''(t)\\ &+\psi_j(x)\,h_j(t) + \psi_{j+2}(x)\,h_{j+2}(t) + \psi_{j+4}(x)\,k_j(t)\,. \end{aligned} \tag{5.13}$$

Then the right hand sides (5.13) of the above non-homogeneous wave equations $S\left[v_j(t,x)\right] = g_j(t,x)$ and the initial functions $\overset{*}{v}_j(x)$, $\overset{**}{v}_j(x)$ (5.10) are expanded into the series wrt $Z_{j,\mu}(x)$

$$g_j(t,x) = \sum_{\mu=1}^{\infty} g_{j,\mu}(t)\,Z_{j,\mu}(x)\,,\quad \overset{*}{v}_j(x) = \sum_{\mu=1}^{\infty}\overset{*}{v}_{j,\mu}\,Z_{j,\mu}(x)\,,\quad \overset{**}{v}_j(x) = \sum_{\mu=1}^{\infty}\overset{**}{v}_{j,\mu}\,Z_{j,\mu}(x)\,,$$

where the expanded form of $g_{j,\mu}(t)$ reads

$$\begin{aligned} g_{j,\mu}(t) = &-\phi_{j,\mu}\,h_j''(t) \quad - \phi_{j+2,\mu}\,h_{j+2}''(t) \quad - \phi_{j+4,\mu}\,k_j''(t)\\ &+ a_j\,\phi^\star_{j,\mu}\,h_j(t) + a_j\,\phi^\star_{j+2,\mu}\,h_{j+2}(t) + a_j\,\phi^\star_{j+4,\mu}\,k_j(t)\,, \end{aligned} \tag{5.14}$$

and the subcoefficients $\phi_{j+\nu,\mu}$, $\phi^\star_{j+\nu,\mu}$ are as follows

$$\left\|Z_{j,\mu}\right\|_j^2\,\phi_{j+\nu,\mu} = \left(\phi_{j+\nu}, Z_{j,\mu}\right)_j\,,\qquad \left\|Z_{j,\mu}\right\|_j^2\,\phi^\star_{j+\nu,\mu} = \left(\phi''_{j+\nu}, Z_{j,\mu}\right)_j\,. \tag{5.15}$$

Substituting the ansatze for the solutions to the IBVPA$_j$ (5.11), (5.12)

$$v_j(t,x) = \sum_{\mu=1}^{\infty} O_{j,\mu}(t)\,Z_{j,\mu}(x) \tag{5.16}$$

into the IBVPA$_j$ yields to the Cauchy problems wrt the coefficient functions $O_{j,\mu}(t)$

$$\left\{\begin{array}{l} O''_{j,\mu}(t) + \omega_\mu^2\,O_{j,\mu}(t) = g_{j,\mu}(t)\,, \qquad t\in(0,T]\,,\\ \left.\begin{array}{l} O'_{j,\mu}(0) = \overset{**}{v}_{j,\mu}\\ O_{j,\mu}(0) = \overset{*}{v}_{j,\mu}\end{array}\right\}, \end{array}\right. \qquad \mu\in\mathbb{N}\,. \tag{5.17}$$

The resulting expressions for the coefficients can be readily presented in the convolution form as follows

$$O_{j,\mu}(t) = \overset{*}{v}_{j,\mu} \cos\big(\omega_\mu t\big) + \omega_\mu^{-1} \overset{**}{v}_{j,\mu} \sin\big(\omega_\mu t\big) + \omega_\mu^{-1} g_{j,\mu}(t) * \sin\big(\omega_\mu t\big) . \tag{5.18}$$

Finally, the ansatze (5.6) give the solutions to the IBVP$_j$

$$\begin{cases} u_j(t,x) = \displaystyle\sum_{\mu=1}^{\infty} O_{j,\mu}(t)\, Z_{j,\mu}(x) \\ \qquad\qquad + \phi_j(x)\, h_j(t) + \phi_{j+2}(x)\, h_{j+2}(t) + \phi_{j+4}(x)\, k_j(t) . \end{cases} \tag{5.19}$$

We need further the first order partial derivatives of the solutions (5.19) and the flux, calculated here in advance as follows

$$\begin{cases} \dfrac{\partial u_j(t,x)}{\partial t} = \displaystyle\sum_{\mu=1}^{\infty} O'_{j,\mu}(t)\, Z_{j,\mu}(x) \\ \qquad\qquad + \phi_j(x)\, h'_j(t) + \phi_{j+2}(x)\, h'_{j+2}(t) + \phi_{j+4}(x)\, k'_j(t) , \end{cases} \tag{5.20}$$

$$\begin{cases} \dfrac{\partial u_j(t,x)}{\partial x} = \displaystyle\sum_{\mu=1}^{\infty} O_{j,\mu}(t)\, Z'_{j,\mu}(x) \\ \qquad\qquad + \phi'_j(x)\, h_j(t) + \phi'_{j+2}(x)\, h_{j+2}(t) + \phi'_{j+4}(x)\, k_j(t) , \end{cases} \tag{5.21}$$

$$\begin{cases} f_j(t,x) = \displaystyle\sum_{\mu=1}^{\infty} O_{j,\mu}(t)\, \Psi_{j,\mu}(x) \\ \qquad\qquad + \varphi_j(x)\, h_j(t) + \varphi_{j+2}(x)\, h_{j+2}(t) + \varphi_{j+4}(x)\, k_j(t) , \end{cases} \tag{5.22}$$

where $\Psi_{j,\mu}(x) = a_j\, Z'_{j,\mu}(x)$ are the fluxes of $Z_{j,\mu}(x)$, and the first order differentiation of the coefficient functions (5.18)

$$O'_{j,\mu}(t) = -\omega_\mu\, \overset{*}{v}_{j,\mu} \sin\big(\omega_\mu t\big) + \overset{**}{v}_{j,\mu} \cos\big(\omega_\mu t\big) + g_{j,\mu}(t) * \cos\big(\omega_\mu t\big) \tag{5.23}$$

is performed accounting for the formula

$$\big(p(t) * q(t)\big)' = p(t)\, q(0) + p(t) * q'(t) . \tag{5.24}$$

5.3. Matching the solutions to the IBVP$_1$ and the IBVP$_2$

Matching the obtained solutions (5.19) to the IBVP$_1$ (1.14), (1.15) and IBVP$_2$ (1.16), (1.17) of Sect. 1 follows the procedure:

a) to substitute $u_j(t,x)$ into the matching conditions (1.19) on p. 93, as follows

$$\begin{cases} \displaystyle\sum_{\mu=1}^{\infty} O_{1,\mu}(t)\, Z_{1,\mu}(0) + \phi_1(0)\, h_1(t) + \phi_3(0)\, h_3(t) + \phi_5(0)\, k_1(t) \\ \displaystyle = \sum_{\mu=1}^{\infty} O_{2,\mu}(t)\, Z_{2,\mu}(0) + \phi_2(0)\, h_2(t) + \phi_4(0)\, h_4(t) + \phi_6(0)\, k_2(t)\,, \end{cases} \tag{5.25}$$

$$\begin{cases} \displaystyle\sum_{\mu=1}^{\infty} O_{1,\mu}(t)\, \Psi_{1,\mu}(0) + \varphi_1(0)\, h_1(t) + \varphi_3(0)\, h_3(t) + \varphi_5(0)\, k_1(t) \\ \displaystyle = \sum_{\mu=1}^{\infty} O_{2,\mu}(t)\, \Psi_{2,\mu}(0) + \varphi_2(0)\, h_2(t) + \varphi_4(0)\, h_4(t) + \varphi_6(0)\, k_2(t)\,; \end{cases} \tag{5.26}$$

b) to account for: 1) the values $Z_{j,\mu}(0)=1$, $\Psi_{j,\mu}(0)=a_j Z'_{j,\mu}(0)=a_j\,\omega_\mu$, and 2) the boundary conditions (5.8) imposed on the blending functions $\phi_{j+m}(x)$ and their fluxes $\varphi_{j+m}(x)=a_j\,\phi'_{j+m}(x)$, to obtain the following equations wrt $h_{j+2}(t)$ and $k_j(t)$

$$\begin{cases} \displaystyle\sum_{\mu=1}^{\infty} O_{1,\mu}(t) + h_3(t) = \sum_{\mu=1}^{\infty} O_{2,\mu}(t) + h_4(t)\,, \\ \qquad\qquad a_1 k_1(t) = \qquad\qquad a_2\, k_2(t)\,, \end{cases} \qquad t\in[0,T]\,. \tag{5.27}$$

Since the system (5.27) is incomplete (it involves two equations wrt four unknown functions $h_{j+2}(t)$ and $k_j(t)$), we supply it with the following condition

$$\frac{\partial f_1(t,0)}{\partial x} = \frac{\partial f_2(t,0)}{\partial x}\,, \qquad t\in[0,T]\,, \tag{5.28}$$

being the flux continuous differentiability on the dividing segment (the flux (3.22) of the solution (3.21) obtained in Sect. 3 this condition holds). Then:

a) substituting $u_j(t,x)$ (5.19) into the above condition

$$\begin{cases} \displaystyle\sum_{\mu=1}^{\infty} O_{1,\mu}(t)\, \Psi'_{1,\mu}(0) + \psi_1(0)\, h_1(t) + \psi_3(0)\, h_3(t) + \psi_5(0)\, k_1(t) \\ \displaystyle = \sum_{\mu=1}^{\infty} O_{2,\mu}(t)\, \Psi'_{2,\mu}(0) + \psi_2(0)\, h_2(t) + \psi_4(0)\, h_4(t) + \psi_6(0)\, k_2(t)\,; \end{cases} \tag{5.29}$$

b) accounting for: 1) the values $\Psi'_{j,\mu}(0)=-a_j\,\omega_\mu^2$; 2) the boundary conditions imposed on $\psi_{j+m}(x)=\varphi'_{j+m}(x)$, gives one more equation wrt $h_{j+2}(t)$ and $k_j(t)$

$$a_1\sum_{\mu=1}^{\infty}\omega_\mu^2\, O_{1,\mu}(t) + a_1 b_3\, h_3(t) = a_2\sum_{\mu=1}^{\infty}\omega_\mu^2\, O_{2,\mu}(t) + a_2 b_4\, h_4(t)\,, \quad t\in[0,T]\,. \tag{5.30}$$

Now the difference between the number of unknown functions $h_{j+2}(t)$, $k_j(t)$ and the number of equations (5.27), (5.30) is equal to one, so we need one more additional equation to obtain a complete system to find the required functions.

In contrast to the local equations (5.27), (5.30), valid on the dividing segment, we can choose: *a*) the nonlocal total energy rate equation (2.2)

$$\sum_{j=1}^{2}\Big[\Omega_j'(t)+\Pi_j'(t)\Big]=\mathrm{A}(t)\,,\quad t\in[0,T]\,,\tag{5.31}$$

or *b*) the nonlocal total energy equation (2.3)

$$\sum_{j=1}^{2}\Big[\Omega_j(t)+\Pi_j(t)-\Omega_j(0)-\Pi_j(0)\Big]=\int_0^t \mathrm{A}(t)\,,\quad t\in[0,T]\,,\tag{5.32}$$

as the required additional equation: 1) both (5.31), (5.32) are composed of those respective equations (2.4) and (2.5) for the parts of the string; 2) the kinetic and potential energy are presented as

$$\Omega_j(t)=\frac{1}{2}\int_{K_j}\left[\frac{\partial u_j(t,x)}{\partial t}\right]^2\mathrm{d}x\,,\qquad \Pi_j(t)=\frac{a_j}{2}\int_{K_j}\left[\frac{\partial u_j(t,x)}{\partial x}\right]^2\mathrm{d}x\,,\tag{5.33}$$

3) the power of the external forces acting on the ends of the 'string' reads

$$\mathrm{A}(t)=a_2\frac{\partial u_2(t,+1)}{\partial x}\,h_2'(t)-a_1\frac{\partial u_1(t,-1)}{\partial x}\,h_1'(t)\,,\tag{5.34}$$

and 4) the partial derivatives in (5.33), (5.34) are those given by (5.20), (5.21).

Thus, we have obtained the system involving four equations wrt four unknown functions $h_{j+2}(t)$ and $k_j(t)$, three equations of the system are given by (5.27), (5.30), whereas the fourth is one of (5.31), (5.32).

The second equation of (5.27) is a trivial algebraic relation, whereas the first one is a linear integro-differential equation of convolution type, its expanded form (due to the resulting expression (5.18) for $O_{j,\mu}(t)$) reads

$$\left\{\begin{aligned}
&\quad -\underline{\phi}_1(t)*h_1''(t)\quad -\underline{\phi}_3(t)*h_3''(t)\quad -\underline{\phi}_5(t)*k_1''(t)+\overset{*}{\underline{v}}_1(t)+\overset{**}{\underline{v}}_1(t)\\
&+a_1\,\underline{\phi}_1^{\star}(t)*h_1(t)+a_1\,\underline{\phi}_3^{\star}(t)*h_3(t)+a_1\,\underline{\phi}_5^{\star}(t)*k_1(t)+h_3(t)\\
&=-\underline{\phi}_2(t)*h_2''(t)\quad -\underline{\phi}_4(t)*h_4''(t)\quad -\underline{\phi}_6(t)*k_2''(t)+\overset{*}{\underline{v}}_2(t)+\overset{**}{\underline{v}}_2(t)\\
&+a_2\,\underline{\phi}_2^{\star}(t)*h_2(t)+a_2\,\underline{\phi}_4^{\star}(t)*h_4(t)+a_2\,\underline{\phi}_6^{\star}(t)*k_2(t)+h_4(t)\,,
\end{aligned}\right.\tag{5.35}$$

where the once underlined functions of t are determined by the following series

$$\left\{\begin{aligned}
&\overset{*}{\underline{v}}_j(t)=\sum_{\mu=1}^{\infty}\qquad \overset{*}{v}_{j,\mu}\cos\big(\omega_\mu t\big),\qquad \underline{\phi}_{j+\nu}(t)=\sum_{\mu=1}^{\infty}\omega_\mu^{-1}\,\phi_{j+\nu,\mu}\,\sin\big(\omega_\mu t\big)\,,\\
&\overset{**}{\underline{v}}_j(t)=\sum_{\mu=1}^{\infty}\omega_\mu^{-1}\,\overset{**}{v}_{j,\mu}\sin\big(\omega_\mu t\big),\qquad \underline{\phi}_{j+\nu}^{\star}(t)=\sum_{\mu=1}^{\infty}\omega_\mu^{-1}\,\phi_{j+\nu,\mu}^{\star}\,\sin\big(\omega_\mu t\big)\,.
\end{aligned}\right.\tag{5.36}$$

The same is true for the equation (5.30) written in its expanded form as follows

$$\left\{\begin{aligned} & - a_1\,\underline{\underline{\phi}}_1(t) * h_1''(t) - a_1\,\underline{\underline{\phi}}_3(t) * h_3''(t) - a_1\,\underline{\underline{\phi}}_5(t) * k_1''(t) + a_1\,\overset{*}{\underline{\underline{v}}}_1(t) + a_1\,\overset{**}{\underline{\underline{v}}}_1(t) \\ & + a_1^2\,\underline{\underline{\phi}}^\star_1(t) * h_1(t) + a_1^2\,\underline{\underline{\phi}}^\star_3(t) * h_3(t) + a_1^2\,\underline{\underline{\phi}}^\star_5(t) * k_1(t) + \qquad\quad b_3 a_1\,h_3(t) \\ & = \\ & - a_2\,\underline{\underline{\phi}}_2(t) * h_2''(t) - a_2\,\underline{\underline{\phi}}_4(t) * h_4''(t) - a_2\,\underline{\underline{\phi}}_6(t) * k_2''(t) + a_2\,\overset{*}{\underline{\underline{v}}}_2(t) + a_2\,\overset{**}{\underline{\underline{v}}}_2(t) \\ & + a_2^2\,\underline{\underline{\phi}}^\star_2(t) * h_2(t) + a_2^2\,\underline{\underline{\phi}}^\star_4(t) * h_4(t) + a_2^2\,\underline{\underline{\phi}}^\star_6(t) * k_2(t) + \qquad\quad b_4 a_2\,h_4(t)\,, \end{aligned}\right. \tag{5.37}$$

where the twice underlined functions of t are determined by the following series

$$\left\{\begin{aligned} & \overset{*}{\underline{\underline{v}}}_j(t) = \sum_{\mu=1}^{\infty} \omega_\mu^2\,\overset{*}{v}_{j,\mu} \cos\big(\omega_\mu t\big), \qquad \underline{\underline{\phi}}_{j+\nu}(t) = \sum_{\mu=1}^{\infty} \omega_\mu\;\phi_{j+\nu,\mu}\,\sin\big(\omega_\mu t\big)\,, \\ & \overset{**}{\underline{\underline{v}}}_j(t) = \sum_{\mu=1}^{\infty} \omega_\mu\;\overset{**}{v}_{j,\mu} \sin\big(\omega_\mu t\big), \qquad \underline{\underline{\phi}}^\star_{j+\nu}(t) = \sum_{\mu=1}^{\infty} \omega_\mu\;\phi^\star_{j+\nu,\mu}\,\sin\big(\omega_\mu t\big)\,. \end{aligned}\right. \tag{5.38}$$

5.4. Finding the image functions $H_{j+2}(\tau)$

Applying the Laplace transformation (4.1) to the linear integro-differential equations (5.35) – (5.38) wrt the origins $h_{j+2}(t)$ and $k_j(t)$ yields to the following linear algebraic non-homogeneous equations wrt the images $H_{j+2}(\tau)$ and $K_j(\tau)$

$$\left\{\begin{aligned} & \Big[1 - \hat{\underline{\Phi}}_4(\tau)\Big]\,H_4(\tau) - \quad \Big[1 - \hat{\underline{\Phi}}_3(\tau)\Big]\,H_3(\tau) = \quad \underline{D}_2(\tau) - \quad \underline{D}_1(\tau)\,, \\ a_2 & \Big[b_4 - \hat{\underline{\underline{\Phi}}}_4(\tau)\Big]\,H_4(\tau) - a_1\Big[b_3 - \hat{\underline{\underline{\Phi}}}_3(\tau)\Big]\,H_3(\tau) = a_2\,\underline{\underline{D}}_2(\tau) - a_1\underline{\underline{D}}_1(\tau)\,, \end{aligned}\right. \tag{5.39}$$

where the right hand sides are calculated using the following generic functions

$$\begin{aligned} D_j(\tau) = & - V_j(\tau) + \hat{\Phi}_{j+0}(\tau)\,H_j(\tau) - \Phi_{j+0}(\tau)\,\big[h_{j+0}(0)\,\tau + h'_{j+0}(0)\big] \\ & \qquad\qquad\qquad\qquad - \Phi_{j+2}(\tau)\,\big[h_{j+2}(0)\,\tau + h'_{j+2}(0)\big] \\ & \quad + \hat{\Phi}_{j+4}(\tau)\,K_j(\tau) - \Phi_{j+4}(\tau)\,\big[k_j \quad (0)\,\tau + k'_j \quad (0)\big]\,. \end{aligned} \tag{5.40}$$

For obtaining the once $\underline{D}_j(\tau)$ and twice $\underline{\underline{D}}_j(\tau)$ underlined functions on the rhs of (5.39) one should replace all the functions of τ in (5.40) with the corresponding once and twice underlined functions of τ, determined by the following series

$$\begin{aligned} & \underline{\Phi}_{j+m}(\tau) = \sum_{\mu=1}^{\infty} \frac{\phi_{j+m,\mu}}{\tau^2 + \omega_\mu^2}\,, \qquad \underline{\underline{\Phi}}_{j+m}(\tau) = \sum_{\mu=1}^{\infty} \omega_\mu^2\,\frac{\phi_{j+m,\mu}}{\tau^2 + \omega_\mu^2}\,, \\ & \underline{\Phi}^\star_{j+m}(\tau) = \sum_{\mu=1}^{\infty} \frac{\phi^\star_{j+m,\mu}}{\tau^2 + \omega_\mu^2}\,, \qquad \underline{\underline{\Phi}}^\star_{j+m}(\tau) = \sum_{\mu=1}^{\infty} \omega_\mu^2\,\frac{\phi^\star_{j+m,\mu}}{\tau^2 + \omega_\mu^2}\,, \end{aligned} \tag{5.41}$$

$$\begin{aligned}
\underline{V}_j(\tau) &= \sum_{\mu=1}^{\infty} \frac{\overset{*}{v}_{j,\mu}\,\tau + \overset{**}{v}_{j,\mu}}{\tau^2 + \omega_\mu^2}\,, & \hat{\underline{\Phi}}_{j+m}(\tau) &= \tau^2\,\underline{\Phi}_{j+m}(\tau) - a_j\,\underline{\Phi}^{\star}_{j+m}(\tau)\,,\\
\underline{\underline{V}}_j(\tau) &= \sum_{\mu=1}^{\infty} \omega_\mu^2\,\frac{\overset{*}{v}_{j,\mu}\,\tau + \overset{**}{v}_{j,\mu}}{\tau^2 + \omega_\mu^2}\,, & \hat{\underline{\underline{\Phi}}}_{j+m}(\tau) &= \tau^2\,\underline{\underline{\Phi}}_{j+m}(\tau) - a_j\,\underline{\underline{\Phi}}^{\star}_{j+m}(\tau)\,.
\end{aligned} \tag{5.42}$$

Note that: a) the images $K_j(\tau)$ are placed on the right hand side of (5.39), as if the former images were known, to solve (5.39) wrt the images $H_{j+2}(\tau)$ in an iterative manner; b) the second equation of (5.27) is not used explicitly to retain (5.39) in symmetric form wrt $K_j(\tau)$.

Invoking the Cramer rule yields to the unique solution to (5.39) as follows

$$H_3(\tau) = \frac{\Delta_3(\tau)}{\Delta_0(\tau)}\,, \qquad H_4(\tau) = \frac{\Delta_4(\tau)}{\Delta_0(\tau)}\,, \tag{5.43}$$

where the determinants are

$$\Delta_0(\tau) = \begin{vmatrix} \Big[1 - \hat{\underline{\Phi}}_3(\tau)\Big] & \Big[1 - \hat{\underline{\Phi}}_4(\tau)\Big] \\ a_1\Big[b_3 - \hat{\underline{\underline{\Phi}}}_3(\tau)\Big] & a_2\Big[b_4 - \hat{\underline{\underline{\Phi}}}_4(\tau)\Big] \end{vmatrix}, \tag{5.44}$$

$$\Delta_3(\tau) = \begin{vmatrix} \Big[1 - \hat{\underline{\Phi}}_4(\tau)\Big] & \Big[\quad \underline{D}_2(\tau) - \quad \underline{D}_1(\tau)\Big] \\ a_2\Big[b_4 - \hat{\underline{\underline{\Phi}}}_4(\tau)\Big] & \Big[a_2\,\underline{\underline{D}}_2(\tau) - a_1\underline{\underline{D}}_1(\tau)\Big] \end{vmatrix}, \tag{5.45}$$

$$\Delta_4(\tau) = \begin{vmatrix} \Big[1 - \hat{\underline{\Phi}}_3(\tau)\Big] & \Big[\quad \underline{D}_2(\tau) - \quad \underline{D}_1(\tau)\Big] \\ a_1\Big[b_3 - \hat{\underline{\underline{\Phi}}}_3(\tau)\Big] & \Big[a_2\,\underline{\underline{D}}_2(\tau) - a_1\underline{\underline{D}}_1(\tau)\Big] \end{vmatrix}. \tag{5.46}$$

We present the extended expression for the determinant of the system (5.39)

$$\begin{aligned}
\Delta_0(\tau) = a_1\Big(&b_3 - \hat{\underline{\underline{\Phi}}}_3(\tau) - b_1\,\hat{\underline{\Phi}}_4(\tau) + \hat{\underline{\underline{\Phi}}}_3(\tau)\,\hat{\underline{\Phi}}_4(\tau)\Big)\\
- a_2\Big(&b_4 - \hat{\underline{\underline{\Phi}}}_4(\tau) - b_2\,\hat{\underline{\Phi}}_3(\tau) + \hat{\underline{\Phi}}_3(\tau)\,\hat{\underline{\underline{\Phi}}}_4(\tau)\Big)\,,
\end{aligned} \tag{5.47}$$

since this expression is important both for: a) solving the system and b) choosing the inversion method for the images $H_{j+2}(\tau)$.

5.5. Finding the functions $k_j(t)$

Substituting the known expressions (5.20), (5.21) into (5.33) yields to the following expressions for the kinetic and potential energy for both parts of the string

$$\begin{aligned}
2\,\Omega_j(t) &= \Omega_{j,0}\,k_j'^{2}(t) + 2\,\Omega_{j,1}(t)\,k_j'(t) + \Omega_{j,2}(t)\,,\\
a_j^{-1} 2\,\Pi_j(t) &= \Pi_{j,0}\,k_j^{2}\,(t) + 2\,\Pi_{j,1}(t)\,k_j(t) + \Pi_{j,2}(t)\,,
\end{aligned} \tag{5.48}$$

where the leading coefficients are constant: $\Omega_{j,0} = \|\phi_{j+4}\|_j^2$, $\Pi_{j,0} = \|\phi'_{j+4}\|_j^2$, and the other coefficients read

$$\begin{aligned} \Omega_{j,1}(t) &= \sum_{\mu=1}^{\infty} \left(Z_{j,\mu}, \phi_{j+4}\right)_j O'_{j,\mu}(t) + \sum_{\nu=0}^{2} \left(\phi_{j+\nu}, \phi_{j+4}\right)_j h'_{j+\nu}(t)\,, \\ \Pi_{j,1}(t) &= \sum_{\mu=1}^{\infty} \left(Z'_{j,\mu}, \phi'_{j+4}\right)_j O_{j,\mu}(t) + \sum_{\nu=0}^{2} \left(\phi'_{j+\nu}, \phi'_{j+4}\right)_j h_{j+\nu}(t)\,, \end{aligned} \tag{5.49}$$

$$\begin{aligned} \Omega_{j,2}(t) &= \sum_{\mu=1}^{\infty} \|Z_{j,\mu}\|_j^2\, O'^2_{j,\mu}(t) + 2\sum_{\nu=0}^{2}\left(\sum_{\mu=1}^{\infty} \left(Z_{j,\mu}, \phi_{j+\nu}\right)_j O'_{j,\mu}(t)\right) h'_{j+\nu}(t) \\ &\quad + \sum_{\nu=0}^{2} \|\phi_{j+\nu}\|_j^2\, h'^2_{j+\nu}(t) + 2\left(\phi_j, \phi_{j+2}\right)_j h'_j(t)\, h'_{j+2}(t)\,, \\ \Pi_{j,2}(t) &= \sum_{\mu=1}^{\infty} \|Z'_{j,\mu}\|_j^2\, O^2_{j,\mu}(t) + 2\sum_{\nu=0}^{2}\left(\sum_{\mu=1}^{\infty} \left(Z'_{j,\mu}, \phi'_{j+\nu}\right)_j O_{j,\mu}(t)\right) h_{j+\nu}(t) \\ &\quad + \sum_{\nu=0}^{2} \|\phi'^2_{j+\nu}\|_j^2\, h^2_{j+\nu}(t) + 2\left(\phi'_j, \phi'_{j+2}\right)_j h_j(t)\, h_{j+2}(t)\,, \end{aligned} \tag{5.50}$$

where the inner products in $\mathscr{L}_{2,j}$ are defined in (5.5); whereas substituting(5.21) into (5.34) is performed straightforwardly and is not presented here.

Note that: *a*) the kinetic and potential energy (5.48) are presented as dependent on $k'_j(t)$ and $k_j(t)$, whereas the functions $h_{j+2}(t)$ and their derivatives are 'hidden' in the expressions of the coefficients (5.49), (5.50), as if $h_{j+2}(t)$ were known, to solve the total energy equation (5.32) or the total energy rate equation (5.31) wrt $k_j(t)$ in an iterative manner; *b*) the second equation of (5.27) is not used explicitly to retain both energy equations, (5.32) and (5.31), in symmetric form wrt $k_j(t)$ and their derivatives.

The total energy equation (5.32), (5.48) – (5.50) is a nonlinear second order integro-differential equation: *a*) the second order derivatives of the required functions $h_j(t)$, $k_j(t)$ are involved in $O_{j,\mu}(t)$ (5.18) and $O'_{j,\mu}(t)$ (5.24) through the convolution terms $g_{j,\mu}(t) * \sin\left(\omega_\mu t\right)$ and $g_{j,\mu}(t) * \cos\left(\omega_\mu t\right)$ respectively, where $g_{j,\mu}(t)$ are given in (5.14); and *b*) nonlinearity stems from the products and squares of the first order derivatives of the required functions outside the convolution terms.

The total energy rate equation (5.31), (5.48) – (5.50) is a nonlinear second order integro-differential equation as well as (5.32), but, in contrast to the latter, it: *a*) involves second order derivatives of the required functions $h_j(t)$, $k_j(t)$ outside the convolution terms; *b*) is linear wrt the former derivatives.

6. Conclusions

We have considered three approaches to solve the IBVPO (the original IBVP) for the composite string with piece-wise constant elastic properties based on:

a) SV applied directly to the IBVPO, not followed by matching, since the former is build-in into SV;

b) LT applied to the associated IBVPs posed for both parts of the string with constant properties, then solving the transformed IBVPs and matching the solutions to the transformed IBVPs;

c) SV applied to the associated IBVPs and followed by matching the solutions to the IBVPs, matching involves applying LT to three matching conditions and solving an ordinary differential equation for one matching condition.

In cases *a*) and *b*) matching needs two local conditions, being continuity of the solution and its flux; whereas in case *c*) matching needs two more conditions, one of which is local, being continuous differentiability of the flux, and the other is non-local, being the energy equation. Both cases *b*) and *c*) need applying the procedures of the inverse Laplace transformation, being sometimes quite sophisticated. Therefore, final comparing the cases *b*) and *c*) will be possible after completing the procedures of the inversion and will be presented in the next publication on the subject.

References

1. V. L. Borsch, P. I. Kogut, G. Leugering, *On an initial boundary-value problem for 1D hyperbolic equation with interior degeneracy: series solutions with the continuously differentiable fluxes*, Journal of Optimization, Differential Equations, and their Applications (JODEA), **28** (1) (2020), 1 – 42.
2. V. L. Borsch, P. I. Kogut, *The exact bounded solution to an initial boundary value problem for 1D hyperbolic equation with interior degeneracy. I. Separation of Variables*, Journal of Optimization, Differential Equations and Their Applications (JODEA), **28** (2) (2020), 2 – 20.
3. V. L. Borsch, P. I. Kogut, *Solutions to a simplified initial boundary value problem for 1D hyperbolic equation with interior degeneracy*, Journal of Optimization, Differential Equations, and their Applications (JODEA), **29** (1) (2021), 1 – 31.
4. V. L. Borsch, P. I. Kogut, *Can a finite degenerate 'string' hear itself? The exact solution to a simplified IBVP*, Journal of Optimization, Differential Equations, and their Applications (JODEA), **30** (1) (2022), 89 – 121.
5. V. L. Borsch, P. I. Kogut, *Can a finite degenerate 'string' hear itself? Numerical Solutions to a Simplified IBVP*, Journal of Optimization, Differential Equations, and their Applications (JODEA), **31** (1) (2023), 95 – 100.
6. B. M. Budak, A. A. Samarskii, A. N. Tikhonov, *A Collection of Problems on Mathematical Physics*, Pergamon, London, 1964.
7. G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation*, Springer, NY, 1974.
8. A. N. Tikhonov, A. A. Samarskii, *Equations of Mathematical Physics*, Dover Publications, Inc., NY, 1963.

Received 08.09.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 115–124, DOI 10.15421/1423013

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

THE ANALYTICAL VIEW OF SOLUTION OF THE FIRST BOUNDARY VALUE PROBLEM FOR THE NONLINEAR EQUATION OF HEAT CONDUCTION WITH DEVIATION OF THE ARGUMENT

Yaroslav M. Drin[*], Iryna I. Drin[†], Svitlana S. Drin[‡], [§]

Abstract. In this article, for the first time, the first boundary value problem for the equation of thermal conductivity with a variable diffusion coefficient and with a nonlinear term, which depends on the sought function with the deviation of the argument, is solved. For such equations, the initial condition is set on a certain interval. Physical and technical reasons for delays can be transport delays, delays in information transmission, delays in decision-making, etc. The most natural are delays when modeling objects in ecology, medicine, population dynamics, etc. Features of the dynamics of vehicles in different environments (water, land, air) can also be taken into account by introducing a delay. Other physical and technical interpretations are also possible, for example, the molecular distribution of thermal energy in various media (solid bodies, liquids, etc.) is modeled by heat conduction equations. The Green's function of the first boundary value problem is constructed for the nonlinear equation of heat conduction with a deviation of the argument, its properties are investigated, and the formula for the solution is established.

Key words: heat nonlinear equation, boundary value problem, Green's function; deviation argument.

2010 Mathematics Subject Classification: 35K61.

Communicated by Prof. A. V. Plotniov

1. Introduction

Thermal conductivity is the molecular distribution of thermal energy in various solids, liquids and gases due to the difference in temperature and due to the fact that the particles are in direct contact with each other. The process of heat conduction was first described by Jean Baptiste Joseph Fourier (1768 - 1830) in 1807 in the work "Equations with partial derivatives for heat conduction in solids". A description of the results of other scientists who studied and developed this theory is presented by T.N. Narasimhan [1]. Based on different criteria, models of heat conduction processes are divided into two groups of models

[*] Department of Mathematical Problems of Management and Cybernetics, Institute of Physical, Technical and Computer Sciences, Yuriy Fedkovych Chernivtsi National University, 2, Kotsyubinsky av., Chernivtsi, 58012, Ukraine, y.drin@chnu.edu.ua

[†] Department of Finance, Accounting and Taxation, Chernivtsi Trade and Economic Institute of the State Trade University, 7, Tsentralna Square, Chernivtsi, 58000, Ukraine, iryna.drin@gmail.com

[‡] Department of Mathematics, National University of Kyiv-Mohyla Academy, 2 Skovorody vul., 04070 Kyiv, Ukraine, svitlana.drin@ukma.edu.ua

[§] Department of Statistics, School of Business, Örebro University, 2 Studentgatan st., 70182 Örebro, Sweden, svitlana.drin@oru.se

using integral and fractional order derivatives. In this paper the solution of the first boundary value problem for the heat conduction equation with the variable diffusion coefficient and deviation of the argument is found.

If the evolution of the concentration of impurities, point defects, and the temperature field is studied, then the corresponding transfer coefficients are not constant values. Non-stationary models of one-dimensional heat conduction are described by the equation of heat conduction [2]. Different methods of solving this problem are described in [3], [4]. Applied aspects of such problems are described in [5], [6].

Processes with spatially dependent transmission coefficients or a desired thermal field are well studied and sufficiently describe processes in heterogeneous and nonlinear media [2], applied problems for modelling and research, whose transmission coefficients depend on time change, are also described here. At the same time, physically adequate modelling of thermal processes often requires their investigation in a semi-limited region [7].

In our paper, for the first time, we consider the first boundary value problem for a non-homogeneous nonlinear equation with a deviation of the argument and a variable diffusion coefficient in a semi-bounded domain, which generalizes the corresponding problem for [9].

2. Statement of the First Boundary Problem

Let $a > 0, h > 0$ be real numbers; $x \in R^+, t \in R^+$, are independent variables; $f, \varphi, \mu, D > 0$ are known continuous functions; $u(x,t)$ is the desired function that describes the evolution of the system defined on the semi-axis $x \in R^+$for all $t \in R^+$. We will study the problem

$$u_t = D(t)u_{xx} + f(x,t,u(x,I_h(t))), x > 0, t > h, \tag{2.1}$$

$$u(x,t)|_{0 \le t \le h} = \varphi(x,t), x \ge 0 \tag{2.2}$$

$$u(0,t) = \mu(t), t \ge h \tag{2.3}$$

which is the first boundary value problem, where the functions $\varphi(x,t) \in C(R^+ \times \{0 \le t \le h\})$ is initial function, $\mu(t) \in C\left(R^+_n\right)$ is boundary function, $R^+_h \equiv \{t; t \ge h\}$, $R^+ \equiv \{x; x \ge 0\})$, $f(x,t,u) \in C\left(R^+ \times R^+_h \times R\right)$ is the inhomogeneity of equation (2.1) is well known. If a smooth solution of the problem (2.1)–(2.3) is sought up to the limit, then the initial and marginal functions must be consistent $\varphi(0,h) = \mu(h)$.

3. The Steps Method

Let $x \in R^+$, then $u(x,I_h(t)) = \varphi(x,t)$, and from (2.1)–(2.3) we get the problem:

$$u_t = D(t)u_{xx} + f(x,t,\varphi(x,t)), \quad x > 0,\ h < t < 2h, \tag{3.1}$$

$$u(x,t)|_{t=h} = \varphi(x,h), x \ge 0, \tag{3.2}$$

$$u(0,t) = \mu(t), t \ge h \tag{3.3}$$

with the conditions of agreed $\varphi(0,h) = \mu(h)$.

We will solve a problem (3.1)–(3.3) in the form of sum of three functions

$$u(x,t) = u_1(x,t) + u_2(x,t) + u_3(x,t), \tag{3.4}$$

where $u_i, 1 \leq i \leq 3$, respectively, take into account the influence only initial condition, the boundary condition and the inhomogeneity of the, that is, they are the solutions of such problems.

Problem 1. Find a function $u_1(x,t)$ that satisfies the conditions

$$\frac{\partial u(x,t)}{\partial t} = D(t)\frac{\partial^2 u(x,t)}{\partial x^2}, x > 0, h < t < 2h, \tag{3.5}$$

$$u(x,h) = \varphi(x,h), x \geq 0, \tag{3.6}$$

$$u(0,t) = 0, h \leq t \leq 2h, \tag{3.7}$$

moreover, $\varphi(0,h) = u(0,h) = 0$ is a condition of agreement.

Problem 2. Find a function $u_2(x,t)$ that satisfies equation (3.5) and conditions

$$u(x,h) = 0, x \geq 0, \tag{3.8}$$

$$u(0,t) = \mu(t), h \leq t \leq 2h, \tag{3.9}$$

moreover, $\mu(h) = 0$ is a condition of agreement.

Problem 3. Find the function $u_3(x,t)$ that satisfies equation (2.1) and conditions (3.7), (3.8), which are agreed.

3.1. Solving problem 1

Let's expand the domain of definition of equation (3.5) and the initial condition to $x \in R, h \leq t \leq 2h$ and solve it by separating of variables method $(u_1(x,t) = X(x)T(t)$. and after rearrangement in (3.5) and separation of variables, we obtain that $T(t) = C(\lambda)e^{-\lambda^2 I_h(t)}$, $I_h(t) = \int_h^t D(\tau)\, d\tau$, $X(x) = e^{i\lambda x}$, where λ is the variable separation parameter. Then the solution is $u_1(x,t,\lambda) = C(\lambda)e^{-\lambda^2 I_h(t)+i\lambda x}$, $\lambda \in R$ and to take into. account all $\lambda \in R$ we create a function

$$u_1(x,t) = \int_{-\infty}^{\infty} C(\lambda)e^{-\lambda^2 I_h(t)+i\lambda x} d\lambda, x \in R, \quad h \leq t \leq 2h,$$

which satisfies condition (3.6). Then we get that

$$\varphi(x,h) = \int_{-\infty}^{\infty} C(\lambda)e^{i\lambda x}\, d\lambda,$$

$$C(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(\xi,h)e^{-i\lambda\xi}\, d\xi, \quad \lambda \in R,$$

$$u_1(x,t) = \int_{-\infty}^{\infty} \frac{1}{2\pi}\left\{\int_{-\infty}^{\infty} e^{-\lambda^2 I_h(t)+i\lambda(x-\xi)} d\lambda\right\} \varphi(\xi,h)\, d\xi.$$

The inner integral calculated

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\lambda^2 I_h(t)+i\lambda(x-\xi)}\, d\lambda = \frac{1}{2\sqrt{\pi I_h(t)}} e^{-\frac{(x-\xi)^2}{4I_h(t)}}$$

is denoted by $G(x-\xi; I_h(t))$ and is the fundamental solution of equation (3.5). Then

$$u_1(x,t) = \int_{-\infty}^{\infty} G(x-\xi; I_h(t))\varphi(\xi,h)\, d\xi, \quad x \in R, \quad h < t < 2h. \tag{3.10}$$

We use formula (3.10) to construct a solution to problem 1. For this, instead of equation (3.5), we consider equation

$$\frac{\partial U(x,t)}{\partial t} = \frac{\partial^2 U(x,t)}{\partial x^2}, \quad x \in R, t > h \tag{3.11}$$

with conditions (3.6), (3.7), extending in condition (3.6) the initial function $\varphi(x,h)$ for $x < 0$ undefined, and we set the condition (3.7) as follows:

$$U(x,h) = \Psi(x,h) = \begin{cases} \varphi(x,h), & x \geq 0, \\ -\varphi(-x,h), & x < 0 \end{cases} \tag{3.12}$$

Then, according to formula (3.10), the solution of problem (3.11), (3.12), (3.7) is

$$U(x,t) = \int_{-\infty}^{\infty} \{G(x-\xi; I_h(t)) - G(x+\xi; I_h(t))\}\varphi(\xi,h)\, d\xi.$$

In the integral where $\xi < 0$, we replaced $\xi = -\xi$. Simplifying the difference of the exponents included in the expression for the function G, we obtain that

$$u_1(x,t) = \frac{1}{\sqrt{\pi I_h(t)}} \int_0^{\infty} \varphi(x,h) e^{-\frac{x^2+\xi^2}{4I_h(t)}} \operatorname{sh} \frac{x\xi}{2I_h(t)}\, d\xi,$$

where $x > 0, h < t < 2h$. Using the method of mathematical induction, we prove that in case $x \geq 0, kh < t < (k+1)h$ the solution to problem 1 takes the form

$$u_1(x,t) = \frac{1}{\sqrt{\pi I_{kh}(t)}} \int_0^{\infty} \varphi(\xi,kh) e^{-\frac{x^2+\xi^2}{4I_{kh}(t)}} \operatorname{sh} \frac{x\xi}{2I_{kh}(t)}\, d\xi. \tag{3.13}$$

Let's mark

$$G_1(x,y,I_{kh}(t)) = \frac{1}{\sqrt{\pi I_{kh}(t)}} e^{-\frac{x^2+y^2}{4I_{kh}(t)}} sh \frac{xy}{2I_{kh}(t)} \tag{3.14}$$

$x \geq 0$, $y > 0$, $kh < t < (k+1)h$, $k \in N$.

Definition 3.1. A function $G_1(x,y,I_{kh}(t))$ is called a Green's function of problem (2.1), (2.2), (2.3) if it satisfies the following conditions:

1. the function $G_1(x, y, I_{kh}(t))$ is continuous on x, y, t, continuously differentiable on t and twice continuously differentiable on x, y when $x > 0, y > 0, kh < t < (k+1)h, k \in \mathrm{N}$, and possibly with the exception in the point $x = y, t = kh$;

2. the function $G_1(x, y, I_{kh}(t))$ by variables x and y satisfies the equation $\frac{\partial G_1}{\partial t} = \frac{D(t)\partial^2 G_1}{\partial x^2}$ everywhere except in the points $x = y, t = kh, k \in \mathrm{N}$;

3. the function $G_1(x, y, I_{kh}(t))$ satisfies the boundary condition $G_1(0, y, I_{kh}(t)) = 0$.

The Green's function satisfying this definition is constructed above and takes the form (3.14)

$$G_1(x, y, I_{kh}(t)) = G_1(y, x, I_{kh}(t)).$$

3.2. Properties of the solution of the problem 1

Given that

$$G_1(x, y, I_{kh}(t)) = \frac{1}{2\sqrt{\pi I_{kh}(t)}} \left\{ e^{-\frac{(x-\xi)^2}{4I_{kh}(t)}} - e^{-\frac{(x+\xi)^2}{4I_{kh}(t)}} \right\}$$

we get from (3.13), when $|\varphi(\xi, kh)| \leq M$,

$$\begin{aligned} |u_1(x,t)|) &\leq M \frac{1}{2\sqrt{\pi I_{kh}(t)}} \left\{ \int_0^\infty e^{-\frac{(x-\xi)^2}{4I_{kh}(t)}} \, d\xi - \int_0^\infty e^{-\frac{(x+\xi)^2}{4I_{kh}(t)}} \, d\xi \right\} \\ &\equiv M \left\{ I_1 - I_2 \right\}. \end{aligned}$$

In the integral I_1 we will do replacement $\alpha = \frac{\xi - x}{2\sqrt{I_{kh}(t)}}$, and in the integral I_2 $\alpha = \frac{\xi + x}{2\sqrt{I_{kh}(t)}}$. Then

$$I_1 = 2 \int_{-z}^\infty e^{-\alpha^2} \, d\alpha, I_2 = 2 \int_z^\infty e^{-\alpha^2} \, d\alpha$$

where $z = \frac{x}{2\sqrt{I_{kh}(t)}}$ and we get an estimate

$$|u_1(x,t)| \leq M \operatorname{erf} \left(-\frac{x}{2\sqrt{I_{kh}(t)}} \right), \tag{3.15}$$

$x > 0,\ kh < t < (k+1)h$

So, the following theorem is proved.

Theorem 3.1. *If there exists a number $M > 0$ such that the initial function $\varphi(x, kh)$ is bounded when $x > 0,\ h > 0,\ k \in \mathrm{N},\ |\varphi(x, kh)| \leq M$, then the function $u_1(x,t)$ (3.13) when $x > 0,\ kh < t < (k+1)h$ is also bounded and the estimate (3.15) is true for it.*

If $\varphi(\xi, kh) = \varphi_0$, where φ_0 is a number, then

$$u_1(x,t) = \varphi_0 \operatorname{erf}\left(\frac{x}{2\sqrt{I_{kh}(t)}}\right)$$

$x > 0, kh < t < (k+1)h, \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left\{-\xi^2\right\} d\xi$ is the error function.

By direct verification, it is possible to make sure that the Green's function (3.14) satisfies the homogeneous heat conduction equation (item 2 of the definition). When formally differentiating the function (3.13) under the sign of the integral, we obtain expressions

$$\frac{1}{(I_{kh}(t))^r} \int_0^{\infty} \varphi(\xi, kh)|x \pm y|^m e^{-\frac{(x\pm y)^2}{4I_{kh}(t)}} \, dy,$$

$x > 0$, $kh < t < (k + \underline{1})h$, where integrable functions are majored by an expression of the form $M|\xi|^m e^{-\xi^2}$ that is integrable on the entire numerical axis. This ensures uniform convergence of the integrals obtained after differentiation under the sign of the integral. Then the Poisson integral (3.13) is a continuous function, differentiable of arbitrary order with respect to x and t when $x > 0$, $kh < t < (k+1)h$, $k \in \mathrm{N}$, bounded with a bounded initial function, satisfying the homogeneous heat conduction equation (3.5), since the Green's function (3.14) satisfies equation (3.5). The implementation of the initial condition (3.6) and the boundary condition (3.7) is carried out. Let us prove the uniqueness theorem of the solution to problem 1.

Theorem 3.2. *Let there be a number $M > 0$ such that in the domain $x \geq 0$ and $kh \leq t \leq (k+1)h, k \in N$ the functions $u_1(x,t)$ and $u_2(x,t)$ are bounded, that is $|u_i(x,t)| < M, \quad i = 1,2,$ satisfy the equation* (3.5) *and condition*

$$u_1(x, kh) = u_2(x, kh), \quad x \geq 0, k \in N,$$

then

$$u_1(x,t) = u_2(x,t), \quad x \geq 0, kh \leq t \leq (k+1)h$$

Consider the function

$$v(x,t) = u_1(x,t) - u_2(x,t),$$

which is continuous, equation (3.5), bounded by

$$|v(x,t)| \leq |u_1(x,t)| + |u_2(x,t)| < 2M,$$
$$x \geq 0, \quad kh \leq t \leq (k+1)h, \quad v(x,kh) = 0.$$

Consider the domain $0 \leq x \leq L, kh \leq t \leq (k+1)h$, where L is a real number and a function

$$V(x,t) = \frac{4M}{L^2}\left(\frac{x^2}{2} + (I_{kh}(t))\right)$$

for which

$$\frac{\partial V}{\partial t} = \frac{4M}{L^2}, \frac{\partial V}{\partial x} = \frac{4Mx}{L^2}, \frac{\partial^2 V}{\partial x^2} = \frac{4M}{L^2}$$

and which satisfies the thermal conductivity equation (3.5), as well as

$$\begin{aligned} &V(x, kh) \geq v(x, kh) = 0, \\ &V(\pm L, t) \geq 2M \geq |v(\pm L, t)| \end{aligned} \tag{3.16}$$

For each limited region $0 \leq x \leq L$, $kh \leq t \leq (k+1)h$, $k \in N$, the principle of the maximum value is true. The functions $\underline{u} = -V(x, t)$, $u = v(x, t)$, $\bar{u} = V(x, t)$, taking into account (3.17), we obtain that

$$-\frac{4M}{L^2}\left(\frac{x^2}{2} + I_{kh}(t)\right) \leq v(x, t) \leq \frac{4M}{L^2}\left(\frac{x^2}{2} + I_{kh}(t)\right). \tag{3.17}$$

We fix (x, t) and use the fact that L is arbitrary and can be increased indefinitely. Passing to the limit at $L \to \infty$, we obtain that $v(x, t) \equiv 0$ for $x \geq 0, kh \leq t \leq (k+1)h$. Theorem 2 is proved.

Therefore, the following theorem is true.

Theorem 3.3. *If* $|\varphi(x, h)| \leq M$, $\quad x \geq 0$, $\quad M > 0\ h > 0$, *then the solution of problem* (3.5), (3.6), (3.7) *exists, is unique and is determined by formula* (3.13).

3.3. Solving the problems 2 and 3

It is necessary to solve equation (3.5) when the zero initial condition (3.8) and the general boundary condition (3.9) are met. First, let's solve the auxiliary problem of cooling a heated rod, at the boundary of which a constant zero temperature is maintained. Then, for equation (3.5), the Cauchy condition and the boundary condition are given as follows:

$$V_1(x, t_0) = T, v_1(0, t) = 0, x > 0, t > h.$$

Then, according to formula (3.13), we get that

$$\bar{v} = T \operatorname{erf}\left(\frac{x}{2\sqrt{I_{t_0}(t)}}\right), x \geq 0, t > t_0, \tag{3.18}$$

Let $\mu(t) = \mu_0 \equiv$ *const* in condition (3.9). Then, according to (3.18), the function

$$\bar{v} = \mu \operatorname{erf}\left(\frac{x}{2\sqrt{I_{t_0}(t)}}\right), x \geq 0, t > t_0,$$

is a solution of problem (3.5), (3.8), (3.9). Then the function

$$v(x, t) = \mu_0 - \bar{v}(x, t) = \mu_0\left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{I_{t_0}(t)}}\right)\right], \quad x > 0, t > 0. \tag{3.19}$$

We denote the expression in parentheses of formula (3.19) by $U\left(x, I_{t_0}(t)\right)$, which makes sense when $t > t_0$. If for $t < t_0$ the value of this function is extended by zero, then this definition is consistent with the zero value of the function at $t = t_0$. The limit value of this function at $x = 0$ is a step function equal to zero at $t < t_0$ and equal to 1 at $t > t_0$. The constructed function is often found in applications and is an auxiliary link in constructing the solution to problem 2.

The second auxiliary task is to find a solution of the equation (3.5) under the following conditions:

$$v\left(x, t_0\right) = 0, x \geq 0,$$
$$v(0,t) \equiv \mu(t) = \left\{ \begin{array}{c} \mu_0, t_0 < t < t_1, \\ 0, t > t_1. \end{array} \right\}.$$

|

It is directly verified that $V(x,t) = \mu_0\left[U\left(x, I_{t_0}(t)\right) - U\left(x, I_{t_1}(t)\right)\right]$, $x \geq 0, t > t_0$. If

$$\mu(t) = \begin{cases} \mu_0, t_0 < t \leq t_1, \\ \mu_1, t_1 < t \leq t_2, \\ \cdots\cdots\cdots & , \\ \mu_{n-1}, t_{n-2} < t \leq t_{n-1}, \\ \mu_{n-1}, t_{n-1} < t \leq t_n, \end{cases}$$

and then the solution of the corresponding problem can be written in the form

$$u(x,t) = \sum_{i=0}^{n-2} \mu_i \left[U\left(x, I_{t_i}(t)\right) - U\left(x, I_{t_n}(t)\right)\right] + \mu_{n-1} U\left(x, I_{t_{n-1}}(t)\right).$$

Using the theorem on finite increments, we get

$$u(x,t) = \sum_{i=0}^{n-2} \mu_i \frac{\partial}{\partial t} U(x, I_\tau(t))\Bigg|_{\tau=\tau_i} + \mu_{n-1} U\left(x, I_{t_n}(t)\right), \tag{3.20}$$

where $x \geq 0$, $t_i \leq \tau_i \leq t_{i+1}$.

The approximate solution of problem 2 can be obtained by formula (3.20), if replace the function $\mu(t)$ with a piecewise-constant function.

Heading to the limit when the interval of constancy of the auxiliary function decreases, we obtain that the limit of the sum (3.20) will take the form

$$\int_0^t \frac{\partial U}{\partial t}(x, I_\tau(t)) \mu(\tau)\, d\tau$$

because when $x \geq 0$, we have

$$\lim_{t - t_{n-1} \to 0} \mu_{n-1} U\left(x, I_{t_{n-1}}(t)\right) = 0.$$

If we consider

$$\frac{\partial U}{\partial t}(x,t) = -2\frac{\partial G}{\partial x}(x,0,t) = 2\frac{\partial G}{\partial \xi}\bigg|_{\xi=0}$$

then we will get the final result

$$u_2(x,t) = \frac{1}{2\sqrt{\pi}}\int_{kh}^{t} \frac{x}{[I_\tau(t)]^{3/2}} \times \exp\left\{-\frac{x^2}{4I_\tau(t)}\right\}\mu(\tau)\,d\tau, \tag{3.21}$$

$x > 0$, $kh \le t \le (k+1)h$.

The solution of problem 3 using the Green's function (3.14) can be written in the form of a Poisson integral

$$u_3(t,x) = \int_{kh}^{t} d\tau \int_{0}^{\infty} f(y,\tau)G_1(x,y,I_{kh}(t))\,dy \tag{3.22}$$

$x > 0$, $kh \le t \le (k+1)h$, $k \in \mathrm{N}$, for the existence of which the function $f(x,t)$ must be such that the improper integral in formula (3.22) coincides.

So, the following theorem is proved.

Theorem 3.4. *The solution of problem* (3.5), (3.8), (3.9) *is determined by formula* (3.22). *The solution of problem* (3.1), (3.2), (3.3) *is determined by formula* (3.4), *where the terms* u_1, u_2, u_3 *are the solutions of problems 1, 2 and 3 respectively.*

The first, second and third initial-boundary problems for the heat conduction equation with inversion of the argument and $D(t) \equiv a^2 > 0$, constant are considered in [9], [10], [11].

References

1. T.N. Narasimhan, *Fourier's heat transfer equations: History, influence and connections*, *Review ws of Geophysics*, **37** (1) (1999), 151–172.
2. L.C. Evans, *Partial Differential Equations*, *Grad. Stud. Math., Amer. Math. Soc.*, Providence, RI, **19** (2010), 44-65.
3. Ian N. Sneddon, *Elements of Partial Differential Equations*, *Dover Books of Mathematics*, 2006, 352 p.
4. A.B. Tayler, *Mathematical Models in Applied Mechanics*, *Oxford*, 2008, 288 p.
5. I. Kaur, Y. Mishin, W. Gust, *Fundamentals of Grain and Interphase Boundary Diffusion*, *John Wiley & Sons Ltd*, Chichester, 1995, 512 p.
6. J. Biazar, Z. Ayati, *An Approximation to the Solution of Parabolic Equation by Adomian Decomposition Method and Comparing the Result with Crank-Nicolson Method*, *International Mathematical Journal*, **39** (2006), 1925–1933.
7. Cole Kevin D., Beck James V., Haji-Sheikh A., Litkouhi Bahan, *Heat conduction using Green's functions, Series in Computational and Physical Processes in Mechanics and Thermal Sciences (2nd ed.)*, Boca Raton, FL: CRC Press, 2011.
8. R.K.M. Thambynayagam, *The Diffusion Handbook: Applied Solutions for Engineers*, *McGraw-Hill Professional*, 2011.

9. Y.M. Drin, I.I. Drin, S.S. Drin, Y.P. Stetsko, *The first boundary value problem for the nonlinear equation of heat conduction with deviation of the argument*, in Proc. The 12th International Conference on Electronics, Communications and Computing, 20-21 October, 2022, Chisinau, Republic of Moldova.

10. Y.M. Drin, I.I. Drin, R.Y. Drin, *The analytical view of solution of the second boundary value problem for the nonlinear equation of heat conduction with deviation of the argument*, in Proc. The Eleven International Conference on "Informatics and computer technique problems" (PICT - 2022), 10-13 October 2022, Chernivtsi, Ukraine, 11–18.

11. Ya. Drin, I. Drin, R. Drin, *The third initial-boundary value problem for the nonlinear equation of heat conduction with deviation of the argument*, 2022 International Conference on Innovative Solutions in Software Engineering (ICISSE), Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine, Nov. 29-30, 2022, 282–285

Received 10.08.2023

JOURNAL OF OPTIMIZATION, DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS (JODEA)
Volume **31**, Issue **2**, December **2023**, pp. 125–173, DOI 10.15421/1423014

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

QUALITATIVE ANALYSIS OF AN OPTIMAL SPARSE CONTROL PROBLEM FOR QUASI-LINEAR PARABOLIC EQUATION WITH VARIABLE ORDER OF NONLINEARITY

Ciro D'Apice[*], Peter Kogut[†], Rosanna Manzo[‡]

Abstract. In this work, we study a sparse optimal control problem involving a quasi-linear parabolic equation with variable order of nonlinearity as a state equation and with a pointwise control constraints. We show that in the case if the cost functional contains the terminal term of the tracking type, the proposed optimal control problem is ill-posed, in general. In view of this, we provide a sufficiently mild relaxation of the proposed problem and establish the existence of optimal solutions for the relaxed version. Using the compensated compactness technique and the consept of variational convergence of minimization problems, we study the attainability of optimal pairs to the relaxed problem by optimal solutions of the special approximating problems. We also discuss the optimality conditions for approximating problems and provide their substantiation.

Key words: Weak solution, parabolic equation, variable order of nonlinearity, noncoercive problem, compensated compactness technique..

2010 Mathematics Subject Classification: 35K20, 49J20, 35D30, 35K92.

Communicated by Prof. O. M. Stanzhytskyi

1. Introduction

1.1. Motivation

Over the past few decades, the role of optical satellite multi-band images in remote sensing of the Earth surface has been increasingly contributing to many agricultural monitoring services. In spite of the fact that optical images have a high resolution and are easily captured by low-cost cameras, the real-life satellite images frequently suffer from different types of noise, blur, and other atmosphere artifacts , which greatly reduce the effective information is such images. Hence, removing noise is a crucial step for image quality improvement in image processing task. In the last decades, models based on partial differential equations (PDEs) have been widely used in the image de-noising problems. Since 1990s, originated from the pioneering work of Perona and Malikl [51], many different

[*] Dipartimento di Scienze Aziendali - Management and Innovation Systems, University of Salerno, 132, Via Giovanni Paolo II, Fisciano, SA, Italy (cdapice@unisa.it)

[†] Department of Mathematical Analysis and Optimization, Oles Honchar Dnipro National University, Gagarin av., 72, 49010 Dnipro, Ukraine, EOS Data Analytics Ukraine, Gagarin av., 103a, Dnipro, Ukraine (p.kogut@i.ua, peter.kogut@eosda.com)

[‡] Dipartimento di Scienze Politiche e della Comunicazione, University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy (rmanzo@unisa.it)

models have been proposed to separate noise from the noisy images. Without being too exhaustive, we refer to [1, 14, 16–18, 21, 22, 47, 54] for a wide variety of different variational models related to the image denoising problems.

However, since the noise, edges, and texture are high-frequency components, it is difficult to distinguish them in the process of denoising, and, as a result, the denoised images could inevitably lose some details. This problems becomes much more difficult if the original image is contaminated by an impulse noise. In view of this, we mainly focus on those approaches where the denoising problem can be stated in the form of some optimal control problem with special class of controls simulating the presence of both the white Gaussian additive noise n and the noise v with a strong impulsive nature which the Gaussian model fails to describe (see, for instance, [2, 13, 48]). In this case the observed image can be represented as $f = u + v + n$, and the question is how to separate a true image u eliminating both Gaussian noise n and impulse noise v from f.

1.2. Statement of the problem

Inspired in the work [2], the first goal of this paper is to analyze the consistency and well-posedness of the following optimal control problem (OCP):

$$\text{Minimize } J(v,u) = \|v\|^2_{L^2(0,T;L^1(\Omega))} + \frac{\mu}{2}\int_\Omega |u(T) - f_0|^2\,dx \tag{1.1}$$

subject to the following constraints

$$\frac{\partial u}{\partial t} - \operatorname{div}\left(|R_\eta \nabla u|^{p_u(t,x)-2} R_\eta \nabla u\right) = \kappa\,(f - u - v) \ \text{ in } \ Q_T := (0,T)\times\Omega, \tag{1.2}$$

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

$$u(0,\cdot) = f_0(\cdot) \quad \text{in} \ \ \Omega, \tag{1.4}$$

$$v_a(x) \leqslant v(t,x) \leqslant v_b(x), \quad \text{a.e. in} \ \ Q_T. \tag{1.5}$$

Here, $\Omega \subset \mathbb{R}^2$ is a bounded simple-connected open set with a sufficiently smooth boundary $\partial\Omega$, $T > 0$ is a positive value, $\kappa \in \mathbb{R}$ is a given positive parameter, $f \in L^2(\Omega)$, $f_0 \in L^2(\Omega)$ and $v_a, v_b \in L^2(\Omega)$, $v_a(x) \leqslant v_b(x)$ a.e. in Ω, are given distributions,

$$\|v\|^2_{L^2(0,T;L^1(\Omega))} = \int_0^T \left(\int_\Omega |v|\,dx\right)^2 dx \tag{1.6}$$

is the so-called directional sparsity term, $R_\eta : L^1(\Omega;\mathbb{R}^2) \to L^1(\Omega;\mathbb{R}^2)$ is a linear bounded operator, and the exponent $p_u : Q_T \to \mathbb{R}$ is defined by the rule

$$p_u(t,x) := 1 + g\left(\frac{1}{h}\int_{t-h}^t \left|\left(\nabla G_\sigma * \widetilde{u}(\tau,\cdot)\right)(x)\right|\,d\tau\right), \quad \forall\,(t,x) \in Q_T, \tag{1.7}$$

where $g : [0,\infty) \to (0,1]$ is a continuous non-increasing function such that $g(0) =$

1 and $g(s) > 0$ for all $s > 0$ with $\lim\limits_{s\to\infty} g(s) = 0$,

$$|g(s) - g(y)| \leqslant C_g|s-y|, \quad \forall s, y \in [0,\infty) \text{ with some constant } C_g > 0, \tag{1.8}$$

$$G_\sigma(x) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^2} \exp\left(-\frac{|x|^2}{2\sigma^2}\right), \quad \sigma > 0, \tag{1.9}$$

$$\left(G_\sigma * \widetilde{u}(t,\cdot)\right)(x) = \int_{\mathbb{R}^2} G_\sigma(x-y)\widetilde{u}(t,y)\,dy, \tag{1.10}$$

$\widetilde{u}$ denotes zero extension of u from Q_T to $\mathbb{R}\times\mathbb{R}^2$, and $h > 0$ and $\sigma > 0$ are given small positive values.

In particular, the function g in (1.7) can be defined in the form of the Cauchy law

$$\begin{array}{c} g(s) = \delta + \frac{a^2(1-\delta)}{a^2+s^2}, \quad \forall\, s \in [0,+\infty) \\ \text{with an appropriate } a > 0 \text{ and } 0 < \delta \ll 1. \end{array} \tag{1.11}$$

Moreover, it will be shown further that, for each function u with properties $u \in L^1(Q_T) \cap L^\infty(0,T;L^2(\Omega))$, there exists a positive value $\delta > 0$ such that $p_u(t,x) \in [p^-, p^+] \subset (1,2]$ almost everywhere in Q_T with $p^- = 1+\delta$ and $p^+ = 2$.

We can indicate here a few main characteristic features of the addressed OCP (1.1)–(1.5). The first one is a special character of the linear operator R_η. In fact, this operators plays the role of the so-called Directional Total Variation along a given vector field. In practice, having some vector field $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, we determine this operator as follows:

$$R_\eta \nabla u = \left[I - \eta^2\,\theta\otimes\theta\right]\nabla u, \quad \forall\, u \in W^{1,1}(\Omega),$$

where $\eta \in (0,1)$ is a given threshold. So, $R_\eta\nabla u$ can be reduced to $(1-\eta^2)\nabla v$ if the gradient $\nabla u(t,x)$ at this point is co-linear to θ, and to $\nabla u(t,x)$ provided $\nabla u(t,x)$ is orthogonal to θ. In other words, this operator impose some anisotropy effect in the standard diffusivity of u.

The second characteristic point of OCP (1.1)–(1.5) is related to the variable character of the exponent $p = p(t,x)$. As follows from representation (1.7) this characteristic depends not only on (t,x) but also on $u(t,x)$. So, in contrast to the recent paper [19], where the authors study the solvability issues for the nonlinear parabolic equation having nonstandard growth condition with respect to the gradient and with well predefined variable exponent, the function $p_u(t,x)$ in (1.2) is unknown a priori and strictly depends on the current solution of the initial-boundary value problem (IBVP) (1.2)–(1.4). It is worth also to emphasize that we do not assume here that the dependency $u \mapsto p_u$ is local whereas it is the crucial assumption in the most of existing publications (see for instance [9,12]). The next difficulty in the analysis of this IBVP relies that its weak formulation cannot be written as equality in terms of duality in a fixed Banach space (for the details we refer to [23]). In fact, we show that each weak solution to the IBVP (1.2)–(1.4) lives in the corresponding 'personal' functional space, and, in view of our assumptions on the structure of exponent $p_u(t,x)$, the problem (1.2)–(1.4) can admit the weak solutions that may not possess the usual properties of solutions to parabolic equations. In particular, it would be rather questionable assertion that a weak solution to the above is unique, belong to the space $C([0,T];L^2(\Omega))$, and satisfies the standard energy equality.

It is well-known that the variable character of exponent p causes a gap between the monotonicity and coercivity conditions. Because of this gap, the problem (1.1)–(1.4) can be termed an optimal control problem for the quasi-linear parabolic equations with nonstandard growth conditions, and it can be viewed as a generalization of the evolutional version of $p(t,x)$-Laplacian equation

$$\frac{\partial u}{\partial t} = \operatorname{div}\left(|\nabla u|^{p(t,x)-2}\nabla u\right) \tag{1.12}$$

with an exponent that depends only on t and x. During the last decades equation (1.12) was intensively studied by many authors. There is extensive literature devoted to equation (1.12). We limit ourselves by referring here to the following ones [9, 10, 15, 50, 52, 58] which provide an excellent insight to the theory of evolutional $p(t,x)$-Laplacian equations.

Albeit PDEs with variable nonlinearity are rather interesting from the purely mathematical point of view as was mentioned before, their study is often motivated by various applications where the problem (1.2)–(1.4), or some special cases of it, appear in the most natural way [2, 3, 14, 21]. It was recently shown that the model (1.2)–(1.4) naturally appears as the Euler-Lagrange equation in the problem of restoration of cloud contaminated satellite optical images [27]. Moreover, the above mentioned problem can be considered as a model for the deblurring and denoising of multi-spectral satellite images. In particular, this model has been proposed in [28, 43] in order to avoid the blurring of edges and other localization problems presented by linear diffusion models in images processing. We also refer to [40], where the authors study some optimal control problems associated with a special case of the model (1.2)–(1.4) and show that, in contrast to the case of the problem (1.2)–(1.4), the proposed in [40] class of optimal control problems is well posed.

It is also worth to notice that the model (1.2)–(1.4) can be considered as a natural generalization of the well-know Perona-Malik model [51]. In spite of the fact that Perona-Malik model reduces the diffusivity of color in places having higher likelihood of being edges, its major defect is that this model is ill-posed and there are no results of existence and its consistency (see [40]). To overcome this problem it has been proposed to modify this model by applying a Gaussian filter on the gradient (we can refer to the pioneering works [7, 20]).

The next characteristic feature of OCP (1.1)–(1.5) is that this control problem is formulated with $L^1(\Omega;L^2(0,T))$ control cost functional (together with some additional pointwise control constraints). Because of this the resulting optimal control may have directional sparsity, i.e., its support is a constant in time and the control v is identically zero on some parts of the domain Ω.

All of this leads us to the followings conclusion: OCP (1.1)–(1.5) is sufficiently challenging and its consistency is an open question. In fact, it will be shown in the next sections that because of the variable character of exponent p and its dependence on t and x, we can lose the continuity of the mapping $t \mapsto \|u(t,\cdot)\|_{L^2(\Omega)}$. Hence, the cost functional (1.1) is not well-defined and, as a result, we can assert that the OCP (1.1)–(1.5) is ill-posed, in general. Because of this, the original OCP requires some relaxation and approximations.

1.3. Organization of the paper

The paper is organized as follows. In Section 2 we give some preliminaries and introduce the main assumptions on the structure of the operator R_η and the variable exponent $p_u(t,x)$. We also give here the main auxiliary results concerning the Orlicz spaces, Sobolev-Orlicz spaces with variable exponent, weighted energy space, and convergence of fluxes to flux. In Section 3 we focus on the solvability issues for IBVP (1.2)–(1.4). With that in mind we follows the indirect approach using the technique of passing to the limit in some special approximation scheme. In this section we show that the IBVP (1.2)–(1.4) admits at least one weak solutions that can be attained by the solutions of more regular Caushy-Neumann problem for quasi-linear parabolic equations. In Section 4 we propose rather mild scheme of relaxation for the original OCP, and show that at each level of relaxation the corresponding OCP is well-posed and admits at least one solution. The questions attainability of the solutions to the relaxed problems are the subject of Section 5. In fact, in this section we introduce the family of OCPs for the special class of parabolic equations

$$\frac{\partial u}{\partial t} - \varepsilon\Delta u - \operatorname{div} A_u^\varepsilon(t,x,\nabla u) + \kappa u = \kappa(f - v) \quad \text{in} \quad Q_T := (0,T)\times\Omega,$$

where the flux $A_u^\varepsilon(t,x,\nabla u)$ we define as follows

$$A_u^\varepsilon(t,x,\nabla u) := \left(|R_\eta\nabla u| + \varepsilon\right)^{p_u(t,x)-2} R_\eta\nabla u.$$

We show that due to this approximation, some optimal solutions to the relaxed OCP can be attained in an appropriate topology by the solutions of the proposed family of OCPs.

The last Section 6 is devoted to the deriving of some optimality conditions for approximating OCPs and their substantiation.

2. Main Assumptions and Preliminaries

Let $\Omega \subset \mathbb{R}^2$ be a bounded connected open set with a sufficiently smooth boundary $\partial\Omega$, and let $T > 0$ be a given value. We suppose that the unit outward normal $\nu = \nu(x)$ is well-defined for a.e. $x \in \partial\Omega$, where a.e. means here with respect to the 1-dimensional Hausdorff measure $\mathcal{H}^1$. We set $Q_T = (0,T)\times\Omega$. For any measurable subset $D \subset \Omega$ we denote by $|D|$ its 2-dimensional Lebesgue measure $\mathcal{L}^2(D)$. We denote its closure by $\overline{D}$ and its boundary by ∂D.

For vectors $\xi \in \mathbb{R}^2$ and $\eta \in \mathbb{R}^2$, $(\xi, \eta) = \xi^t \eta$ denotes the standard vector inner product in $\mathbb{R}^2$, where ^{t} stands for the transpose operator. The norm $|\xi|$ is the Euclidean norm given by $|\xi| = \sqrt{(\xi, \xi)}$. We also make use of the following notation $\operatorname{diam}\Omega = \sup_{x,y\in\Omega} |x - y|$.

2.1. Functional Spaces

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 X and X', respectively.

For given $1 \leqslant p \leqslant +\infty$, the space $L^p(\Omega;\mathbb{R}^2)$ is defined by

$$L^p(\Omega;\mathbb{R}^2) = \left\{ f : \Omega \to \mathbb{R}^2 \ : \ \|f\|_{L^p(\Omega;\mathbb{R}^2)} < +\infty \right\},$$

where $\|f\|_{L^p(\Omega;\mathbb{R}^2)} = \left(\int_\Omega |f(x)|^p\, dx\right)^{1/p}$ for $1 \leqslant p < +\infty$. The inner product of two functions f and g in $L^p(\Omega;\mathbb{R}^2)$ with $p \in [1,\infty)$ is given by

$$(f,g)_{L^p(\Omega;\mathbb{R}^2)} = \int_\Omega (f(x), g(x))\, dx = \int_\Omega \sum_{k=1}^{2} f_k(x) g_k(x)\, dx.$$

We denote by $C_c^\infty(\mathbb{R}^2)$ the locally convex space of all infinitely differentiable functions with compact support in $\mathbb{R}^2$. We recall here some functional spaces that will be used throughout this paper. We define the Banach space $W^{1,p^-}(\Omega)$ with $p^- > 1$ as the closure of $C_c^\infty(\mathbb{R}^2)$ with respect to the norm

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

We denote by $\left(W^{1,p^-}(\Omega)\right)'$ the dual space of $W^{1,p^-}(\Omega)$. Let us remark that in this case the embedding $L^2(\Omega) \hookrightarrow \left(W^{1,p^-}(\Omega)\right)'$ is continuous.

Given a real separable Banach space X, we will denote by $C([0,T];X)$ the space of all continuous functions from $[0,T]$ into X. We recall that a function $u : [0,T] \to X$ is said to be Lebesgue measurable if there exists a sequence $\{u_k\}_{k\in\mathbb{N}}$ of step functions (i.e., $u_k = \sum_{j=1}^{n_k} a_j^k \chi_{A_j^k}$ for a finite number n_k of Borel subsets $A_j^k \subset [0,T]$ and with $a_j^k \in X$) converging to u almost everywhere with respect to the Lebesgue measure in $[0,T]$.

Then for $1 \leqslant p < \infty$, $L^p(0,T;X)$ is the space of all measurable functions $u : [0,T] \to X$ such that

$$\|u\|_{L^p(0,T;X)} = \left(\int_0^T \|u(t)\|_X^p\, dt \right)^{\frac{1}{p}} < \infty,$$

while $L^\infty(0,T;X)$ is the space of measurable functions such that

$$\|u\|_{L^\infty(0,T;X)} = \sup_{t\in[0,T]} \|u(t)\|_X < \infty.$$

This choice makes $L^p(0,T;X)$ a Banach space and guarantees that its dual can be identified with $L^{p'}(0,T;X')$, where $p'=p/(p-1)$ and X' is the dual space to X. In particular, for functions $f\in L^2(0,T;L^1(\Omega))$ the continuous Minkowski inequality (see [55, p.499]) yields $f\in L^1(0,T;L^2(\Omega))$ and moreover

$$\begin{aligned}\|f\|_{L^2(0,T;L^1(\Omega))}&:=\left(\int_0^T\left(\int_\Omega |f|\,dx\right)^2 dx\right)^{1/2}\\&\leqslant \int_\Omega\left(\int_0^T |f|^2\,dt\right)^{1/2} dx=:\|f\|_{L^1(0,T;L^2(\Omega))}.\end{aligned}$$

Hence, we have $L^2(0,T;L^1(\Omega))\hookrightarrow L^1(0,T;L^2(\Omega))$. The full presentation of this topic can be found in [29].

2.2. Variable Exponent

Let $u\in L^1(0,T;L^1(\Omega))\cap L^\infty(0,T;L^2(\Omega))$ be a given function. We associate with $u:Q_T\mapsto\mathbb{R}$ the exponent $p_u:Q_T\to\mathbb{R}$ defined by the rule (1.7).

Since $G_\sigma\in C^\infty(\mathbb{R}^2)$, it follows from (1.7) and from absolute continuity of the Lebesgue integral that $1<p_u(t,x)\leqslant 2$ in Q_T and $p_u\in C^1([0,T];C^\infty(\mathbb{R}^2))$ even if u is just an absolutely integrable function in Q_T. Moreover, for each $t\in[0,T]$, $p_u(t,x)\approx 1$ in those places of Ω where some discontinuities are present in $u(t,\cdot)$, and $p_u(t,x)\approx 2$ in places where $u(t,x)$ is smooth or contains homogeneous features. In view of this, $p_u(t,x)$ can be interpreted as a characteristic of the sparse texture of the function u.

The following result plays a crucial role in the sequel (for comparison, we refer to [41, Lemma 2.1]).

Lemma 2.1. *Let $\{u_k\}_{k\in\mathbb{N}}\subset L^1(0,T;L^1(\Omega))\cap L^\infty(0,T;L^2(\Omega))$ be a sequence of measurable functions such that each element of this sequence is extended by zero outside of Q_T and*

$$\begin{gathered}\sup_{k\in\mathbb{N}}\|u_k\|_{L^\infty(0,T;L^2(\Omega))}<+\infty,\\ u_k\to u\ \ \textit{weakly in } L^1(0,T;L^1(\Omega))\ \textit{for some}\ \ u\in L^1(0,T;L^1(\Omega)).\end{gathered}\tag{2.1}$$

Let

$$\left\{p_{u_k}=1+g\left(\frac{1}{h}\int_{t-h}^t |(\nabla G_\sigma*\widetilde{u}_k(\tau,\cdot))|\,d\tau\right)\right\}_{k\in\mathbb{N}}$$

be the corresponding sequence of variable exponents. Then there exist constants $C>0$ and $\delta\in(0,1)$ depending on Ω, G, g, $\sup_{k\in\mathbb{N}}\|u_k\|_{L^\infty(0,T;L^2(\Omega))}$, and

$\sup_{k\in\mathbb{N}}\|u_k\|_{L^1(0,T;L^1(\Omega))}$ *such that*

$$p^- := 1+\delta \leqslant p_{u_k}(t,x) \leqslant p^+ := 2, \quad \forall\,(t,x)\in Q_T,\ \forall\,k\in\mathbb{N}, \tag{2.2}$$

$$\{p_{u_k}(\cdot)\}\subset\mathfrak{S}=\left\{ q\in C^{0,1}(Q_T)\ \middle|\ \begin{array}{c} |q(t,x)-q(s,y)|\leqslant C\left(|x-y|+|t-s|\right), \\ \forall\,(t,x),(s,y)\in\overline{Q_T}, \\ 1<p^-\leqslant q(\cdot,\cdot)\leqslant p^+ \ \textit{in}\ \overline{Q_T}. \end{array}\right\} \tag{2.3}$$

$$\begin{gathered} p_{u_k}\to p_u = 1+g\left(\frac{1}{h}\int_{t-h}^t \left|\left(\nabla G_\sigma * \widetilde{u}(\tau,\cdot)\right)(\cdot)\right|\,d\tau\right) \\ \textit{uniformly in}\ \overline{Q_T}\ \textit{as}\ k\to\infty. \end{gathered} \tag{2.4}$$

Proof. Since the sequence $\{u_k\}_{k\in\mathbb{N}}$ is uniformly bounded in $L^1(0,T;L^1(\Omega))$ and the Gaussian filter kernel G_σ is smooth, it follows that

$$\begin{aligned} \frac{1}{h}\int_{t-h}^t \Big|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(x)\Big|\,d\tau &\leqslant \frac{1}{h}\int_{t-h}^t\left(\int_\Omega |\nabla G_\sigma(x-y)|\,|\widetilde{u}_k(\tau,y)|\,dy\right)d\tau \\ &\leqslant \|G_\sigma\|_{C^1(\overline{\Omega-\Omega})}\frac{1}{h}\|u_k\|_{L^1(0,T;L^1(\Omega))}, \\ 2\geqslant p_{u_k}(t,x) = 1+g&\left(\int_{t-h}^t \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(x)\right|\,d\tau\right) \\ &\geqslant 1+g\left(\|G_\sigma\|_{C^1(\overline{\Omega-\Omega})}\tfrac{1}{h}\sup_{k\in\mathbb{N}}\|u_k\|_{L^1(0,T;L^1(\Omega))}\right), \\ &\qquad \forall\,(t,x)\in Q_T, \end{aligned}$$

where

$$\begin{aligned} \|G_\sigma\|_{C^1(\overline{\Omega-\Omega})} &= \max_{\substack{z=x-y \\ x\in\overline{\Omega},y\in\overline{\Omega}}}\Big[|G_\sigma(z)|+|\nabla G_\sigma(z)|\Big] \\ &= \frac{e^{-1}}{\left(\sqrt{2\pi}\sigma\right)^2}\left[1+\frac{1}{\sigma^2}\operatorname{diam}\Omega\right]. \end{aligned} \tag{2.5}$$

Then L^1-boundedness of $\{u_k\}_{k\in\mathbb{N}}$ guarantees the existence of a positive value $\delta\in(0,1)$ such that $p_{u_k}(t,x)\geqslant 1+\delta$. Hence, the estimate (2.2) holds true for all $k\in\mathbb{N}$.

Moreover, as follows from (1.8) and the relations

$$\begin{aligned} &\big|p_{u_k}(t,x)-p_{u_k}(t,y)\big| \\ &\leqslant C_g\left|\int_{t-h}^t \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(x)\right|\,d\tau - \int_{t-h}^t \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right|\,d\tau\right| \\ &\leqslant C_g\int_0^T \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(x)-\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right|\,d\tau \\ &\leqslant C_g\int_0^T\int_\Omega |u(\tau,z)|\,dz\,d\tau\max_{z\in\Omega}|\nabla G_\sigma(x-z)-\nabla G_\sigma(y-z)| \\ &= C_g\gamma_1\max_{z\in\Omega}|\nabla G_\sigma(x-z)-\nabla G_\sigma(y-z)|, \quad \forall\,x,y\in\overline{\Omega} \end{aligned} \tag{2.6}$$

with $\gamma_1 = \sup_{k\in\mathbb{N}} \|u_k\|_{L^1(0,T;L^1(\Omega))}$, and from smoothness of the function $\nabla G_\sigma(\cdot)$, there exists a positive constant $C_G > 0$ independent of k such that, for each $t \in [0,T]$, we have the following estimate

$$|p_{u_k}(t,x) - p_{u_k}(t,y)| \leqslant \gamma_1 C_g C_G |x-y|, \quad \forall\, x,y \in \overline{\Omega}.$$

Arguing in a similar manner, we see that

$$\begin{aligned}
&\big|p_{u_k}(t,y) - p_{u_k}(s,y)\big| \\
&\leqslant C_g \left| \int_{t-h}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right| \, d\tau - \int_{s-h}^{s} \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right| \, d\tau \right| \\
&\leqslant C_g \left| \int_{s}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right| \, d\tau - \int_{s-h}^{t-h} \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(y)\right| \, d\tau \right| \\
&\leqslant 2\gamma_1\gamma_2 C_g \|G_\sigma\|_{C^1(\overline{\Omega-\Omega})} |t-s|, \quad \forall\, t,s \in [0,T], \qquad (2.7)
\end{aligned}$$

where $\gamma_2 = \sup_{k\in\mathbb{N}} \|u_k\|_{L^\infty(0,T;L^2(\Omega))}$.

As a result, utilizing the estimates (2.6)–(2.7), and setting

$$C := C_g \gamma_1 \left(1 + 2\gamma_2 \|G_\sigma\|_{C^1(\overline{\Omega-\Omega})}\right), \qquad (2.8)$$

we see that

$$\begin{aligned}
|p_{u_k}(t,x) - p_{u_k}(s,y)| &\leqslant |p_{u_k}(t,x) - p_{u_k}(t,y)| + |p_{u_k}(t,y) - p_{u_k}(s,y)| \\
&\leqslant C\left[|x-y| + |t-s|\right], \\
&\forall\, (t,x),(s,y) \in \overline{Q_T} := [0,T] \times \overline{\Omega}.
\end{aligned} \qquad (2.9)$$

Thus, $\{p_{u_k}\} \subset \mathfrak{S}$. Since $\max_{(t,x)\in\overline{Q_T}} |p_{u_k}(t,x)| \leqslant p^+$ and each element of the sequence $\{p_{u_k}\}_{k\in\mathbb{N}}$ has the same modulus of continuity, it follows that this sequence is uniformly bounded and equi-continuous. Hence, by Arzelà–Ascoli Theorem the sequence $\{p_{u_k}\}_{k\in\mathbb{N}}$ is relatively compact with respect to the strong topology of $C(\overline{Q_T})$. Taking into account the estimate (2.9) and the fact that the set $\mathfrak{S}$ is closed with respect to the uniform convergence and

$$\frac{1}{h}\int_{t-h}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}_k(\tau,\cdot)\right)(x)\right| \, d\tau \to \frac{1}{h}\int_{t-h}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}(\tau,\cdot)\right)(x)\right| \, d\tau$$
$$\text{as } k\to\infty,\ \forall\,(t,x)\in Q_T$$

by definition of the weak convergence in $L^1(0,T;L^1(\Omega))$, we deduce: $p_{u_k} \to p_u$ uniformly in $\overline{Q_T}$ as $k\to\infty$, where

$$p_u(t,x) = 1 + g\left(\frac{1}{h}\int_{t-h}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}(\tau,\cdot)\right)(\cdot)\right| \, d\tau\right)$$

in Q_T. The proof is complete. □

2.3. Anisotropic Diffusion Tensor

Let $E_I \in W^{1,1}(\Omega)$ be a given function. Then for each $\lambda \in \mathbb{R}$ the upper level set of E_I can be defined as follows

$$Z_\lambda(E_I) = \{E_I \geqslant \lambda\} := \{x \in \Omega \ : \ E_I(x) \geqslant \lambda\}\,.$$

It was proven in [8] that for each function $E_I \in W^{1,1}(\Omega)$ its upper level sets $Z_\lambda(E_I)$ are sets of finite perimeter. So, the boundaries of level sets can be described by a countable family of Jordan curves with finite length, i.e., by continuous maps from the circle into the plane $\mathbb{R}^2$ without crossing points. As a result, at almost all points of almost all level sets of $E_I \in W^{1,1}(\Omega)$ we can define a unit normal vector $\theta(x)$. This vector field formally satisfies the following relations

$$(\theta, \nabla E_I) = |\nabla E_I| \quad \text{and} \quad |\theta| \leqslant 1 \text{ a.e. in } \Omega.$$

In the sequel, we will refer to θ as the vector field of unit normals to the topographic map of a function E_I. In fact, this vector field can be defined by the rule $\theta(x) = \frac{\nabla U(t,x)}{|\nabla U(t,x)|}$ with $t > 0$ small enough, where $U(t,x)$ is a solution the following initial-boundary value problem

$$\frac{\partial U}{\partial t} = \operatorname{div}\left(\frac{\nabla U}{|\nabla U| + \delta}\right), \quad t \in (0,+\infty),\ x \in \Omega, \tag{2.10}$$

$$U(0,x) = E_I(x), \quad x \in \Omega, \tag{2.11}$$

$$\frac{\partial U(0,x)}{\partial \nu} = 0, \quad t \in (0,+\infty),\ x \in \partial\Omega \tag{2.12}$$

with a relaxed version of the $1D$-Laplace operator in the principle part of (2.10). Here, $\delta > 0$ is a sufficiently small positive value and it can be chosen as in (1.8).

Let $\eta \in (0,1)$ be a given threshold. For the simplicity, we set $\eta = 1 - \delta$. Then, we associate with the vector field $\theta : \Omega \to \mathbb{R}^2$ the following linear operator $R_\eta : \mathbb{R}^2 \to \mathbb{R}^2$:

$$R_\eta \nabla v := \nabla v - \eta^2\,(\theta, \nabla v)\,\theta = \left[I - \eta^2 \theta \otimes \theta\right] \nabla v, \quad \forall\, v \in W^{1,1}(\Omega). \tag{2.13}$$

In fact, this operator can be interpreted as the Directional Total Variation of v along the vector field θ (see [16] for the details).

Remark 2.1. In practice, the function E_I is usually associated with the spectral energy for a smoothed version $I = [I_1, I_2, I_3]^t \in L^2(\Omega; \mathbb{R}^3)$ of the original color image which is presumably has been corrupted by some noise. The standard rule for that is the following one

$$E_I(x) := \alpha_1 I_1(x) + \alpha_2 I_2(x) + \alpha_3 I_3(x), \quad \forall\, x \in \Omega,$$

with $\alpha_1 = 0.114$, $\alpha_2 = 0.587$, and $\alpha_3 = 0.299$.

As for the operator $R_\eta : \mathbb{R}^2 \to \mathbb{R}^2$, in this case it accumulates the structural prior information about the spectral energy E_I. Indeed, let us assume that $x \in \Omega$

is a point in which E_I is not expected to change drastically in any direction, i.e. x is not close to a discontinuity or rapid change in the known structure of E_I. In this case, R_η can be represented as a unit matrix. So, at this point we obviously have $R_\eta \nabla v \approx \nabla v$.

On the other hand, if we consider a point that is close to a discontinuity of E_I, then $R_\eta \nabla v$ reduces to $(1-\eta^2)\nabla v$ if the gradient $\nabla v(t,x)$ at this point is co-linear to θ, and to $\nabla v(t,x)$ provided $\nabla v(t,x)$ is orthogonal to θ. So, this operator does not enforce gradients of v in the direction θ. Moreover, the following two-side estimate

$$(1-\eta^2)|\nabla v|^2 \leqslant |(\nabla v, R_\eta \nabla v)| \leqslant |\nabla v|^2, \quad \text{a.e. in } Q_T \tag{2.14}$$

holds for each $v \in L^\infty(0,T;W^{1,1}(\Omega))$. We also make use of the following observation: since $|\xi|^2 \leqslant \left(\xi, \left[I-\eta^2\theta\otimes\theta\right]^{-1}\xi\right) \leqslant (1-\eta^2)^{-1}|\xi|^2$ and

$$(1-\eta^2)|\nabla v|^2 \leqslant \left(R_\eta\nabla v, \left[I-\eta^2\theta\otimes\theta\right]^{-1} R_\eta\nabla v\right) \leqslant (1-\eta^2)^{-1}|R_\eta\nabla v|^2,$$
$$|R_\eta\nabla v|^2 \leqslant \left(R_\eta\nabla v, \left[I-\eta^2\theta\otimes\theta\right]^{-1} R_\eta\nabla v\right) \leqslant |\nabla v|^2$$

it follows that

$$(1-\eta^2)|\nabla v| \leqslant |R_\eta \nabla v| \leqslant |\nabla v|, \quad \text{a.e. in } Q_T. \tag{2.15}$$

2.4. On Orlicz Spaces

Let $w \in L^1(0,T;L^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$ be a given function. Let $p_w : Q_T \to \mathbb{R}$ be the corresponding variable exponent which is defined by the rule (1.7). Then

$$1 < p^- \leqslant p_w(t,x) \leqslant p^+ < \infty \ \text{ a.e. in } \ Q_T \tag{2.16}$$

(see Lemma 2.1), where the constants p^- and p^+ are given by (2.2). Let $p'_w(t,x) = \frac{p_w(t,x)}{p_w(t,x)-1}$ be the corresponding conjugate exponent. It is clear that

$$2 = \underbrace{\frac{p^+}{p^+-1}}_{(p^+)'} \leqslant p'_w(t,x) \leqslant \underbrace{\frac{p^-}{p^--1}}_{(p^-)'} = \frac{p^-}{\delta} \quad \textit{a.e. in } Q_T, \tag{2.17}$$

where $(p^+)'$ and $(p^-)'$ stand for the conjugates of constant exponents. Denote by $L^{p_w(\cdot)}(Q_T)$ the set of all measurable functions $f : Q_T \to \mathbb{R}$ such that the modular is finite, i.e.

$$\rho_{p_w(t,x)}(f) := \int_{Q_T} |f(t,x)|^{p_w(t,x)}\, dxdt < \infty. \tag{2.18}$$

Equipped with the Luxembourg norm

$$\|f\|_{L^{p_w(\cdot)}(Q_T)} = \inf\left\{\lambda > 0 \ : \ \int_{Q_T} |\lambda^{-1} f(t,x)|^{p_w(t,x)}\, dxdt \leqslant 1\right\}. \tag{2.19}$$

$L^{p_w(\cdot)}(Q_T)$ becomes a Banach space (see [24, 30] for the details). The space $L^{p_w(\cdot)}(Q_T)$ is a sort of Musielak-Orlicz space that can be denoted by generalised Lebesgue space, because many of its properties are inherited from the classical Lebesgue spaces. In particular, the two-sides inequality (2.16) implies that $L^{p_w(\cdot)}(Q_T)$ is reflexive, separable, and the set $C_0^\infty(Q_T)$ is dense in $L^{p_w(\cdot)}(Q_T)$. Moreover, under condition (2.16), $L^\infty(Q_T) \cap L^{p_w(\cdot)}(Q_T)$ is also dense in $L^{p_w(\cdot)}(Q_T)$.

Its dual can be identified with $L^{p'_w(\cdot)}(Q_T)$ and, therefore, any continuous functional $F = F(f)$ on $L^{p_w(\cdot)}(Q_T)$ has the form (see [58, Lemma 13.2])

$$F(f) = \int_{Q_T} fg\,dxdt, \quad \text{with } \ g \in L^{p'_w(\cdot)}(Q_T).$$

Since the relation between the modular (2.18) and the norm (2.19) that is not so direct as in the classical Lebesgue spaces, it can be proved, from its definitions in (2.18) and (2.19), that

$$\begin{aligned}\min\left\{\|f\|^{p^-}_{L^{p_w(\cdot)}(Q_T)}, \|f\|^{p^+}_{L^{p_w(\cdot)}(Q_T)}\right\} &\leqslant \rho_{p_w(t,x)}(f)\\ &\leqslant \max\left\{\|f\|^{p^-}_{L^{p_w(\cdot)}(Q_T)}, \|f\|^{p^+}_{L^{p_w(\cdot)}(Q_T)}\right\},\\ \min\left\{\rho^{\frac{1}{p^-}}_{p_w(t,x)}(f), \rho^{\frac{1}{p^+}}_{p_w(t,x)}(f)\right\} &\leqslant \|f\|_{L^{p_w(\cdot)}(Q_T)}\\ &\leqslant \max\left\{\rho^{\frac{1}{p^-}}_{p_w(t,x)}(f), \rho^{\frac{1}{p^+}}_{p_w(t,x)}(f)\right\}. \end{aligned}\tag{2.20}$$

When proving some estimates the following consequence of (2.20) is very useful,

$$\begin{gathered}\|f\|^{p^-}_{L^{p_w(\cdot)}(Q_T)} - 1 \leqslant \int_{Q_T} |f(t,x)|^{p_w(t,x)}\,dxdt \leqslant \|f\|^{p^+}_{L^{p_w(\cdot)}(Q_T)} + 1,\\ \forall\, f \in L^{p_w(\cdot)}(Q_T),\end{gathered}\tag{2.21}$$

$$\begin{gathered}\|f_k - f\|_{L^{p_w(\cdot)}(Q_T)} \to 0 \quad \Longleftrightarrow \quad \int_{Q_T} |f_k(t,x) - f(t,x)|^{p_w(t,x)}\,dxdt \to 0\\ \text{as } k \to \infty.\end{gathered}\tag{2.22}$$

Moreover, if $f \in L^{p_w(\cdot)}(Q_T)$ then

$$\|f\|_{L^{p^-}(Q_T)} \leqslant (1 + T|\Omega|)^{1/p^-}\,\|f\|_{L^{p_w(\cdot)}(Q_T)},\tag{2.23}$$

$$\begin{gathered}\|f\|_{L^{p_w(\cdot)}(Q_T)} \leqslant (1 + T|\Omega|)^{1/(p^+)'}\,\|f\|_{L^{p^+}(Q_T)},\\ (p^+)' = \frac{p^+}{p^+ - 1},\ \forall\, f \in L^{p^+}(Q_T),\end{gathered}\tag{2.24}$$

(see, for instance, [24, 30, 57]).

In generalised Lebesgue spaces, there holds a version of Young's inequality,

$$|fg| \leqslant \varepsilon \frac{|f|^{p_w(\cdot)}}{p_w(\cdot)} + C(\varepsilon)\frac{|g|^{p'_w(\cdot)}}{p'_w(\cdot)},$$

valid for some positive constant $C(\varepsilon)$ and any $\varepsilon > 0$.

The following result can be viewed as an analogous of the Hölder inequality in Lebesgue spaces with variable exponents (for the details we refer to [24, 30]).

Proposition 2.1. If $f \in L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)$ and $g \in L^{p'_w(\cdot)}(Q_T;\mathbb{R}^2)$, then $(f,g) \in L^1(Q_T)$ and

$$\int_{Q_T} (f,g)\, dxdt \leqslant 2\|f\|_{L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)}\|g\|_{L^{p'_w(\cdot)}(Q_T;\mathbb{R}^2)}. \tag{2.25}$$

As a consequence of (2.25), we have, for a bounded domain $Q_T = (0,T)\times\Omega$ and $p_w(\cdot)$ satisfying to (2.16), the following continuous imbedding

$$L^{p_w(\cdot)}(Q_T) \hookrightarrow L^{r(\cdot)}(Q_T) \ \text{ whenever } \ p_w(t,x) \geqslant r(t,x) \ \text{ for a.e. } (t,x)\in Q_T. \tag{2.26}$$

Let $\{p_k\}_{k\in\mathbb{N}} \subset C^{0,\widehat{\delta}}(\overline{Q_T})$, with some $\widehat{\delta}\in(0,1]$, be a given sequence of exponents. Hereinafter in this subsection we assume that

$$\begin{array}{c} p, p_k \in C^{0,\widehat{\delta}}(\overline{Q_T}) \ \text{ for } k = 1,2,\dots, \text{ and} \\ p_k(\cdot) \to p(\cdot) \ \text{ uniformly in } \overline{Q_T} \text{ as } k\to\infty. \end{array} \tag{2.27}$$

We associate with this sequence the another one $\left\{f_k \in L^{p_k(\cdot)}(Q_T)\right\}_{k\in\mathbb{N}}$. The characteristic feature of this set of functions is that each element f_k lives in the corresponding Orlicz space $L^{p_k(\cdot)}(Q_T)$. So, we have a sequence in the scale of spaces $\left\{L^{p_k(\cdot)}(Q_T)\right\}_{k\in\mathbb{N}}$. We say that the sequence $\left\{f_k \in L^{p_k(\cdot)}(Q_T)\right\}_{k\in\mathbb{N}}$ is bounded if

$$\limsup_{k\to\infty} \int_{Q_T} |f_k(t,x)|^{p_k(t,x)}\, dxdt < +\infty. \tag{2.28}$$

Definition 2.1. A bounded sequence $\left\{f_k \in L^{p_k(\cdot)}(Q_T)\right\}_{k\in\mathbb{N}}$ is weakly convergent in the variable Orlicz space $L^{p_k(\cdot)}(Q_T)$ to a function $f \in L^{p(\cdot)}(Q_T)$, where $p \in C^{0,\delta}(\overline{Q_T})$ is the limit of $\{p_k\}_{k\in\mathbb{N}} \subset C^{0,\widehat{\delta}}(\overline{Q_T})$ in the uniform topology of $C(\overline{Q_T})$, if

$$\lim_{k\to\infty}\int_{Q_T} f_k\varphi\, dxdt = \int_{Q_T} f\varphi\, dxdt, \quad \forall\,\varphi \in C_c^\infty(\mathbb{R}\times\mathbb{R}^2). \tag{2.29}$$

For our further analysis, we make use of the following result concerning the lower semicontinuity property of the variable $L^{p_k(\cdot)}$-norm with respect to the weak convergence in $L^{p_k(\cdot)}(Q_T)$ (for the proof, we refer to [23, Lemma 3.1], see also [58, Lemma 13.3] and [41, Lemma 2.1] for comparison).

Proposition 2.2. If the sequence of exponents $\{p_k\}_{k\in\mathbb{N}}$ satisfies condition (2.16), $p_k \to p$ as $k\to\infty$ a.e.in Q_T, and a bounded sequence $\left\{f_k \in L^{p_k(\cdot)}(Q_T)\right\}_{k\in\mathbb{N}}$ converges weakly in $L^{p^-}(Q_T)$ to f, then $f \in L^{p(\cdot)}(Q_T)$, $f_k \rightharpoonup f$ in variable $L^{p_k(\cdot)}(Q_T)$, and

$$\liminf_{k\to\infty}\int_{Q_T} |f_k(t,x)|^{p_k(t,x)}\, dxdt \geqslant \int_{Q_T} |f(t,x)|^{p(t,x)}\, dxdt. \tag{2.30}$$

We recall also the inequality which is classical in the theory of p-Laplace equations: if $1 < p \leqslant 2$ then, for all $\xi, \eta \in \mathbb{R}^N$, the following estimate holds true

$$(p-1)|\xi-\eta|^2 \leqslant \left(\left[|\xi|^{p-2}\xi - |\eta|^{p-2}\eta\right], \xi-\eta\right)\left(|\xi|^p + |\eta|^p\right)^{\frac{2-p}{p}}.$$

2.5. On Weighted Energy Space with Variable Exponent

Let $R_\eta : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear operator defined by the rule (2.13) and associated with some vector field $\theta \in L^\infty(\Omega;\mathbb{R}^2)$. Let $w \in C([0,T];L^2(\Omega))$ be a given function. We define the weighted energy space $W_w(Q_T)$ as the set of all functions $u(t,x)$ such that

$$\begin{gathered} u \in L^2(Q_T), \quad u(t,\cdot) \in W^{1,1}(\Omega) \ \text{ for a.e.} t \in [0,T], \\ \int_{Q_T} |R_\eta \nabla u|^{p_w(t,x)}\, dxdt < +\infty. \end{gathered} \tag{2.31}$$

We equip $W_w(Q_T)$ with the norm

$$\|u\|_{W_w(Q_T)} = \|u\|_{L^2(Q_T)} + \|R_\eta \nabla u\|_{L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)}, \tag{2.32}$$

where the second term on the right-hand side is the norm of the vector-valued function $R_\eta \nabla u(t,x)$ in the Orlicz space $L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)$. Due to the estimate (2.14), we see that $W_w(Q_T)$, equipped with the norm (2.32), is a reflexive Banach space. Moreover, due to the fact that the exponent $p_w : Q_T \to \mathbb{R}$ is Lipschitz continuous, the smooth functions are dense in the weighted Sobolev-Orlicz space $W_w(Q_T)$ (see [4]). So, $W_w(Q_T)$ can be considered as the closure of the set $\left\{\varphi \in C^\infty(\overline{Q}_T)\right\}$ with respect to the norm $\|\cdot\|_{W_w(Q_T)}$.

2.6. On the weak convergence of fluxes to flux

Let us consider the following collection of parabolic equations of monotone type

$$\frac{\partial u_k}{\partial t} - \operatorname{div} A_k(t,x,\nabla u_k) = f, \quad (t,x) \in Q_T, \tag{2.33}$$

where $f \in L^2(\Omega)$ and $k = 1,2,\ldots$. Let u_k be a solution of (2.33) for a given $k \in \mathbb{N}$ and this solution is understood in the sense of distributions. Assume that $A_k(\cdot,\cdot,\xi) \to A(\cdot,\cdot,\xi)$ as $k \to \infty$ pointwise a.e. with respect to the first two arguments and for all $\xi \in \mathbb{R}^N$.

A typical situation arising in the study of most optimization problems and which is of fundamental importance in many others areas of nonlinear analysis, can be stated as follows: suppose it is known that a solution $u_k \in L^2(0,T;W^{1,p^-}(\Omega))$ of (2.33) and the corresponding flow $w_k = A_k(\cdot,\cdot,\nabla u_k) \in L^{(p^+)'}(Q_T;\mathbb{R}^N)$ converge weakly, namely,

$$\begin{gathered} u_k \rightharpoonup u \ \text{ in } \ L^2(0,T;W^{1,p^-}(\Omega)), \quad w_k \rightharpoonup w \ \text{ in } \ L^{(p^+)'}(Q_T;\mathbb{R}^N), \\ 1 < p^- < p^+, \ (p^+)' = \frac{p^+}{p^+-1}. \end{gathered}$$

The main question is whether a flux converges to a flux, i.e., whether the equality for the limit elements $A(t,x,\nabla u) = w$ holds. The situation is not trivial because the function $A(\cdot,\cdot,v)$ is nonlinear in v and the weak convergence $v_k \rightharpoonup v$ is far from sufficient to derive the limit relation $A_k(\cdot,\cdot,v_k) \rightharpoonup A(\cdot,\cdot,v)$. So, the important problem is to show that $w = A(\cdot,\cdot,\nabla u)$, although the validity of this equality is by no means obvious at this stage. The conditions (first of all, on the exponents p^- and p^+) under which the answer to the above question is affirmative, have been obtained by Zhikov and Pastukhova in their celebrated paper [60].

Theorem 2.1. *Assume that the following conditions are satisfied:*

(C1) $A_k(t,x,\xi)$ *and* $A(t,x,\xi)$ *are* $\mathbb{R}^N$*-valued Carathéodory functions, that is, these functions are continuous in* $\xi \in \mathbb{R}^N$ *for a.e.* $(t,x) \in Q_T$ *and measurable with respect to* $(t,x) \in Q_T$ *for each* $\xi \in \mathbb{R}^N$*;*

(C2) $\Big(A_k(t,x,\xi) - A_k(t,x,\zeta), \xi - \zeta\Big) \geqslant 0$, $A_k(t,x,0) = 0 \quad \forall \xi,\zeta \in \mathbb{R}^N$ *and for a.e.* $(t,x) \in Q_T$*;*

(C3) $|A_k(t,x,\xi)| \leqslant c(|\xi|) < \infty$ *and* $\lim_{k\to\infty} A_k(t,x,\xi) = A(t,x,\xi)$ *for all* $\xi \in \mathbb{R}^N$ *and for a.e.* $(t,x) \in Q_T$*;*

(C4) $u_k \rightharpoonup u$ *in* $L^{p^-}(0,T;W^{1,p^-}(\Omega))$, $p^- > 1$, *and* $\{u_k\}_{k\in\mathbb{N}}$ *are bounded in* $L^\infty(0,T;L^2(\Omega))$*;*

(C5) $w_k = A_k(t,x,\nabla u_k) \rightharpoonup w$ *in* $L^{(p^+)'}(Q_T;\mathbb{R}^N)$, $p^+ > 1$*;*

(C6) $u_k \in L^{p^+}(0,T;W^{1,p^+}(\Omega))$ *for all* $k \in \mathbb{N}$, *and* $\sup_{k\in\mathbb{N}} \| (w_k, \nabla u_k) \|_{L^1(Q_T)} < \infty$*;*

(C7) $1 < p^- < p^+ < 2p^-$.

Then the flux $A_k(t,x,\nabla u_k)$ *weakly converges in the Lebesgue space* $L^{(p^+)'}(Q_T;\mathbb{R}^N)$ *to the flux* $A(t,x,\nabla u)$.

For our further analysis, we make also use of the following well-known results.

Lemma 2.2 ([57])**.** *Let* Ψ *be a class of integrands* $F(t,x,\xi)$ *that are convex with respect to* $\xi \in \mathbb{R}^N$*, measurable with respect to* $(t,x) \in Q_T$*, and satisfy the estimate*

$$c_1|\xi|^{p^-} \leqslant F(t,x,\xi) \leqslant c_2|\xi|^{p^+}, \quad 1 < p^- \leqslant p^+ < \infty,\ c_1,c_2 > 0.$$

Suppose that F_k *and* F *belong to the class* Ψ *and the following condition holds:*

$$\lim_{k\to\infty} F_k(t,x,\xi) = F(t,x,\xi) \quad \textit{for a.e. } (t,x) \in Q_T \textit{ and any } \xi \in \mathbb{R}^N.$$

Then the following lower semicontinuity property is valid:
if $v_k \rightharpoonup v$ *in* $L^1(Q_T;\mathbb{R}^N)$ *then*

$$\liminf_{k\to\infty} \int_{Q_T} F_k(t,x,v_k)\,dxdt \geqslant \int_{Q_T} F(t,x,v)\,dxdt. \tag{2.34}$$

Lemma 2.3 ([59]). *Let $A_k(t,x,\xi)$ and $A(t,x,\xi)$ be $\mathbb{R}^N$-valued Carathéodory functions with properties (C1)–(C3). Assume that*

$$v_k \rightharpoonup v \quad \text{and} \quad w_k = A_k(t,x,v_k) \rightharpoonup w \quad \text{in } L^1(Q_T;\mathbb{R}^N) \text{ as } k \to \infty,$$

and $(w,v) \in L^1(Q_T)$. Then

$$\liminf_{k\to\infty} \int_{Q_T} (A_k(t,x,v_k),v_k)\; dxdt \geqslant \int_{Q_T} (w,v)\; dxdt. \tag{2.35}$$

Lemma 2.4 ([4]). *Let ε be a small parameter which varies within a strictly decreasing sequence of positive numbers converging to 0. Assume that the following conditions*

$$(i)\quad p_\varepsilon, p \in C(\overline{Q_T}), \quad p_\varepsilon \to p \ \text{ in } C(\overline{Q_T}) \text{ as } \varepsilon \to 0,$$

$$(ii)\quad v_\varepsilon \in L^1(Q_T;\mathbb{R}^N), \quad \int_{Q_T} \left[|v_\varepsilon|^{p_\varepsilon} + \varepsilon|v_\varepsilon|^{p^+}\right] dxdt \leqslant K < \infty \ \text{ for each } \varepsilon > 0,$$

$$(iii)\quad |v_\varepsilon|^{p_\varepsilon-2}v_\varepsilon + \varepsilon|v_\varepsilon|^{p^+-2}v_\varepsilon \rightharpoonup z \ \text{ in } L^{(p^+)'}(Q_T;\mathbb{R}^N),$$
$$(p^+)' = p^+/(p^+-1) \ \text{ as } \varepsilon \to 0$$

hold true with some p^- and p^+ such that $1 < p^- \leqslant p_\varepsilon(t,x) \leqslant p^+ < \infty$ for all $\varepsilon > 0$ and $(t,x) \in Q_T$. Then $z \in L^{p'(\cdot)}(Q_T;\mathbb{R}^N)$.

3. Existence Result for a Class of Parabolic Equations with Variable Nonlocal Exponent

The main object of our consideration in this section is the following initial-boundary value problem (IBVP)

$$\frac{\partial u}{\partial t} - \operatorname{div} A_u(t,x,\nabla u) + \kappa u = \kappa(f - v) \quad \text{in } \ Q_T, \tag{3.1}$$

$$\partial_\nu u = 0 \quad \text{on } \ (0,T)\times\partial\Omega, \tag{3.2}$$

$$u(0,\cdot) = f_0 \quad \text{in } \ \Omega. \tag{3.3}$$

Here,

$$A_w(t,x,\nabla u) := |R_\eta \nabla u|^{p_w(t,x)-2} R_\eta \nabla u, \tag{3.4}$$

the exponent $p_w : Q_T \to (1,2]$ is given by the rule (1.7), the linear operator R_η is defined in (2.13), ∂_ν stands for the outward normal derivative, $f \in L^2(Q_T)$ and $f_0 \in L^2(\Omega)$ are given distributions, $v \in \mathcal{V}_{ad}$ stands for the control, and the class of admissible controls $\mathcal{V}_{ad}$ is defined as

$$\mathcal{V}_{ad} = \left\{ v \in L^2(0,T;L^1(\Omega)) \ : \ v_a(x) \leqslant v(t,x) \leqslant v_b(x), \text{ a.e. in } \ Q_T \right\}. \tag{3.5}$$

As follows from (3.4), (2.14), and Lemma 2.1, for each fixed function $w \in C([0,T];L^2(\Omega))$, the mapping $(t,x,\xi) \mapsto A_w(t,x,\xi)$ is a Carathéodory vector function, that is, $A_w(t,x,\xi)$ is continuous in $\xi \in \mathbb{R}^2$ and is measurable with respect to (t,x) for each $\xi \in \mathbb{R}^2$. Moreover, the following monotonicity, coerciveness and boundedness conditions hold for a.e. $(t,x) \in Q_T$ [58]:

$$\Big(A_w(t,x,\xi) - A_w(t,x,\zeta), \xi - \zeta\Big) \geqslant 0, \quad \forall\, \xi, \zeta \in \mathbb{R}^2, \tag{3.6}$$

$$\begin{aligned}(A_w(t,x,\xi),\xi) &= |R_\eta \xi|^{p_w(t,x)-2}\left(R_\eta\xi, R_\eta^{-1} R_\eta \xi\right)\\ &\overset{\text{by (2.15)}}{\geqslant} \left(1-\eta^2\right)^{p_w(t,x)} |\xi|^{p_w(t,x)} \geqslant \left(1-\eta^2\right)^2 |\xi|^{p_w(t,x)},\ \forall\, \xi \in \mathbb{R}^2,\end{aligned} \tag{3.7}$$

$$|A_w(t,x,\xi)|^{p_w'(t,x)} \leqslant |\xi|^{p_w(t,x)}, \quad \forall\, \xi \in \mathbb{R}^2, \tag{3.8}$$

However, in general, the principle operator $-\operatorname{div} A_u(t,x,\nabla u) + \kappa u$ provides an example of a strongly non-linear, non-monotone, and non-coercive operator in divergence form.

It is worth mentioning here that if the exponent $p = p(t,x)$ is a given function (i.e., it does not depend on the unknown solution u) and $p \in C^{0,\delta}(\overline{Q_T})$, with some $\delta \in (0,1]$, then for every $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, and $v \in \mathcal{V}_{ad}$, problem (3.1), (3.3) with $R_\eta = I$ and with zero Dirichlet boundary conditions (instead of the Neumann one (3.2)) admits a weak solution $u \in C([0,T];L^2(\Omega))$ such that $\int_{Q_T} |\nabla u|^{p(t,x)}\,dxdt < +\infty$ (see, e.g. [10, Ch.4]). In this case the time derivative of the weak solution is a distribution u_t which may not belong to any Lebesgue space $L^s(Q_T)$ with $s > 1$. Moreover, the issue of uniqueness for the weak solutions remains, apparently, an open question for nowadays [35, Chapter III].

As for the case of Dirichlet problem for the equation (3.1) with $p_u(t,x)$ given by (1.7), its regularity (see Lemma 2.1) is insufficient for the convergence of the sequence of Galerkin's approximations to a weak solution. To overcome this difficulty, it was recently proposed in [11] to construct the strong solutions with the extra regularity property $u_t \in L^2(Q_T)$. However, the existence of a strong solution and its uniqueness to the Dirichlet problem for the equation (3.1) has been proven in [11] if only the following condition for the range of the exponent $p_u(t,x)$ holds true

$$\frac{2N}{2+N} < p^- \leqslant p^+ < 2, \quad \text{where } \ N = \dim\, \Omega.$$

Since the fulfillment of this condition is rather questionable in our case (see Lemma 2.1), our prime interest in this section is to study the solvability issues for Cauchy-Neumann initial-boundary value problem (3.1)–(3.3) with $p_u(t,x)$ given by (1.7). We recall that a challenging feature of the equation (3.1) is that it cannot be interpreted as a duality relation in a fixed Banach space. Because of this, we can not write down the weak formulation of (3.1)–(3.3) as some equality in terms of duality. In particular, sequences of solutions u_k to this problem that correspond to different exponents p_{u_k}, belong to possible distinct Sobolev spaces. Mainly because of this, we specify the notion of weak solution as follows:

Definition 3.1. We say that, for given $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $v \in \mathcal{V}_{ad}$, a function u is a weak solution to the problem (3.1)–(3.3) if $u \in W_u(Q_T)$, i.e.,

$$\begin{gathered} u \in L^2(Q_T),\ u(t,\cdot) \in W^{1,1}(\Omega) \ \text{ for a.e. } t \in [0,T], \\ \int_{Q_T} |R_\eta \nabla u|^{p_u(t,x)}\, dxdt < +\infty, \end{gathered} \tag{3.9}$$

and the integral identity

$$\begin{aligned} \int_{Q_T} \left(-u\frac{\partial \varphi}{\partial t} + (A_u(t,x,\nabla u), \nabla \varphi) + \kappa u \varphi \right) dxdt \\ = \kappa \int_{Q_T} (f-v)\varphi\, dxdt + \int_\Omega f_0 \varphi|_{t=0}\, dx \end{aligned} \tag{3.10}$$

holds true for any function $\varphi \in \Phi$, where $\Phi = \left\{ \varphi \in C^\infty(\overline{Q}_T) \ : \ \varphi|_{t=T} = 0 \right\}$.

To clarify the sense in which the initial value $u(0,\cdot) = f_0$ is assumed for the weak solutions, we give the following assertion (for the proof, we refer to [41, Proposition 2.2]).

Proposition 3.1. Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $v \in \mathcal{V}_{ad}$ be given distributions. Let $u \in W_u(Q_T)$ be a weak solution to the problem (3.1)–(3.3) in the sense of Definition 3.1. Then, for any $\eta \in C^\infty(\overline{\Omega})$, the scalar function $h(t) = \int_\Omega u(t,x)\eta(x)\, dx$ belongs to $W^{1,1}(0,T)$ and $h(0) = \int_\Omega f_0(x)\eta(x)\, dx$.

Let us show that the problem (3.1)–(3.3) admits at least one weak solution. With that in mind, we make use of the perturbation technique and a classical fixed point theorem of Schauder [49] (we refer to [25, 33, 39, 42, 45] where the similar technique has been used).

We begin with the following auxiliary results. Following result is crucial in this section.

Theorem 3.1. *For given functions $w \in L^\infty(0,T;L^2(\Omega))$ and $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, let the exponent $p_w : Q_T \to \mathbb{R}$ and the linear operator R_η be defined by the rules* (1.7) *and* (2.13)*, respectively. Then there exists a positive constant Λ such that, for a.a $(t,x) \in Q_T$ and for $\varepsilon > 0$ small enough, the following inequality holds true*

$$(A^\varepsilon_w(t,x,\nabla u), \nabla u) \geqslant \begin{cases} \Lambda |\nabla u|^{p_w(t,x)}, & \textit{if } \ |\nabla u| \geqslant 1, \\ \Lambda\left(|\nabla u|^{p_w(t,x)} - 4 \right), & \textit{if } \ |\nabla u| < 1, \end{cases} \quad \textit{a.e. in } \ Q_T, \tag{3.11}$$

where ε is a small positive value and

$$A^\varepsilon_w(t,x,\nabla u) := \left(|R_\eta \nabla u| + \varepsilon \right)^{p_w(t,x)-2} R_\eta \nabla u.$$

Proof. Taking into account that (see (2.15))

$$(1-\eta^2)|\zeta| \leqslant |R_\eta \zeta| \leqslant |\zeta|, \quad \forall\, \zeta \in \mathbb{R}^2,\ \forall\,(t,x) \in Q_T,$$

we make use of the following chain of inequalities

$$\begin{aligned}(A^\varepsilon_w(t,x,\nabla u),\nabla u) &= (|R_\eta \nabla u| + \varepsilon)^{p_w(t,x)-2}\,(R_\eta \nabla u, \nabla u)\\ &\overset{\text{by (2.14)}}{\geqslant} \left(1-\eta^2\right)\frac{|R_\eta \nabla u|^2}{(|R_\eta \nabla u|+\varepsilon)^{2-p_w(t,x)}}\\ &\overset{\text{by (2.15)}}{\geqslant} \left(1-\eta^2\right)^3 \frac{|\nabla u|^2}{(|\nabla u|+\varepsilon)^{2-p_w(t,x)}} \quad \text{a.e. in } \ Q_T. \end{aligned} \tag{3.12}$$

To deduce the proof, it remains to distinguish two cases $|\nabla u| \geqslant 1$ and $\nabla u| < 1$ (see Lemma 1 in [56, Lemma 1]). As a result, we see that, for all $\varepsilon > 0$,

$$\begin{aligned}(A^\varepsilon_w(t,x,\nabla u),\nabla u) &\geqslant \frac{\left(1-\eta^2\right)^3}{2^{2-p_w(t,x)}}|\nabla u|^{p_w(t,x)}\\ &\geqslant \frac{\left(1-\eta^2\right)^3}{2}|\nabla u|^{p^-}, \quad \text{if } \ |\nabla u| \geqslant 1. \end{aligned} \tag{3.13}$$

At the same time, if $|\nabla u| < 1$, then we get

$$\begin{aligned}(A^\varepsilon_w(t,x,\nabla u),\nabla u) &= \Big(A^\varepsilon_w(t,x,\nabla u), \left[I - \eta^2\theta\otimes\theta\right]^{-1} R_\eta \nabla\Big)\\ &\geqslant \left(1-\eta^2\right)|\nabla u|^2\,(1+|\nabla u|)^{p_w(t,x)-2}\\ &= \left(1-\eta^2\right)(|\nabla u|+1-1)^2\,(1+|\nabla u|)^{p_w(t,x)-2}\\ &\geqslant \left(1-\eta^2\right)\Big(|\nabla u|^{p_w(t,x)} - 2\,(1+|\nabla u|)^{p_w(t,x)-1}\Big)\\ &\geqslant \left(1-\eta^2\right)\Big(|\nabla u|^{p_w(t,x)} - 4\Big) \quad \text{a.e. in } \ Q_T. \end{aligned} \tag{3.14}$$

□

Theorem 3.2. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, and $v \in \mathcal{V}_{ad}$ be given distributions, and $\theta \in L^\infty(\Omega;\mathbb{R}^2)$ is some vector field. Then, for each positive value $\varepsilon > 0$, the Cauchy-Neumann problem*

$$\frac{\partial u}{\partial t} - \varepsilon\Delta u - \operatorname{div} A^\varepsilon_u(t,x,\nabla u) + \kappa u = \kappa(f-v) \quad \textit{in} \ \ Q_T := (0,T)\times\Omega, \tag{3.15}$$

$$\partial_\nu u = 0 \quad \textit{on} \ \ (0,T)\times\partial\Omega, \tag{3.16}$$

$$u(0,\cdot) = f_0 \quad \textit{in} \ \ \Omega, \tag{3.17}$$

has a weak solution $u_\varepsilon \in C([0,T];L^2(\Omega)) \cap L^2(0,T;W^{1,2}(\Omega))$ verifying (3.15)–(3.17) in the sense of distributions.

Proof. We introduce the space

$$W(0,T)=\left\{w\in L^2(0,T;W^{1,2}(\Omega)),\ \frac{dw}{dt}\in L^2(0,T;\left[W^{1,2}(\Omega)\right]')\right\}.$$

This space is a Hilbert space with respect to the graph norm. Let us fix an arbitrary function $w\in W(0,T)\cap L^\infty(0,T;L^2(\Omega))$ such that

$$\left.\begin{aligned}\|w\|_{L^2(0,T;W^{1,2}(\Omega))}&\leqslant C_1,\\ \|w\|_{L^\infty(0,T;L^2(\Omega))}&\leqslant\sqrt{2\kappa}C_1,\\ \|\tfrac{\partial w}{\partial t}\|_{L^2(0,T;(W^{1,2}(\Omega))')}&\leqslant C_3,\\ w(0,\cdot)&=f_0\ \text{ in }\ \Omega,\end{aligned}\right\}\tag{3.18}$$

with

$$\begin{aligned}C_1&=\sqrt{\|f-v\|^2_{L^2(Q_T)}+\frac{2}{\kappa}\|f_0\|^2_{L^2(\Omega)}}\ ,\\ C_2&=\left(\Lambda^{-1}\kappa C_1^2+5\right)^{1/p^-},\end{aligned}$$

where the constant Λ comes from inequality (3.11), and C_3 is defined in (3.31).

We divide the proof onto several steps.

Step 1. Let us associate with w the following variational problem: Find $u=U_\varepsilon(w)\in W(0,T)$ satisfying

$$\begin{aligned}\left\langle\frac{\partial u(t)}{\partial t},\psi\right\rangle&+\int_\Omega\left[\varepsilon\left(\nabla u(t),\nabla\psi\right)+\left(A^\varepsilon_w(t,x,\nabla u(t)),\nabla\psi\right)+\kappa u(t)\psi\right]dx\\ &=\kappa\int_\Omega\left(f(t)-v(t)\right)\psi\,dx,\quad\forall\,\psi\in W^{1,2}(\Omega)\quad\text{a.e. in }[0,T],\end{aligned}\tag{3.19}$$

$$u(0)=f_0.\tag{3.20}$$

Since

$$\text{the condition } w\in W(0,T)\text{ implies } w\in C([0,T];L^2(\Omega))\tag{3.21}$$

(see [29, Chapter XVIII]), it follows from Lemma 2.1 that the corresponding exponent

$$p_w:=1+g\left(\frac{1}{h}\int_{t-h}^t\left|(\nabla G_\sigma*w(\tau,\cdot))\right|^2\,d\tau\right)$$

is such that $p_w\in C^{0,1}(Q_T)$ and $1<p^-\leqslant q(\cdot,\cdot)\leqslant p^+$ in $\overline{Q_T}$, with $p^+=2$ and $p^-=1+\delta$, where (see the proof of Lemma 2.1)

$$\delta=g\left(\|G_\sigma\|_{C^1(\overline{\Omega-\Omega})}\frac{1}{h}\|w\|_{L^1(0,T;L^1(\Omega))}\right).$$

Taking into account that the anisotropic diffusion tensor R_η satisfies the two-side inequality (2.14), it is easy to deduce that, for a given value $\varepsilon>0$, the principle operator $B : L^2(0,T;W^{1,2}(\Omega)) \to L^2(0,T;\big(W^{1,2}(\Omega)\big)')$, defined by the rule

$$\langle Bu, q\rangle = \int_{Q_T} (\varepsilon\nabla u + A^\varepsilon_w(t,x,\nabla u), \nabla q)\; dxdt + \kappa \int_{Q_T} uq\, dxdt,$$

is coercive, monotone, and hemicontinuous, where the hemicontinuity property means the continuity of the scalar function

$$\begin{aligned} z(\lambda) &= \langle B(u+\lambda q), \varphi\rangle \\ &= \int_{Q_T} (\varepsilon(\nabla u + \lambda\nabla q) + A^\varepsilon_w(t,x,\nabla u + \lambda\nabla q), \nabla\varphi)\; dxdt \\ &\quad + \kappa \int_{Q_T} (u+\lambda q)\varphi\, dxdt, \quad \forall\, u, q, \varphi \in L^2(0,T;W^{1,2}(\Omega)) \end{aligned}$$

at the point $\lambda = 0$. Since A^ε_w is a Carathéodory functions, this property can be easily derived with the Lebesgue theorem and the following estimate

$$\begin{aligned} &|A^\varepsilon_w(t,x,\nabla u + \lambda\nabla q)||\nabla\varphi| \\ &\quad\leqslant \frac{1}{p'_w(t,x)}|A^\varepsilon_w(t,x,\nabla u+\lambda\nabla q)|^{p'_w(t,x)} + \frac{1}{p_w(t,x)}|\nabla\varphi|^{p_w(t,x)} \\ &\quad\overset{\text{by (2.2)}}{\leqslant} \frac{c_0}{2}\left(|\nabla u + \lambda\nabla q| + \varepsilon\right)^{p_w(t,x)-2} |\nabla u + \lambda\nabla q|^2 + \frac{1}{p^-}|\nabla\varphi|^{p_w(t,x)} \\ &\quad\leqslant c_1\left(|\nabla u|^{p_w(t,x)} + |\nabla q|^{p_w(t,x)} + 1\right) + \frac{1}{p^-}|\nabla\varphi|^{p_w(t,x)} \in L^1(Q_T). \end{aligned} \tag{3.22}$$

Hence, by the classical results on parabolic equations [46] (see also results of Alkhutov and Zhikov [4, 5]), we deduce that the problem (3.19)–(3.20) has a unique weak solution $U_\varepsilon(w) \in W(0,T)$ in the sense of distributions. Since the integral identity (3.19) is valid for all test functions $\psi = \psi(t,x)$ which are stepwise with respect to variable t, it follows that this identity remains true for all $\psi \in L^2(0,T;W^{1,2}(\Omega))$, and hence for all $\psi \in W^{1,2}(Q_T)$ such that $\psi(T,\cdot) = 0$. So, after integration by parts, one can easily deduces from (3.19) that the solution $U_\varepsilon(w)$ satisfies both the integral identity

$$\begin{aligned} &\int_{Q_T} \left(-U_\varepsilon(w)\frac{\partial\varphi}{\partial t} + (\varepsilon\nabla U_\varepsilon(w) + A^\varepsilon_w(t,x,\nabla U_\varepsilon(w)), \nabla\varphi) + \kappa U_\varepsilon(w)\varphi\right) dxdt \\ &\qquad\qquad = \kappa\int_{Q_T} (f-v)\varphi\, dxdt + \int_\Omega f_0\varphi|_{t=0}\, dx \quad \forall\, \varphi \in \Phi \end{aligned} \tag{3.23}$$

and the energy equality

$$\begin{aligned} &\frac{1}{2}\int_\Omega U^2_\varepsilon(w)\, dx \\ &\quad + \int_0^t \int_\Omega \left(\varepsilon|\nabla U_\varepsilon(w)|^2 + (A^\varepsilon_w(s,x,\nabla U_\epsilon(w)), \nabla U_\varepsilon(w)) + \kappa U^2_\varepsilon(w)\right) dxds \\ &\quad = \kappa\int_0^t \int_\Omega (f-v)U_\varepsilon(w)\, dxds + \int_\Omega f_0^2\, dx, \quad \forall\, t \in [0,T], \end{aligned} \tag{3.24}$$

where, in view of (3.21), the first term in (3.24) is well defined for each $t \in [0, T]$.

Step 2. Using (3.24), we see that

$$\begin{aligned}\frac{1}{2}\int_\Omega U_\varepsilon^2(w)\,dx + \kappa\int_0^t\int_\Omega U_\varepsilon^2(w)\,dxds&\\ \leqslant \frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)} + \frac{\kappa}{2}\|U_\epsilon(w)\|^2_{L^2(Q_T)} + \|f_0\|^2_{L^2(\Omega)}.&\end{aligned}$$

From this, (3.24), and (3.11), we derive the following estimates:

$$\begin{aligned}
&\|U_\varepsilon(w)\|^2_{L^2(Q_T)} \leqslant \|f-v\|^2_{L^2(Q_T)} + \frac{2}{\kappa}\|f_0\|^2_{L^2(\Omega)} =: C_1^2, && (3.25)\\
&\|\nabla U_\varepsilon(w)\|_{L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)} \overset{\text{by (2.21)}}{\leqslant} \left(\int_{Q_T}|\nabla U_\varepsilon(w)|^{p_w(t,x)}\,dxdt + 1\right)^{1/p^-}\\
&\qquad \leqslant \left(\Lambda^{-1}\left(\|f_0\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)} + \frac{\kappa}{2}\|U_\epsilon(w)\|^2_{L^2(Q_T)}\right) + 5\right)^{1/p^-}\\
&\qquad \overset{\text{by (3.25)}}{\leqslant} \left(\Lambda^{-1}\kappa C_1^2 + 5\right)^{1/p^-} =: C_2, && (3.26)\\
&\|U_\varepsilon(w)\|_{L^\infty(0,T;L^2(\Omega))} \leqslant \sqrt{2\left(\|f_0\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)} + \frac{\kappa}{2}\|U_\varepsilon(w)\|^2_{L^2(Q_T)}\right)}\\
&\qquad \leqslant \sqrt{2\kappa}C_1, && (3.27)\\
&\|\nabla U_\varepsilon(w)\|_{L^2(Q_T;\mathbb{R}^2)} \leqslant \frac{1}{\sqrt{\varepsilon}}\sqrt{\|f_0\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)} + \frac{\kappa}{2}\|U_\varepsilon(w)\|^2_{L^2(Q_T)}}\\
&\qquad \leqslant \sqrt{\frac{\kappa}{\varepsilon}}C_1, && (3.28)\\
&\int_{Q_T}|A^\varepsilon_w(t,x,\nabla U_\varepsilon(w))|^{p'_w(t,x)}\,dxdt\\
&\qquad \overset{\text{by (3.25)}}{\leqslant} \left(\sqrt{2}\right)^{p'_w(t,x)}\int_{Q_T}\left(|\nabla U_\varepsilon(w)| + \varepsilon\right)^{p_w(t,x)}\,dxdt\\
&\qquad \overset{\text{by (3.26)}}{<} +\infty. && (3.29)
\end{aligned}$$

We also notice that there exists a constant $C_3 > 0$ such that

$$\begin{aligned}\left|\left\langle\frac{\partial U_\varepsilon(w)}{\partial t},\psi\right\rangle\right| \overset{\text{by (3.19)}}{\leqslant} &\sqrt{\varepsilon}\|\nabla U_\varepsilon(w)\|_{L^2(Q_T;\mathbb{R}^2)}\|\nabla\psi\|_{L^2(Q_T;\mathbb{R}^2)}\\ &+ 2\|A^\varepsilon_w(t,x,\nabla U_\varepsilon(w))\|_{L^{p'_w(\cdot)}(Q_T;\mathbb{R}^2)}\|\nabla\psi\|_{L^{p_w(\cdot)}(Q_T;\mathbb{R}^2)}\\ &+ \kappa\|U_\varepsilon\|_{L^2(Q_T)}\|\psi\|_{L^2(Q_T)} + \kappa\|f-v\|_{L^2(Q_T)}\|\psi\|_{L^2(Q_T)}\end{aligned}$$

$$\begin{aligned}
&\overset{\text{by (2.24)}}{\leqslant} \Big[\sqrt{\varepsilon}\|\nabla U_\varepsilon(w)\|_{L^2(Q_T;\mathbb{R}^2)} + \kappa\|U_\varepsilon\|_{L^2(Q_T)} + \kappa\|f-v\|_{L^2(Q_T)}\Big]\\
&\quad \times \|\psi\|_{L^2(0,T;W^{1,2}(\Omega))}\\
&\quad + \left(1 + \int_{Q_T} |A^\varepsilon_w(t,x,\nabla U_\varepsilon(w))|^{p'_w(t,x)}\,dxdt\right)^{1/2}\\
&\quad \times (1+T|\Omega|)^{1/2}\,\|\psi\|_{L^2(Q_T)}\\
&\overset{\text{by (3.25)–(3.29)}}{\leqslant} C_3\|\psi\|_{L^2(0,T;W^{1,2}(\Omega))}, \quad \forall\,\psi \in L^2(0,T;W^{1,2}(\Omega)). \qquad (3.30)
\end{aligned}$$

Hence,

$$\left\|\frac{\partial U_\varepsilon(w)}{\partial t}\right\|_{L^2(0,T;(W^{1,2}(\Omega))')} \leqslant C_3. \qquad (3.31)$$

Taking into account these estimates, we introduce the following subset W_0 of the space $W(0,T)$

$$W_0 = \left\{ z \in W(0,T) \;\middle|\; \begin{array}{c} \|z\|_{L^2(0,T;W^{1,2}(\Omega))} \leqslant \left(1+\sqrt{\frac{\kappa}{\varepsilon}}\right) C_1,\\ \|z\|_{L^\infty(0,T;L^2(\Omega))} \leqslant \sqrt{2\kappa}C_1,\\ \|\frac{\partial z}{\partial t}\|_{L^2(0,T;(W^{1,2}(\Omega))')} \leqslant C_3,\\ z(0,\cdot) = f_0 \end{array} \right\}$$

In view of estimates (3.25)–(3.31) and condition (3.18), it is clear that $w \in W_0$ and, hence, U_ϵ can be interpreted as a mapping from W_0 into W_0. Moreover, we see that W_0 is a nonempty, convex, and weakly compact subset of $W(0,T)$. Moreover, in view of the fact that the embedding of $W^{1,2}(\Omega)$ in $L^2(\Omega)$ is compact, a refinement of Aubin's lemma (see, e.g. [53, Section 8, Corollary 4] ensures that any bounded subset of $W(0,T)$ is relatively compact in $L^2(Q_T)$. So, in order to apply the Schauder fixed-point theorem, it remains to show that the mapping U_ε is weakly continuous from W_0 into W_0. As a result, the Schauder fixed-point theorem will provide the existence of element u_ε in W_0 such that $u_\varepsilon = U_\epsilon(u_\varepsilon)$.

Step 3. Let $\{w_j\}_{j\in\mathbb{R}}$ be a sequence in W_0 converging weakly in W_0 to some $w \in W_0$. Setting $u_{\varepsilon,j} = U_\varepsilon(w_j)$ and utilizing the weak compactness of the set W_0 and the Aubin's lemma, we see that $\{u_{\varepsilon,j}\}_{j\in\mathbb{R}}$ contains a subsequence such that

$$\begin{aligned}
u_{\varepsilon,j} &\rightharpoonup u_\varepsilon \quad \text{weakly in } L^2(0,T;W^{1,2}(\Omega)), && (3.32)\\
u_{\varepsilon,j} &\to u_\varepsilon \quad \text{strongly in } L^2(0,T;L^2(\Omega)), && (3.33)\\
\frac{\partial u_{\varepsilon,j}}{\partial t} &\rightharpoonup \frac{\partial u_\varepsilon}{\partial t} \quad \text{weakly in } L^2(0,T;\big(W^{1,2}(\Omega)\big)'), && (3.34)\\
u_{\varepsilon,j} &\to u_\varepsilon \quad \text{strongly in } L^2(0,T;L^2(\Omega)) \text{ and a.e. in } Q_T, && (3.35)\\
\frac{\partial u_{\varepsilon,j}}{\partial x_i} &\rightharpoonup \frac{\partial u_\varepsilon}{\partial x_i} \quad \text{weakly in } L^2(0,T;L^2(\Omega)), && (3.36)\\
w_j &\to w \quad \text{strongly in } L^2(0,T;L^2(\Omega)). && (3.37)
\end{aligned}$$

Then Lemma 2.1 implies that

$$p_{w_j}(t,x) \to p_w(t,x) \quad \text{uniformly in } \overline{Q_T} \text{ as } j \to \infty. \tag{3.38}$$

Moreover, taking into account that

$$\begin{aligned}
\|A^\varepsilon_{w_j}(t,x,\nabla u_{\varepsilon,j})\|^{(p^+)'}_{L^{(p^+)'}(Q_T;\mathbb{R}^2)} &\overset{\text{by (2.23)}}{\leqslant} (1+T|\Omega|)\, \|A^\varepsilon_{w_j}(t,x,\nabla u_{\varepsilon,j})\|^{(p^+)'}_{L^{p'_{w_j}(\cdot)}(Q_T;\mathbb{R}^2)} \\
&\overset{\text{by (2.21)}}{\leqslant} (1+T|\Omega|)\left(1+\int_{Q_T} |A_{w_j}(t,x,\nabla u_{\varepsilon,j})|^{p'_{w_j}(t,x)}\,dxdt\right) \\
&\leqslant C\left(1+\int_{Q_T} |\nabla u_{\varepsilon,j}|^{p_{w_j}}\,dxdt\right) \overset{\text{by (3.26)}}{<} \infty,
\end{aligned} \tag{3.39}$$

we deduce from (3.28) and (2.17) that the sequence $\left\{\varepsilon\nabla u_{\varepsilon,j} + A^\varepsilon_{w_j}(t,x,\nabla u_{\varepsilon,j})\right\}_{j\in\mathbb{R}}$ is bounded in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$. Hence, we can suppose that there exists an element $z \in L^{(p^+)'}(Q_T;\mathbb{R}^2)$ such that

$$\varepsilon\nabla u_{\varepsilon,j} + A^\varepsilon_{w_j}(t,x,\nabla u_{\varepsilon,j}) \rightharpoonup z \quad \text{weakly in } L^{(p^+)'}(Q_T;\mathbb{R}^2) \text{ as } j \to \infty. \tag{3.40}$$

Utilizing this fact together with the properties

$$\begin{aligned}
u_{\varepsilon,j} \rightharpoonup u_\varepsilon \ \text{ in } L^{p^-}(0,T;W^{1,p^-}(\Omega)) \text{ with } p^- = 1+\delta \text{ (by (3.32))}, \\
\{u_{\varepsilon,j}\}_{j\in\mathbb{N}} \ \text{ are bounded in } L^\infty(0,T;L^2(\Omega)) \text{ (by (3.27))}, \\
u_{\varepsilon,j} \in L^{p^+}(0,T;W^{1,p^+}(\Omega))\ \forall\, j\in\mathbb{N} \ \text{ by (3.28)}, \\
\sup_{j\in\mathbb{N}} \|\left(A^\varepsilon_{w_j}(t,x,\nabla u_{\varepsilon,j}), \nabla u_{\varepsilon,j}\right)\|_{L^1(Q_T)} < \infty \text{ (by (3.22))},
\end{aligned} \tag{3.41}$$

and taking into account that $1 < 1+\delta = p^- < p^+ = 2 < 2p^-$, we deduce from Theorem 2.1 that the flow $A^\varepsilon_{w_j}(t,x,\nabla u_{\epsilon,j})$ weakly converges in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ to the flow $A^\varepsilon_w(t,x,\nabla u_\varepsilon)$, i.e., $z = A^\varepsilon_w(t,x,\nabla u_\varepsilon)$.

Then we can pass to the limit in relations (3.19)–(3.20) with $u = u_{\varepsilon,j}$ and $w = w_j$ as $j \to \infty$. This yields

$$\begin{aligned}
\left\langle \frac{\partial u_\varepsilon(t)}{\partial t}, \psi \right\rangle + \int_\Omega [\varepsilon\,(\nabla u_\varepsilon(t), \nabla\psi) + (A^\varepsilon_w(t,x,\nabla u_\varepsilon(t)), \nabla\psi) + \kappa u_\varepsilon(t)\psi]\,dx \\
= \kappa \int_\Omega (f(t) - v(t))\,\psi\,dx, \quad \forall\,\psi \in W^{1,2}(\Omega) \quad \text{a.e. in } [0,T], \quad (3.42) \\
u_\varepsilon(0) = f_0, \quad (3.43)
\end{aligned}$$

i.e., $u_\varepsilon = U_\varepsilon(w)$. Moreover, since variational problem (3.42)–(3.43) has a unique solution, it follows that the entire sequence $\{u_{\varepsilon,j}\}_{j\in\mathbb{R}}$ converges weakly in $W(0,T)$ to $u_\varepsilon = U_\varepsilon(w)$.

Thus, the mapping $U_\varepsilon : W_0 \mapsto W_0$ is weakly continuous and, hence, by the Schauder fixed point theorem, u_ε is a weak solution of the perturbed problem (3.15)–(3.17).

To the end of this proof, let us make use of the following observation: if u_ε is a weak solution to (3.15)–(3.17), then arguing as at the Step 1 and using the integration by parts formula, it is easily to deduce from (3.19) that u_ε satisfies the integral identity

$$\int_{Q_T} \left(-u_\varepsilon \frac{\partial \varphi}{\partial t} + \varepsilon \left(\nabla u_\varepsilon, \nabla \varphi \right) + \left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon), \nabla \varphi \right) + \kappa u_\varepsilon \varphi \right) dxdt \\ = \kappa \int_{Q_T} (f-v)\varphi \, dxdt + \int_\Omega f_0 \varphi|_{t=0} \, dx \quad \forall\, \varphi \in \Phi. \tag{3.44}$$

□

Let us specify some extra properties of the weak solutions u_ε, given by Theorem 3.2.

Corollary 3.1. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $v \in \mathcal{V}_{ad}$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $\varepsilon > 0$ be given. Let $u_\varepsilon \in W(0,T)$ be a weak solution of* (3.15)–(3.17) *in the sense of distributions given by Theorem 3.2. Then $u_\varepsilon \in W_{u_\varepsilon}(Q_T)$ and the following energy equality holds*

$$\frac{1}{2}\int_\Omega u_\varepsilon^2 \, dx + \int_0^t \int_\Omega \left(\varepsilon |\nabla u_\varepsilon|^2 + \left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon), \nabla u_\varepsilon \right) + \kappa u_\varepsilon^2 \right) dxds \\ = \kappa \int_0^t \int_\Omega (f-v) u_\varepsilon \, dxds + \int_\Omega f_0^2 \, dx \quad \forall\, t \in [0,T]. \tag{3.45}$$

Proof. Taking into account the definition of the space $W_{u_\varepsilon}(Q_T)$ (see (3.9)), let us show that

$$\int_{Q_T} |R_\eta \nabla u_\varepsilon|^{p_{u_\varepsilon}(t,x)} \, dxdt < +\infty \quad \text{for a.e. } t \in [0,T].$$

Since u_ε is a solution of (3.15)–(3.17) given by Theorem 3.2, it follows that there exists a sequence $\{u_{\varepsilon,j}\}_{j\in\mathbb{R}} \in W_0$ with properties (3.32)–(3.36) and such that $u_{\varepsilon,j} = U_\varepsilon(u_{\varepsilon,j-1})$, for $j = 2,3,\dots$. Moreover, this sequence possesses the properties (3.40)–(3.41). Hence, $\nabla u_{\varepsilon,j} \in L^1(Q_T;\mathbb{R}^2)$ by (3.32), and

$$\int_{Q_T} |\nabla u_{\varepsilon,j}|^{p_{u_{\varepsilon,j}}} \, dxdt = \frac{1}{\Lambda}\int_{Q_T} \Lambda |\nabla u_{\varepsilon,j}|^{p_{u_{\varepsilon,j}}} \, dxdt > \infty \\ \leqslant \frac{1}{\Lambda} \sup_{j\in\mathbb{N}} \int_{Q_T} \left(\left| \left(A^\varepsilon_{u_{\varepsilon,j-1}}(t,x,\nabla u_{\varepsilon,j}), \nabla u_{\varepsilon,j} \right) \right| + 4 \right) dxdt \\ \overset{\text{by (3.41)}}{<} \infty. \tag{3.46}$$

Therefore,

$$\int_{Q_T} |R_\eta \nabla u_{\varepsilon,j}|^{p_{u_{\varepsilon,j}}(t,x)}\, dxdt \leqslant \int_{Q_T} |\nabla u_{\varepsilon,j}|^{p_{u_{\varepsilon,j}}(t,x)}\, dxdt < +\infty. \tag{3.47}$$

Then, by Proposition 2.2 and property $(3.41)_1$, we have:

$$\int_{Q_T} |R_\eta \nabla u_\varepsilon|^{p_{u_\varepsilon}(t,x)}\, dxdt \leqslant \liminf_{j\to\infty} \int_{Q_T} |R_\eta \nabla u_{\varepsilon,j}|^{p_{u_{\varepsilon,j}}(t,x)}\, dxdt < \infty.$$

Thus, $u_\varepsilon \in W_{u_\varepsilon}(Q_T)$.

It remains to prove the energy equality (3.45). Since u_ε is in $W(0,T)$ and the set of test functions $C^\infty([0,T];C_c^\infty(\mathbb{R}^N))$ is dense in $L^2(0,T;W^{1,2}(\Omega))$, it follows that there exists a sequence $\{\varphi_j\}_{j\in\mathbb{N}} \subset C^\infty([0,T];C_c^\infty(\mathbb{R}^N))$ such that

$$\varphi_j \to u_\varepsilon \quad \text{in } L^2(0,T;W^{1,2}(\Omega)) \quad \text{as } j\to\infty. \tag{3.48}$$

Taking into account that, for each $j\in\mathbb{N}$, the integral identity

$$\begin{aligned}&\int_0^t \int_\Omega \left[\varepsilon\left(\nabla u_\varepsilon(t), \nabla\varphi_j(t)\right) + \left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon(t)), \nabla\varphi_j(t)\right) + \kappa u_\varepsilon(t)\varphi_j(t)\right] dxdt \\ &+\int_0^t \left\langle \frac{\partial u_\varepsilon(t)}{\partial t}, \varphi_j(t)\right\rangle dt = \kappa \int_0^t\int_\Omega \left(f(t)-v(t)\right)\varphi_j(t)\, dxdt, \quad \forall\, j\in\mathbb{N},\ \forall\, t\in[0,T]\end{aligned} \tag{3.49}$$

holds true, we can pass to the limit in (3.49) as $j\to\infty$. To do so, we notice that

$$\begin{aligned}\int_0^t \int_\Omega &\left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon(t)), \nabla\varphi_j(t)\right) dxdt \\ &= \int_0^t\int_\Omega \left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon(t)), \nabla u_\varepsilon(t)\right) dxdt \\ &\quad + \int_0^t\int_\Omega \left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon(t)), \nabla\varphi_j(t) - \nabla u_\varepsilon(t)\right) dxdt,\end{aligned}$$

where $\nabla\varphi_j - \nabla u_\varepsilon \to 0$ a.e. in Q_T by (3.48), and

$$\left|\left(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon), \nabla\varphi_j - \nabla u_\varepsilon\right)\right| \leqslant c_1|\nabla u_\varepsilon|^{p_{u_\varepsilon}} + \frac{1}{p^-}|\nabla\varphi_j - \nabla u_\varepsilon|^{p_{u_\varepsilon}} + c_1 \in L^1(Q_T)$$

by (3.22). Hence, by the Lebesgue dominated theorem, the limit passage in (3.49) leads to the equality

$$\begin{aligned}&\int_0^t \int_\Omega \left[\varepsilon\left(\nabla u_\varepsilon(t), \nabla u_\varepsilon(t)\right) + \left(A^\varepsilon_{w}(t,x,\nabla u_\varepsilon(t)), \nabla u_\varepsilon(t)\right) + \kappa u^2_\varepsilon(t)\right] dxdt \\ &\quad+\int_0^t \left\langle \frac{\partial u_\varepsilon(t)}{\partial t}, u_\varepsilon(t)\right\rangle dt = \kappa \int_0^t\int_\Omega \left(f(t)-v(t)\right)u_\varepsilon(t)\, dxdt, \quad \forall\, t\in[0,T].\end{aligned} \tag{3.50}$$

Thus, to obtain the energy equality (3.45), it remains to apply the integration by parts formula. □

For our further analysis, we make use of the following result.

Lemma 3.1. *Let* $\{u_\varepsilon\}_{\varepsilon\to 0}\subset W(0,T)$ *be a sequence such that*

$$\sup_{\varepsilon\to 0}\left(\varepsilon\int_{Q_T}|\nabla u_\varepsilon|^2\,dxdt\right)<+\infty. \tag{3.51}$$

Then $\varepsilon\nabla u_\varepsilon\rightharpoonup 0$ *in* $L^2(Q_T;\mathbb{R}^N)$.

Proof. Let $\varphi\in C_0^\infty(Q_T)$ be an arbitrary vector-function. Then

$$\left|\int_{Q_T}(\varepsilon\nabla u_\varepsilon,\varphi)\,dxdt\right|\leqslant\sqrt{\varepsilon}\left(\int_{Q_T}\varepsilon|\nabla u_\varepsilon|^2\,dxdt\right)^{1/2}\left(\int_{Q_T}|\varphi|^2\,dxdt\right)^{1/2}.$$

Hence, the sequence $\{\varepsilon\nabla u_\varepsilon\}_{\varepsilon\to 0}$ is bounded in $L^2(Q_T;\mathbb{R}^N)$. As a result, we have

$$\left|\int_{Q_T}(\varepsilon\nabla u_\varepsilon,\varphi)\,dxdt\right|\overset{\text{by (3.51)}}{\leqslant}C\sqrt{\varepsilon}\left(\int_{Q_T}\varepsilon|\nabla u_\varepsilon|^2\,dxdt\right)^{1/2}\leqslant\widehat{C}\sqrt{\varepsilon}\to 0.$$

The proof is complete. □

We are now in a position to prove the main result of this section.

Theorem 3.3. *Let* $f\in L^2(Q_T)$, $f_0\in L^2(\Omega)$, *and* $\theta\in L^\infty(\Omega;\mathbb{R}^2)$ *be given distributions. Then, for each* $v\in\mathcal{V}_{ad}$, *the initial-boundary value problem* (3.1)–(3.3) *admits at least one weak solution* $u\in W_u(Q_T)$.

Proof. Let ε be a small parameter which varies within a strictly decreasing sequence of positive numbers converging to 0. Let $\{u_\varepsilon\in W(0,T)\}_{\varepsilon\to 0}$ be a sequence of weak solutions to the approximating problem (3.15)–(3.17) given by Theorem 3.2. Then, for each $\varepsilon>0$, u_ε satisfies the energy equality (3.45). Hence, we can deduce from (3.45) the following estimates

$$\sup_{\varepsilon>0}\|u_\varepsilon\|^2_{L^2(Q_T)}\leqslant C_1^2=\|f-v\|^2_{L^2(Q_T)}+\frac{2}{\kappa}\|f_0\|^2_{L^2(\Omega)}, \tag{3.52}$$

$$\begin{aligned}\sup_{\varepsilon>0}\|\nabla u_\varepsilon\|_{L^{p_{u_\varepsilon}(\cdot)}(Q_T;\mathbb{R}^2)}&\overset{\text{by (2.21)}}{\leqslant}\sup_{\varepsilon>0}\left(\int_{Q_T}|\nabla u_\varepsilon|^{p_{u_\varepsilon}(t,x)}\,dxdt+1\right)^{1/p^-}\\&\overset{\text{by (3.45)}}{\leqslant}\sup_{\varepsilon>0}\left(\Lambda^{-1}\left(\|f_0\|^2_{L^2(\Omega)}+\frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)}+\frac{\kappa}{2}\|u_\varepsilon\|^2_{L^2(Q_T)}\right)+5\right)^{1/p^-}\\&\overset{\text{by (3.52)}}{\leqslant}C_2=\left(\Lambda^{-1}\kappa C_1^2+5\right)^{1/p^-},\end{aligned} \tag{3.53}$$

$$\begin{aligned}\sup_{\varepsilon>0}\|\nabla u_\varepsilon\|_{L^{p^-}(Q_T;\mathbb{R}^2)} &\overset{\text{by (2.23)}}{\leqslant} \sup_{\varepsilon>0}\left(1+T|\Omega|\right)^{1/p^-}\|\nabla u_\varepsilon\|_{L^{p_{u_\varepsilon}(\cdot)}(Q_T;\mathbb{R}^N)}\\ &\overset{\text{by (3.53)}}{\leqslant} \left(1+T|\Omega|\right)^{1/p^-}C_2, \end{aligned}\tag{3.54}$$

$$\begin{aligned}\sup_{\varepsilon>0}\|u_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} &\leqslant \sup_{\varepsilon>0}\sqrt{2\left(\|f_0\|^2_{L^2(\Omega)}+\frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)}+\frac{\kappa}{2}\|u_\varepsilon\|^2_{L^2(Q_T)}\right)}\\ &\leqslant \sqrt{2\kappa}C_1,\end{aligned}\tag{3.55}$$

$$\begin{aligned}\|\nabla u_\varepsilon\|_{L^2(Q_T;\mathbb{R}^2)} &\leqslant \frac{1}{\sqrt{\varepsilon}}\sqrt{|f_0\|^2_{L^2(\Omega)}+\frac{\kappa}{2}\|f-v\|^2_{L^2(Q_T)}+\frac{\kappa}{2}\|u_\varepsilon\|^2_{L^2(Q_T)}}\\ &\leqslant \sqrt{\frac{\kappa}{\varepsilon}}C_1.\end{aligned}\tag{3.56}$$

Taking this into account, we see that the sequence $\{u_\varepsilon\}_{\varepsilon\to 0}$ is bounded in the spaces $L^\infty(0,T;L^2(\Omega))$ and $L^{p^-}(0,T;W^{1,p^-}(\Omega))$. Therefore, there exists an element

$$u\in L^{p^-}(0,T;W^{1,p^-}(\Omega))\cap L^\infty(0,T;L^2(\Omega))\tag{3.57}$$

such that, up to a subsequence, $u_\varepsilon \rightharpoonup u$ in $L^{p^-}(0,T;W^{1,p^-}(\Omega))$ as $\varepsilon\to 0$. Moreover, the uniform boundedness of the fluxes $\left\{A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon)\right\}_{\varepsilon\to 0}$ in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ with respect to $\varepsilon>0$ implies that this sequence is sequentially weakly compact in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ (for arguments see (3.39)). Hence, we may admit the existence of a vector-function w such that $w_\varepsilon = A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon)\rightharpoonup w$ in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ as $\varepsilon\to 0$. Arguing as in the proof of Theorem 3.2, it can be shown that $\sup_{\varepsilon\to 0}\|\left(w_\varepsilon,\nabla u_\varepsilon\right)\|_{L^1(Q_T)}<\infty$. As a result, Theorem 2.1 implies that the flow $A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon)$ weakly converges in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ to the flow $w=A_u(t,x,\nabla u)$. Then, utilizing Lemma 3.1, we see that the passage to the limit in the integral identity (3.44) leads to a similar identity for equation (3.1). It remains to take into account Lemma 2.4 and relations (3.46)–(3.47) in order to deduce that $\int_{Q_T}|R_\eta\nabla u|^{p_u(t,x)}\,dxdt<+\infty$. Thus, u is an element of the space $W_u(Q_T)$ and, as a consequence, u is a weak solution to the problem (3.1)–(3.3). □

Before proceeding further, it is worth to notice that the uniqueness of weak solutions to the perturbed problem (3.15)–(3.17) and, hence, for the original one (3.1)–(3.3), seems to be an open question. In view of this, we adopt the following concept:

Definition 3.2. We say that a weak solution $u\in W_u(Q_T)$ to the problem (3.1)–(3.3), for given distributions $f\in L^2(Q_T)$, $f_0\in L^2(\Omega)$, $\theta\in L^\infty(\Omega;\mathbb{R}^2)$, and $v\in\mathcal{V}_{ad}$, is W_0-attainable if there exists a sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ converging to zero as $n\to\infty$ and such that

$$\left.\begin{array}{r}u_n\rightharpoonup u\ \text{ in }\ L^{p^-}(0,T;W^{1,p^-}(\Omega)),\\ A^{\varepsilon_n}_{u_{n-1}}(t,x,\nabla u_n)\rightharpoonup A_u(t,x,\nabla u)\ \text{ in }\ L^{(p^+)'}(Q_T;\mathbb{R}^2)\end{array}\right.\ \text{as } n\to\infty,\tag{3.58}$$

where, for each $n \in \mathbb{N}$, u_n is the weak solution to the following perturbed problem

$$\frac{\partial u}{\partial t} - \varepsilon_n \Delta u - \operatorname{div} A^{\varepsilon_n}_{u_{n-1}}(t,x,\nabla u) + \kappa u = \kappa(f-v) \quad \text{in } \ Q_T, \tag{3.59}$$

$$\partial_\nu u = 0 \quad \text{on } \ (0,T)\times\partial\Omega, \tag{3.60}$$

$$u(0,\cdot) = f_0 \quad \text{in } \ \Omega. \tag{3.61}$$

We can supplement the result on Theorem 3.3 with the following assertions.

Corollary 3.2. *Let $u \in W_u(Q_T)$ be a weak solution to the problem* (3.1)–(3.3) *that has been obtained as a cluster point of the weak solutions $\{u_\epsilon \in W(0,T)\}$ to the approximating problems* (3.15)–(3.17)*. Then the following energy inequality*

$$\frac{1}{2}\int_\Omega u^2\,dx + \int_0^t\int_\Omega \left((A_u(t,x,\nabla u),\nabla u) + \kappa u^2\right)\,dxdt \leqslant \kappa\int_0^t\int_\Omega (f-v)u\,dxdt + \int_\Omega f_0^2\,dx \tag{3.62}$$

holds true for almost all $t \in [0,T]$.

Proof. To deduce this inequality, we make use of the estimate (3.31) and the celebrated Aubin's lemma. As a result, we can supplement the properties (3.58) by the following one: $u_\varepsilon \to u$ in $L^2(0,T;L^2(\Omega))$ as $\varepsilon \to 0$. So, without loss of generality, we can suppose that $u_\varepsilon(t,x) \to u(t,x)$ almost everywhere in Q_T. Hence,

$$\|u_\varepsilon(t,\cdot)\|^2_{L^2(\Omega)} \to \|u(t,\cdot)\|^2_{L^2(\Omega)} \text{ for a.a } \ t \in [0,T]. \tag{3.63}$$

Taking this fact into account and passing to the limit in relation (3.45) as $\varepsilon \to 0$ using the weak convergence $u_\varepsilon \rightharpoonup u$ in $L^{p^-}(0,T;W^{1,p^-}(\Omega))$, Lemma 2.3, and the weak convergence of fluxes to flux (Theorem 2.1), we arrive at the announced inequality (3.62). □

Remark 3.1. It is worth to emphasize that Theorem 3.3 can be now specified as follows: For given $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $v \in \mathcal{V}_{ad}$, the initial-boundary value problem (3.1)–(3.3) admits at least one W_0-attainable weak solution $u \in W_u(Q_T)$ for which the energy inequality (3.62) holds true for all $t \in [0,T]$. Moreover, as follows from estimates (3.52)–(3.55), this solution in bounded in $L^{p^-}(0,T;W^{1,p^-}(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$. However, the mapping $t \mapsto \|u(t,\cdot)\|_{L^2(\Omega)}$ is not necessary continuous. Because of that the second term $\frac{1}{2}\int_\Omega |u(T)-f_0|^2\,dx$ in the cost functional is not well defined (see Proposition 3.1 for the details). This means that the original OCP (1.1)–(1.5) is generally ill posed, and some its relaxation is required.

4. Setting of the Relaxed Optimal Control Problem and Existence Result

As was pointed out in the previous section, the operator $-\operatorname{div} A_u(t,x,\nabla u)+\kappa u$ provides an example of a non-linear operator in divergence form which is neither monotone nor coercive. In this case (see Theorem 3.3) the initial-boundary value problem (3.1)–(3.3) admits a W_0-attainable weak solution that satisfies the energy inequality (3.62). However, it is unknown whether under some admissible control $v\in\mathcal{V}_{ad}$ this solution is unique and belongs to the space $C([0,T];L^2(\Omega))$. Moreover, it is an open question whether all weak solutions to (3.1)–(3.3) satisfy energy inequality (3.62) that plays a crucial role for derivation of a priori estimates like (3.52)–(3.55).

Our prime interest in this section is to study the existence issues for the following relaxed version of the original optimal control problem

$$\begin{gathered}\text{Minimize } J(v,u)=\|v\|^2_{L^2(0,T;L^1(\Omega))}+\frac{\mu}{2\omega}\int_{T-\omega}^{T}\|u(t,\cdot)-f_0(\cdot)\|^2_{L^2(\Omega)}\,dt\\ \text{subject to the constraints (1.2)–(1.4), (3.5),}\end{gathered}\tag{4.1}$$

where ω is a small positive value such that $T-\omega\gg 0$, $f\in L^2(\Omega)$, $f_0\in L^2(\Omega)$, $v\in\mathcal{V}_{ad}$, and $\theta\in L^\infty(\Omega;\mathbb{R}^2)$ are given distributions.

In image processing, the distributions $f\in L^2(\Omega)$ and $f_0\in L^2(\Omega)$ are usually related to some noise-corrupted image. For instance, $f\in L^2(\Omega)$ is the original gray-scale image with noise, whereas $f_0\in L^2(\Omega)$ is the pre-denoised image by applying a median filter to f. In this case $\theta\in L^\infty(\Omega;\mathbb{R}^2)$ can be stood for the vector field of unit normals to the topographic map of a smoothed version of function f_0. So, instead of E_I in (2.11), we can take

$$\left(G_\sigma * f_0\right)(x)=\int_\Omega G_\sigma(x-y)f_0(y)\,dy,\quad \forall\, x\in\Omega.$$

We say that $(v.u)$ is a feasible pair to OCP (4.1) if:

$$\left.\begin{gathered}v\in\mathcal{V}_{ad},\quad u\in W_u(Q_T),\quad J(v,u)<+\infty,\\ (v,u)\text{ are related by integral identity (3.10) and energy inequality (3.62),}\\ \text{and } u \text{ is a } W_0\text{-attainable weak solution to (3.1)–(3.3) for the given } v.\end{gathered}\right\}\tag{4.2}$$

Let $\Xi\subset L^2(Q_T)\times W_u(Q_T)$ be the set of all feasible solutions to the problem (4.1). Then Theorem 3.3 implies that $\Xi\neq\emptyset$. Since the structure of the set Ξ and its main topological properties are unknown, we begin with the following observation.

Theorem 4.1. *For given distributions $f\in L^2(Q_T)$, $f_0\in L^2(\Omega)$, and $\theta\in L^\infty(\Omega;\mathbb{R}^2)$, the set Ξ is sequentially closed with respect to the weak topology of $L^2(Q_T)\times L^{p^-}(0,T;W^{1,p^-}(\Omega))$.*

Proof. Let $\{(v_k,u_k)\}_{k\in\mathbb{N}}\subset\Xi$ be a sequence such that

$$v_k\rightharpoonup v\ \text{ in }\ L^2(Q_T),\quad u_k\rightharpoonup u\ \text{ in }\ L^{p^-}(0,T;W^{1,p^-}(\Omega)). \tag{4.3}$$

Since the set $\mathcal{V}_{ad}$ is convex and closed, it follows by Mazur's theorem that $\mathcal{V}_{ad}$ is sequentially closed with respect to the weak topology of $L^2(Q_T)$. Therefore, $v\in\mathcal{V}_{ad}$. Our aim is to show that $(v,u)\in\Xi$. We will do it in several steps.

Step 1. By the initial assumptions, for each $k\in\mathbb{N}$, the pair (v_k,u_k) satisfy the energy inequality (3.62), and u_k is a W_0-attainable weak solution of (3.1)–(3.3). Hence, we may suppose that there exists a sequence $\{u_{k,n}\}_{n\in\mathbb{N}}\subset W(0,T)$ such that $\{u_{k,n}\}_{n\in\mathbb{N}}$ are the weak solutions (in the sense of distributions) of (3.59)–(3.61) with $\varepsilon_n=1/n$ and $v=v_k$. and

$$u_{k,n}\rightharpoonup u_k\ \text{ in }\ L^{p^-}(0,T;W^{1,p^-}(\Omega)),\quad \text{as } n\to\infty, \tag{4.4}$$

$$A^{1/n}_{u_{k,n-1}}(t,x,\nabla u_{k,n})\rightharpoonup A_{u_k}(t,x,\nabla u_k)\ \text{ in }\ L^{(p^+)'}(Q_T;\mathbb{R}^2)\quad \text{as } n\to\infty, \tag{4.5}$$

Moreover, the fact that the energy equality

$$\frac{1}{2}\int_\Omega u_{k,n}^2\,dx+\int_0^t\int_\Omega\left(\frac{1}{n}|\nabla u_{k,n}|^2+\left(A^{1/n}_{u_{k,n-1}}(t,x,\nabla u_{k,n}),\nabla u_{k,n}\right)+\kappa u_{k,n}^2\right)dxdt$$
$$=\kappa\int_{Q_T}(f-v_k)u_{k,n}\,dxdt+\int_\Omega f_0^2\,dx,\quad\forall\,t\in[0,T] \tag{4.6}$$

holds true for all $n,k\in\mathbb{N}$, implies the boundedness of the sequence $\{u_{k,k}\}_{k\in\mathbb{N}}$ in the space $L^{p^-}(0,T;W^{1,p^-}(\Omega))\cap L^\infty(0,T;L^2(\Omega))$. Hence, combining this fact with (4.4) and (4.3), we deduce

$$u_{k,k}\rightharpoonup u\ \text{ in }\ L^{p^-}(0,T;W^{1,p^-}(\Omega)),\quad \text{as } k\to\infty, \tag{4.7}$$

$$u_{k,k}\rightharpoonup u\ \text{ in }\ L^2(0,T;L^2(\Omega)),\quad \text{as } k\to\infty. \tag{4.8}$$

Step 2. Utilizing the energy equality (4.6) and arguing as in (3.25)–(3.27), we can derive the following a priori estimates

$$\|u_{k,k}\|^2_{L^2(Q_T)}\leqslant 2\|f\|^2_{L^2(Q_T)}+2\sup_{k\in\mathbb{N}}\|v_k\|^2_{L^2(Q_T)}+\frac{2}{\kappa}\|f_0\|^2_{L^2(\Omega)}=:S_1^2, \tag{4.9}$$

$$\begin{aligned}&\|\nabla u_{k,k}\|^{p^-}_{L^{p_{u_{k,k-1}}(\cdot)}(Q_T;\mathbb{R}^2)}\\&\quad\leqslant\Lambda^{-1}\left(\|f_0\|^2_{L^2(\Omega)}+\frac{\kappa}{2}\|f-v_k\|^2_{L^2(Q_T)}+\frac{\kappa}{2}\|u_{k,k}\|^2_{L^2(Q_T)}\right)+5\\&\quad\leqslant\Lambda^{-1}\left(2\|f_0\|^2_{L^2(\Omega)}+2\kappa\|f\|^2_{L^2(Q_T)}+2\kappa\|v_k\|^2_{L^2(Q_T)}\right)+5\\&\quad=\frac{\Lambda^{-1}}{\kappa}S_1^2+5=:S_2^{p^-},\end{aligned} \tag{4.10}$$

$$\|\nabla u_{k,k}\|_{L^{p^-}(Q_T;\mathbb{R}^2)}\leqslant(1+T|\Omega|)^{1/p^-}S_2, \tag{4.11}$$

$$\|u_{k,k}\|_{L^\infty(0,T;L^2(\Omega))} \leqslant \sqrt{2\left(\|f_0\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|f - v_k\|^2_{L^2(Q_T)} + \frac{\kappa}{2}\|u_{k,k}\|^2_{L^2(Q_T)}\right)}$$
$$\leqslant \sqrt{\frac{2}{\kappa}} S_1, \tag{4.12}$$
$$\|\nabla u_{k,k}\|_{L^2(Q_T;\mathbb{R}^2)} \leqslant \sqrt{k}\sqrt{\|f_0\|^2_{L^2(\Omega)} + \kappa\|f - v_k\|_{L^2(Q_T)}\|u_{k,k}\|_{L^2(Q_T)}}$$
$$\overset{\text{by (4.9)}}{\leqslant} \sqrt{\frac{k}{\kappa}} S_1. \tag{4.13}$$

for all $k \in \mathbb{N}$, where

$$\sup_{k\in\mathbb{N}} \|v_k\|_{L^2(Q_T)} \leqslant \sqrt{T}\|v_b\|_{L^2(\Omega)} < +\infty. \tag{4.14}$$

Let us show that the following asymptotic property

$$\frac{1}{k}\nabla u_{k,k} \rightharpoonup 0 \quad \text{in} \;\; L^2(Q_T;\mathbb{R}^2) \tag{4.15}$$

holds true.

Indeed, for any vector-valued test function $\varphi \in C_0^\infty(Q_T)$, we have

$$\left|\int_{Q_T} \left(\frac{1}{k}\nabla u_{k,k}, \varphi\right) dxdt\right| \leqslant \frac{1}{\sqrt{k}} \left(\int_{Q_T} \frac{1}{k}|\nabla u_{k,k}|^2\, dxdt\right)^{1/2} \left(\int_{Q_T} |\varphi|^2\, dxdt\right)^{1/2}.$$

Hence, the sequence $\left\{\frac{1}{k}\nabla u_{k,k}\right\}_{k\in\mathbb{N}}$ is bounded in $L^2(Q_T;\mathbb{R}^2)$. As a result, we obtain

$$\left|\int_{Q_T} \left(\frac{1}{k}\nabla u_{k,k}, \varphi\right) dxdt\right| \overset{\text{by (4.13)}}{\leqslant} S_1 \frac{1}{\sqrt{k\kappa}} \left(\int_{Q_T} |\varphi|^2\, dxdt\right)^{1/2} \to 0 \quad \text{as} \;\; k \to \infty.$$

Step 3. At this step we prove that the flux $\frac{1}{k}\nabla u_{k,k} + A^{1/k}_{u_{k,k-1}}(t, x, \nabla u_{k,k})$ weakly converges in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$ to the flux $A_u(t, x, \nabla u)$ as $k \to \infty$. With that in mind, we show that all preconditions (C1)–(C7) of Theorem 2.1 are fulfilled.

To begin with, we notice that the conclusion, similar to (4.7), can be also made with respect to the sequence $\{u_{k,k-1}\}_{k\in\mathbb{N}}$. Then Lemma 2.1 implies that

$$p_{u_{k,k-1}}(t, x) \to p_u(t, x) \text{ uniformly in } \overline{Q_T} \text{ as } k \to \infty. \tag{4.16}$$

Moreover, we deduce from (3.8) and (4.10) that the sequence

$$\left\{\frac{1}{k}\nabla u_{k,k} + A^{1/k}_{u_{k,k-1}}(t, x, \nabla u_{k,k})\right\}_{k\in\mathbb{R}}$$

is bounded in $L^{(p^+)'}(Q_T;\mathbb{R}^2)$. Hence, we can suppose that there exists an element $z \in L^{(p^+)'}(Q_T;\mathbb{R}^2)$ such that

$$\frac{1}{k}\nabla u_{k,k} + A^{1/k}_{u_{k,k-1}}(t, x, \nabla u_{k,k}) \rightharpoonup z \quad \text{weakly in } L^{(p^+)'}(Q_T;\mathbb{R}^2) \text{ as } k \to \infty. \tag{4.17}$$

We also make use of the following observation: the sequence

$$\left\{\frac{1}{k}|\nabla u_{k,k}|^2 + \left(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}), \nabla u_{k,k}\right)\right\}_{k\in\mathbb{N}} \tag{4.18}$$

is uniformly bounded in $L^1(Q_T)$. Indeed, this inference is a direct consequence of estimates (4.13), (4.10), and (3.22). Utilizing this fact together with the properties (4.16), (4.17), (3.6), (4.7) and

$$u_{k,k} \in L^{p^+}(0,T;W^{1,p^+}(\Omega))\ \forall\, k\in\mathbb{N} \ \text{ by (4.9),(4.13),}$$

and taking into account that $1 < 1+\delta = p^- < p^+ = 2 < 2p^-$, we see that all preconditions of Theorem 2.1 hold true. Hence, in view of the property (4.15), the assertion (4.17) can be rewritten as follows

$$\frac{1}{k}\nabla u_{k,k} + A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}) \rightharpoonup A_u(t,x,\nabla u)$$
$$\text{weakly in } L^{(p^+)'}(Q_T;\mathbb{R}^2) \text{ as } k\to\infty. \tag{4.19}$$

Step 4. The standard formulation of the Aubin-Lions lemma states that if U is a bounded set in $L^p(0,T;X)$ and $\partial U/\partial t = \{\partial u/\partial t\ :\ u\in U\}$ is bounded in $L^r(0,T;Y)$, $r \geqslant 1$, then U is relatively compact in $L^p(0,T;B)$, under the conditions that

$$X \hookrightarrow B \ \text{ compactly,} \quad B\hookrightarrow Y \text{ continuously.}$$

Setting $U = \{u_{k,k}\}_{k\in\mathbb{N}}$, we deduce from (4.9)–(4.12) that

$$\{u_{k,k}\}_{k\in\mathbb{N}} \text{ is bounded in } L^{p^-}(0,T;W^{1,p^-}(\Omega)\cap L^2(\Omega)). \tag{4.20}$$

Since, by the Sobolev embedding Theorem, $W^{1,p^-}(\Omega) \hookrightarrow L^{p^-}(\Omega)$ compactly, it follows from the Lebesgue dominated Theorem that the following embeddings are compact as well

$$W^{1,p^-}(\Omega)\cap L^2(\Omega) \hookrightarrow L^2(\Omega), \quad L^2(\Omega)\hookrightarrow \left(W^{1,2}(\Omega)\right)' \text{ (by the duality arguments).} \tag{4.21}$$

Further, having in mind the fact that for each $k\in\mathbb{N}$, the functions $u_{k,k}$ are the solutions in $W(0,T)$ of the variational problem

$$\left\langle \frac{\partial u_{k,k}(t)}{\partial t}, \varphi\right\rangle_{(W^{1,2}(\Omega))';W^{1,2}(\Omega)} + \int_\Omega \left[\frac{1}{k}\left(\nabla u_{k,k}(t),\nabla\varphi\right)\right] dx$$
$$+ \int_\Omega \left[\left(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}(t)),\nabla\varphi\right) + \kappa u_{k,k}(t)\varphi\right] dx$$
$$= \kappa\int_\Omega (f(t)-v_k(t))\varphi\, dx, \quad \forall\,\varphi\in W^{1,2}(\Omega) \quad \text{a.e. in } [0,T], \tag{4.22}$$
$$u_{k,k}(0) = f_0. \tag{4.23}$$

we derive from this the following estimate

$$\begin{aligned}
\left|\left\langle \frac{\partial u_{k,k}}{\partial t}, \varphi \right\rangle\right| &\leqslant \frac{1}{k}\|\nabla u_{k,k}\|_{L^2(Q_T;\mathbb{R}^2)}\|\nabla\varphi\|_{L^2(Q_T;\mathbb{R}^2)} \\
&\quad + 2\|A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k})\|_{L^{p'_{u_{k,k-1}}(\cdot)}(Q_T;\mathbb{R}^2)}\|\nabla\varphi\|_{L^{p_{u_{k,k-1}}(\cdot)}(Q_T;\mathbb{R}^2)} \\
&\quad + \kappa\|u_{k,k}\|_{L^2(Q_T)}\|\varphi\|_{L^2(Q_T)} + \kappa\|f - v_k\|_{L^2(Q_T)}\|\varphi\|_{L^2(Q_T)} \\
&\leqslant \text{(by (4.9)–(4.13))} \\
&\leqslant \left[\frac{1}{\sqrt{\kappa}}S_1 + \kappa S_1 + \kappa\|f\|_{L^2(Q_T)} + \kappa\sup_{k\in\mathbb{N}}\|v_k\|_{L^2(Q_T)}\right]\|\varphi\|_{L^2(0,T;W^{1,2}(\Omega))} \\
&\quad + \left(1 + \int_{Q_T} |A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k})|^{p'_{u_{k,k-1}}(t,x)}\,dxdt\right)^{1/2}(1+T|\Omega|)^{1/2}\|\varphi\|_{L^2(Q_T)} \\
&\overset{\text{by (4.10),(3.8)}}{\leqslant} \text{const}\|\varphi\|_{L^2(0,T;W^{1,2}(\Omega))}, \quad \forall\,\varphi \in L^2(0,T;W^{1,2}(\Omega)).
\end{aligned}$$

Hence,

$$\left\|\frac{\partial u_{k,k}}{\partial t}\right\|_{L^2(0,T;(W^{1,2}(\Omega))')} < +\infty. \tag{4.24}$$

Utilizing this fact together with (4.20) and (4.21), we deduce from the Aubin-Lions lemma that the set $U = \{u_{k,k}\}_{k\in\mathbb{N}}$ is relatively compact in $L^{p^-}(0,T;L^2(\Omega))$. Hence, we can suppose that $u_{k,k} \to u$ strongly in $L^{p^-}(0,T;L^2(\Omega))$ as $k \to \infty$. Since, U is bounded in $L^\infty(0,T;L^2(\Omega))$, it leads to the conclusion

$$u_{k,k} \to u \ \text{ strongly in } \ L^2(0,T;L^2(\Omega)), \quad \text{as } k \to \infty. \tag{4.25}$$

Step 5. At this stage we show that the limit pair (v,u) is related by the integral identity (3.10). First we notice that $u_{k,k}$ is a weak solution (in the sense of distributions) of (3.59)–(3.61) with $n = k$, $\varepsilon_n = 1/k$ and $v = v_k$. Hence, $u_{k,k}$ satisfies the integral identity

$$\begin{aligned}
\int_{Q_T}\left(-u_{k,k}\frac{\partial\varphi}{\partial t} + \frac{1}{k}(\nabla u_{k,k},\nabla\varphi) + \left(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}),\nabla\varphi\right) + \kappa u_{k,k}\varphi\right)dxdt \\
= \kappa\int_{Q_T}(f - v_k)\varphi\,dxdt + \int_\Omega f_0\varphi|_{t=0}\,dx \quad \forall\,\varphi\in\Phi.
\end{aligned} \tag{4.26}$$

Then, utilizing the properties (4.19), (4.7), and (4.3), and passing to the limit in (4.26) as $k \to \infty$, we immediately arrive at the announced identity (3.10).

Step 6. In order to show that the limit pair (v,u) satisfies the energy inequality (3.62), we have to realize the limit passage as $k \to \infty$ in the relation (see [41])

$$\begin{aligned}
\frac{1}{2}\int_\Omega u^2_{k,k}\,dx + \int_0^t\int_\Omega\left(\frac{1}{k}|\nabla u_{k,k}|^2 + \left(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}),\nabla u_{k,k}\right) + \kappa u^2_{k,k}\right)dxdt \\
= \kappa\int_{Q_T}(f - v_k)u_{k,k}\,dxdt + \int_\Omega f_0^2\,dx \quad \forall\,t\in[0,T].
\end{aligned} \tag{4.27}$$

that can be viewed as the energy equality for the weak solutions of the problem (3.59)–(3.61) with $n = k$, $\varepsilon_n = 1/k$ and $v = v_k$. To this end, we notice that the strong convergence in (4.25) implies the pointwise convergence

$$u^2_{k,k}(t,\cdot) \to u^2(t,\cdot) \quad \text{a.e. in } Q_T.$$

Then, in view of estimate (4.12), we have (by the Lebesgue dominated Theorem) the strong convergence $u^2_{k,k}(t,\cdot) \to u^2(t,\cdot)$ in $L^1(\Omega)$ for a.a. $t \in (0,T)$, and, therefore,

$$\frac{1}{2}\lim_{k\to\infty}\int_\Omega u^2_{k,k}(t,x)\,dx = \frac{1}{2}\int_\Omega u^2(t,x)\,dx \quad \text{for a.a. } t \in (0,T). \tag{4.28}$$

Moreover, taking into account that the $L^2(Q_T)$-norm is continuous with respect to the strong convergence (4.25), we see that

$$\lim_{k\to\infty}\int_0^t\int_\Omega u^2_{k,k}\,dxdt = \int_0^t\int_\Omega u^2\,dxdt. \tag{4.29}$$

We also notice that due to the properties (4.15), (4.19), and (4.7), we have

$$\nabla u_{k,k} \rightharpoonup \nabla u \quad\text{and}\quad A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}) \rightharpoonup A_u(t,x,\nabla u) \quad\text{in } L^1(Q_T;\mathbb{R}^2) \text{ as } k\to\infty.$$

Since $(A_u(t,x,\nabla u),\nabla u) \in L^1(Q_T)$ (see (3.22)), it follows from Lemma 2.3 (see also Proposition 2.2) that

$$\begin{aligned}
\lim_{k\to\infty}\int_0^t\int_\Omega \Big[\frac{1}{k}|\nabla u_{k,k}|^2 &+ \Big(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}),\nabla u_{k,k}\Big)\Big]\,dxdt\\
&\geqslant \lim_{k\to\infty}\int_0^t\int_\Omega \Big[\frac{1}{k}|\nabla u_{k,k}|^2\Big]\,dxdt\\
&\quad + \liminf_{k\to\infty}\int_0^t\int_\Omega \Big(A^{1/k}_{u_{k,k-1}}(t,x,\nabla u_{k,k}),\nabla u_{k,k}\Big)\,dxdt\\
&\overset{\text{by (4.13)}}{\geqslant} \int_0^t\int_\Omega (A_u(t,x,\nabla u),\nabla u)\,dxdt.
\end{aligned} \tag{4.30}$$

So, in order to pass to the limit in (4.27), it remains to notice that the term

$$\int_{Q_T}(f - v_k)u_{k,k}\,dxdt$$

is the product of weakly and strongly convergent sequences in $L^2(0,T;L^2(\Omega))$. As a result, we have

$$\lim_{k\to\infty}\int_{Q_T}(f - v_k)u_{k,k}\,dxdt = \int_{Q_T}(f - v)u\,dxdt. \tag{4.31}$$

Thus, utilizing the obtained collection of properties (see (4.28), (4.29), (4.30), and (4.31)), we can pass to the limit in (4.27) as $k \to \infty$. As a result, we arrive at the energy inequality (3.62).

Step 7. To conclude the proof, it remains to notice that, due to the properties (3.9), that were established at the previous steps, we have: $J(v,u) < +\infty$ and $u \in W_u(Q_T)$. Moreover, in this case the sequence $\{u_{k,k}\}_{k\in\mathbb{N}}$ satisfies all requirements that were mentioned in Definition 3.2. Hence, $u \in W_u(Q_T)$ is a W_0-attainable weak solution to the problem (3.1)–(3.3). The proof is complete. □

Taking this result into account, it is easy to show that the original optimal control problem (4.1) has a solution. Indeed, this issue immediately follows from Theorem 4.1 and the facts that the set of feasible solutions Ξ is bounded in $L^2(Q_T) \times L^{p^-}(0,T;W^{1,p^-}(\Omega))$ (see estimates (4.9)–(4.13), (4.14)), and the objective functional $J(v,u)$ is lower semicontinuous with respect to the weak topology of $L^2(Q_T) \times \Big(L^{p^-}(0,T;W^{1,p^-}(\Omega)) \cap L^\infty(0,T;L^2(\Omega))\Big)$. So, as a direct consequence, we can finalize this inference as follows:

Theorem 4.2. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $v_a, v_b \in L^2(\Omega)$, $v_a(x) \leqslant v_b(x)$ a.e. in Ω, be given distributions, and let $\kappa > 0$ and $\mu > 0$ be some constants. Then, for each $0 < \omega < T$ the optimal control problem* (4.1) *admits at least one solution $(v^0, u^0) \in \Xi$.*

5. Approximation of the Relaxed OCP

Let ε, as usual, be a small parameter which varies within a strictly decreasing sequence of positive numbers converging to 0. In order to find out whether some optimal pairs to the original OCP (4.1) can be attained in an appropriate topology, we make use the basic ideas coming from the perturbation theory and variational convergence of minimization problems [26,39,42]. With that in mind, we introduce the following family of perturbed OCPs

$$\text{Minimize} \quad J_\varepsilon(v,u) \quad \text{provided} \quad (v,u) \in \Xi_\varepsilon, \tag{5.1}$$

where

$$J_\varepsilon(v,u) = \|v\|^2_{L^2(0,T;L^1(\Omega))} + \frac{\mu}{2\omega}\int_{T-\omega}^{T}\int_\Omega |u(t,x) - f_0(x)|^2\,dxdt \tag{5.2}$$

and $\Xi_\varepsilon \subset L^2(0,T;L^1(\Omega)) \times L^2(0,T;W^{1,p^-}(\Omega))$ stands for the set of feasible solutions which we define as follows: $(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon$ if

$$\left\{\begin{array}{c} v_\varepsilon \in \mathcal{V}_{ad}, \quad u_\varepsilon \in W_{u_\varepsilon}(Q_T), \quad J_\varepsilon(v_\varepsilon, u_\varepsilon) < +\infty, \\ u_\varepsilon \in C([0,T];L^2(\Omega)) \cap L^2(0,T;W^{1,2}(\Omega)) \text{ is a } W_0\text{-attainable} \\ \text{weak solution of the problem (3.15)–(3.17) for the given } v_\varepsilon. \end{array}\right\} \tag{5.3}$$

Taking into account Theorem 3.2 and arguing as in Theorem 4.2, it can be proven the following result.

Theorem 5.1. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, and $v_a, v_b \in L^2(\Omega)$, $v_a(x) \leqslant v_b(x)$ a.e. in Ω, be given distributions, and let $\kappa > 0$ and $\mu > 0$ be some constants. Then, for each $\varepsilon > 0$ and $0 < \omega < T$, there exists at least one solution $(v^0_\varepsilon, u^0_\varepsilon)$ of optimal control problem (5.1).*

The primary goal of this section is to show that the relaxed OCP (4.1) can be successfully approximated by the OCPs (5.1). In means that there is a pair $(v^0, u^0) \in \Xi$ such that

$$J(v^0,u^0) = \inf_{(v,u)\in\Xi} J(v,u),$$
$$\lim_{\varepsilon\to 0} J(v^0_\varepsilon, u^0_\varepsilon) = \lim_{\varepsilon\to 0} \inf_{(v,u)\in\Xi_\varepsilon} J_\varepsilon(v,u) = J(v^0,u^0),$$
$$(v^0_\varepsilon, u^0_\varepsilon) \to (v^0,u^0) \text{ as } \varepsilon \to 0 \text{ in some appropriate topology.}$$

We begin with a couple of auxiliaries lemmas.

Lemma 5.1. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega;\mathbb{R}^2)$ be given distributions. Let $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon\to 0}$ be a sequence of feasible pairs such that $\{v_\varepsilon\}_{\varepsilon\to 0}$ is bounded in $L^2(Q_T)$. Then there exists a constant $C > 0$ such that*

$$\sup_{\varepsilon\to 0} \Big[\|u_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} + \|\nabla u_\varepsilon\|_{L^{p^-}(Q_T;\mathbb{R}^2)} \Big] \leqslant C. \tag{5.4}$$

Proof. The fact that $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon\to 0}$ is a collection of feasible pairs to the corresponding problems (5.1), implies (see Theorem 3.2 and Corollary 3.1) that, for each $\varepsilon > 0$, they are related by the integral identity

$$\int_{Q_T} \left(-u_\varepsilon \frac{\partial \varphi}{\partial t} + \varepsilon\,(\nabla u_\varepsilon, \nabla\varphi) + \big(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon), \nabla\varphi\big) + \kappa u_\varepsilon \varphi \right) dxdt$$
$$= \kappa \int_{Q_T} (f - v_\varepsilon)\varphi\, dxdt + \int_\Omega f_0 \varphi|_{t=0}\, dx \quad \forall\, \varphi \in \Phi. \tag{5.5}$$

and satisfy the energy equality

$$\frac{1}{2}\int_\Omega u^2_\varepsilon\, dx + \int_0^t \int_\Omega \big(\varepsilon|\nabla u_\varepsilon|^2 + \big(A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon), \nabla u_\varepsilon\big) + \kappa u^2_\varepsilon\big)\, dxds$$
$$= \kappa \int_0^t \int_\Omega (f - v_\varepsilon) u_\varepsilon\, dxds + \int_\Omega f_0^2\, dx \quad \text{for a.a } \ t \in [0,T]. \tag{5.6}$$

Then arguing as in the proof of Theorem 3.3, we deduce from (5.6)

$$\left.\begin{array}{c} \sup_{\varepsilon>0} \|u_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leqslant \sqrt{2\kappa} C_1, \\ \sup_{\varepsilon>0} \|\nabla u_\varepsilon\|_{L^{p^-}(Q_T;\mathbb{R}^2)} \leqslant (1 + T|\Omega|)^{1/p^-} \left(\Lambda^{-1}\kappa C_1^2 + 5\right)^{1/p^-}, \\ \sup_{\varepsilon>0} \|\nabla u_\varepsilon\|_{L^2(Q_T;\mathbb{R}^2)} \leqslant \frac{1}{\sqrt{\varepsilon\kappa}} C_1 \end{array}\right\} \tag{5.7}$$

with $C_1 = \|f\|_{L^2(Q_T)} + \frac{2}{\kappa}\|f_0\|_{L^2(\Omega)} + \sup_{\varepsilon>0}\|v_\varepsilon\|_{L^2(Q_T)}$. As a result, we arrive at the estimate (5.4). □

Taking this result into account and arguing as in Theorem 4.1, it can be shown the weak $L^2(Q_T)$-compactness of admissible controls for the perturbed OCPs (5.1) implies some compactness properties for the corresponding sequence of feasible solutions.

Lemma 5.2. *Let $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon\to 0}$ be a sequence of feasible pairs to the OCPs (5.1). Assume that $v_\varepsilon \rightharpoonup v$ in $L^2(Q_T)$. Then, for given $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, and $\theta \in L^\infty(\Omega;\mathbb{R}^2)$, we have*

$$u_\varepsilon \to u \quad \textit{strongly in} \quad L^2(0,T;L^2(\Omega)), \tag{5.8}$$

$$u_\varepsilon \rightharpoonup u \quad \textit{weakly in} \quad L^{p^-}(0,T;W^{1,p^-}(\Omega)), \tag{5.9}$$

$$\varepsilon\nabla u_\varepsilon \rightharpoonup 0 \quad \textit{weakly in} \quad L^2(Q_T;\mathbb{R}^2), \tag{5.10}$$

$$p_{u_\varepsilon}(t,x) \to p_u(t,x) \ \textit{uniformly in} \ \overline{Q_T}, \tag{5.11}$$

$$\varepsilon\nabla u_\varepsilon + A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon) \rightharpoonup A_u(t,x,\nabla u) \quad \textit{weakly in} \ L^{(p^+)'}(Q_T;\mathbb{R}^2), \tag{5.12}$$

where $(v,u) \in \Xi$.

Proof. Since for each $\varepsilon > 0$ the function u_ε is a W_0-attainable weak solution to the problems (3.15)–(3.17) with $v = v_\varepsilon$, it follows that we can utilize the a priori estimates (3.52)–(3.56). Hence, the existence of element u with properties (5.9)–(5.12) follows from Theorem 3.3. To establish the fact that the pair (v,u) is feasible to the relaxed OCP 4.1, we can apply the arguments of the proof of Theorem 4.1. It remains to notice that in order to deduce the strong convergence property (5.8), it is enough to take into account the boundedness of the sequence $\{u_\varepsilon\}_{\varepsilon\to 0}$ in $L^{p^-}(0,T;W^{1,p^-}(\Omega))$ (see (5.9)), the estimate

$$\begin{aligned}
&\left|\left\langle \frac{\partial u_\varepsilon}{\partial t}, \varphi\right\rangle\right| \leqslant \varepsilon\|\nabla u_\varepsilon\|_{L^2(Q_T;\mathbb{R}^2)}\|\nabla\varphi\|_{L^2(Q_T;\mathbb{R}^2)}\\
&\quad + 2\|A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon)\|_{L^{p'_{u_\varepsilon}(\cdot)}(Q_T;\mathbb{R}^2)}\|\nabla\varphi\|_{L^{p_{u_\varepsilon}(\cdot)}(Q_T;\mathbb{R}^2)}\\
&\quad + \kappa\|u_\varepsilon\|_{L^2(Q_T)}\|\varphi\|_{L^2(Q_T)} + \kappa\|f - v_\varepsilon\|_{L^2(Q_T)}\|\varphi\|_{L^2(Q_T)}\\
&\leqslant \text{(by (5.7))}\\
&\leqslant \Big[\text{const} + \kappa\|f\|_{L^2(Q_T)} + \kappa\sup_{k\in\mathbb{N}}\|v_k\|_{L^2(Q_T)}\Big]\|\varphi\|_{L^2(0,T;W^{1,2}(\Omega))}\\
&\quad + \left(1 + \int_{Q_T}|A^\varepsilon_{u_\varepsilon}(t,x,\nabla u_\varepsilon)|^{p'_{u_\varepsilon}(t,x)}\,dxdt\right)^{1/2}(1+T|\Omega|)^{1/2}\|\varphi\|_{L^2(Q_T)}\\
&\leqslant \text{const}\|\varphi\|_{L^2(0,T;W^{1,2}(\Omega))}, \quad \forall\,\varphi \in L^2(0,T;W^{1,2}(\Omega)).
\end{aligned}$$

and then (5.8) immediately follows from the Aubin-Lions lemma. □

The main question we are going to discuss the convergence of minima of (5.1) to minima of (4.1) as ε tends to zero. In other words, our aim is to show that some optimal solutions to (4.1) can be approximated by the solutions of (5.1). To this end, we make use of the basic results of the variational convergence of minimization problems [36–38, 42, 44]. We begin with some preliminaries.

Lemma 5.3. *Let* $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon\to 0}$ *be a sequence of feasible pairs such that* $\{v_\varepsilon\}_{\varepsilon\to 0}$ *is bounded in* $L^2(Q_T)$. *Then there exists a function* $u \in W_u(Q_T)$ *with properties* (5.8)–(5.12) *such that*

$$(v, u) \in \Xi \quad \textit{and} \quad J(v, u) \leqslant \liminf_{\varepsilon\to 0} J_\varepsilon(v_\varepsilon, u_\varepsilon). \tag{5.13}$$

Proof. This assertion immediately follows from Lemma 5.2 the lower semicontinuity of the $L^2(Q_T)$ and $L^2(0, T; L^1(\Omega))$-norms with respect to weak convergence in $L^2(Q_T) \times L^{p^-}(0, T; W^{1,p^-}(\Omega))$. As a result, we have

$$\lim_{\varepsilon\to 0} \|v_\varepsilon\|^2_{L^2(0,T;L^1(\Omega))} \geqslant \|v\|^2_{L^2(0,T;L^1(\Omega))},$$
$$\lim_{\varepsilon\to 0} \int_{T-\omega}^{T} \|u_\varepsilon(t, \cdot) - f_0\|^2_{L^2(\Omega)}\, dt = \int_{T-\omega}^{T} \|u(t, \cdot) - f_0\|^2_{L^2(\Omega)}\, dt.$$

□

Before proceeding further, we note that, the initial-boundary value problem (3.15)–(3.17) may have a non-unique solution under a fixed control (see Theorem 3.2). In view of this we define the binary relation $\langle L; \Xi_\varepsilon\rangle$ on each of the sets Ξ_ε following the rule: $(v_\varepsilon, u_\varepsilon)\, L\, (\widehat{v}_\varepsilon, \widehat{u}_\varepsilon)$ if and only if $v_\varepsilon = \widehat{v}_\varepsilon$ a.e. in Q_T. It is easily seen that $\langle L; \Xi_\varepsilon\rangle$ is an equivalence relation. So, hereinafter we will not distinguish the triplets belonging to the same class of equivalence.

Lemma 5.4. *For every class of equivalence* $\Xi/L(v)$ *with* $v \in \mathcal{V}_{ad}$ *there can be found a pair* $(v, u) \in \Xi/L(v)$ *and a sequence* $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ *with properties* (5.8)–(5.12) *such that*

$$v_\varepsilon \rightharpoonup v \quad \textit{in} \quad L^2(Q_T), \quad \textit{and} \quad J(v, u) \geqslant \liminf_{\varepsilon\to 0} J_\varepsilon(v_\varepsilon, u_\varepsilon). \tag{5.14}$$

Proof. Let $v^* \in \mathcal{V}_{ad}$ be an arbitrary admissible control. Let

$$\Xi/L(v^*) = \{(v^*, u) \in \Xi\}$$

be the corresponding class on equivalence.

We define the sequence $\{(v_\varepsilon, u_\varepsilon) \in \Xi_\varepsilon\}_{\varepsilon>0}$ as follows $v_\varepsilon \equiv v^*$ for each $\varepsilon > 0$, and u_ε is a weak solution to the problem (3.1)–(3.3) with $v = v^*$ in the sense of Definition 3.1. Then arguing as in Lemma 5.3 and Theorems 3.2 and 3.3, it can be shown that there exists a W_0-attainable solution $u^* \in W_{u^*}(Q_T)$ to the problem (3.1)–(3.3) such that $(v^*, u^*) \in \Xi$ and $u_\varepsilon \to u^*$ strongly in $L^2(0, T; L^2(\Omega))$ (see properties (5.8)–(5.12)). It is clear now that $(v^*, u^*) \in \Xi/L$ and the following relations

$$\lim_{\varepsilon\to 0} \|v_\varepsilon\|^2_{L^2(0,T;L^1(\Omega))} = \|v^*\|^2_{L^2(0,T;L^1(\Omega))},$$
$$\lim_{\varepsilon\to 0} \int_{T-\omega}^{T} \|u_\varepsilon(t, \cdot) - f_0\|^2_{L^2(\Omega)}\, dt = \int_{T-\omega}^{T} \|u^*(t, \cdot) - f_0\|^2_{L^2(\Omega)}\, dt$$

hold true. From this (5.14) follows. □

We are now in a position to prove the main result of this section.

Theorem 5.2. *Let $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $v_a, v_b \in L^2(\Omega)$, and $\theta \in L^\infty(\Omega;\mathbb{R}^2)$ be given distributions. Let $\left\{(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ be a sequence of optimal pairs to the corresponding OCPs* (5.1)*. Then there exists a pair $(v^0, u^0) \in \Xi$ such that, up to a subsequence, $(v^0_\varepsilon, u^0_\varepsilon) \to (v^0, u^0)$ in the sense of convergences* (5.8)–(5.12) *and*

$$\inf_{(v,u)\in\Xi_\varepsilon} J(v,u) = J(v^0,u^0) = \lim_{\varepsilon\to 0} J_\varepsilon(v^0_\varepsilon, u^0_\varepsilon) = \lim_{\varepsilon\to 0} \inf_{(v,u)\in\Xi_\varepsilon} J_\varepsilon(v,u). \tag{5.15}$$

Proof. Let $\left\{(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ be a given sequence of optimal pairs to the OCPs (5.1). Taking into account the definition of the set of admissible controls $\mathcal{V}_{ad}$ and the a priori estimates (3.52)–(3.56), it can be show that (we refer to Lemma 5.2 for the details) the sequence $\left\{(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ is relatively compact with respect to the convergence (5.8)–(5.12). So, we may suppose that there exist a subsequence $\left\{(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}) \in \Xi_{\varepsilon_k}\right\}_{k\in\mathbb{N}}$ of the sequence of optimal solutions and a pair (v^*, u^*), such that $(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}) \longrightarrow (v^*, u^*)$ as $\varepsilon_k \to 0$ in the sense of (5.8)–(5.12). Then, by Lemma 5.2, we deduce that $(v^*, y^*) \in \Xi$, and

$$\begin{aligned}\liminf_{k\to\infty} \min_{(v,v)\in\Xi_{\varepsilon_k}} J_{\varepsilon_k}(v,u) &= \liminf_{k\to\infty} J_{\varepsilon_k}(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}) \\ &\geqslant J(v^*,u^*) \geqslant \min_{(v,u)\in\Xi} J(v,u) = J(v^0,u^0),\end{aligned} \tag{5.16}$$

where (v^0, u^0) is an optimal pair to (5.1).

Let $\Xi/L(v^0) = \left\{(v^0, u) \in \Xi\right\}$ be the corresponding class on equivalence. It is clear that $(v^0, u^0) \in \Xi/L(v^0)$. Since u^0 is a W_0-attainable weak solution to the problems (3.15)–(3.17) with $v = v^0$ (in fact, (v^0, u^0) is the limit of a minimizing sequence), it follows from Lemma 5.4 that there exists a sequence $\left\{(v^0, \widehat{u}_\varepsilon) \in \Xi_\varepsilon\right\}_{\varepsilon>0}$ with properties (5.8)–(5.12) such that

$$J(v^0,u^0) \geqslant \limsup_{\varepsilon\to 0} J_\varepsilon(v^0, \widehat{u}_\varepsilon).$$

Using this fact, we have

$$\begin{aligned}\min_{(v,u)\in\Xi} J(v,u) &= J(v^0,u^0) \\ &\geqslant \limsup_{\varepsilon\to 0} J_\varepsilon(v^0, \widehat{u}_\varepsilon) \geqslant \limsup_{\varepsilon\to 0} \min_{(v,u)\in\Xi_\varepsilon} J_\varepsilon(v,u) \\ &\geqslant \limsup_{k\to\infty} \min_{(v,u)\in\Xi_{\varepsilon_k}} J_{\varepsilon_k}(v,u) = \limsup_{k\to\infty} J_{\varepsilon_k}(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}).\end{aligned} \tag{5.17}$$

From this and (5.16) we deduce that

$$\liminf_{k\to\infty} J_{\varepsilon_k}(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}) \geqslant \limsup_{k\to\infty} J_{\varepsilon_k}(v^0_{\varepsilon_k}, u^0_{\varepsilon_k}).$$

Then, combining (5.16) and (5.17), we get

$$J(v^*, u^*) = J(v^0, u^0) = \min_{(v,u)\in\,\Xi} J(v, u) = \lim_{k\to\infty} \min_{(v,u)\,\in\,\Xi_{\varepsilon_k}} J_{\varepsilon_k}(v, u). \tag{5.18}$$

Using these relations and the fact that the problem(5.4) has a nonempty set of solutions, we may suppose that $v^* = v^0$ and u^* together with u^0 belong to the same class of equivalence $\Xi/L(v^0)$. Since equality (5.18) holds for the limits of all subsequences of $\left\{(v^0_\varepsilon, u^0_\varepsilon)\right\}_{\varepsilon>0}$, it follows that these limits coincide and, therefore, (v^0, u^0) is the limit of the whole sequence $\left\{(v^0_\varepsilon, u^0_\varepsilon)\right\}_{\varepsilon>0}$. Then, using the same argument for the entire sequence of minimizers, we have

$$\begin{aligned}\liminf_{\varepsilon\to0} \min_{(v,u)\in\,\Xi_\varepsilon} J_\varepsilon(v, u) &= \liminf_{\varepsilon\to0} J_\varepsilon(v^0_\varepsilon, u^0_\varepsilon) \geqslant J(v^0, u^0) = \min_{(v,u)\in\,\Xi} J(v, u)\\ &\geqslant \limsup_{\varepsilon\to0} J_\varepsilon(v^0, \widehat{u}_\varepsilon) \geqslant \limsup_{\varepsilon\to0} \min_{(v,u)\in\,\Xi_\varepsilon} J_\varepsilon(v, u)\\ &= \limsup_{\varepsilon\to0} J_\varepsilon(v^0_\varepsilon, u^0_\varepsilon)\end{aligned}$$

and this concludes the proof. □

6. On Optimality Conditions for Approximating OCPs

For the sake of simplicity, we assume that the function $g : [0,\infty) \to (0, 1]$ in (1.7) is defined by the rule

$$g(s) = \delta + \frac{a^2(1-\delta)}{a^2 + s^2}, \quad \forall\, s \in [0, +\infty), \tag{6.1}$$

where $0 < \delta \ll 1$ is a given threshold, and $a \in [0,\infty)$ is a tuning parameters. We also assume that the variance σ in the Gaussian kernel G_σ (see (1.9)) is small enough. It means that, for each function $u \in W^{1,1}(\Omega)$, its texture index $p_u(t, x)$ can be approximately defined as

$$p_u(t, x) = 1 + g(|\nabla u|). \tag{6.2}$$

It is worth to notice that from practical point of view, the above mentioned simplification is not very strong because we can always suppose that in this case the restriction of exponent $p_u(t, x)$ given by (6.2) on any grid will be almost the same the restriction on this grid of the Lipschitz-continuous function $\widehat{p}_u(t, x) := 1 + g\left(\frac{1}{h}\int_{t-h}^{t} \left|\left(\nabla G_\sigma * \widetilde{u}(\tau,\cdot)\right)(x)\right|\, d\tau\right)$ provided $\sigma > 0$ and $h > 0$ are small enough.

Further, for each $\varepsilon > 0$, we associate with the OCP (5.1) the following Lagrange functional

$$\begin{aligned}\mathcal{L}_\varepsilon\left(v, u, \lambda, z\right) := \lambda J_\varepsilon(v, u) + \int_{Q_T} \left[-\varepsilon\Delta u - \operatorname{div} A^\varepsilon_u(t, x, \nabla u)\right] z\, dxdt\\ + \int_{Q_T} \left[\frac{\partial u}{\partial t} + \kappa(u - f + v)\right] z\, dxdt, \end{aligned}\tag{6.3}$$

where $\lambda \in \mathbb{R}_+$, $z \in L^2(0,T;W^{1,p^-}(\Omega))$, and

$$J_\varepsilon(v,u) = \|v\|^2_{L^2(0,T;L^1(\Omega))} + \frac{\mu}{2\omega}\int_{T-\omega}^{T}\int_\Omega |u(t,x) - f_0(x)|^2\,dxdt, \tag{6.4}$$

$$A^\varepsilon_u(t,x,\nabla u) := \left(|R_\eta \nabla u| + \varepsilon\right)^{p_u(t,x)-2} R_\eta \nabla u. \tag{6.5}$$

Let $(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\varepsilon$ be an optimal pair to the problem (5.1). To characterize this solution, we make use of the celebrated Ioffe-Tikhomirov theorem [34]. With that in mind, let us show that the mappings

$$u \mapsto J_\varepsilon(v,u), \tag{6.6}$$

$$u \mapsto \frac{\partial u}{\partial t} - \varepsilon\Delta u + \kappa(u - f + v), \tag{6.7}$$

$$u \mapsto -\operatorname{div} A^\varepsilon_u(t,x,\nabla u) \tag{6.8}$$

are continuously differentiable in some neighborhood of the point u_ε. Since this property is obviously true for the mappings (6.6)–(6.7), we establish it for the $u \mapsto -\operatorname{div} A^\varepsilon_u(t,x,\nabla u)$.

Let $F'(u)[h]$ stands for the directional derivative of a functional $F : X \to Y$ at the point $u \in X$ along a vector $h \in X$, i.e.,

$$F'(u)[h] = \lim_{\sigma\to 0}\frac{F(u+\sigma h) - F(u)}{\sigma}.$$

Then, arguing as in [6], we can obtain the following result:

Proposition 6.1. Let $p : Q_T \to [p^-, p^+] \subset (1,2]$, with $p^\pm = \text{const}$, be a given exponent and let

$$\widetilde{F}_1(u) = -\operatorname{div}\left[\left(|\nabla u| + \varepsilon\right)^{p(x)-2}\nabla u\right], \quad \forall\, u \in W_u(Q_T).$$

Then, for each $u \in W_u(Q_T)$, we have

$$\widehat{F}_1'(u)[h] = \widehat{\mathbb{F}}_{11}(u)[h] + \widehat{\mathbb{F}}_{12}(u)[h], \tag{6.9}$$

where

$$\widehat{\mathbb{F}}_{11}(u)[h] = -\operatorname{div}\left[\left(|\nabla u| + \varepsilon\right)^{p-2}\nabla h\right],$$

$$\widehat{\mathbb{F}}_{12}(u)[h] = -\operatorname{div}\left[(p-2)\left(|\nabla u| + \varepsilon\right)^{p-4}\nabla u\,(\nabla u, \nabla h)\right].$$

Then, applying the similar arguments and taking into account the representation (2.13), we can generalize the previous proposition as follows.

Proposition 6.2. Let $p : Q_T \to [p^-, p^+] \subset (1,2]$, with $p^\pm = \text{const}$, be a given exponent and let

$$F_1(u) = -\operatorname{div}\left[\left(|R_\eta \nabla u| + \varepsilon\right)^{p(x)-2} R_\eta \nabla u\right], \quad \forall\, u \in W_u(Q_T).$$

Then, for each $u \in W_u(Q_T)$, we have

$$F_1'(u)[h] = \mathbb{F}_{11}(u)[h] + \mathbb{F}_{12}(u)[h] + \mathbb{F}_{13}(u)[h] + \mathbb{F}_{14?}(u)[h], \tag{6.10}$$

where

$$\mathbb{F}_{11}(u)[h] = -\operatorname{div}\left[(|R_\eta \nabla u| + \varepsilon)^{p-2}\,\nabla h\right], \tag{6.11}$$

$$\mathbb{F}_{12}(u)[h] = -\operatorname{div}\left[(p-2)\,(|R_\eta \nabla u| + \varepsilon)^{p-4}\,R_\eta \nabla u\,(\nabla u, \nabla h)\right], \tag{6.12}$$

$$\mathbb{F}_{13}(u)[h] = \eta^2 \operatorname{div}\left[(|R_\eta \nabla u| + \varepsilon)^{p-2}\,(\theta \otimes \theta)\,\nabla h\right], \tag{6.13}$$

$$\begin{aligned}\mathbb{F}_{14}(u)[h] &= \eta^2 \operatorname{div}\left[(p-2)\,(|R_\eta \nabla u| + \varepsilon)^{p-4}\,R_\eta \nabla u\,(\theta, \nabla h)\,\theta\right] \\ &= \eta^2 \operatorname{div}\left[(p-2)\,(|R_\eta \nabla u| + \varepsilon)^{p-4}\,R_\eta \nabla u\,(\theta \otimes \theta)\,\nabla h\right].\end{aligned} \tag{6.14}$$

Proposition 6.3. Let $(v,u) \in \Xi_\varepsilon$ be a given feasible solution, let

$$p[\nabla u] := 1 + \delta + \frac{a^2(1-\delta)}{a^2 + |\nabla u|^2},$$

and let

$$F_2(u) = -\operatorname{div}\left(|R_\eta \nabla q|^{p[\nabla u]-2} R_\eta \nabla q\right), \quad \forall\, u \in L^2(0,T; W^{1,1+\delta}(\Omega)),$$

where $q \in L^2(0,T; W^{1,p[\nabla u]}(\Omega))$ is a given function. Then, for each element $q \in L^2(0,T; W^{1,p[\nabla u]}(\Omega))$ and for all $h \in L^2(0,T; W^{1,2}(\Omega))$, we have

$$F_2'(u)[h] = -\operatorname{div}\left(|R_\eta \nabla q|^{p[\nabla u]-2} \frac{2a^2(1-\delta)\log\left(|R_\eta \nabla q|\right)}{(a^2 + |\nabla u|^2)^2}\,(R_\eta \nabla q \otimes \nabla u)\,\nabla h\right). \tag{6.15}$$

Proof. The representation (6.15) immediately follows from definition of the directional derivative. □

Utilizing the representations (6.10)–(6.15), we see that

$$\begin{aligned}\left[I_\varepsilon(v_\varepsilon^0, u_\varepsilon^0)\right]_u'[h] &= \lambda \frac{\mu}{\omega} \int_{T-\omega}^{T} \int_\Omega \left(u_\varepsilon^0(t,x) - f_0(x)\right) h(t,x)\,dxdt \\ &\quad + \int_{Q_T} \left[\frac{\partial h}{\partial t} + \kappa h\right] z\,dxdt \\ &\quad + \int_{Q_T} \left[-\varepsilon \Delta h + \sum_{i=1}^{4} \mathbb{F}_{1i}(u_\varepsilon^0)[h]\right] z\,dxdt \\ &\quad + \int_{Q_T} F_2'(u_\varepsilon^0)[h]\Big|_{q=u_\varepsilon^0} z\,dxdt, \quad \forall\, h \in .L^2(0,T; W^{1,2}(\Omega)).\end{aligned} \tag{6.16}$$

Taking into account that

$$\begin{aligned}&\int_{Q_T} F_2'(u_\varepsilon^0)[h]z\,dxdt\\ &= \int_{Q_T} |R_\eta \nabla u_\varepsilon^0|^{p[\nabla u_\varepsilon^0]-2} \frac{2a^2(1-\delta)\log\left(|R_\eta \nabla u_\varepsilon^0|\right)}{(a^2+|\nabla u_\varepsilon^0|^2)^2} \left(R_\eta \nabla u_\varepsilon^0 \otimes \nabla u_\varepsilon^0\right)(\nabla z, \nabla h)\,dxdt,\end{aligned}$$

we have

$$\begin{aligned}|R_\eta \nabla u_\varepsilon^0|^{p[\nabla u_\varepsilon^0]-2} &\frac{\log\left(|R_\eta \nabla u_\varepsilon^0|\right)}{(a^2+|\nabla u_\varepsilon^0|^2)^2} \left|\left(R_\eta \nabla u_\varepsilon^0 \otimes \nabla u_\varepsilon^0\right)\right| |(\nabla z, \nabla h)|\\ &\leqslant (1-\eta^2)^{-2} |R_\eta \nabla u_\varepsilon^0|^{p[\nabla u_\varepsilon^0]} \frac{\log\left(|R_\eta \nabla u_\varepsilon^0|\right)}{(a^2+|\nabla u_\varepsilon^0|^2)^2} |\nabla z||\nabla h|,\end{aligned}$$

where

$$\begin{aligned}|R_\eta \nabla u_\varepsilon^0|^2 \frac{|\log\left(|R_\eta \nabla u_\varepsilon^0|\right)|}{(a^2+|R_\eta \nabla u_\varepsilon^0|^2)^2} |\nabla z||\nabla h| &\leqslant \frac{|\log\left(|R_\eta \nabla u_\varepsilon^0|\right)|}{a^2+|R_\eta \nabla u_\varepsilon^0|^2} |\nabla z||\nabla h|\\ &\leqslant \mathrm{const}|\nabla z||\nabla h|, \quad \text{as } |\nabla u_\varepsilon^0| \to \infty,\\ |R_\eta \nabla u_\varepsilon^0|^2 \frac{|\log\left(|R_\eta \nabla u_\varepsilon^0|\right)|}{(a^2+|R_\eta \nabla u_\varepsilon^0|^2)^2} &\leqslant \frac{1}{a^4} |R_\eta \nabla u_\varepsilon^0|^2 |\log\left(|R_\eta \nabla u_\varepsilon^0|\right)| < +\infty\\ &\text{as } |\nabla u_\varepsilon^0| \to 0\end{aligned}$$

by the L'Hôpital's rule.

Thus, from this we can deduce that

$$|R_\eta \nabla u_\varepsilon^0|^2 \frac{|\log\left(|R_\eta \nabla u_\varepsilon^0|\right)|}{(a^2+|R_\eta \nabla u_\varepsilon^0|^2)^2} \in L^\infty(Q_T)$$

and there exists a constant $M>0$ such that

$$\begin{aligned}\left|\int_{Q_T} F_2'(u_\varepsilon)[h]z\,dxdt\right| &\leqslant 2a^2 \frac{(1-\delta)}{(1-\eta^2)^2} \left\| |R_\eta \nabla u_\varepsilon^0|^2 \frac{|\log\left(|R_\eta \nabla u_\varepsilon^0|\right)|}{(a^2+|R_\eta \nabla u_\varepsilon^0|^2)^2} \right\|_{L^\infty(\Omega)}\\ &\quad\times \int_{Q_T} |R_\eta \nabla u_\varepsilon^0|^{p(|\nabla u_\varepsilon^0|)-2} |\nabla z||\nabla h|\,dxdt\\ &\leqslant \mathrm{const} \int_{Q_T} |\nabla u_\varepsilon^0|^{p(|\nabla u_\varepsilon^0|)-2} |\nabla z||\nabla h|\,dxdt\\ &\leqslant \mathrm{const} \left(\int_{Q_T} |\nabla u_\varepsilon^0|^{p(|\nabla u_\varepsilon^0|)-2} |\nabla z|^2\,dxdt\right)^{1/2}\\ &\quad\times \left(\int_{Q_T} |\nabla u_\varepsilon^0|^{p(|\nabla u_\varepsilon^0|)-2} |\nabla h|^2\,dxdt\right)^{1/2}\\ &\leqslant \mathrm{const}\, \|z\|_{L^2(0,T;H^{p^-,u_\varepsilon^0}(\Omega))} \|h\|_{L^2(0,T;H^{p^-,u_\varepsilon^0}(\Omega))}\\ &\leqslant M\|h\|_{L^2(0,T;W^{1,2}(\Omega))} \|z\|_{L^2(0,T;H^{p^-,u_\varepsilon^0}(\Omega))}\,. \qquad (6.17)\end{aligned}$$

Here, $H^{p^-,u^0_\varepsilon}(\Omega)$ stands for the weighted Sobolev space which is defined as a completeness of $C^\infty_c(\mathbb{R}^2)$ with respect to the norm

$$\|z\|_{H^{p^-,u^0_\varepsilon}(\Omega)} = \int_\Omega \left[z^2 + \left(1 + |\nabla u^0_\varepsilon|\right)^{p^- -2} |\nabla z|^2\right] dx.$$

It is easy to check that $H^{p^-,u^0_\varepsilon}(\Omega)$ is a Hilbert space with the inner product

$$(z_1, z_2)_{H^{p^-,u^0_\varepsilon}(\Omega)} = \int_\Omega \left[z_1 z_2 + \left(1 + |\nabla u^0_\varepsilon|\right)^{p^- -2} (\nabla z_1, \nabla z_2)\right] dx.$$

Moreover, since $p^- = 1 + \delta << 2$, it follows from the estimates

$$\int_\Omega \left(1 + |\nabla u^0_\varepsilon|\right)^{p^- -2} |\nabla z|^2 \, dx \leqslant \int_\Omega |\nabla z|^2 \, dx,$$

$$\begin{aligned}\int_\Omega |\nabla y|^{p^-} \, dx &= \int_\Omega \frac{|\nabla y|^{p^-}}{\left(1 + |\nabla u^0_\varepsilon|\right)^{\frac{p^-(2-p^-)}{2}}} \left(1 + |\nabla u^0_\varepsilon|\right)^{\frac{p^-\left(2-p^-\right)}{2}} \, dx \\ &\leqslant \left(\int_\Omega \left(1 + |\nabla u^0_\varepsilon|\right)^{p^- -2} |\nabla y|^2 \, dx\right)^{\frac{p^-}{2}} \left(\int_\Omega \left(1 + |\nabla u^0_\varepsilon|\right)^{p^-} \, dx\right)^{\frac{2-p^-}{2}},\end{aligned}$$

which hold true for each $z \in H^{p^-,u^0_\varepsilon}(\Omega)$ and $y \in W^{1,p^-}(\Omega)$, that

$$H^1(\Omega) \hookrightarrow H^{p^-,u^0_\varepsilon}(\Omega) \hookrightarrow W^{1,p^-}(\Omega)$$

with continuous embeddings.

Thus, in view of estimate (6.17), it is clear that, the directional derivative $F_2'(u)[h]$ with its representation (6.15) is the Gâteaux derivative of the operator F_2 and moreover, this derivative is a strongly continuous and bounded mapping. Arguing as in [6], it can be shown that, in fact, the mapping $u \mapsto \mathcal{L}_\varepsilon(v, u, \lambda, z)$ is continuously differentiable in some neighborhood of the optimal pair $(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\varepsilon$. Thus, in order to derive optimality conditions for the approximating OCP (5.1), it remains to repeat all arguments from [32] (see Section 2.8.2). As a result, we deduce that $\lambda = 1$ in (6.3), and the following result holds true.

Theorem 6.1. *Let for given distributions $f \in L^2(Q_T)$, $f_0 \in L^2(\Omega)$, $\theta \in L^\infty(\Omega; \mathbb{R}^2)$, and $v_a, v_b \in L^2(\Omega)$, and for given values of the small parameters $\varepsilon > 0$ and $\omega > 0$, $(v^0_\varepsilon, u^0_\varepsilon) \in \Xi_\varepsilon\varepsilon$ is an optimal pair to the OCP* (5.1). *Then there exists a unique $z_\varepsilon \in L^2(0, T; H^{p^-,u^0_\varepsilon}(\Omega))$ such that $\dot{z}_\varepsilon \in L^2(0, T; \left[H^{p^-,u^0_\varepsilon}(\Omega)\right]')$ and*

$$\left.\begin{aligned}\frac{\partial u^0_\varepsilon}{\partial t} - \varepsilon\Delta u^0_\varepsilon - \operatorname{div}\left[\left(|R_\eta \nabla u^0_\varepsilon| + \varepsilon\right)^{p_{u^0_\varepsilon}(t,x)-2} R_\eta \nabla u^0_\varepsilon\right] + \kappa u^0_\varepsilon& \\ = \kappa(f - v^0_\varepsilon) \quad \text{in} \quad Q_T := (0, T) \times \Omega,& \\ \partial_\nu u^0_\varepsilon = 0 \quad \text{on} \quad (0, T) \times \partial\Omega,& \\ u^0_\varepsilon(0, \cdot) = f_0 \quad \text{in} \quad \Omega,&\end{aligned}\right\} \qquad (6.18)$$

$$\left.\begin{aligned}
&-\frac{\partial z_\varepsilon}{\partial t} - \varepsilon\Delta z_\varepsilon - \operatorname{div}\left[\left(|R_\eta \nabla u^0_\varepsilon| + \varepsilon\right)^{p_{u^0_\varepsilon}(t,x)-2} \nabla z_\varepsilon\right] + \kappa z_\varepsilon\\
&\qquad - \operatorname{div}\left[(p_{u^0_\varepsilon}(t,x) - 2)\left(|R_\eta \nabla u^0_\varepsilon| + \varepsilon\right)^{p_{u^0_\varepsilon}(t,x)-4} R_\eta \nabla u^0_\varepsilon \left(\nabla u^0_\varepsilon, \nabla z_\varepsilon\right)\right]\\
&\qquad + \eta^2 \operatorname{div}\left[\left(|R_\eta \nabla u^0_\varepsilon| + \varepsilon\right)^{p_{u^0_\varepsilon}(t,x)-2} (\theta\otimes\theta) \nabla z_\varepsilon\right]\\
&\qquad + \eta^2 \operatorname{div}\left[(p_{u^0_\varepsilon}(t,x) - 2)\left(|R_\eta \nabla u^0_\varepsilon| + \varepsilon\right)^{p_{u^0_\varepsilon}(t,x)-4} R_\eta \nabla u^0_\varepsilon (\theta\otimes\theta) \nabla z_\varepsilon\right]\\
&\qquad - \operatorname{div}\left(|R_\eta \nabla u^0_\varepsilon|^{p_{u^0_\varepsilon}(t,x)-2} \frac{2a^2(1-\delta)\log\left(|R_\eta \nabla u^0_\varepsilon|\right)}{\left(a^2 + |\nabla u^0_\varepsilon|^2\right)^2} \left(R_\eta \nabla u^0_\varepsilon \otimes \nabla u^0_\varepsilon\right) \nabla z_\varepsilon\right)\\
&\qquad = -\frac{\mu}{\omega}(u^0_\varepsilon - f_0)\chi_{[T-\omega,T]}(t) \quad in \ \ Q_T := (0,T)\times\Omega,\\
&\ \partial_\nu z_\varepsilon = 0 \quad on \ \ (0,T)\times\partial\Omega,\\
&z_\varepsilon(T,\cdot) = 0 \quad in \ \ \Omega,
\end{aligned}\right\} \tag{6.19}$$

$$\int_0^T \left[\int_\Omega \frac{2v^0_\varepsilon}{|v^0_\varepsilon| + \varepsilon} \int_\Omega |v^0_\varepsilon|\, dx + \kappa z_\varepsilon \left(v - v^0_\varepsilon\right)\, dx\right] dt \geqslant 0, \quad \forall\, v \in \mathcal{V}_{ad}. \tag{6.20}$$

Here, $\chi_{[T-\omega,T]}(t)$ stands for the characteristic function of the set $[T-\omega, T]$.

Finally, it is worth to notice that the elliptic operator in the principle part of the system 7.15b is coercive, monotone, and hemicontinuous for each $\varepsilon > 0$. Hence, by the classical results of the theory of linear PDE [31], a weak solution of the adjoint system (6.19) is unique in the weighted space

$$\left\{z \in z_\varepsilon \in L^2(0,T; H^{p^-,u^0_\varepsilon}(\Omega)), \quad \dot{z}_\varepsilon \in L^2(0,T; \left[H^{p^-,u^0_\varepsilon}(\Omega)\right]')\right\}.$$

References

1. S. Amat, S. Busquier, J. Ruiz, J.C. Trillo S. Zouaoui, *On some new variational problems for image denoising*, *Mathematical Methods in the Applied Sciences*, Special Issue: New Advances for Computational and Mathematical Methods in scientific problems, **42**(17) (2019), 5881–5897.
2. L. Afraites, A. Hadri, A. Laghrib, *A denoising model adapted for impulse and Gaussian noises using a constrained-PDE*, Inverse Problems, **36**(2) (2020), Id:025006.
3. 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, **16**(4) (2022), 827–870.
4. Yu.A. Alkhutov, V.V. Zhikov, *Existence theorems for solutions of parabolic equations with variable order of nonlinearity*, Proceedings of the Steklov Institute of Mathematics, **270** (2010), 15–26.
5. Yu.A. Alkhutov, V.V. Zhikov, *Existence and uniqueness theorems for solutions of parabolic equations with a variable nonlinearity exponent*, Sbornik : Mathematics, **205** (3) (2014), 307–318.
6. A.B. Al'shin, M.O. Korpusov, A.G. Sveshnikov, *Blow-up in nonlinear Sobolev type equations.* Walter de Gruyter GmbH & Co. KG, Berlin/New York, 2011.

7. 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.

8. L. Ambrosio, V. Caselles, S. Masnou and J. M. Morel, *The connected components of sets of finite perimeter*, European Journal of Math., **3** (2001), 39–92.

9. B. Andreianov, M. Bendahmane, S. Ouaro, *Structural stability for nonlinear elliptic problems of the $p(x)$- and $p(u)$-laplacian kind*, 2009, HAL Id: hal-00363284.

10. S. Antontsev, S. Shmarev, *Evolution PDEs with Nonstandard Growth Conditions: Existence, Uniqueness, Localization, Blow-up*, Atlantis Studies in Differential Equations, Vol. 4, Atlantis Press, 2015.

11. S. Antontsev, S. Shmarev, *On a class of nonlocal evolution equations with the $p[u(x,t)]$-Laplace operator*, *Nonlinear Analysis: Real World Applications*, **56** (2020), Id 103165, 1–23.

12. S. Antontsev, V. Zhikov, *Higher integrability for parabolic equations of $p(x,t)$-Laplacian type*, *Advances in Differential Equations*, **10** (9) (2005), 1053–1080.

13. T. Barbu, G. Marinoschi, *Image denoising by a nonlinear control technique*, Int. *Journal of Control*, **90** (5) (2017), 1005–1017.

14. P. Blomgren, T.F. Chan, P. Mulet, C. Wong, *Total variation image restoration: Numerical methods and extensions*, In *Proceedings of the IEEE International Conference on Image Processing*, **III** IEEE (1997), 384–387.

15. M. Bokalo, *Initial-boundary value problems for anisotropic parabolic equations with variable exponents of the nonlinearity in unbounded domains with conditions at infinity*, Journal of Optimization, Differential Equations and Their Applications *(JODEA)*, **30** (1) (2022), 98–121.

16. L. Bungert, D.A. Coomes, M.J. Ehrhardt, J. Rasch, R. Reisenhofer, R. & C.-B. Schönlieb, *Blind image fusion for hyperspectral imaging with the directional total variation*, Inverse Problems, **34** (4) (2018), Article 044003.

17. L. Bungert, M.J. Ehrhardt, *Robust Image Reconstruction with Misaligned Structural Information*, *IEEE Access*, **8** (2020), 222944–222955.

18. L. Calatroni, J.C. De Los Reyes, C.B. Schönlieb, *Infimal convolution of data discrepancies for mixed noise removal*, *SIAM Journal on Imaging Sciences*, **10** (3) (2017), 1196–1233.

19. A. Charkaoui, H. Fahim, N.E. Alaa, *Nonlinear parabolic equation having nonstandard growth condition with respect to the gradient and variable exponent*, *Opuscula Math.*, **41** (1) (2021), 25–53.

20. 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.

21. Y. Chen, S. Levine, M. Rao, *Variable exponent, linear growth functionals in image restoration*, *SIAM J. of Appl. Math.*, **66** (4) (2006), 1383–1406.

22. Y. Chen, S. Levine, J. Stanich, *Image Restoration via Nonstandard Diffusion*, Technical Report, Duquesne University, (2004).

23. M. Chipot, H.B. de Oliveira, *Some results on the $p(u)$-Laplacian problem*, *Mathematische Annalen*, **375** (2019), 283–306.

24. D.V. Cruz-Uribe, A. Fiorenza, *Variable Lebesgue Spaces: Foundations and Harmonic Analysis*. Birkhäuser, New York, 2013.

25. C. D'Apice, U. De Maio, P.I. 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.

26. C. D'Apice, U. De Maio, P. I. Kogut, *Gap phenomenon in homogenization of parabolic optimal control problems*, *IMA Journal of Mathematical Control and Information* (Cambridge University), **25** (2008), 461–480.

27. C. D'Apice, P.I. Kogut, R. Manzo, M.V. Uvarov, *Variational Model with Nonstandard Growth Conditions for Restoration of Satellite Optical Images Using Synthetic Aperture Radar*, *Europian Journal of Applies Math.*, **34** (1) (2023), 77–105

28. C. D'Apice, P.I. Kogut, R. Manzo, M.V. Uvarov, *On Variational Problem with Nonstandard Growth Conditions and Its Applications to Image Processing*, *Proceeding of the 19th International Conference of Numerical Analysis and Applied Mathematics*, ICNAAM 2021, 20?26 September 2021, Location: Rhodes, Greece.

29. R. Dautray, J.L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol.5, Springer-Verlag, Berlin Heidelberg, 1985.

30. L. Diening, P. Harjulehto, P. Hästö, M. Ru̇ẑiĉk, *Lebesgue and Sobolev Spaces with Variable Exponents*. Springer, New York, 2011.

31. L.C. Evans, *Partial Differential Equations*. AMS, Graduate Studies in Mathematics, New York, **19**, 2010.

32. A.V. Fursikov, *Optimal Control of Distributed Systems. Theore and Application*, AMS, Providence, 1999.

33. T. Horsin, P. 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.

34. A.D. Ioffe, V.M. Tikhomirov, *Theory of Extremal Problems*. North-Holland, Amsterdam, 1979.

35. D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic, New York, 1980.

36. P.I. Kogut *On approximation of an optimal boundary control problem for linear elliptic equation with unbounded coefficients*, *Discrete and Continuous Dynamical Systems, Series A*, **34**(5) (2014), 2105–2133.

37. P.I. Kogut *Variational S-convergence of minimization problems. Part I. Definitions and basic properties*, *Problemy Upravleniya i Informatiki (Avtomatika)*, **5** (1996), 29–42.

38. P.I. Kogut *S-convergence of the conditional optimization problems and its variational properties*, *Problemy Upravleniya i Informatiki (Avtomatika)*, **4** (1997), 64–79.

39. P.I. Kogut, *On optimal and quasi-optimal controls in coefficients for multidimensional thermistor problem with mixed Dirichlet-Neumann boundary conditions*, *Control and Cybernetics*, **48**(1) (2019), 31–68.

40. P. Kogut, Ya. Kohut, R. Manzo, *Fictitious Controls and Approximation of an Optimal Control Problem for Perona-Malik Equation*, *Journal of Optimization, Differential Equations and Their Applications (JODEA)*, **30** (1) (2022), 42–70.

41. P. Kogut, Ya. Kohut, N. Parfinovych, *Solvability issues for some noncoercive and nonmonotone parabolic equations arising in the image denoising problems*, *Journal of Optimization, Differential Equations and Their Applications (JODEA)*, **30** (2) (2022), 19–48.

42. P.I. Kogut, O.P. Kupenko, *Approximation Methods in Optimization of Nonlinear Systems*, De Gruyter Series in Nonlinear Analysis and Applications 32, Walter de Gruyter GmbH, Berlin, Boston, 2019.

43. P.I. Kogut, O.P. Kupenko, N.V. Uvarov, *On increasing of resolution of satellite images via their fusion with imagery at higher resolution*, J. of Optimization, Differential Equations and Their Applications (JODEA), **29**(1) (2021), 54–78.

44. P.I. Kogut, L. Leugering, *On S-homogenization of an optimal control problem with control and state constraints*, Zeitschrift fur Analysis und ihre Anwendung, **20** (2) (2001) 395–429.

45. 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.

46. 0.A. Ladyshenskaya, V.A. Solonnikov, N.N. Ural'tseva, *Linear and Quasi linear Equation of Parabolic Type.* American Mathematical Society, Providence, RI, 1968.

47. F. Li, Zh. Li, L. Pi, *Variable exponent functionals in image restoration*, Applied Mathematics and Computation, **216** (2010), 870–882.

48. G. Marinoschi, *Dual Variational Approach to Nonlinear Duffusion Equations*, Series: Progress in Nonlinear Differential Equations and Their Applications, **102**, Birkhäuser, Switzerland, 2023.

49. L. Nirenberg, *Topics in Nonlinear Analysis*, Lecture Notes, New York University, New York, 1974.

50. S.Ouaro, N. Sawadogo, *Structural Stability of Nonlinear Elliptic p(u)-Laplacian Problem with Robin Type Boundary Condition.* In: *Studies in Evolution Equations and Related Topics. STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health*, Springer, Cham, 2021, (2021), 69–111.

51. P. Perona, J. Malik, *Scale-space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Machine Intelligence, **12** (1990), 161–192.

52. V. Rădulescu, D. Repovš, *Partial differential equations with variable exponents: variational methods and qualitative analysis*, CRC Press, Boca Raton, London, New York, 2015.

53. J Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. pura Appl., **146** (1987), 65–96.

54. C.-B. Schönlieb, *Partial Differential Equations Methods for Image Inpainting*, Cambridge University Press, 2015.

55. A. Defant, K. Floret, *Tensor Norms and Operator Ideals*, North-Holland Math. Stud., 176, North-Holland, Amsterdam, 1993.

56. A. Tolksdorf, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Equations, **51** (1984), 126–150.

57. V.V. Zhikov, *Solvability of the thre-dimensional thermistor problem*, Proceedings of the Steklov Institute of Mathematics **281** (2008), 98–111.

58. V.V. Zhikov, *On variational problems and nonlinear elliptic equations with nonstandard growth conditions*, Journal of Mathematical Sciences, **173**:5 (2011), 463–570.

59. V.V. Zhikov, *On the weak convergence of fluxes to a flux*, Doklady Mathematics,, **81**:1 (2010), 58–62.

60. V.V. Zhikov, S.E. Pastukhova, *Lemmas on compensated compactness in elliptic and parabolic equations*, Proceedings of the Steklov Institute of Mathematics, **270** (2010), 104–131.

Received 17.12.2023

Journal of Optimization, Differential Equations and Their Applications

Volume 31 **Issue 2** **December 2023**

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 31 **Issue 2** **December 2023**

ISSN 2617-0108

03102

9 772617 010000

CONTENTS