Proceedings of EQUADIFF 2017 Conference

Editors:

Karol Mikula, Daniel Ševčovič, and Jozef Urbán

Published by Slovak University of Technology, 2017

SPEKTRUM STU

This volume contains Proceedings of EQUADIFF 2017 Conference held in Bratislava, Slovakia, July 24-28, 2017, and organized by Slovak University of Technology in Bratislava, in cooperation with Comenius University in Bratislava, Union of Slovak Mathematicians and Physicists, Slovak Mathematical Society and Algoritmy:SK, ltd.

© 2017, Published by Slovak University of Technology, SPEKTRUM STU Publishing

Karol Mikula, Daniel Ševčovič, and Jozef Urbán,
Editors of Proceedings of Equadiff 2017 Conference

ISBN 978-80-227-4757-8

Preface

The Equadiff is a series of biannual conferences on mathematical analysis, numerical approximation and applications of differential equations. It is held in rotation in the Czech Republic, Slovakia and Western Europe. The last Equadiff (Equadiff 14 in the Czecho-Slovak series) was organized in Bratislava, Slovakia, July 24-28, 2017 by the Slovak University of Technology, in cooperation with Comenius University, Union of Slovak Mathematicians and Physicists, Slovak Mathematical Society and Algoritmy:SK, ltd.

During the last decades the Equadiff has clearly developed into the world platform for international exchange of ideas on all mathematical and numerical aspects of differential equations, ranging from fundamental concepts to applications.

The scientific program of Equadiff 2017 Conference was proposed and prepared by the members International Scientific Programme Committee: Michal Beneš (Czech Technical University, Prague, Czech Republic), Charlie Elliott (University of Warwick, UK), Eduard Feireisl (Czech Academy of Sciences, Prague, Czech Republic), Marek Fila (Comenius University, Bratislava, Slovakia), Raphaele Herbin (University of Aix-Marseille, France), Grzegorz Karch (University of Wroclaw, Poland), Karol Mikula (Slovak University of Technology, Bratislava, Slovakia), Masayasu Mimura (Meiji University, Tokyo, Japan), Mario Ohlberger (University of Münster, Germany), Peter Poláčik (University of Minnesota, Minneapolis, USA), Otmar Scherzer (University of Vienna, Austria), Pavol Quittner (Comenius University, Bratislava, Slovakia), Eiji Yanagida (Tokyo Institute of Technology, Japan). Organizing Committee of the conference consisted of Peter Frolkovič, Angela Handlovičová, Martin Kalina, Karol Mikula, Daniel Ševčovič, Róbert Špir and Peter Struk. The conference was chaired by Karol Mikula and co-chaired by Marek Fila.

Proceedings of Equadiff 2017 Conference contain peer-reviewed contributions of participants of the conference. The proceedings cover a wide range of topics presented by plenary, minisymposia and contributed talks speakers. The scope of papers ranges from ordinary differential equations, differential inclusions and dynamical systems towards qualitative and numerical analysis of partial differential equations, stochastic PDEs and their applications.

In several papers, the authors studied qualitative and numerical properties of solutions to cross-diffusion systems with entropy structure, boundedness and stabilization of solutions in a three-dimensional and two-species chemotaxis-Navier-Stokes system, boundedness of solutions in a fully parabolic chemotaxis system with signal-dependent sensitivity and logistic term. Several authors studied well-posedness of solutions for a mass conserved Allen-Cahn equation with a nonlinear diffusion term, the porous medium equations and nonlinear cross-diffusion systems and efficient linear numerical scheme for solving the Stefan problem.

The authors also investigated qualitative behavior of solutions of the undamped Klein-Gordon equation and entropy of the attractor of the strongly damped wave equation. The conference proceedings contain papers on dynamical models of viscoplasticity and Lyapunov stability in hypoplasticity models. The proceedings further include papers dealing with qualitative properties of solutions for systems of fractional boundary value problems and analysis of inequalities with gradient nonlinearities and fractional Laplacian operators. The proceedings also contain papers dealing with qualitative properties like uniqueness and regularity of solutions for systems of coupled elliptic and parabolic equations.

Several papers are devoted to the numerical analysis of finite element and discrete Galerkin methods for elliptic problems with nonlinear boundary conditions. Applications of theoretical results cover viral infection modelling with diffusion and state-dependent delay, an analysis of a model of suspension flowing down an inclined plane as well as applications of tree-grid and finite stencil numerical methods in computational finance, optimal control and optimal design. Interesting applications of partial differential equations in image segmentation and computational differential geometry can be also found in the proceedings.

We thank all the authors for their interesting contributions to the conference proceedings. We also thank our reviewers for their valuable comments and suggestions which improved quality of presentation of results.

Bratislava, November 30, 2017

Karol Mikula, Daniel Ševčovič, and Jozef Urbán

Editors of Proceedings of Equadiff 2017 Conference

Table of Contents

Proceedings of EQUADIFF 2017
pp. 1–10

POSITIVE SOLUTIONS FOR A SYSTEM OF FRACTIONAL BOUNDARY VALUE PROBLEMS

JOHNNY HENDERSON* AND RODICA LUCA†

Abstract. We investigate the existence and multiplicity of positive solutions for a system of nonlinear Riemann-Liouville fractional differential equations with nonnegative nonlinearities which can be nonsingular or singular functions, subject to multi-point boundary conditions that contain fractional derivatives.

Key words. Riemann-Liouville fractional differential equations, multi-point boundary conditions, positive solutions, existence

AMS subject classifications. 34A08, 34B15, 45G15

1. Introduction. We consider the system of nonlinear ordinary fractional differential equations

$$(S) \qquad \begin{cases} D^{\alpha}_{0+}u(t)+f(t,v(t))=0, \ \ t\in(0,1), \\ D^{\beta}_{0+}v(t)+g(t,u(t))=0, \ \ t\in(0,1), \end{cases}$$

with the multi-point boundary conditions

$$(BC) \qquad \begin{cases} u^{(j)}(0)=0, \ j=0,\dots,n-2; \ \ D^{p_1}_{0+}u(t)|_{t=1}=\displaystyle\sum_{i=1}^{N}a_iD^{q_1}_{0+}u(t)|_{t=\xi_i}, \\ \\ v^{(j)}(0)=0, \ j=0,\dots,m-2; \ \ D^{p_2}_{0+}v(t)|_{t=1}=\displaystyle\sum_{i=1}^{M}b_iD^{q_2}_{0+}v(t)|_{t=\eta_i}, \end{cases}$$

where α, $\beta\in\mathbf{R}$, $\alpha\in(n-1,n]$, $\beta\in(m-1,m]$, n, $m\in\mathbf{N}$, n, $m\geq 3$, p_1, p_2, q_1, $q_2\in\mathbf{R}$, $p_1\in[1,n-2]$, $p_2\in[1,m-2]$, $q_1\in[0,p_1]$, $q_2\in[0,p_2]$, ξ_i, $a_i\in\mathbf{R}$ for all $i=1,\dots,N$ ($N\in\mathbf{N}$), $0<\xi_1<\cdots<\xi_N\leq 1$, η_i, $b_i\in\mathbf{R}$ for all $i=1,\dots,M$ ($M\in\mathbf{N}$), $0<\eta_1<\cdots<\eta_M\leq 1$, and D^{k}_{0+} denotes the Riemann-Liouville derivative of order k (for $k=\alpha$, β, p_1, p_2, q_1, q_2).

Under sufficient conditions on functions f and g, which can be nonsingular or singular in the points $t=0$ and/or $t=1$, we study the existence and multiplicity of positive solutions of problem $(S)-(BC)$. We use some theorems from the fixed point index theory (from [1] and [27]) and the Guo-Krasnosel'skii fixed point theorem (see [9]). By a positive solution of problem $(S)-(BC)$ we mean a pair of functions $(u,v)\in C([0,1];\mathbf{R}_+)\times C([0,1];\mathbf{R}_+)$ ($\mathbf{R}_+=[0,\infty)$) satisfying (S) and (BC) with $u(t)>0$ and $v(t)>0$ for all $t\in(0,1]$. The system (S) with the boundary conditions

$$(\widetilde{BC}) \qquad \begin{cases} u^{(j)}(0)=0, \ j=0,\dots,n-2; \ \ u(1)=\int_0^1 u(s)\,dH(s), \\ v^{(j)}(0)=0, \ j=0,\dots,m-2; \ \ v(1)=\int_0^1 v(s)\,dK(s), \end{cases}$$

*Department of Mathematics, Baylor University, Waco, Texas, 76798-7328 USA (Johnny_Henderson@baylor.edu).

†Department of Mathematics, Gh. Asachi Technical University, Iasi 700506, Romania (rluca@math.tuiasi.ro).

where the integrals from $(\widetilde{BC})$ are Riemann-Stieltjes integrals, has been investigated in [10]. The existence, multiplicity and nonexistence of positive solutions for the system (S) and the corresponding one with some positive parameters, namely the system

$$(S') \qquad \begin{cases} D^{\alpha}_{0+}u(t) + \lambda f(t, u(t), v(t)) = 0, \ \ t \in (0,1), \\ D^{\beta}_{0+}v(t) + \mu g(t, u(t), v(t)) = 0, \ \ t \in (0,1), \end{cases}$$

subject to coupled boundary conditions

$$(BC') \qquad \begin{cases} u^{(j)}(0) = 0, \ j = 0, \dots, n-2; \ \ u(1) = \int_0^1 v(s)\, dH(s), \\ v^{(j)}(0) = 0, \ j = 0, \dots, m-2; \ \ v(1) = \int_0^1 u(s)\, dK(s), \end{cases}$$

were studied in [11], [12], [13], [14], [16], [19], where the nonlinearities f and g are nonnegative or sign-changing functions. Fractional differential equations describe many phenomena in several fields of engineering and scientific disciplines such as physics, biophysics, chemistry, biology (for example, the primary infection with HIV), economics, control theory, signal and image processing, thermoelasticity, aerodynamics, viscoelasticity, electromagnetics and rheology (see [2], [3], [4], [5], [6], [7], [8], [17], [18], [20], [21], [22], [23], [24], [25], [26]). Fractional differential equations are also regarded as a better tool for the description of hereditary properties of various materials and processes than the corresponding integer order differential equations.

The paper is organized as follows. In Section 2, we present some auxiliary results which investigate a nonlocal boundary value problem for fractional differential equations, and give the properties of the Green functions associated to our problem. Section 3 contains the existence and multiplicity results for the positive solutions of problem $(S) - (BC)$ in the nonsingular case, and Section 4 presents the existence results in the singular case. Finally, in Section 5 we give two examples which support our main results.

2. Auxiliary results. We present here the definitions of Riemann-Liouville fractional integral and Riemann-Liouville fractional derivative, and some auxiliary results from [15] that will be used to prove our main results.

Definition 2.1 *The (left-sided) fractional integral of order $\alpha > 0$ of a function $f : (0, \infty) \to \mathbf{R}$ is given by*

$$(I^{\alpha}_{0+}f)(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\, ds, \ \ t > 0,$$

provided the right-hand side is pointwise defined on $(0, \infty)$, where $\Gamma(\alpha)$ is the Euler gamma function defined by $\Gamma(\alpha) = \displaystyle\int_0^{\infty} t^{\alpha-1}e^{-t}\, dt$, $\alpha > 0$.

Definition 2.2 *The Riemann-Liouville fractional derivative of order $\alpha \geq 0$ for a function $f : (0, \infty) \to \mathbf{R}$ is given by*

$$(D^{\alpha}_{0+}f)(t) = \left(\frac{d}{dt}\right)^n \left(I^{n-\alpha}_{0+}f\right)(t) = \frac{1}{\Gamma(n-\alpha)} \left(\frac{d}{dt}\right)^n \int_0^t \frac{f(s)}{(t-s)^{\alpha-n+1}}\, ds, \ \ t > 0,$$

where $n = \lfloor \alpha \rfloor + 1$, provided that the right-hand side is pointwise defined on $(0, \infty)$.

The notation $\lfloor \alpha \rfloor$ stands for the largest integer not greater than α. If $\alpha = m \in \mathbf{N}$ then $D^m_{0+}f(t) = f^{(m)}(t)$ for $t > 0$, and if $\alpha = 0$ then $D^0_{0+}f(t) = f(t)$ for $t > 0$.

We consider now the fractional differential equation

$$D_{0+}^{\alpha}u(t)+x(t)=0,\ \ 0<t<1, \tag{2.1}$$

with the multi-point boundary conditions

$$u^{(j)}(0)=0,\ \ j=0,\dots,n-2;\ \ D_{0+}^{p_1}u(t)\big|_{t=1}=\sum_{i=1}^{N}a_iD_{0+}^{q_1}u(t)\big|_{t=\xi_i}, \tag{2.2}$$

where $\alpha\in(n-1,n]$, $n\in\mathbf{N}$, $n\geq 3$, a_i, $\xi_i\in\mathbf{R}$, $i=1,\dots,N$ ($N\in\mathbf{N}$), $0<\xi_1<\cdots<\xi_N\leq 1$, p_1, $q_1\in\mathbf{R}$, $p_1\in[1,n-2]$, $q_1\in[0,p_1]$, and $x\in C(0,1)\cap L^1(0,1)$. We denote by $\Delta_1=\frac{\Gamma(\alpha)}{\Gamma(\alpha-p_1)}-\frac{\Gamma(\alpha)}{\Gamma(\alpha-q_1)}\sum_{i=1}^{N}a_i\xi_i^{\alpha-q_1-1}$.

LEMMA 2.1. *([15]) If $\Delta_1\neq 0$, then the function $u\in C[0,1]$ given by*

$$u(t)=\int_0^1 G_1(t,s)x(s)\,ds,\ \ t\in[0,1], \tag{2.3}$$

is solution of problem (2.1)-(2.2), where

$$G_1(t,s)=g_1(t,s)+\frac{t^{\alpha-1}}{\Delta_1}\sum_{i=1}^{N}a_ig_2(\xi_i,s),\ \ \forall\,(t,s)\in[0,1]\times[0,1], \tag{2.4}$$

and

$$\begin{aligned} g_1(t,s)&=\frac{1}{\Gamma(\alpha)}\begin{cases} t^{\alpha-1}(1-s)^{\alpha-p_1-1}-(t-s)^{\alpha-1},\ \ 0\leq s\leq t\leq 1,\\ t^{\alpha-1}(1-s)^{\alpha-p_1-1},\ \ 0\leq t\leq s\leq 1,\end{cases}\\ g_2(t,s)&=\frac{1}{\Gamma(\alpha-q_1)}\begin{cases} t^{\alpha-q_1-1}(1-s)^{\alpha-p_1-1}-(t-s)^{\alpha-q_1-1},\\ \qquad 0\leq s\leq t\leq 1,\\ t^{\alpha-q_1-1}(1-s)^{\alpha-p_1-1},\ \ 0\leq t\leq s\leq 1.\end{cases}\end{aligned} \tag{2.5}$$

LEMMA 2.2. *([15]) The functions g_1 and g_2 given by (2.5) have the properties:*

a) $g_1(t,s)\leq h_1(s)$ for all t, $s\in[0,1]$, where
$h_1(s)=\frac{1}{\Gamma(\alpha)}(1-s)^{\alpha-p_1-1}(1-(1-s)^{p_1})$, $s\in[0,1]$;

b) $g_1(t,s)\geq t^{\alpha-1}h_1(s)$ for all t, $s\in[0,1]$;

c) $g_1(t,s)\leq\frac{t^{\alpha-1}}{\Gamma(\alpha)}$, for all t, $s\in[0,1]$;

d) $g_2(t,s)\geq t^{\alpha-q_1-1}h_2(s)$ for all t, $s\in[0,1]$, where
$h_2(s)=\frac{1}{\Gamma(\alpha-q_1)}(1-s)^{\alpha-p_1-1}(1-(1-s)^{p_1-q_1})$, $s\in[0,1]$;

e) $g_2(t,s)\leq\frac{1}{\Gamma(\alpha-q_1)}t^{\alpha-q_1-1}$ for all t, $s\in[0,1]$;

f) The functions g_1 and g_2 are continuous on $[0,1]\times[0,1]$; $g_1(t,s)\geq 0$, $g_2(t,s)\geq 0$ for all t, $s\in[0,1]$; $g_1(t,s)>0$, $g_2(t,s)>0$ for all t, $s\in(0,1)$.

LEMMA 2.3. *([15]) Assume that $a_i\geq 0$ for all $i=1,\dots,N$ and $\Delta_1>0$. Then the function G_1 given by (2.4) is a nonnegative continuous function on $[0,1]\times[0,1]$ and satisfies the inequalities:*

a) $G_1(t,s)\leq J_1(s)$ for all t, $s\in[0,1]$, where $J_1(s)=h_1(s)+\frac{1}{\Delta_1}\sum_{i=1}^{N}a_ig_2(\xi_i,s)$, $s\in[0,1]$;

b) $G_1(t,s)\geq t^{\alpha-1}J_1(s)$ for all t, $s\in[0,1]$;

c) $G_1(t,s)\leq\sigma_1t^{\alpha-1}$, for all t, $s\in[0,1]$, where $\sigma_1=\frac{1}{\Gamma(\alpha)}+\frac{1}{\Delta_1\Gamma(\alpha-q_1)}$ $\times\sum_{i=1}^{N}a_i\xi_i^{\alpha-q_1-1}$.

LEMMA 2.4. *([15]) Assume that $a_i \geq 0$ for all $i = 1, \ldots, N$, $\Delta_1 > 0$, $x \in C(0,1) \cap L^1(0,1)$ and $x(t) \geq 0$ for all $t \in (0,1)$. Then the solution u of problem (2.1)-(2.2) given by (2.3) satisfies the inequality $u(t) \geq t^{\alpha-1} u(t')$ for all t, $t' \in [0,1]$.*

We can also formulate similar results as Lemmas 2.1-2.4 for the fractional boundary value problem

$$D_{0+}^{\beta} v(t) + y(t) = 0, \ \ 0 < t < 1, \tag{2.6}$$

$$v^{(j)}(0) = 0, \ \ j = 0, \ldots, m-2; \ \ D_{0+}^{p_2} v(t)\big|_{t=1} = \sum_{i=1}^{M} b_i D_{0+}^{q_2} v(t)\big|_{t=\eta_i}, \tag{2.7}$$

where $\beta \in (m-1, m]$, $m \in \mathbf{N}$, $m \geq 3$, b_i, $\eta_i \in \mathbf{R}$, $i = 1, \ldots, M$ ($M \in \mathbf{N}$), $0 < \eta_1 < \cdots < \eta_M \leq 1$, p_2, $q_2 \in \mathbf{R}$, $p_2 \in [1, m-2]$, $q_2 \in [0, p_2]$, and $y \in C(0,1) \cap L^1(0,1)$.

We denote by Δ_2, g_3, g_4, G_2, h_3, h_4, J_2 and σ_2 the corresponding constants and functions for problem (2.6)-(2.7) defined in a similar manner as Δ_1, g_1, g_2, G_1, h_1, h_2, J_1 and σ_1, respectively. More precisely, we have

$$\Delta_2 = \frac{\Gamma(\beta)}{\Gamma(\beta-p_2)} - \frac{\Gamma(\beta)}{\Gamma(\beta-q_2)} \sum_{i=1}^{M} b_i \eta_i^{\beta-q_2-1},$$

$$g_3(t,s) = \frac{1}{\Gamma(\beta)} \begin{cases} t^{\beta-1}(1-s)^{\beta-p_2-1} - (t-s)^{\beta-1}, & 0 \leq s \leq t \leq 1, \\ t^{\beta-1}(1-s)^{\beta-p_2-1}, & 0 \leq t \leq s \leq 1, \end{cases}$$

$$g_4(t,s) = \frac{1}{\Gamma(\beta-q_2)} \begin{cases} t^{\beta-q_2-1}(1-s)^{\beta-p_2-1} - (t-s)^{\beta-q_2-1}, & 0 \leq s \leq t \leq 1, \\ t^{\beta-q_2-1}(1-s)^{\beta-p_2-1}, & 0 \leq t \leq s \leq 1, \end{cases}$$

$$G_2(t,s) = g_3(t,s) + \frac{t^{\beta-1}}{\Delta_2} \sum_{i=1}^{M} b_i g_4(\eta_i, s), \ \ \forall\, (t,s) \in [0,1] \times [0,1],$$

$$h_3(s) = \frac{1}{\Gamma(\beta)} (1-s)^{\beta-p_2-1} (1-(1-s)^{p_2}), \ \ s \in [0,1],$$

$$h_4(s) = \frac{1}{\Gamma(\beta-q_2)} (1-s)^{\beta-p_2-1} (1-(1-s)^{p_2-q_2}), \ \ s \in [0,1],$$

$$J_2(s) = h_3(s) + \frac{1}{\Delta_2} \sum_{i=1}^{M} b_i g_4(\eta_i, s), \ \ s \in [0,1],$$

$$\sigma_2 = \frac{1}{\Gamma(\beta)} + \frac{1}{\Delta_2 \Gamma(\beta-q_2)} \sum_{i=1}^{M} b_i \eta_i^{\beta-q_2-1}.$$

The inequalities from Lemmas 2.3 and 2.4 for the functions G_2 and v are the following $G_2(t,s) \leq J_2(s)$, $G_2(t,s) \geq t^{\beta-1} J_2(s)$, $G_2(t,s) \leq \sigma_2 t^{\beta-1}$, for all t, $s \in [0,1]$, and $v(t) \geq t^{\beta-1} v(t')$ for all t, $t' \in [0,1]$.

The proofs of our results in the nonsingular case are based on the following fixed point index theorems. Let E be a real Banach space, $P \subset E$ a cone, "$\leq$" the partial ordering defined by P and θ the zero element in E. For $\varrho > 0$, let $B_\varrho = \{u \in E, \ \|u\| < \varrho\}$ be the open ball of radius ϱ centered at 0, and its boundary $\partial B_\varrho = \{u \in E, \ \|u\| = \varrho\}$.

THEOREM 2.5. *([1]) Let $A : \bar{B}_\varrho \cap P \to P$ be a completely continuous operator which has no fixed point on $\partial B_\varrho \cap P$. If $\|Au\| \leq \|u\|$ for all $u \in \partial B_\varrho \cap P$, then $i(A, B_\varrho \cap P, P) = 1$.*

THEOREM 2.6. *([1]) Let $A : \bar{B}_\varrho \cap P \to P$ be a completely continuous operator. If there exists $u_0 \in P \setminus \{\theta\}$ such that $u - Au \neq \lambda u_0$, for all $\lambda \geq 0$ and $u \in \partial B_\varrho \cap P$, then $i(A, B_\varrho \cap P, P) = 0$.*

THEOREM 2.7. *([27]) Let $A : \bar{B}_\varrho \cap P \to P$ be a completely continuous operator which has no fixed point on $\partial B_\varrho \cap P$. If there exists a linear operator $L : P \to P$ and $u_0 \in P \setminus \{\theta\}$ such that*

$$i) \ u_0 \leq Lu_0, \ \ ii) \ Lu \leq Au, \ \ \forall\, u \in \partial B_\varrho \cap P,$$

then $i(A, B_\varrho \cap P, P) = 0$.

We also present the Guo-Krasnosel'skii fixed point theorem (see [9]) that we will use in the proofs of our main results in the singular case.

THEOREM 2.8. *Let X be a Banach space and let $C \subset X$ be a cone in X. Assume Ω_1 and Ω_2 are bounded open subsets of X with $0 \in \Omega_1 \subset \overline{\Omega_1} \subset \Omega_2$ and let $\mathcal{A} : C \cap (\overline{\Omega_2} \setminus \Omega_1) \to C$ be a completely continuous operator such that, either*

i) $\|\mathcal{A}u\| \leq \|u\|$, $u \in C \cap \partial\Omega_1$, and $\|\mathcal{A}u\| \geq \|u\|$, $u \in C \cap \partial\Omega_2$, or

ii) $\|\mathcal{A}u\| \geq \|u\|$, $u \in C \cap \partial\Omega_1$, and $\|\mathcal{A}u\| \leq \|u\|$, $u \in C \cap \partial\Omega_2$.

Then $\mathcal{A}$ has a fixed point in $C \cap (\overline{\Omega_2} \setminus \Omega_1)$.

3. The nonsingular case. In this section, we investigate the existence and multiplicity of positive solutions for problem $(S) - (BC)$ under various assumptions on nonsingular functions f and g.

We present the assumptions that we shall use in the sequel.

$(H1)$ $\alpha, \beta \in \mathbf{R}$, $\alpha \in (n-1, n]$, $\beta \in (m-1, m]$, $n, m \in \mathbf{N}$, $n, m \geq 3$, $p_1, p_2, q_1, q_2 \in \mathbf{R}$, $p_1 \in [1, n-2]$, $p_2 \in [1, m-2]$, $q_1 \in [0, p_1]$, $q_2 \in [0, p_2]$, $\xi_i \in \mathbf{R}$, $a_i \geq 0$ for all $i = 1, \ldots, N$ $(N \in \mathbf{N})$, $0 < \xi_1 < \cdots < \xi_N \leq 1$, $\eta_i \in \mathbf{R}$, $b_i \geq 0$ for all $i = 1, \ldots, M$ $(M \in \mathbf{N})$, $0 < \eta_1 < \cdots < \eta_M \leq 1$, $\Delta_1 = \frac{\Gamma(\alpha)}{\Gamma(\alpha-p_1)} - \frac{\Gamma(\alpha)}{\Gamma(\alpha-q_1)} \sum_{i=1}^{N} a_i \xi_i^{\alpha-q_1-1} > 0$, $\Delta_2 = \frac{\Gamma(\beta)}{\Gamma(\beta-p_2)} - \frac{\Gamma(\beta)}{\Gamma(\beta-q_2)} \sum_{i=1}^{M} b_i \eta_i^{\beta-q_2-1} > 0$.

$(H2)$ The functions f, $g : [0,1] \times \mathbf{R}_+ \to \mathbf{R}_+$ are continuous and $f(t,0) = g(t,0) = 0$ for all $t \in [0,1]$.

If the pair of functions $(u, v) \in C[0,1] \times C[0,1]$ is a solution of the nonlinear integral system

$$(3.1) \qquad \begin{cases} u(t) = \displaystyle\int_0^1 G_1(t,s) f\left(s, \int_0^1 G_2(s,\tau) g(\tau, u(\tau))\, d\tau\right) ds, \quad t \in [0,1], \\ v(t) = \displaystyle\int_0^1 G_2(t,s) g(s, u(s))\, ds, \quad t \in [0,1], \end{cases}$$

then it is a solution of problem $(S) - (BC)$.

We consider the Banach space $X = C[0,1]$ with supremum norm $\|\cdot\|$ and define the cone $P \subset X$ by $P = \{u \in X, \ u(t) \geq 0, \ \forall\, t \in [0,1]\}$.

We also define the operators $\mathcal{A} : P \to X$ by

$$(\mathcal{A}u)(t) = \int_0^1 G_1(t,s) f\left(s, \int_0^1 G_2(s,\tau) g(\tau, u(\tau))\, d\tau\right) ds, \ t \in [0,1], \ u \in P,$$

and $\mathcal{B} : P \to X$, $\mathcal{C} : P \to X$ by

$$(\mathcal{B}u)(t) = \int_0^1 G_1(t,s) u(s)\, ds, \ \ (\mathcal{C}u)(t) = \int_0^1 G_2(t,s) u(s)\, ds, \ \ t \in [0,1], \ u \in P.$$

Under the assumptions $(H1)$ and $(H2)$ it is easy to see that $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are completely continuous from P to P. Thus we will investigate the existence and multiplicity of fixed points u of operator $\mathcal{A}$, which together with v given in (3.1) will be solutions of problem $(S) - (BC)$.

Using Theorems 2.5-2.6 and some similar arguments as those used in the proofs of Theorems 3.1-3.3 from [10], we obtain for our problem $(S) - (BC)$ the following results.

THEOREM 3.1. *Assume that $(H1) - (H2)$ hold. If the functions f and g also satisfy the conditions*

$(H3)$ *There exist positive constants* $p \in (0,1]$ *and* $c \in (0,1)$ *such that*

$$i)\ f_\infty^i = \liminf_{u\to\infty} \inf_{t\in[c,1]} \frac{f(t,u)}{u^p} \in (0,\infty]; \quad ii)\ g_\infty^i = \lim_{u\to\infty} \inf_{t\in[c,1]} \frac{g(t,u)}{u^{1/p}} = \infty,$$

$(H4)$ *There exists positive constants* $\beta_1,\ \beta_2 > 0$ *with* $\beta_1\beta_2 \geq 1$ *such that*

$$i)\ f_0^s = \limsup_{u\to 0^+} \sup_{t\in[0,1]} \frac{f(t,u)}{u^{\beta_1}} \in [0,\infty); \quad ii)\ g_0^s = \lim_{u\to 0^+} \sup_{t\in[0,1]} \frac{g(t,u)}{u^{\beta_2}} = 0,$$

then the problem $(S)-(BC)$ *has at least one positive solution* $(u(t),v(t)),\ t\in[0,1]$.

THEOREM 3.2. *Assume that* $(H1)-(H2)$ *hold. If the functions* f *and* g *also satisfy the conditions*

$(H5)$ *There exist positive constants* $\alpha_1,\ \alpha_2 > 0$ *with* $\alpha_1\alpha_2 \leq 1$ *such that*

$$i)\ f_\infty^s = \limsup_{u\to\infty} \sup_{t\in[0,1]} \frac{f(t,u)}{u^{\alpha_1}} \in [0,\infty); \quad ii)\ g_\infty^s = \lim_{u\to\infty} \sup_{t\in[0,1]} \frac{g(t,u)}{u^{\alpha_2}} = 0,$$

$(H6)$ *There exists* $c \in (0,1)$ *such that*

$$i)\ f_0^i = \liminf_{u\to 0^+} \inf_{t\in[c,1]} \frac{f(t,u)}{u} \in (0,\infty]; \quad ii)\ g_0^i = \lim_{u\to 0^+} \inf_{t\in[c,1]} \frac{g(t,u)}{u} = \infty,$$

then the problem $(S)-(BC)$ *has at least one positive solution* $(u(t),v(t)),\ t\in[0,1]$.

THEOREM 3.3. *Assume that* $(H1)-(H3)$ *and* $(H6)$ *hold. If the functions* f *and* g *also satisfy the condition*

$(H7)$ *For each* $t\in[0,1]$, $f(t,u)$ *and* $g(t,u)$ *are nondecreasing with respect to* u, *and there exists a constant* $N_0 > 0$ *such that*

$$f\left(t, m_0 \int_0^1 g(s,N_0)\,ds\right) < \frac{N_0}{m_0}, \ \ \forall\, t\in[0,1],$$

where $m_0 = \max\{K_1,K_2\}$, $K_1 = \max\limits_{s\in[0,1]} J_1(s)$, $K_2 = \max\limits_{s\in[0,1]} J_2(s)$ *and* $J_1,\ J_2$ *are defined in Section 2,*

then the problem $(S)-(BC)$ *has at least two positive solutions* $(u_1(t),v_1(t))$, $(u_2(t),v_2(t)),\ t\in[0,1]$.

4. The singular case. In this section we study the existence of positive solutions for our problem $(S)-(BC)$ under various assumptions on functions f and g which may be singular at $t=0$ and/or $t=1$.

The basic assumptions used here are the following.

$(A1)$ $\equiv (H1)$,

$(A2)$ The functions $f,\ g\in C((0,1)\times\mathbf{R}_+,\mathbf{R}_+)$ and there exist $\widetilde{p}_i\in C((0,1),\mathbf{R}_+)$, $\widetilde{q}_i\in C(\mathbf{R}_+,\mathbf{R}_+)$, $i=1,2$, with $0<\int_0^1 \widetilde{p}_i(t)\,dt<\infty$, $i=1,2$, $\widetilde{q}_1(0)=0$, $\widetilde{q}_2(0)=0$ such that

$$f(t,x)\leq \widetilde{p}_1(t)\widetilde{q}_1(x), \ \ g(t,x)\leq \widetilde{p}_2(t)\widetilde{q}_2(x), \ \ \forall\, t\in(0,1), \ \ x\in\mathbf{R}_+.$$

We consider the Banach space $X=C([0,1])$ with supremum norm and define the cone $P\subset X$ by $P=\{u\in X,\ u(t)\geq 0,\ \forall\, t\in[0,1]\}$. We also define the operator $\widetilde{\mathcal{A}}:P\to X$ by

$$(\widetilde{\mathcal{A}}u)(t) = \int_0^1 G_1(t,s) f\left(s, \int_0^1 G_2(s,\tau) g(\tau,u(\tau))\,d\tau\right) ds, \ \ t\in[0,1], \ \ u\in P.$$

Using Theorem 2.8 and similar arguments as those used in the proofs of Lemma 4.1 and Theorems 4.1-4.2 from [10], we obtain for our problem $(S)-(BC)$ the following results.

LEMMA 4.1. *Assume that* $(A1)-(A2)$ *hold. Then* $\widetilde{\mathcal{A}}: P \to P$ *is completely continuous.*

THEOREM 4.2. *Assume that* $(A1)-(A2)$ *hold. If the functions* f *and* g *also satisfy the conditions*

$(A3)$ *There exist* $\alpha_1,\ \alpha_2 \in (0,\infty)$ *with* $\alpha_1\alpha_2 \leq 1$ *such that*

$$i)\ q_{1\infty}^s = \limsup_{x\to\infty} \frac{\widetilde{q}_1(x)}{x^{\alpha_1}} \in [0,\infty);\quad ii)\ q_{2\infty}^s = \lim_{x\to\infty} \frac{\widetilde{q}_2(x)}{x^{\alpha_2}} = 0,$$

$(A4)$ *There exist* $\beta_1,\ \beta_2 \in (0,\infty)$ *with* $\beta_1\beta_2 \leq 1$ *and* $c \in (0,1/2)$ *such that*

$$i)\ \widetilde{f}_0^i = \liminf_{x\to 0^+} \inf_{t\in[c,1-c]} \frac{f(t,x)}{x^{\beta_1}} \in (0,\infty];\quad ii)\ \widetilde{g}_0^i = \lim_{x\to 0^+} \inf_{t\in[c,1-c]} \frac{g(t,x)}{x^{\beta_2}} = \infty,$$

then the problem $(S)-(BC)$ *has at least one positive solution* $(u(t),v(t)),\ t\in[0,1]$.

THEOREM 4.3. *Assume that* $(A1)-(A2)$ *hold. If the functions* f *and* g *also satisfy the conditions*

$(A5)$ *There exist* $r_1,\ r_2 \in (0,\infty)$ *with* $r_1r_2 \geq 1$ *such that*

$$i)\ q_{10}^s = \limsup_{x\to 0^+} \frac{\widetilde{q}_1(x)}{x^{r_1}} \in [0,\infty);\quad ii)\ q_{20}^s = \lim_{x\to 0^+} \frac{\widetilde{q}_2(x)}{x^{r_2}} = 0,$$

$(A6)$ *There exist* $l_1,\ l_2 \in (0,\infty)$ *with* $l_1l_2 \geq 1$ *and* $c \in (0,1/2)$ *such that*

$$i)\ \widetilde{f}_\infty^i = \liminf_{x\to\infty} \inf_{t\in[c,1-c]} \frac{f(t,x)}{x^{l_1}} \in (0,\infty];\ ii)\ \widetilde{g}_\infty^i = \lim_{x\to\infty} \inf_{t\in[c,1-c]} \frac{g(t,x)}{x^{l_2}} = \infty,$$

then the problem $(S)-(BC)$ *has at least one positive solution* $(u(t),v(t)),\ t\in[0,1]$.

As against to the positive solutions obtained in the paper [10], in this paper, by using Lemma 2.4, we deduce that the fixed points u of operators $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ together with v given in (3.1) satisfy the conditions $u(t)>0$ and $v(t)>0$ for all $t\in(0,1]$, that is the pairs (u,v) are positive solutions of problem $(S)-(BC)$ in the sense of definition from Section 1.

5. Examples. Let $n=3$, $m=5$, $\alpha=\frac{5}{2}$, $\beta=\frac{17}{4}$, $p_1=1$, $q_1=\frac{1}{2}$, $p_2=\frac{7}{3}$, $q_2=\frac{3}{2}$, $N=2$, $M=1$, $\xi_1=\frac{1}{3}$, $\xi_2=\frac{2}{3}$, $a_1=2$, $a_2=\frac{1}{2}$, $\eta_1=\frac{1}{2}$, $b_1=4$.

We consider the system of fractional differential equations

$$(S_0)\qquad \begin{cases} D_{0+}^{5/2}u(t)+f(t,v(t))=0,\ \ t\in(0,1),\\ D_{0+}^{17/4}v(t)+g(t,u(t))=0,\ \ t\in(0,1),\end{cases}$$

with the multi-point boundary conditions

$$(BC_0)\qquad \begin{cases} u(0)=u'(0)=0,\ u'(1)=2D_{0+}^{1/2}u(t)|_{t=\frac{1}{3}}+\frac{1}{2}D_{0+}^{1/2}u(t)|_{t=\frac{2}{3}},\\ v(0)=v'(0)=v''(0)=v'''(0)=0,\ D_{0+}^{7/3}v(t)|_{t=1}=4D_{0+}^{3/2}v(t)|_{t=\frac{1}{2}}.\end{cases}$$

We have $\Delta_1=\frac{6-3\sqrt{\pi}}{4}\approx 0.17065961>0$, $\Delta_2=\frac{\Gamma(17/4)}{\Gamma(23/12)}-\frac{2^{1/4}\Gamma(17/4)}{\Gamma(11/4)}\approx 2.43672831>0$. So assumptions $(H1)$ and $(A1)$ are satisfied.

Besides we deduce

$$g_1(t,s)=\frac{1}{\Gamma(5/2)}\begin{cases} t^{3/2}(1-s)^{1/2}-(t-s)^{3/2}, & 0\le s\le t\le 1,\\ t^{3/2}(1-s)^{1/2}, & 0\le t\le s\le 1,\end{cases}$$
$$g_2(t,s)=\begin{cases} t(1-s)^{1/2}-(t-s), & 0\le s\le t\le 1,\\ t(1-s)^{1/2}, & 0\le t\le s\le 1,\end{cases}$$
$$g_3(t,s)=\frac{1}{\Gamma(17/4)}\begin{cases} t^{13/4}(1-s)^{11/12}-(t-s)^{13/4}, & 0\le s\le t\le 1,\\ t^{13/4}(1-s)^{11/12}, & 0\le t\le s\le 1,\end{cases}$$
$$g_4(t,s)=\frac{1}{\Gamma(11/4)}\begin{cases} t^{7/4}(1-s)^{11/12}-(t-s)^{7/4}, & 0\le s\le t\le 1,\\ t^{7/4}(1-s)^{11/12}, & 0\le t\le s\le 1.\end{cases}$$

Then we obtain

$$G_1(t,s)=g_1(t,s)+\frac{t^{3/2}}{\Delta_1}\left(2g_2\left(\frac{1}{3},s\right)+\frac{1}{2}g_2\left(\frac{2}{3},s\right)\right),$$
$$G_2(t,s)=g_3(t,s)+\frac{4t^{13/4}}{\Delta_2}g_4\left(\frac{1}{2},s\right),$$
$$h_1(s)=\frac{4}{3\sqrt{\pi}}s(1-s)^{1/2},\ h_3(s)=\frac{1}{\Gamma(17/4)}(1-s)^{11/12}(1-(1-s)^{7/3}),$$
$$J_1(s)=\frac{4}{3\sqrt{\pi}}s(1-s)^{1/2}+\frac{1}{\Delta_1}\left(2g_2\left(\frac{1}{3},s\right)+\frac{1}{2}g_2\left(\frac{2}{3},s\right)\right)$$
$$=\begin{cases}\frac{4}{3\sqrt{\pi}}s(1-s)^{1/2}+\frac{1}{2\Delta_1}\left[2(1-s)^{1/2}+5s-2\right], & 0\le s<\frac{1}{3},\\ \frac{4}{3\sqrt{\pi}}s(1-s)^{1/2}+\frac{1}{6\Delta_1}\left[6(1-s)^{1/2}+3s-2\right], & \frac{1}{3}\le s<\frac{2}{3},\\ \frac{4}{3\sqrt{\pi}}s(1-s)^{1/2}+\frac{1}{\Delta_1}(1-s)^{1/2}, & \frac{2}{3}\le s\le 1.\end{cases}$$
$$J_2(s)=\frac{1}{\Gamma(17/4)}(1-s)^{11/12}(1-(1-s)^{7/3})+\frac{4}{\Delta_2}g_4\left(\frac{1}{2},s\right)$$
$$=\begin{cases}\frac{1}{\Gamma(17/4)}(1-s)^{11/12}(1-(1-s)^{7/3})+\frac{2^{1/4}}{\Delta_2\Gamma(11/4)}[(1-s)^{11/12}-(1-2s)^{7/4}],\\ \qquad 0\le s<\frac{1}{2},\\ \frac{1}{\Gamma(17/4)}(1-s)^{11/12}(1-(1-s)^{7/3})+\frac{2^{1/4}}{\Delta_2\Gamma(11/4)}(1-s)^{11/12}, \quad \frac{1}{2}\le s\le 1.\end{cases}$$

Example 1. We consider the functions

$$f(t,u)=a(u^{\alpha_0}+u^{\beta_0}),\ \ g(t,u)=b(u^{\gamma_0}+u^{\delta_0}),\ \ t\in[0,1],\ \ u\ge 0,$$

where $\alpha_0>1$, $0<\beta_0<1$, $\gamma_0>2$, $0<\delta_0<1$, $a,b>0$. We have $K_1=\max_{s\in[0,1]}J_1(s)\approx 4.01249183$, $K_2=\max_{s\in[0,1]}J_2(s)\approx 0.22467674$. Then $m_0=\max\{K_1,K_2\}=K_1$. The functions $f(t,u)$ and $g(t,u)$ satisfy the assumption $(H2)$. Besides, they are nondecreasing with respect to u, for any $t\in[0,1]$, and for $p=1/2$ and $c\in(0,1)$ the assumptions $(H3)$ and $(H6)$ are satisfied; indeed we obtain

$$f_\infty^i=\lim_{u\to\infty}\frac{a(u^{\alpha_0}+u^{\beta_0})}{u^{1/2}}=\infty,\ \ g_\infty^i=\lim_{u\to\infty}\frac{b(u^{\gamma_0}+u^{\delta_0})}{u^2}=\infty,$$
$$f_0^i=\lim_{u\to 0^+}\frac{a(u^{\alpha_0}+u^{\beta_0})}{u}=\infty,\ \ g_0^i=\lim_{u\to 0^+}\frac{b(u^{\gamma_0}+u^{\delta_0})}{u}=\infty.$$

We take $N_0=1$ and then $\int_0^1 g(s,1)\,ds=2b$ and $f(t,2bm_0)=a[(2bm_0)^{\alpha_0}+(2bm_0)^{\beta_0}]$. If $a[(2bm_0)^{\alpha_0}+(2bm_0)^{\beta_0}]<\frac{1}{m_0}\ \Leftrightarrow\ a\left[m_0^{\alpha_0+1}(2b)^{\alpha_0}+m_0^{\beta_0+1}(2b)^{\beta_0}\right]<1$, then the

assumption $(H7)$ is satisfied. For example, if $\alpha_0 = 3/2$, $\beta_0 = 1/3$, $b = 1/2$ and $a < \dfrac{1}{m_0^{5/2} + m_0^{4/3}}$ (e.g. $a \leq 0.0258$), then the above inequality is satisfied. By Theorem 3.3, we deduce that the problem $(S_0) - (BC_0)$ has at least two positive solutions.

Example 2. We consider the functions

$$f(t,x) = \frac{x^a}{t^{\zeta_1}(1-t)^{\rho_1}}, \;\; g(t,x) = \frac{x^b}{t^{\zeta_2}(1-t)^{\rho_2}},$$

with $a, b > 1$ and $\zeta_1, \rho_1, \zeta_2, \rho_2 \in (0,1)$. Here $f(t,x) = \widetilde{p}_1(t)\widetilde{q}_1(x)$ and $g(t,x) = \widetilde{p}_2(t)\widetilde{q}_2(x)$, where

$$\widetilde{p}_1(t) = \frac{1}{t^{\zeta_1}(1-t)^{\rho_1}}, \;\; \widetilde{p}_2(t) = \frac{1}{t^{\zeta_2}(1-t)^{\rho_2}}, \;\; \widetilde{q}_1(x) = x^a, \;\; \widetilde{q}_2(x) = x^b.$$

We have $0 < \int_0^1 \widetilde{p}_1(s)\,ds < \infty$, $0 < \int_0^1 \widetilde{p}_2(s)\,ds < \infty$, so the functions f and g satisfy the assumption $(A2)$.

In $(A5)$, for $r_1 < a, \;\; r_2 < b$ and $r_1 r_2 \geq 1$, we obtain

$$\lim_{x\to 0^+} \frac{\widetilde{q}_1(x)}{x^{r_1}} = 0, \;\; \lim_{x\to 0^+} \frac{\widetilde{q}_2(x)}{x^{r_2}} = 0.$$

In $(A6)$, for $l_1 < a, \;\; l_2 < b$, $l_1 l_2 \geq 1$ and $c \in (0, 1/2)$, we have

$$\lim_{x\to\infty} \inf_{t\in[c,1-c]} \frac{f(t,x)}{x^{l_1}} = \infty, \;\; \lim_{x\to\infty} \inf_{t\in[c,1-c]} \frac{g(t,x)}{x^{l_2}} = \infty.$$

For example, if $a = 3/2$, $b = 2$, $r_1 = 1$, $r_2 = 3/2$, $l_1 = 1$, $l_2 = 3/2$, the above conditions are satisfied. Then, by Theorem 4.3, we deduce that the problem $(S_0) - (BC_0)$ has at least one positive solution.

REFERENCES

[1] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Review, 18 (1976), pp. 620–709.

[2] A. A. M. Arafa, S. Z. Rida, M. Khalil, *Fractional modeling dynamics of HIV and CD4$^+$ T-cells during primary infection*, Nonlinear Biomed. Phys., 6 (1) (2012), pp. 1–7.

[3] D. Baleanu, K. Diethelm, E. Scalas, J. J. Trujillo, *Fractional Calculus Models and Numerical Methods. Series on Complexity, Nonlinearity and Chaos*, World Scientific, Boston, 2012.

[4] K. Cole, *Electric conductance of biological systems*, in: Proc. Cold Spring Harbor Symp. Quant. Biol., Col Springer Harbor Laboratory Press, New York, 1993, pp. 107–116.

[5] S. Das, *Functional Fractional Calculus for System Identification and Controls*, Springer, New York, 2008.

[6] Y. Ding, H. Ye, *A fractional-order differential equation model of HIV infection of CD4$^+$ T-cells*, Math. Comp. Model., 50 (2009), pp. 386–392.

[7] V. Djordjevic, J. Jaric, B. Fabry, J. Fredberg, D. Stamenovic, *Fractional derivatives embody essential features of cell rheological behavior*, Ann. Biomed. Eng., 31 (2003), pp. 692–699.

[8] Z. M. Ge, C. Y. Ou, *Chaos synchronization of fractional order modified Duffing systems with parameters excited by a chaotic signal*, Chaos Solitons Fractals, 35 (2008), pp. 705–717.

[9] D. Guo, V. Lakshmikantham, *Nonlinear Problems in Abstract Cones*, Academic Press, New York, 1988.

[10] J. Henderson, R. Luca, *Existence and multiplicity of positive solutions for a system of fractional boundary value problems*, Bound. Value Probl., 2014:60 (2014), pp. 1–17.

[11] J. Henderson, R. Luca, *Positive solutions for a system of fractional differential equations with coupled integral boundary conditions*, Appl. Math. Comput., 249 (2014), pp. 182–197.
[12] J. Henderson, R. Luca, *Nonexistence of positive solutions for a system of coupled fractional boundary value problems*, Bound. Value Probl., 2015:138 (2015), pp. 1–12.
[13] J. Henderson, R. Luca, *Positive solutions for a system of semipositone coupled fractional boundary value problems*, Bound. Value Probl., 2016(61) (2016), pp. 1–23.
[14] J. Henderson, R. Luca, *Boundary Value Problems for Systems of Differential, Difference and Fractional Equations. Positive Solutions*, Elsevier, Amsterdam, 2016.
[15] J. Henderson, R. Luca, *Existence of positive solutions for a singular fractional boundary value problem*, Nonlinear Anal. Model. Control, 22(1) (2017), pp. 99–114.
[16] J. Henderson, R. Luca, A. Tudorache, *On a system of fractional differential equations with coupled integral boundary conditions*, Fract. Calc. Appl. Anal., 18(2) (2015), pp. 361–386.
[17] A. A. Kilbas, H. M. Srivastava, J. J. Trujillo, *Theory and Applications of Fractional Differential Equations*, North-Holland Mathematics Studies, 204, Elsevier Science B.V., Amsterdam, 2006.
[18] J. Klafter, S. C. Lim, R. Metzler (Eds.), *Fractional Dynamics in Physics*, Singapore, World Scientific, 2011.
[19] R. Luca, A. Tudorache, *Positive solutions to a system of semipositone fractional boundary value problems*, Adv. Difference Equ., 2014(179) (2014), pp. 1–11.
[20] R. Metzler, J. Klafter, *The random walks guide to anomalous diffusion: a fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.
[21] M. Ostoja-Starzewski, *Towards thermoelasticity of fractal media*, J. Therm. Stress., 30 (2007), pp. 889–896.
[22] I. Podlubny, *Fractional Differential Equations*, Academic Press, San Diego, 1999.
[23] Y.Z. Povstenko, *Fractional Thermoelasticity*, New York, Springer, 2015.
[24] J. Sabatier, O. P. Agrawal, J. A. T. Machado (Eds.), *Advances in Fractional Calculus: Theoretical Developments and Applications in Physics and Engineering*, Springer, Dordrecht, 2007.
[25] S. G. Samko, A. A. Kilbas, O. I. Marichev, *Fractional Integrals and Derivatives. Theory and Applications*, Gordon and Breach, Yverdon, 1993.
[26] I. M. Sokolov, J. Klafter, A. Blumen, *A fractional kinetics*, Phys. Today, 55 (2002), pp. 48–54.
[27] Y. Zhou, Y. Xu, *Positive solutions of three-point boundary value problems for systems of nonlinear second order ordinary differential equations*, J. Math. Anal. Appl., 320 (2006), pp. 578–590.

Proceedings of EQUADIFF 2017
pp. 11–20

BOUNDEDNESS AND STABILIZATION IN A THREE-DIMENSIONAL TWO-SPECIES CHEMOTAXIS-NAVIER-STOKES SYSTEM WITH COMPETITIVE KINETICS*

MISAKI HIRATA, SHUNSUKE KURIMA, MASAAKI MIZUKAMI, TOMOMI YOKOTA†

Abstract. This paper is concerned with the two-species chemotaxis-Navier–Stokes system with Lotka–Volterra competitive kinetics

$$\begin{cases} (n_1)_t + u\cdot\nabla n_1 = \Delta n_1 - \chi_1\nabla\cdot(n_1\nabla c) + \mu_1 n_1(1-n_1-a_1n_2) & \text{in } \Omega\times(0,\infty),\\ (n_2)_t + u\cdot\nabla n_2 = \Delta n_2 - \chi_2\nabla\cdot(n_2\nabla c) + \mu_2 n_2(1-a_2n_1-n_2) & \text{in } \Omega\times(0,\infty),\\ c_t + u\cdot\nabla c = \Delta c - (\alpha n_1+\beta n_2)c & \text{in } \Omega\times(0,\infty),\\ u_t + (u\cdot\nabla)u = \Delta u + \nabla P + (\gamma n_1+\delta n_2)\nabla\Phi, \quad \nabla\cdot u = 0 & \text{in } \Omega\times(0,\infty) \end{cases}$$

under homogeneous Neumann boundary conditions and initial conditions, where Ω is a bounded domain in $\mathbb{R}^3$ with smooth boundary. Recently, in the 2-dimensional setting, global existence and stabilization of classical solutions to the above system were first established. However, the 3-dimensional case has not been studied: Because of difficulties in the Navier–Stokes system, we can not expect existence of classical solutions to the above system. The purpose of this paper is to obtain global existence of weak solutions to the above system, and their eventual smoothness and stabilization.

Key words. chemotaxis, Navier–Stokes, Lotka–Volterra, large-time behaviour

AMS subject classifications. 35B40, 35K55, 35Q30, 92C17

1. Introduction. This paper deals with the following two-species chemotaxis-Navier–Stokes system with Lotka–Volterra competitive kinetics:

$$\begin{cases} (n_1)_t + u\cdot\nabla n_1 = \Delta n_1 - \chi_1\nabla\cdot(n_1\nabla c) + \mu_1 n_1(1-n_1-a_1n_2) & \text{in } \Omega\times(0,\infty),\\ (n_2)_t + u\cdot\nabla n_2 = \Delta n_2 - \chi_2\nabla\cdot(n_2\nabla c) + \mu_2 n_2(1-a_2n_1-n_2) & \text{in } \Omega\times(0,\infty),\\ c_t + u\cdot\nabla c = \Delta c - (\alpha n_1+\beta n_2)c & \text{in } \Omega\times(0,\infty),\\ u_t + \kappa(u\cdot\nabla)u = \Delta u + \nabla P + (\gamma n_1+\delta n_2)\nabla\Phi, \quad \nabla\cdot u = 0 & \text{in } \Omega\times(0,\infty),\\ \partial_\nu n_1 = \partial_\nu n_2 = \partial_\nu c = 0, \quad u = 0 & \text{on } \partial\Omega\times(0,\infty),\\ n_1(\cdot,0) = n_{1,0}, \quad n_2(\cdot,0) = n_{2,0}, \quad c(\cdot,0) = c_0, \quad u(\cdot,0) = u_0 & \text{in } \Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^3$ with smooth boundary $\partial\Omega$ and ∂_ν denotes differentiation with respect to the outward normal of $\partial\Omega$; $\kappa = 1$, $\chi_1, \chi_2, a_1, a_2 \geq 0$ and $\mu_1, \mu_2, \alpha, \beta, \gamma, \delta > 0$ are constants; $n_{1,0}, n_{2,0}, c_0, u_0, \Phi$ are known functions satisfying

$$0 < n_{1,0}, n_{2,0} \in C(\overline{\Omega}), \quad 0 < c_0 \in W^{1,q}(\Omega), \quad u_0 \in D(A^\theta), \tag{1.2}$$

$$\Phi \in C^{1+\lambda}(\overline{\Omega}) \tag{1.3}$$

for some $q > 3$, $\theta \in (\frac{3}{4}, 1)$, $\lambda \in (0,1)$ and A denotes the realization of the Stokes operator under homogeneous Dirichlet boundary conditions in the solenoidal subspace $L^2_\sigma(\Omega)$ of $L^2(\Omega)$.

*This work was supported by JSPS Research Fellowships for Young Scientists, No. 17J00101 and Grant-in-Aid for Scientific Research (C), No. 16K05182.

†Department of Mathematics, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan (yokota@rs.kagu.tus.ac.jp).

In the mathematical point of view, difficulties of this problem are mainly caused by the chemotaxis terms $-\chi_1\nabla\cdot(n_1\nabla c)$, $-\chi_2\nabla\cdot(n_2\nabla c)$, the competitive kinetics $\mu_1 n_1(1-n_1-a_1n_2)$, $\mu_2 n_2(1-a_2n_1-n_2)$ and the Navier–Stokes equation which is the fourth equation in (1.1). In the case that $n_2=0$, global existence of weak solutions, and their eventual smoothness and stabilization were shown in [5]. On the other hand, in the case that $n_2\neq 0$ and $\Omega\subset\mathbb{R}^2$, global existence and boundedness of classical solutions to (1.1) have been attained ([4]). Moreover, in the case that $\kappa=0$ in (1.1), which namely means that the fourth equation in (1.1) is the *Stokes* equation, global existence and stabilization can be found in [2]; in the case that $\kappa=0$ in (1.1) and that $-(\alpha n_1+\beta n_2)c$ is replaced with $+\alpha n_1+\beta n_2-c$, global existence and boundedness of classical solutions to the Keller–Segel-Stokes system and their asymptotic behaviour are found in [3].

As we mentioned above, global classical solutions are found in (1.1) in the 2-dimensional setting and the case that $\kappa=0$. However, global existence of solutions in 3-dimensional setting has not been attained. Thus the main purposes of this paper is to obtain global existence of solutions to (1.1) in the case that $\Omega\subset\mathbb{R}^3$. Nevertheless, because of the difficulties of the Navier–Stokes equation, we can not expect global existence of *classical solutions* to (1.1) in the 3-dimensional case. Therefore our goal is to obtain global existence of *weak solutions* to (1.1) in the following sense.

Definition 1.1. *A quadruple (n_1,n_2,c,u) is called a (global) weak solution of (1.1) if*

$$\begin{aligned} n_1,n_2 &\in L^2_{\mathrm{loc}}([0,\infty);L^2(\Omega))\cap L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty);W^{1,\frac{4}{3}}(\Omega)),\\ c &\in L^2_{\mathrm{loc}}([0,\infty);W^{1,2}(\Omega)),\\ u &\in L^2_{\mathrm{loc}}([0,\infty);W^{1,2}_{0,\sigma}(\Omega)) \end{aligned}$$

and for all $T>0$ the identities

$$\begin{aligned} &-\int_0^\infty\!\!\int_\Omega n_1\varphi_t-\int_\Omega n_{1,0}\varphi(\cdot,0)-\int_0^\infty\!\!\int_\Omega n_1u\cdot\nabla\varphi\\ &\quad=-\int_0^\infty\!\!\int_\Omega\nabla n_1\cdot\nabla\varphi+\chi_1\int_0^\infty\!\!\int_\Omega n_1\nabla c\cdot\nabla\varphi+\mu_1\int_0^\infty\!\!\int_\Omega n_1(1-n_1-a_1n_2)\varphi,\\ &-\int_0^\infty\!\!\int_\Omega n_2\varphi_t-\int_\Omega n_{2,0}\varphi(\cdot,0)-\int_0^\infty\!\!\int_\Omega n_2u\cdot\nabla\varphi\\ &\quad=-\int_0^\infty\!\!\int_\Omega\nabla n_2\cdot\nabla\varphi+\chi_2\int_0^\infty\!\!\int_\Omega n_2\nabla c\cdot\nabla\varphi+\mu_2\int_0^\infty\!\!\int_\Omega n_2(1-a_2n_1-n_2)\varphi,\\ &-\int_0^\infty\!\!\int_\Omega c\varphi_t-\int_\Omega c_0\varphi(\cdot,0)-\int_0^\infty\!\!\int_\Omega cu\cdot\nabla\varphi\\ &\quad=-\int_0^\infty\!\!\int_\Omega\nabla c\cdot\nabla\varphi-\int_0^\infty\!\!\int_\Omega(\alpha n_1+\beta n_2)c\varphi,\\ &-\int_0^\infty\!\!\int_\Omega u\cdot\psi_t-\int_\Omega u_0\cdot\psi(\cdot,0)-\int_0^\infty\!\!\int_\Omega u\otimes u\cdot\nabla\psi\\ &\quad=-\int_0^\infty\!\!\int_\Omega\nabla u\cdot\nabla\psi+\int_0^\infty\!\!\int_\Omega(\gamma n_1+\delta n_2)\nabla\psi\cdot\nabla\Phi \end{aligned}$$

hold for all $\varphi\in C_0^\infty(\overline{\Omega}\times[0,\infty))$ and all $\psi\in C^\infty_{0,\sigma}(\Omega\times[0,\infty))$, respectively.

Now the main results read as follows. The first theorem is concerned with global existence of weak solutions to (1.1).

THEOREM 1.2. *Let $\Omega \subset \mathbb{R}^3$ be a bounded smooth domain and let $\chi_1, \chi_2, a_1, a_2 \geq 0$ and $\mu_1, \mu_2, \alpha, \beta, \gamma, \delta > 0$. Assume that $n_{1,0}, n_{2,0}, c_0, u_0$ satisfy (1.2) with some $q > 3$ and $\theta \in (\frac{3}{4}, 1)$ and $\Phi \in C^{1+\lambda}(\overline{\Omega})$ for some $\lambda \in (0,1)$. Then there is a weak solution of (1.1), which can be approximated by a sequence of solutions $(n_{1,\varepsilon}, n_{2,\varepsilon}, c_\varepsilon, u_\varepsilon)$ of (2.1) (see Section 2) in a pointwise manner.*

The second theorem gives eventual smoothness and stabilization.

THEOREM 1.3. *Let the assumption of Theorem 1.2 be satisfied. Then there are $T > 0$ and $\alpha' \in (0,1)$ such that the solution (n_1, n_2, c, u) given by Theorem 1.2 satisfies*

$$n_1, n_2, c \in C^{2+\alpha', 1+\frac{\alpha'}{2}}(\overline{\Omega} \times [T, \infty)), \quad u \in C^{2+\alpha', 1+\frac{\alpha'}{2}}(\overline{\Omega} \times [T, \infty)).$$

Moreover, the solution of (1.1) has the following properties:

(i) *Assume that $a_1, a_2 \in (0,1)$. Then*

$$n_1(\cdot,t) \to N_1, \quad n_2(\cdot,t) \to N_2, \quad c(\cdot,t) \to 0, \quad u(\cdot,t) \to 0 \quad \text{in } L^\infty(\Omega)$$

as $t \to \infty$, where

$$N_1 := \frac{1-a_1}{1-a_1a_2}, \quad N_2 := \frac{1-a_2}{1-a_1a_2}.$$

(ii) *Assume that $a_1 \geq 1 > a_2$. Then*

$$n_1(\cdot,t) \to 0, \quad n_2(\cdot,t) \to 1, \quad c(\cdot,t) \to 0, \quad u(\cdot,t) \to 0 \quad \text{in } L^\infty(\Omega)$$

as $t \to \infty$.

The proofs of the main theorems are based on the arguments in [5]. The strategies for the proofs is to construct energy estimates for the solution $(n_{1,\varepsilon}, n_{2,\varepsilon}, c_\varepsilon, u_\varepsilon)$ of (2.1). In Section 2 we consider the energy function $\mathcal{F}_\varepsilon$ defined as

$$\mathcal{F}_\varepsilon := \int_\Omega n_{1,\varepsilon} \log n_{1,\varepsilon} + \int_\Omega n_{2,\varepsilon} \log n_{2,\varepsilon} + \frac{\chi}{2} \int_\Omega \frac{|\nabla c_\varepsilon|^2}{c_\varepsilon} + k_4 \chi \int_\Omega |u_\varepsilon|^2$$

with some constant $\chi > 0$. Noting that for all $\rho, \xi_i > 0$ there exists $C > 0$ such that

$$\begin{aligned} &\int_\Omega \nabla c_\varepsilon \cdot \nabla n_{i,\varepsilon} \left(\frac{\chi_i}{1+\varepsilon n_{i,\varepsilon}} - \frac{\chi\alpha \text{ (or } \chi\beta)}{1+\varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})} \right) \\ &\leq \rho \int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + \xi_i \int_\Omega \frac{|\nabla n_{i,\varepsilon}|^2}{n_{i,\varepsilon}} + C \int_\Omega n_{i,\varepsilon}^2 \quad (i = 1,2), \end{aligned}$$

which did not appear in the previous work [5], from the estimate for the energy function $\mathcal{F}_\varepsilon$ we obtain global-in-time solvability of approximate solutions. Then we moreover see convergence as $\varepsilon \searrow 0$. Furthermore, in Section 3, according to an argument similar to [4], by putting

$$\mathcal{G}_{\varepsilon,B} := \int_\Omega \left(n_{1,\varepsilon} - N_1 \log \frac{n_{1,\varepsilon}}{N_1} \right) + \int_\Omega \left(n_{2,\varepsilon} - N_2 \log \frac{n_{2,\varepsilon}}{N_2} \right) + \frac{B}{2} \int_\Omega c_\varepsilon^2$$

with suitable constant $B > 0$ and establishing the Hölder estimates for the solution of (1.1) through the estimate for the energy function $\mathcal{G}_{\varepsilon,B}$, we can discuss convergence of $\big(n_1(\cdot,t), n_2(\cdot,t), c(\cdot,t), u(\cdot,t)\big)$ as $t \to \infty$.

2. Proof of Theorem 1.2 (Global existence). We will start by considering an approximate problem with parameter $\varepsilon > 0$, namely:

$$\begin{cases} (n_{1,\varepsilon})_t + u_\varepsilon \cdot \nabla n_{1,\varepsilon} = \Delta n_{1,\varepsilon} - \chi_1 \nabla \cdot \left(\frac{n_{1,\varepsilon}}{1+\varepsilon n_{1,\varepsilon}} \nabla c_\varepsilon \right) + \mu_1 n_{1,\varepsilon}(1 - n_{1,\varepsilon} - a_1 n_{2,\varepsilon}), \\ (n_{2,\varepsilon})_t + u_\varepsilon \cdot \nabla n_{2,\varepsilon} = \Delta n_{2,\varepsilon} - \chi_2 \nabla \cdot \left(\frac{n_{2,\varepsilon}}{1+\varepsilon n_{2,\varepsilon}} \nabla c_\varepsilon \right) + \mu_2 n_{2,\varepsilon}(1 - a_2 n_{1,\varepsilon} - n_{2,\varepsilon}), \\ (c_\varepsilon)_t + u_\varepsilon \cdot \nabla c_\varepsilon = \Delta c_\varepsilon - c_\varepsilon \frac{1}{\varepsilon} \log\big(1 + \varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})\big), \\ (u_\varepsilon)_t + (Y_\varepsilon u_\varepsilon \cdot \nabla) u_\varepsilon = \Delta u_\varepsilon + \nabla P_\varepsilon + (\gamma n_{1,\varepsilon} + \delta n_{2,\varepsilon}) \nabla \Phi, \quad \nabla \cdot u_\varepsilon = 0, \\ \partial_\nu n_{1,\varepsilon}|_{\partial\Omega} = \partial_\nu n_{2,\varepsilon}|_{\partial\Omega} = \partial_\nu c_\varepsilon|_{\partial\Omega} = 0, \quad u_\varepsilon|_{\partial\Omega} = 0, \\ n_{1,\varepsilon}(\cdot,0) = n_{1,0}, \quad n_{2,\varepsilon}(\cdot,0) = n_{2,0}, \quad c_\varepsilon(\cdot,0) = c_0, \quad u_\varepsilon(\cdot,0) = u_0, \end{cases} \tag{2.1}$$

where $Y_\varepsilon = (1 + \varepsilon A)^{-1}$, and provide estimates for its solutions. We first give the following result which states local existence in (1.1).

LEMMA 2.1. *Let $\chi_1, \chi_2, a_1, a_2 \geq 0$, $\mu_1, \mu_2, \alpha, \beta, \gamma, \delta > 0$, and $\Phi \in C^{1+\lambda}(\overline{\Omega})$ for some $\lambda \in (0,1)$ and assume that $n_{1,0}, n_{2,0}, c_0, u_0$ satisfy (1.2) with some $q > 3, \theta \in (\frac{3}{4}, 1)$. Then for all $\varepsilon > 0$ there are $T_{\max,\varepsilon}$ and uniquely determined functions*:

$$\begin{aligned} n_{1,\varepsilon}, n_{2,\varepsilon} &\in C^0(\overline{\Omega} \times [0, T_{\max,\varepsilon})) \cap C^{2,1}(\overline{\Omega} \times (0, T_{\max,\varepsilon})), \\ c_\varepsilon &\in C^0(\overline{\Omega} \times [0, T_{\max,\varepsilon})) \cap C^{2,1}(\overline{\Omega} \times (0, T_{\max,\varepsilon})) \cap L^\infty_{\mathrm{loc}}([0, T_{\max,\varepsilon}); W^{1,q}(\Omega)), \\ u_\varepsilon &\in C^0(\overline{\Omega} \times [0, T_{\max,\varepsilon})) \cap C^{2,1}(\overline{\Omega} \times (0, T_{\max,\varepsilon})), \end{aligned}$$

which together with some $P_\varepsilon \in C^{1,0}(\overline{\Omega} \times (0, T_{\max,\varepsilon}))$ solve (2.1) classically. Moreover, $n_{1,\varepsilon}$, $n_{2,\varepsilon}$ and c_ε are positive and the following alternative holds: $T_{\max,\varepsilon} = \infty$ *or*

$$\|n_{1,\varepsilon}(\cdot,t)\|_{L^\infty(\Omega)} + \|n_{2,\varepsilon}(\cdot,t)\|_{L^\infty(\Omega)} + \|c_\varepsilon(\cdot,t)\|_{W^{1,q}(\Omega)} + \|A^\theta u_\varepsilon(\cdot,t)\|_{L^2(\Omega)} \to \infty \tag{2.2}$$

as $t \nearrow T_{\max,\varepsilon}$.

We next show the following lemma which holds a key for the proof of Theorem 1.2. This lemma derives the estimate for the energy function.

LEMMA 2.2. *For all $\xi_1, \xi_2 \in (0,1)$ and $\chi > 0$ there are $C, \overline{C}, \widetilde{C}, k, \overline{k} > 0$ such that*

$$\mathcal{F}_\varepsilon := \int_\Omega n_{1,\varepsilon} \log n_{1,\varepsilon} + \int_\Omega n_{2,\varepsilon} \log n_{2,\varepsilon} + \frac{\chi}{2} \int_\Omega \frac{|\nabla c_\varepsilon|^2}{c_\varepsilon} + \overline{k}\chi \int_\Omega |u_\varepsilon|^2$$

satisfies

$$\begin{aligned} \frac{d}{dt}\mathcal{F}_\varepsilon \leq & -\frac{\mu_1}{4} \int_\Omega n_{1,\varepsilon}^2 \log n_{1,\varepsilon} - \frac{\mu_2}{4} \int_\Omega n_{2,\varepsilon}^2 \log n_{2,\varepsilon} \\ & - (1-\xi_1) \int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} - (1-\xi_2) \int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}} + C \int_\Omega n_{1,\varepsilon}^2 + \overline{C} \int_\Omega n_{2,\varepsilon}^2 + \widetilde{C} \\ & - k \int_\Omega c_\varepsilon |D^2 \log c_\varepsilon|^2 - k \int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} - k \int_\Omega |\nabla u_\varepsilon|^2 \end{aligned}$$

on $(0, T_{\max,\varepsilon})$ for all $\varepsilon > 0$.

Proof. Noting, the boundedness of $s(1-s)$ and $s(1-\frac{s}{2})\log s$, we have that there exists $C_1>0$ such that

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega n_{1,\varepsilon}\log n_{1,\varepsilon}\\
&=-\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}}+\chi_1\int_\Omega \frac{\nabla c_\varepsilon\cdot\nabla n_{1,\varepsilon}}{1+\varepsilon n_{1,\varepsilon}}\\
&\quad+\mu_1\int_\Omega n_{1,\varepsilon}(1-n_{1,\varepsilon}-a_1n_{2,\varepsilon})\log n_{1,\varepsilon}+\mu_1\int_\Omega n_{1,\varepsilon}(1-n_{1,\varepsilon}-a_1n_{2,\varepsilon})\\
&\le-\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}}+\chi_1\int_\Omega \frac{\nabla c_\varepsilon\cdot\nabla n_{1,\varepsilon}}{1+\varepsilon n_{1,\varepsilon}}-\frac{\mu_1}{2}\int_\Omega n_{1,\varepsilon}^2\log n_{1,\varepsilon}\\
&\quad-\mu_1a_1\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}\log n_{1,\varepsilon}-\mu_1a_1\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}+C_1. \qquad (2.3)
\end{aligned}$$

Similarly, there is $C_2>0$ such that

$$\begin{aligned}
\frac{d}{dt}\int_\Omega n_{2,\varepsilon}\log n_{2,\varepsilon}\le&-\int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}}+\chi_2\int_\Omega \frac{\nabla c_\varepsilon\cdot\nabla n_{2,\varepsilon}}{1+\varepsilon n_{2,\varepsilon}}-\frac{\mu_2}{2}\int_\Omega n_{2,\varepsilon}^2\log n_{2,\varepsilon}\\
&-\mu_2a_2\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}\log n_{2,\varepsilon}-\mu_2a_2\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}+C_2. \qquad (2.4)
\end{aligned}$$

According to an argument similar to that in the proof of [5, Lemma 2.8], there exist $k_1,C_3,C_4>0$ such that

$$\begin{aligned}
\frac{d}{dt}\int_\Omega \frac{|\nabla c_\varepsilon|^2}{c_\varepsilon}\le&-k_1\int_\Omega c_\varepsilon|D^2\log c_\varepsilon|^2-k_1\int_\Omega\frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3}\\
&+C_3+C_4\int_\Omega|\nabla u_\varepsilon|^2-2\int_\Omega\frac{\alpha\nabla c_\varepsilon\cdot\nabla n_{1,\varepsilon}+\beta\nabla c_\varepsilon\cdot\nabla n_{2,\varepsilon}}{1+\varepsilon(\alpha n_{1,\varepsilon}+\beta n_{2,\varepsilon})}. \qquad (2.5)
\end{aligned}$$

Now we let $\overline{k},\eta_1,\eta_2,k$ be constants satisfying $\frac{C_4}{2}-\overline{k}=-\frac{k_1}{4}$, $\eta_1=\frac{\mu_1}{4\overline{k}\chi}$, $\eta_2=\frac{b\mu_2}{4\overline{k}\chi}$ and $k=\frac{\chi k_1}{4}$. Then we have

$$\frac{d}{dt}\int_\Omega|u_\varepsilon|^2=-2\int_\Omega|\nabla u_\varepsilon|^2-2\int_\Omega u_\varepsilon\cdot(Y_\varepsilon u_\varepsilon\cdot\nabla)u_\varepsilon+2\int_\Omega u_\varepsilon\cdot(\gamma n_{1,\varepsilon}+\delta n_{2,\varepsilon})\nabla\Phi.$$

From the Schwarz inequality, the Poincaré inequality, the Young inequality and the fact that $\int_\Omega\varphi^2\le a\int_\Omega\varphi^2\log\varphi+|\Omega|e^{\frac{1}{a}}$ holds for any positive function φ and any $a>0$, there exist $C_5,C_{\eta_1},C_{\eta_2}>0$ such that

$$\begin{aligned}
\gamma\int_\Omega|n_{1,\varepsilon}\nabla\Phi\cdot u_\varepsilon|&\le\gamma\|\nabla\Phi\|_{L^\infty}\left(\int_\Omega n_{1,\varepsilon}^2\right)^{\frac12}\left(\int_\Omega|u_\varepsilon|^2\right)^{\frac12}\\
&\le\gamma\|\nabla\Phi\|_{L^\infty}\left(\int_\Omega n_{1,\varepsilon}^2\right)^{\frac12}\left(C_5\int_\Omega|\nabla u_\varepsilon|^2\right)^{\frac12}\\
&\le\gamma^2C_5\|\nabla\Phi\|_{L^\infty}^2\int_\Omega n_{1,\varepsilon}^2+\frac14\int_\Omega|\nabla u_\varepsilon|^2\\
&\le\frac{\eta_1}{2}\int_\Omega n_{1,\varepsilon}^2\log n_{1,\varepsilon}+\frac{C_{\eta_1}}{2}+\frac14\int_\Omega|\nabla u_\varepsilon|^2
\end{aligned}$$

and

$$\delta \int_\Omega |n_{2,\varepsilon} \nabla \Phi \cdot u_\varepsilon| \le \frac{\eta_2}{2} \int_\Omega n_{2,\varepsilon}^2 \log n_{2,\varepsilon} + \frac{C_{\eta_2}}{2} + \frac{1}{4} \int_\Omega |\nabla u_\varepsilon|^2$$

hold. Therefore we have

$$\frac{d}{dt} \int_\Omega |u_\varepsilon|^2 \le - \int_\Omega |\nabla u_\varepsilon|^2 + \eta_1 \int_\Omega n_{1,\varepsilon}^2 \log n_{1,\varepsilon} + \eta_2 \int_\Omega n_{2,\varepsilon}^2 \log n_{2,\varepsilon} + C_{\eta_1} + C_{\eta_2}. \tag{2.6}$$

Thus a combination of (2.3), (2.4), (2.5) and (2.6) leads to

$$\begin{aligned}
&\frac{d}{dt} \left[\int_\Omega n_{1,\varepsilon} \log n_{1,\varepsilon} + \int_\Omega n_{2,\varepsilon} \log n_{2,\varepsilon} + \frac{\chi}{2} \int_\Omega \frac{|\nabla c_\varepsilon|^2}{c_\varepsilon} + \overline{k} \chi \int_\Omega |u_\varepsilon|^2 \right] \\
&\le \left(\overline{k} \chi \eta_1 - \frac{\mu_1}{2} \right) \int_\Omega n_{1,\varepsilon}^2 \log n_{1,\varepsilon} + \left(\overline{k} \chi \eta_2 - \frac{\mu_2}{2} \right) \int_\Omega n_{2,\varepsilon}^2 \log n_{2,\varepsilon} \\
&\quad - \left(\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} + \int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}} \right) + \left(\frac{\chi}{2} C_4 - \overline{k} \chi \right) \int_\Omega |\nabla u_\varepsilon|^2 \\
&\quad + \int_\Omega \nabla c_\varepsilon \cdot \nabla n_{1,\varepsilon} \left(\frac{\chi_1}{1 + \varepsilon n_{1,\varepsilon}} - \frac{\chi \alpha}{1 + \varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})} \right) \\
&\quad + \int_\Omega \nabla c_\varepsilon \cdot \nabla n_{2,\varepsilon} \left(\frac{\chi_2}{1 + \varepsilon n_{2,\varepsilon}} - \frac{\chi \beta}{1 + \varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})} \right) \\
&\quad - \frac{\chi}{2} k_1 \int_\Omega c_\varepsilon |D^2 \log c_\varepsilon|^2 - \frac{\chi}{2} k_1 \int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + C_1 + C_2 + \frac{\chi}{2} C_3 + \overline{k} \chi (C_{\eta_1} + C_{\eta_2}) \\
&\quad - \mu_1 a_1 \int_\Omega n_{1,\varepsilon} n_{2,\varepsilon} (\log n_{1,\varepsilon} + 1) - \mu_2 a_2 \int_\Omega n_{1,\varepsilon} n_{2,\varepsilon} (\log n_{2,\varepsilon} + 1).
\end{aligned}$$

Here, since $n_{1,\varepsilon}, n_{2,\varepsilon}$ are nonnegative, we can find $C_6, C_7 > 0$ such that

$$\begin{aligned}
&\int_\Omega \nabla c_\varepsilon \cdot \nabla n_{1,\varepsilon} \left(\frac{\chi_1}{1 + \varepsilon n_{1,\varepsilon}} - \frac{\chi \alpha}{1 + \varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})} \right) \\
&\le (\chi_1 + \chi \alpha) \int_\Omega |\nabla c_\varepsilon \cdot \nabla n_{1,\varepsilon}| \\
&\le \frac{\chi k_1}{8 \|c_0\|_{L^\infty}^3} \int_\Omega |\nabla c_\varepsilon|^4 + C_6 \int_\Omega |\nabla n_{1,\varepsilon}|^{\frac{4}{3}} \\
&\le \frac{\chi k_1}{8} \int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + \xi_1 \int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} + C_7 \int_\Omega n_{1,\varepsilon}^2
\end{aligned}$$

and there is $C_8 > 0$ such that

$$\begin{aligned}
&\int_\Omega \nabla c_\varepsilon \cdot \nabla n_{2,\varepsilon} \left(\frac{\chi_2}{1 + \varepsilon n_{2,\varepsilon}} - \frac{\chi \beta}{1 + \varepsilon(\alpha n_{1,\varepsilon} + \beta n_{2,\varepsilon})} \right) \\
&\le \frac{\chi k_1}{8} \int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + \xi_2 \int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}} + C_8 \int_\Omega n_{2,\varepsilon}^2,
\end{aligned}$$

which with the fact that $s\log s \geq -\frac{1}{e}$ $(s>0)$ enables us to obtain

$$\begin{aligned}
&\Big(\overline{k}\chi\eta_1 - \frac{\mu_1}{2}\Big)\int_\Omega n_{1,\varepsilon}^2\log n_{1,\varepsilon} + \Big(\overline{k}\chi\eta_2 - \frac{\mu_2}{2}\Big)\int_\Omega n_{2,\varepsilon}^2\log n_{2,\varepsilon}\\
&- \bigg(\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} + \int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}}\bigg) + \Big(\frac{\chi}{2}C_4 - \overline{k}\chi\Big)\int_\Omega |\nabla u_\varepsilon|^2\\
&+ \int_\Omega \nabla c_\varepsilon\cdot\nabla n_{1,\varepsilon}\bigg(\frac{\chi_1}{1+\varepsilon n_{1,\varepsilon}} - \frac{\chi\alpha}{1+\varepsilon(\alpha n_{1,\varepsilon}+\beta n_{2,\varepsilon})}\bigg)\\
&+ \int_\Omega \nabla c_\varepsilon\cdot\nabla n_{2,\varepsilon}\bigg(\frac{\chi_2}{1+\varepsilon n_{2,\varepsilon}} - \frac{\chi\beta}{1+\varepsilon(\alpha n_{1,\varepsilon}+\beta n_{2,\varepsilon})}\bigg)\\
&- \frac{\chi}{2}k_1\int_\Omega c_\varepsilon|D^2\log c_\varepsilon|^2 - \frac{\chi}{2}k_1\int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + C_1 + C_2 + \frac{\chi}{2}C_3 + \overline{k}\chi(C_{\eta_1}+C_{\eta_2})\\
&- \mu_1 a_1\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}(\log n_{1,\varepsilon}+1) - \mu_2 a_2\int_\Omega n_{1,\varepsilon}n_{2,\varepsilon}(\log n_{2,\varepsilon}+1)\\
&\leq -\frac{\mu_1}{4}\int_\Omega n_{1,\varepsilon}^2\log n_{1,\varepsilon} - \frac{\mu_2}{4}\int_\Omega n_{2,\varepsilon}^2\log n_{2,\varepsilon}\\
&\quad - (1-\xi_1)\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} - (1-\xi_2)\int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{1,\varepsilon}}\\
&\quad - k\int_\Omega |\nabla u_\varepsilon|^2 - k\int_\Omega c_\varepsilon|D^2\log c_\varepsilon| - k\int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + C_7\int_\Omega n_{1,\varepsilon}^2 + C_8\int_\Omega n_{2,\varepsilon}^2 + C_9.
\end{aligned}$$

Therefore we obtain this lemma. □

Proof of Theorem 1.2. Let $\tau = \min\{1, \frac{1}{2}T_{\max,\varepsilon}\}$, $\xi_1,\xi_2\in(0,1)$ and $\chi>0$. Lemma 2.2, the facts that $s^2\log s \geq s\log s - \frac{1}{2e}$ $(s>0)$ and $n_{1,\varepsilon}, n_{2,\varepsilon}, c_\varepsilon > 0$ imply

$$\frac{d}{dt}\mathcal{F}_\varepsilon + \mathcal{F}_\varepsilon \leq C\int_\Omega n_{1,\varepsilon}^2 + \overline{C}\int_\Omega n_{2,\varepsilon}^2 + \widetilde{C}'$$

for some $C, \overline{C}, \widetilde{C}' > 0$. According to [5, Lemma 2.5], there exists $C_1>0$ such that

$$\int_t^{t+\tau}\int_\Omega n_{i,\varepsilon}^2 \leq C_1$$

for all $t\in(0, T_{\max,\varepsilon}-\tau)$ and each $i=1,2$. From the uniform Gronwall type lemma (see e.g., [6, Lemma 3.2]) we can find $C_2>0$ such that

$$\int_\Omega n_{1,\varepsilon}\log n_{1,\varepsilon} + \int_\Omega n_{2,\varepsilon}\log n_{2,\varepsilon} + \frac{\chi}{2}\int_\Omega \frac{|\nabla c_\varepsilon|^2}{c_\varepsilon} + \overline{k}\chi\int_\Omega |u_\varepsilon|^2 \leq C_2 \tag{2.7}$$

for all $t\in(0,T_{\max,\varepsilon})$. Moreover, we have from integration of the differential inequality in Lemma 2.2 over $(t,t+\tau)$ that for all $\xi_1,\xi_2\in(0,1)$ there is $C_3>0$ such that

$$\begin{aligned}
&\frac{\mu_1}{4}\int_t^{t+\tau}\int_\Omega n_{1,\varepsilon}^2\log n_{1,\varepsilon} + \frac{\mu_2}{4}\int_t^{t+\tau}\int_\Omega n_{2,\varepsilon}^2\log n_{2,\varepsilon} + k\int_t^{t+\tau}\int_\Omega c_\varepsilon|D^2\log c_\varepsilon|^2\\
&\quad + (1-\xi_1)\int_t^{t+\tau}\int_\Omega \frac{|\nabla n_{1,\varepsilon}|^2}{n_{1,\varepsilon}} + (1-\xi_2)\int_t^{t+\tau}\int_\Omega \frac{|\nabla n_{2,\varepsilon}|^2}{n_{2,\varepsilon}} \leq C_3
\end{aligned}\tag{2.8}$$

and

$$\int_t^{t+\tau}\int_\Omega \frac{|\nabla c_\varepsilon|^4}{c_\varepsilon^3} + \int_t^{t+\tau}\int_\Omega |\nabla u_\varepsilon|^2 \leq C_3 \tag{2.9}$$

as well as

$$\int_t^{t+\tau}\int_\Omega |\nabla n_{1,\varepsilon}|^{\frac{4}{3}} + \int_t^{t+\tau}\int_\Omega |\nabla n_{2,\varepsilon}|^{\frac{4}{3}} + \int_\Omega |\nabla c_\varepsilon|^2 + \int_t^{t+\tau}\int_\Omega |\nabla c_\varepsilon|^4 + \int_t^{t+\tau}\int_\Omega n_{1,\varepsilon}^2 + \int_t^{t+\tau}\int_\Omega n_{2,\varepsilon}^2 \le C_3 \tag{2.10}$$

for all $t \in [0, T_{\max,\varepsilon} - \tau)$. Now we assume $T_{\max,\varepsilon} < \infty$ for some $\varepsilon > 0$. From (2.7), (2.8), (2.9) and (2.10), we can see that there exists $C_4 > 0$ such that

$$\begin{aligned} \|n_{1,\varepsilon}(\cdot,t)\|_{L^\infty(\Omega)} &\le C_4, \qquad \|n_{2,\varepsilon}(\cdot,t)\|_{L^\infty(\Omega)} \le C_4, \\ \|c_\varepsilon(\cdot,t)\|_{W^{1,q}(\Omega)} &\le C_4, \qquad \|A^\sigma u_\varepsilon(\cdot,t)\|_{L^2(\Omega)} \le C_4 \end{aligned}$$

for all $t \in (0, T_{\max,\varepsilon})$, which is inconsistent with (2.2). Therefore we obtain $T_{\max,\varepsilon} = \infty$ for all $\varepsilon > 0$, which means global existence and boundedness of $(n_{1,\varepsilon}, n_{2,\varepsilon}, c_\varepsilon, u_\varepsilon)$. We next verify convergence of the solution $(n_{1,\varepsilon}, n_{2,\varepsilon}, c_\varepsilon, u_\varepsilon)$. Due to Lemma 2.2 and arguments similar to those in [5], we establish that for all $T > 0$ there is $C_5 > 0$ such that

$$\begin{aligned} \|(n_{1,\varepsilon})_t\|_{L^1((0,T);(W_0^{2,4}(\Omega))^*)} &\le C_5, \qquad \|(n_{2,\varepsilon})_t\|_{L^1((0,T);(W_0^{2,4}(\Omega))^*)} \le C_5, \\ \|(c_\varepsilon)_t\|_{L^2((0,T);(W_0^{1,2}(\Omega))^*)} &\le C_5, \qquad \|(u_\varepsilon)_t\|_{L^2((0,T);(W^{1,3}(\Omega))^*)} \le C_5 \end{aligned} \tag{2.11}$$

for all $\varepsilon > 0$, which together with arguments in [5] implies that there exist a sequence $(\varepsilon_j)_{j\in\mathbb{N}}$ such that $\varepsilon_j \searrow 0$ as $j \to \infty$ and functions n_1, n_2, c, u such that

$$\begin{aligned} n_1, n_2 &\in L^2_{\mathrm{loc}}([0,\infty); L^2(\Omega)) \cap L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty); W^{1,\frac{4}{3}}(\Omega)), \\ c &\in L^2_{\mathrm{loc}}([0,\infty); W^{1,2}(\Omega)), \\ u &\in L^2_{\mathrm{loc}}([0,\infty); W^{1,2}_{0,\sigma}(\Omega)) \end{aligned}$$

and that

$$\begin{aligned} n_{1,\varepsilon} &\to n_1 && \text{in } L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty); L^p(\Omega)) \quad \text{for all } p \in \left[1, \frac{12}{5}\right) \quad \text{and a.e. in } \Omega\times(0,\infty), \\ n_{2,\varepsilon} &\to n_2 && \text{in } L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty); L^p(\Omega)) \quad \text{for all } p \in \left[1, \frac{12}{5}\right) \quad \text{and a.e. in } \Omega\times(0,\infty), \\ c_\varepsilon &\to c && \text{in } C^0_{\mathrm{loc}}([0,\infty); L^p(\Omega)) \quad \text{for all } p \in [1,6) \quad \text{and a.e. in } \Omega\times(0,\infty), \\ u_\varepsilon &\to u && \text{in } L^2_{\mathrm{loc}}([0,\infty); L^p(\Omega)) \quad \text{for all } p \in [1,6) \quad \text{and a.e. in } \Omega\times(0,\infty), \\ c_\varepsilon &\to c && \text{weakly}^* \text{ in } L^\infty(\Omega\times(t,t+1)) \quad \text{for all } t \ge 0, \\ \nabla n_{1,\varepsilon} &\to \nabla n_1 && \text{weakly in } L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty); L^{\frac{4}{3}}(\Omega)), \\ \nabla n_{2,\varepsilon} &\to \nabla n_2 && \text{weakly in } L^{\frac{4}{3}}_{\mathrm{loc}}([0,\infty); L^{\frac{4}{3}}(\Omega)), \\ \nabla c_\varepsilon &\to \nabla c && \text{weakly}^* \text{ in } L^\infty_{\mathrm{loc}}([0,\infty); L^2(\Omega)), \\ \nabla u_\varepsilon &\to \nabla u && \text{weakly in } L^2_{\mathrm{loc}}([0,\infty); L^2(\Omega)), \\ Y_\varepsilon u_\varepsilon &\to u && \text{in } L^2_{\mathrm{loc}}([0,\infty); L^2(\Omega)), \\ n_{1,\varepsilon} &\to n_1 && \text{in } L^2_{\mathrm{loc}}([0,\infty); L^2(\Omega)), \\ n_{2,\varepsilon} &\to n_2 && \text{in } L^2_{\mathrm{loc}}([0,\infty); L^2(\Omega)) \end{aligned} \tag{2.12}$$

as $\varepsilon = \varepsilon_j \searrow 0$. Thus we see that (n_1, n_2, c, u) is a weak solution to (1.1) in the sense of Definition 1.1, which means the end of the proof. □

3. Proof of Theorem 1.3 (Eventual smoothness and stabilization). In this section we will prove Theorem 1.3. The following lemma plays an important role in the proof of Theorem 1.3.

LEMMA 3.1.

(i) *Assume that $a_1, a_2 \in (0,1)$. Then there exists $C > 0$ such that for all $\varepsilon > 0$,*

$$\int_0^\infty\!\!\int_\Omega (n_{1,\varepsilon} - N_1)^2 \le C, \quad \int_0^\infty\!\!\int_\Omega (n_{2,\varepsilon} - N_2)^2 \le C,$$

where $N_1 = \frac{1-a_1}{1-a_1a_2}$, $N_2 = \frac{1-a_2}{1-a_1a_2}$.

(ii) *Assume $a_1 \ge a_2 > 0$. Then there exists $C > 0$ such that for all $\varepsilon > 0$,*

$$\int_0^\infty\!\!\int_\Omega n_{1,\varepsilon}^2 \le C, \quad \int_0^\infty\!\!\int_\Omega (n_{2,\varepsilon} - 1)^2 \le C.$$

Proof. Due to arguments similar to those in [4, Lemmas 4.1–4.4], by using the energy functions

$$\mathcal{G}_{\varepsilon,B} := \int_\Omega \left(n_{1,\varepsilon} - N_1 \log \frac{n_{1,\varepsilon}}{N_1} \right) + \int_\Omega \left(n_{2,\varepsilon} - N_2 \log \frac{n_{2,\varepsilon}}{N_2} \right) + \frac{B}{2} \int_\Omega c_\varepsilon^2$$

in the case that $a_1, a_2 \in (0,1)$, and

$$\mathcal{G}_{\varepsilon,B} := \int_\Omega n_{1,\varepsilon} + \int_\Omega \left(n_{2,\varepsilon} - \log n_{2,\varepsilon} \right) + \frac{B}{2} \int_\Omega c_\varepsilon^2$$

in the case that $a_1 \ge 1 > a_2 > 0$, we can see this lemma. □

Proof of Theorem 1.3. According to an argument similar to that in the proof of [5, Lemmas 3.4 and 3.5], for all $\eta > 0$ and $p \in (1,\infty)$ there are $T > 0$, $\varepsilon_0 > 0$ and $C_1 > 0$ such that for all $t > T$ and $\varepsilon \in (0, \varepsilon_0)$,

$$\|c_\varepsilon(\cdot,t)\|_{L^\infty(\Omega)} < \eta, \quad \|n_{1,\varepsilon}^p(\cdot,t)\|_{L^p(\Omega)} \le C_1, \quad \|n_{2,\varepsilon}^p(\cdot,t)\|_{L^p(\Omega)} \le C_1.$$

We next consider the estimate for u_ε. Since $\nabla \cdot u_\varepsilon = 0$, it follows from the Young inequality, the Poincaré inequality, boundedness of $\nabla\Phi$ and (2.1) that there exists $C_2 > 0$ such that

$$\begin{aligned}
\frac{d}{dt}\int_\Omega |u_\varepsilon|^2 &= -2\int_\Omega |\nabla u_\varepsilon|^2 - 2\int_\Omega u_\varepsilon \cdot (Y_\varepsilon u_\varepsilon \cdot \nabla) u_\varepsilon + 2\int_\Omega u_\varepsilon \cdot (\gamma n_{1,\varepsilon} + \delta n_{2,\varepsilon}) \nabla\Phi \\
&= -2\int_\Omega |\nabla u_\varepsilon|^2 - 2\int_\Omega u_\varepsilon \cdot (Y_\varepsilon u_\varepsilon \cdot \nabla) u_\varepsilon \\
&\quad + 2\gamma \int_\Omega u_\varepsilon \cdot (n_{1,\varepsilon} - n_{1,\infty}) \nabla\Phi + 2\delta \int_\Omega u_\varepsilon \cdot (n_{2,\varepsilon} - n_{2,\infty}) \nabla\Phi \\
&\le -\int_\Omega |\nabla u_\varepsilon|^2 - 2\int_\Omega u_\varepsilon \cdot (Y_\varepsilon u_\varepsilon \cdot \nabla) u_\varepsilon \\
&\quad + C_2 \int_\Omega (n_{1,\varepsilon} - n_{1,\infty})^2 + C_2 \int_\Omega (n_{2,\varepsilon} - n_{2,\infty})^2,
\end{aligned}$$

where $(n_{1,\infty}, n_{2,\infty}) = (N_1, N_2)$ in the case that $a_1, a_2 \in (0,1)$ and $(n_{1,\infty}, n_{2,\infty}) = (0,1)$ in the case that $a_1 \ge 1 > a_2 > 0$. Then, noticing from straightforward calculations

that $\int_\Omega u_\varepsilon \cdot (Y_\varepsilon u_\varepsilon \cdot \nabla) u_\varepsilon = \frac{1}{2}\int_\Omega \nabla \cdot (Y_\varepsilon u_\varepsilon)|u_\varepsilon|^2 = 0$, thanks to Lemma 3.1, we obtain from integration of the above inequality over $(0,\infty)$ that there exists $C_3 > 0$ such that

$$\int_0^\infty \int_\Omega |\nabla u_\varepsilon|^2 \le C_3.$$

According to an argument similar to that in the proof of [5, Lemmas 3.7–3.11], there exist $\alpha' > 0$, $T^* > T$, $C_4 > 0$ such that for all $t > T^*$ there exists $\varepsilon_1 > 0$ such that for all $\varepsilon \in (0, \varepsilon_1)$,

$$\begin{aligned} \|n_{1,\varepsilon}\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, & \|n_{2,\varepsilon}\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, \\ \|c_\varepsilon\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, & \|u_\varepsilon\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4. \end{aligned}$$

Then aided by arguments similar to those in the proofs of [5, Corollary 3.3–Lemma 3.13], from (2.11) there are $\alpha' \in (0,1)$ and $T_0 > 0$ as well as a subsequence $\varepsilon_j \searrow 0$ such that for all $t > T_0$

$$n_{1,\varepsilon} \to n_1, \quad n_{2,\varepsilon} \to n_2, \quad c_\varepsilon \to c, \quad u_\varepsilon \to u \quad \text{in } C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])$$

as $\varepsilon = \varepsilon_j \searrow 0$, and then

$$\begin{aligned} \|n_1\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, & \|n_2\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, \\ \|c\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4, & \|u\|_{C^{1+\alpha',\frac{\alpha'}{2}}(\overline{\Omega}\times[t,t+1])} &\le C_4. \end{aligned} \tag{3.1}$$

Then we obtain

$$n_1, n_2, c, u \in C^{2+\alpha',1+\frac{\alpha'}{2}}(\overline{\Omega}\times[T_0,\infty)).$$

Finally, from (3.1) the solution (n_1, n_2, c, u) of (2.1) constructed in (2.12) fulfills

$$n_1(\cdot,t) \to N_1, \quad n_2(\cdot,t) \to N_2, \quad c(\cdot,t) \to 0, \quad u(\cdot,t) \to 0 \quad \text{in } C^1(\overline{\Omega}) \quad (t \to \infty)$$

in the case that $a_1, a_2 \in (0,1)$, and

$$n_1(\cdot,t) \to 0, \quad n_2(\cdot,t) \to 1, \quad c(\cdot,t) \to 0, \quad u(\cdot,t) \to 0, \quad \text{in } C^1(\overline{\Omega}) \quad (t \to \infty)$$

in the case that $a_1 \ge 1 > a_2 > 0$, which enable us to see Theorem 1.3. □

REFERENCES

[1] N. Bellomo, A. Bellouquid, Y. Tao, and M. Winkler, *Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues*, Math. Models Methods Appl. Sci., 25 (2015), pp. 1663–1763.

[2] X. Cao, S. Kurima, and M. Mizukami, *Global existence and asymptotic behavior of classical solutions for a 3D two-species chemotaxis-Stokes system with competitive kinetics*, arXiv: 1703.01794 [math.AP].

[3] X. Cao, S. Kurima, and M. Mizukami, *Global existence and asymptotic behavior of classical solutions for a 3D two-species Keller–Segel-Stokes system with competitive kinetics*, arXiv: 1706.07910 [math.AP].

[4] M. Hirata, S. Kurima, M. Mizukami, and T. Yokota, *Boundedness and stabilization in a two-dimensional two-species chemotaxis-Navier–Stokes system with competitive kinetics*, J. Differential Equations, 263 (2017), pp. 470–490.

[5] J. Lankeit, *Long-term behaviour in a chemotaxis-fluid system with logistic source*, Math. Models Methods Appl. Sci., 26 (2016), pp. 2071–2109.

[6] Y. Tao and M. Winkler, *Blow-up prevention by quadratic degradation in a two-dimensional Keller–Segel-Navier–Stokes system*, Z. Angew. Math. Phys., 67 (2016), Article 138.

Proceedings of EQUADIFF 2017
pp. 21–28

ON THE SOLUTION SET OF A NONCONVEX NONCLOSED SECOND-ORDER EVOLUTION INCLUSION

AURELIAN CERNEA*

Abstract. We consider a nonconvex and nonclosed second-order evolution inclusion and we prove the arcwise connectedness of the set of its mild solutions.

Key words. set-valued contraction, fixed point, solution set

AMS subject classifications. 34A60

1. Introduction. This paper is concerned with the following problem

$$x'' \in A(t)x + F(t, x, H(t, x)), \quad x(0) = x_0, \quad x'(0) = y_0, \tag{1.1}$$

where X is a real separable Banach space, $\mathcal{P}(X)$ is the family of all subsets of X, $I = [0, T]$, $F(.,.,.) : I \times X^2 \to \mathcal{P}(X)$, $H(.,.) : I \times X \to \mathcal{P}(X)$ and $\{A(t)\}_{t\geq 0}$ is a family of linear closed operators from X into X that generates an evolution system of operators $\{\mathcal{U}(t, s)\}_{t,s\in[0,T]}$. The general framework of evolution operators $\{A(t)\}_{t\geq 0}$ that define problem (1.1) has been developed by Kozak ([10]) and improved by Henriquez ([8]).

When F does not depend on the last variable (1.1) reduces to

$$x'' \in A(t)x + F(t, x), \quad x(0) = x_0, \quad x'(0) = y_0. \tag{1.2}$$

Existence results and qualitative properties of the solutions of problem (1.2) may be found in [1, 2, 3, 4, 8, 9] etc. In all the papers concerned with the set-valued framework, the set-valued map F is assumed to be at least closed-valued. Such an assumption is quite natural in order to obtain good properties of the solution set, but it is interesting to investigate the problem when the right-hand side of the multivalued equation may have nonclosed values.

Following the approach in [12] we consider the problem (1.1), where F and H are closed-valued multifunctions Lipschitzian with respect to the second variable and F is contractive in the third variable. Obviously, the right-hand side of the differential inclusion in (1.1) is in general neither convex nor closed. We prove the arcwise connectedness of the solution set of problem (1.1). The main tool is a result ([11, 12]) concerning the arcwise connectedness of the fixed point set of a class of nonconvex nonclosed set-valued contractions.

We note that similar results for other classes of differential inclusions may be found in our previous papers [5, 6, 7].

The paper is organized as follows: in Section 2 we recall some preliminary results that we use in the sequel and in Section 3 we prove our main result.

2. Preliminaries. Let Z be a metric space with the distance d_Z and let 2^Z be the family of all nonempty closed subsets of Z. For $a \in Z$ and $A, B \in 2^Z$ set

*Faculty of Mathematics and Computer Science, University of Bucharest, Academiei 14, 010014 Bucharest, Romania (acernea@fmi.unibuc.ro).

$d_Z(a,B) = \inf_{b\in B} d_Z(a,b)$ and $d^*_Z(A,B) = \sup_{a\in A} d_Z(a,B)$. Denote by D_Z the Pompeiu-Hausdorff generalized metric on 2^Z defined by

$$D_Z(A,B) = \max\{d^*_Z(A,B), d^*_Z(B,A)\}, \quad A,B \in 2^Z.$$

In what follows, when the product $Z = Z_1 \times Z_2$ of metric spaces $Z_i, i = 1,2$, is considered, it is assumed that Z is equipped with the distance $d_Z((z_1,z_2),(z_1',z_2')) = \sum_{i=1}^2 d_{Z_i}(z_i,z_i')$.

Let X be a nonempty set and let $F : X \to 2^Z$ be a set-valued map from X to Z. The range of F is the set $F(X) = \cup_{x\in X} F(x)$. Let $(X,\mathcal{F})$ be a measurable space. The multifunction $F : X \to 2^Z$ is called measurable if $F^{-1}(\Omega) \in \mathcal{F}$ for any open set $\Omega \subset Z$, where $F^{-1}(\Omega) = \{x \in X; F(x) \cap \Omega \neq \emptyset\}$. Let (X,d_X) be a metric space. The multifunction F is called Hausdorff continuous if for any $x_0 \in X$ and every $\epsilon > 0$ there exists $\delta > 0$ such that $x \in X, d_X(x,x_0) < \delta$ implies $D_Z(F(x),F(x_0)) < \epsilon$.

Let $(T,\mathcal{F},\mu)$ be a finite, positive, nonatomic measure space and let $(X,|.|_X)$ be a Banach space. We denote by $L^1(T,X)$ the Banach space of all (equivalence classes of) Bochner integrable functions $u : T \to X$ endowed with the norm

$$|u|_{L^1(T,X)} = \int_T |u(t)|_X d\mu$$

A nonempty set $K \subset L^1(T,X)$ is called decomposable if, for every $u,v \in K$ and every $A \in \mathcal{F}$, one has

$$\chi_A.u + \chi_{T\setminus A}.v \in K$$

where $\chi_B, B \in \mathcal{F}$ indicates the characteristic function of B.

A metric space Z is called an absolute retract if, for any metric space X and any nonempty closed set $X_0 \subset X$, every continuous function $g : X_0 \to Z$ has a continuous extension $g : X \to Z$ over X. It is obvious that every continuous image of an absolute retract is an arcwise connected space.

In what follows we recall some preliminary results that are the main tools in the proof of our result.

Let $(T,\mathcal{F},\mu)$ be a finite, positive, nonatomic measure space, S a separable Banach space and let $(X,|.|_X)$ be a real Banach space. To simplify the notation we write E in place of $L^1(T,X)$. The proofs of the next two lemmas may be found in [11].

LEMMA 2.1. *Assume that $\phi : S \times E \to 2^E$ and $\psi : S \times E \times E \to 2^E$ are Hausdorff continuous multifunctions with nonempty, closed, decomposable values, satisfying the following conditions*

a) There exists $L \in [0,1)$ such that, for every $s \in S$ and every $u,u' \in E$,

$$D_E(\phi(s,u),\phi(s,u')) \leq L|u-u'|_E.$$

b) There exists $M \in [0,1)$ such that $L + M < 1$ and for every $s \in S$ and every $(u,v),(u',v') \in E \times E$,

$$D_E(\psi(s,u,v),\psi(s,u',v')) \leq M(|u-u'|_E + |v-v'|_E).$$

Set $Fix(\Gamma(s,.)) = \{u \in E; u \in \Gamma(s,u)\}$, where $\Gamma(s,u) = \psi(s,u,\phi(s,u)), (s,u) \in S\times E$. Then

1) For every $s \in S$ the set $Fix(\Gamma(s,.))$ is nonempty and arcwise connected.

2) For any $s_i \in S$, and any $u_i \in Fix(\Gamma(s,.)), i = 1, ..., p$ there exists a continuous function $\gamma : S \to E$ such that $\gamma(s) \in Fix(\Gamma(s,.))$ for all $s \in S$ and $\gamma(s_i) = u_i, i = 1, ..., p$.

LEMMA 2.2. *Let $U : T \to 2^X$ and $V : T \times X \to 2^X$ be two nonempty closed-valued multifunctions satisfying the following conditions*

a) U is measurable and there exists $r \in L^1(T)$ such that $D_X(U(t), \{0\}) \leq r(t)$ for almost all $t \in T$.

b) The multifunction $t \to V(t,x)$ is measurable for every $x \in X$.

c) The multifunction $x \to V(t,x)$ is Hausdorff continuous for all $t \in T$.

Let $v : T \to X$ be a measurable selection from $t \to V(t, U(t))$.

Then there exists a selection $u \in L^1(T, X)$ such that $v(t) \in V(t, u(t)), t \in T$.

In what follows $\{A(t)\}_{t\geq 0}$ is a family of linear closed operators from X into X that genearates an evolution system of operators $\{\mathcal{U}(t,s)\}_{t,s\in I}$. By hypothesis the domain of $A(t)$, $D(A(t))$ is dense in X and is independent of t. The following definition is taken from [8, 10].

DEFINITION 2.3. *A family of bounded linear operators $\mathcal{U}(t,s) : X \to X$, $(t,s) \in \Delta := \{(t,s) \in I \times I; s \leq t\}$ is called an evolution operator of the equation*

$$x''(t) = A(t)x(t) \tag{2.1}$$

if

i) For any $x \in X$, the map $(t,s) \to \mathcal{U}(t,s)x$ is continuously differentiable and

a) $\mathcal{U}(t,t) = 0$, $t \in I$.

b) If $t \in I, x \in X$ then $\frac{\partial}{\partial t}\mathcal{U}(t,s)x|_{t=s} = x$ and $\frac{\partial}{\partial s}\mathcal{U}(t,s)x|_{t=s} = -x$.

ii) If $(t,s) \in \Delta$, then $\frac{\partial}{\partial s}\mathcal{U}(t,s)x \in D(A(t))$, the map $(t,s) \to \mathcal{U}(t,s)x$ is of class C^2 and

a) $\frac{\partial^2}{\partial t^2}\mathcal{U}(t,s)x \equiv A(t)\mathcal{U}(t,s)x$.

b) $\frac{\partial^2}{\partial s^2}\mathcal{U}(t,s)x \equiv \mathcal{U}(t,s)A(t)x$.

c) $\frac{\partial^2}{\partial s \partial t}\mathcal{U}(t,s)x|_{t=s} = 0$.

iii) If $(t,s) \in \Delta$, then there exist $\frac{\partial^3}{\partial t^2 \partial s}\mathcal{U}(t,s)x$, $\frac{\partial^3}{\partial s^2 \partial t}\mathcal{U}(t,s)x$ and

a) $\frac{\partial^3}{\partial t^2 \partial s}\mathcal{U}(t,s)x \equiv A(t)\frac{\partial}{\partial s}\mathcal{U}(t,s)x$ and the map $(t,s) \to A(t)\frac{\partial}{\partial s}\mathcal{U}(t,s)x$ is continuous.

b) $\frac{\partial^3}{\partial s^2 \partial t}\mathcal{U}(t,s)x \equiv \frac{\partial}{\partial t}\mathcal{U}(t,s)A(s)x$.

As an example for equation (2.1) one may consider the problem (e.g., [8])

$$\frac{\partial^2 z}{\partial t^2}(t,\tau) = \frac{\partial^2 z}{\partial \tau^2}(t,\tau) + a(t)\frac{\partial z}{\partial t}(t,\tau), \quad t \in [0,T], \tau \in [0,2\pi],$$

$$z(t,0) = z(t,\pi) = 0, \quad \frac{\partial z}{\partial \tau}(t,0) = \frac{\partial z}{\partial \tau}(t,2\pi),\ t \in [0,T],$$

where $a(.) : I \to \mathbf{R}$ is a continuous function. This problem is modeled in the space $X = L^2(\mathbf{R}, \mathbf{C})$ of 2π-periodic 2-integrable functions from $\mathbf{R}$ to $\mathbf{C}$, $A_1 z = \frac{d^2 z(\tau)}{d\tau^2}$ with domain $H^2(\mathbf{R}, \mathbf{C})$ the Sobolev space of 2π-periodic functions whose derivatives belong to $L^2(\mathbf{R}, \mathbf{C})$. It is well known that A_1 is the infinitesimal generator of strongly continuous cosine functions $C(t)$ on X. Moreover, A_1 has discrete spectrum; namely the spectrum of A_1 consists of eigenvalues $-n^2$, $n \in \mathbf{Z}$ with associated eigenvectors $z_n(\tau) = \frac{1}{\sqrt{2\pi}}e^{in\tau}$, $n \in \mathbf{N}$. The set $\{z_n\}$, $n \in \mathbf{N}$ is an orthonormal basis of X.

In particular, $A_1 z = \sum_{n\in\mathbf{Z}} -n^2 < z, z_n > z_n$, $z \in D(A_1)$. The cosine function is given by $C(t)z = \sum_{n\in\mathbf{Z}} \cos(nt) < z, z_n > z_n$ with the associated sine function $S(t)z = t < z, z_0 > z_0 + \sum_{n\in\mathbf{Z}\setminus\{0\}} \frac{\sin(nt)}{n} < z, z_n > z_n$.

For $t \in I$ define the operator $A_2(t)z = a(t)\frac{dz(\tau)}{d\tau}$ with domain $D(A_2(t)) = H^1(\mathbf{R}, \mathbf{C})$. Set $A(t) = A_1 + A_2(t)$. It has been proved in [10] that this family generates an evolution operator as in Definition 2.3.

Definition 2.4. *A continuous mapping $x(.) \in C(I, X)$ is called a mild solution of problem (1.1) if there exists a (Bochner) integrable function $f(.) \in L^1(I, X)$ such that*

$$f(t) \in F(t, x(t)) \quad a.e.\,(I), \tag{2.2}$$

$$x(t) = -\frac{\partial}{\partial s}\mathcal{U}(t,0)x_0 + \mathcal{U}(t,0)y_0 + \int_0^t \mathcal{U}(t,s)f(s)ds,\ t \in I. \tag{2.3}$$

We shall call $(x(.), f(.))$ a *trajectory-selection pair* of (1.1) if $f(.)$ verifies (2.2) and $x(.)$ is defined by (2.3).

We shall use the following notations for the solution sets of (1.1).

$$\mathcal{S}(x_0, y_0) = \{x(.);\quad x(.) \text{ is a mild solution of (1.1)}\}. \tag{2.4}$$

In order to study problem (1.1) we introduce the following hypothesis.

Hypothesis 2.5. i) There exists an evolution operator $\{\mathcal{U}(t,s)\}_{t,s\in I}$ associated to the family $\{A(t)\}_{t\geq 0}$.

ii) There exist $M, M_0 \geq 0$ such that $|\mathcal{U}(t,s)|_{B(X)} \leq M$, $|\frac{\partial}{\partial s}\mathcal{U}(t,s)| \leq M_0$, for all $(t,s) \in \Delta$.

$F : I \times X \times X \to \mathcal{P}(X)$ and $H : I \times X \to \mathcal{P}(X)$ are two set-valued maps with nonempty closed values, satisfying

iii) The set-valued maps $t \to F(t,u,v)$ and $t \to H(t,u)$ are measurable for all $u, v \in X$.

iv) There exist $l(.) \in L^1(I, \mathbf{R})$ such that, for every $u, u' \in X$,

$$D(H(t,u), H(t,u')) \leq l(t)|u - u'| \quad a.e.\ (I).$$

v) There exist $m(.) \in L^1(I, \mathbf{R})$ and $\theta \in [0,1)$ such that, for every $u, v, u', v' \in X$,

$$D(F(t,u,v), F(t,u',v')) \leq m(t)|u - u'| + \theta|v - v'| \quad a.e.\ (I).$$

vi) There exist $f, g \in L^1(I, \mathbf{R})$ such that

$$d(0, F(t,0,0)) \leq f(t), \quad d(0, H(t,0)) \leq g(t) \quad a.e.\ (I).$$

In what follows $N(t) = \max\{l(t), m(t)\}$, $t \in I$, $N^*(t) = \int_0^t N(s)ds$.

Given $\alpha \in \mathbf{R}$ we denote by L^1 the Banach space of all (equivalence classes of) Lebesgue measurable functions $\sigma : I \to X$ endowed with the norm

$$|\sigma|_1 = \int_0^T e^{-\alpha N^*(t)}|\sigma(t)|dt.$$

3. Main result. Even if the multifunction from the right-hand side of (1.1) has, in general, nonclosed nonconvex values, its solution set $\mathcal{S}(x_0, y_0)$ defined in (2.4) has some meaningful properties, stated in theorem below.

THEOREM 3.1. *Assume that Hypothesis 2.5 is satisfied and let* $\alpha > \frac{2M}{1-\theta}$. *Then*

1) For every $(x_0, y_0) \in X \times X$, *the solution set* $\mathcal{S}(x_0, y_0)$ *is nonempty and arcwise connected in the space C(I,X).*

2) For any $(\xi_i, \mu_i) \in X \times X$ *and any* $x_i \in \mathcal{S}(\xi_i, \mu_i)$, $i = 1, ..., p$, *there exists a continuous function* $s : X \times X \to C(I, X)$ *such that* $s(\xi, \mu) \in \mathcal{S}(\xi, \mu)$ *for any* $(\xi, \mu) \in X \times X$ *and* $s(\xi_i, \mu_i) = x_i, i = 1, ..., p$.

3) The set $\mathcal{S} = \cup_{(\xi,\mu)\in X\times X}\mathcal{S}(\xi, \mu)$ *is arcwise connected in C(I,X).*

Proof. 1) For $(\xi, \mu) \in X \times X$ and $f \in L^1$, set

$$x_{\xi,\mu}(t) = -\frac{\partial}{\partial s}\mathcal{U}(t, 0)\xi + \mathcal{U}(t, 0)\mu + \int_0^t \mathcal{U}(t, s)f(s)ds \tag{3.1}$$

and consider $\lambda : X \times X \to C(I, X)$ defined by $\lambda(\xi, \mu)(t) = -\frac{\partial}{\partial s}\mathcal{U}(t, 0)\xi + \mathcal{U}(t, 0)\mu$.

We prove that the multifunctions $\phi : X\times X\times L^1 \to 2^{L^1}$ and $\psi : X\times X\times L^1\times L^1 \to 2^{L^1}$ given by

$$\phi((\xi, \mu), u) = \{v \in L^1; \quad v(t) \in H(t, x_{\xi,\mu}(t)) \quad a.e.(I)\},$$

$$\psi((\xi, \mu), u, v) = \{w \in L^1; \quad w(t) \in F(t, x_{\xi,\mu}(t), v(t)) \quad a.e.(I)\},$$

$(\xi, \mu) \in X \times X$, $u, v \in L^1$ satisfy the hypotheses of Lemma 2.1.

Since $x_{\xi,\mu}(.)$ is measurable and H satisfies Hypothesis 2.5 iii) and iv), the multifunction $t \to H(t, x_{\xi,\mu}(t))$ is measurable and nonempty closed-valued, it has a measurable selection. Therefore due to Hypothesis 2.5 vi), the set $\phi((\xi, \mu), u)$ is nonempty. The fact that the set $\phi((\xi, \mu), u)$ is closed and decomposable follows by a simple computation. In the same way we obtain that $\psi((\xi, \mu), u, v)$ is a nonempty closed decomposable set.

Pick $((\xi, \mu), u), ((\xi_1, \mu_1), u_1) \in X \times X \times L^1$ and choose $v \in \phi((\xi, \mu), u)$. For each $\varepsilon > 0$ there exists $v_1 \in \phi((\xi_1, \mu_1), u_1)$ such that, for every $t \in I$, one has

$$|v(t) - v_1(t)| \le D(H(t, x_{\xi,\mu}(t)), H(t, x_{\xi_1,\mu_1}(t))) + \varepsilon \le$$
$$l(t)[M_0|\xi - \xi_1| + M|\mu - \mu_1| + M\int_0^t |u(s) - u_1(s)|ds] + \varepsilon.$$

Hence

$$|v - v_1|_1 \le [M_0|\xi - \xi_1| + M|\mu - \mu_1|]\int_0^T e^{-\alpha N^*(t)}l(t)dt + M\int_0^T e^{-\alpha N^*(t)}$$
$$l(t)(\int_0^t |u(s) - u_1(s)|ds)dt + \varepsilon T \le \tfrac{1}{\alpha}[M_0|\xi - \xi_1| + M|\mu - \mu_1|] + \tfrac{M}{\alpha}|u - u_1|_1 + \varepsilon T$$

for any $\varepsilon > 0$.

This implies

$$d_{L^1}(v, \phi((\xi_1, \mu_1), u_1)) \le \frac{1}{\alpha}[M_0|\xi - \xi_1| + M|\mu - \mu_1|] + \frac{M}{\alpha}|u - u_1|_1$$

for all $v \in \phi((\xi, \mu), u)$. Therefore,

$$d^*_{L^1}(\phi((\xi, \mu), u), \phi((\xi_1, \mu_1), u_1)) \le \frac{1}{\alpha}[M_0|\xi - \xi_1| + M|\mu - \mu_1|] + \frac{M}{\alpha}|u - u_1|_1$$

Consequently,

$$D_{L^1}(\phi((\xi,\mu),u),\phi((\xi_1,\mu_1),u_1)) \leq \frac{1}{\alpha}[M_0|\xi-\xi_1|+M|\mu-\mu_1|]+\frac{M}{\alpha}|u-u_1|_1$$

which shows that ϕ is Hausdorff continuous and satisfies the assumptions of Lemma 2.1.

Pick $((\xi,\mu),u,v),((\xi_1,\mu_1),u_1,v_1) \in X\times X\times L^1\times L^1$ and choose $w \in \psi((\xi,\mu),u,v)$. Then, as before, for each $\varepsilon>0$ there exists $w_1 \in \psi((\xi_1,\mu_1),u_1,v_1)$ such that for every $t \in I$

$$|w(t)-w_1(t)| \leq D(F(t,x_{\xi,\mu}(t),v(t)),F(t,x_{\xi_1,\mu_1}(t),v_1(t)))+\varepsilon \leq$$
$$m(t)[M_0|\xi-\xi_1|+M|\mu-\mu_1|+M\int_0^t|u(s)-u_1(s)|ds]+\theta|v(t)-v_1(t)|+\varepsilon.$$

Hence

$$|w-w_1|_1 \leq \tfrac{1}{\alpha}[M_0|\xi-\xi_1|+M|\mu-\mu_1|]+\tfrac{M}{\alpha}|u-u_1|_1+\theta|v-v_1|_1+\varepsilon T$$
$$\leq \tfrac{1}{\alpha}[M_0|\xi-\xi_1|+M|\mu-\mu_1|]+(\tfrac{M}{\alpha}+\theta)(|u-u_1|_1+|v-v_1|_1)+\varepsilon T$$
$$\leq \tfrac{1}{\alpha}[M_0|\xi-\xi_1|+M|\mu-\mu_1|]+(\tfrac{M}{\alpha}+\theta)d_{L^1\times L^1}((u,v),(u_1,v_1))+\varepsilon T.$$

As above, we deduce that

$$D_{L^1}(\psi((\xi,\mu),u,v),\psi((\xi_1,\mu_1),u_1,v_1)) \leq$$
$$\tfrac{1}{\alpha}[M_0|\xi-\xi_1|+M|\mu-\mu_1|]+(\tfrac{M}{\alpha}+\theta)d_{L^1\times L^1}((u,v),(u_1,v_1)).$$

namely, the multifunction ψ is Hausdorff continuous and satisfies the hypothesis of Lemma 2.1.

Define $\Gamma((\xi,\mu),u)=\psi((\xi,\mu),u,\phi((\xi,\mu),u)),((\xi,\mu),u) \in X^2\times L^1$. According to Lemma 2.1, the set $Fix(\Gamma((\xi,\mu),.))=\{u\in L^1; u\in\Gamma((\xi,\mu),u)\}$ is nonempty and arcwise connected in $L^1(I,X)$. Moreover, for fixed $(\xi_i,\mu_i)\in X^2$ and $u_i \in Fix(\Gamma((\xi_i,\mu_i),.)), i=1,...,p$, there exists a continuous function $\gamma: X^2\to L^1$ such that

$$\gamma((\xi,\mu)) \in Fix(\Gamma((\xi,\mu),.)), \quad \forall(\xi,\mu)\in X^2, \tag{3.2}$$

$$\gamma((\xi_i,\mu_i))=u_i, \quad i=1,...,p. \tag{3.3}$$

We shall prove that

$$Fix(\Gamma((\xi,\mu),.))=\{u\in L^1; \quad u(t)\in F(t,x_{\xi,\mu}(t),H(t,x_{\xi,\mu}(t))) \quad a.e.\ (I)\}. \tag{3.4}$$

Denote by $A(\xi,\mu)$ the right-hand side of (3.4). If $u\in Fix(\Gamma((\xi,\mu),.))$ then there is $v\in\phi((\xi,\mu),v)$ such that $u\in\psi((\xi,\mu),u,v)$. Therefore, $v(t)\in H(t,x_{\xi,\mu}(t))$ and

$$u(t)\in F(t,x_{\xi,\mu}(t),v(t))\subset F(t,x_{\xi,\mu}(t),H(t,x_{\xi,\mu}(t))) \quad a.e.\ (I),$$

so that $Fix(\Gamma((\xi,\mu),.))\subset A(\xi,\mu)$.

Let now $u\in A(\xi,\mu)$. By Lemma 2.2, there exists a selection $v\in L^1$ of the multifunction $t\to H(t,x_{\xi,\mu}(t))$ satisfying

$$u(t)\in F(t,x_{\xi,\mu}(t),v(t)) \quad a.e.\ (I).$$

Hence, $v\in\phi((\xi,\mu),v)$, $u\in\psi((\xi,\mu),u,v)$ and thus $u\in\Gamma((\xi,\mu),u)$, which completes the proof of (3.4).

We next note that the function $T : L^1 \to C(I, X)$,

$$T(u)(t) := \int_0^t \mathcal{U}(t, s)u(s)ds$$

is continuous and one has

$$\mathcal{S}(\xi, \mu) = \lambda(\xi, \mu) + T(Fix(\Gamma((\xi, \mu), .))), \quad (\xi, \mu) \in X^2. \tag{3.5}$$

Since $Fix(\Gamma((\xi, \mu), .))$ is nonempty and arcwise connected in L^1, the set $\mathcal{S}(\xi, \mu)$ has the same properties in $C(I, X)$.

2) Let $(\xi_i, \mu_i) \in X^2$ and let $x_i \in \mathcal{S}(\xi_i, \mu_i), i = 1, ..., p$ be fixed. By (3.5) there exists $v_i \in Fix(\Gamma((\xi_i, \mu_i), .))$ such that

$$x_i = \lambda(\xi_i, \mu_i) + T(v_i), \quad i = 1, ..., p.$$

If $\gamma : X^2 \to L^1$ is a continuous function satisfying (3.2) and (3.3) we define, for every $(\xi, \mu) \in X^2$,

$$s(\xi, \mu) = \lambda(\xi, \mu) + T(\gamma(\xi, \mu)).$$

Obviously, the function $s : X \to C(I, X)$ is continuous, $s(\xi, \mu) \in \mathcal{S}(\xi, \mu)$ for all $(\xi, \mu) \in X^2$ and

$$s(\xi_i, \mu_i) = \lambda(\xi_i, \mu_i) + T(\gamma(\xi_i, \mu_i)) = \lambda(\xi_i, \mu_i) + T(v_i) = x_i, \quad i = 1, ..., p.$$

3) Let $x_1, x_2 \in \mathcal{S} = \cup_{(\xi,\mu)\in X^2}\mathcal{S}(\xi, \mu)$ and choose $(\xi_i, \mu_i) \in X^2$, $i = 1, 2$ such that $x_i \in S(\xi_i, \mu_i)$, $i = 1, 2$. From the conclusion of 2) we deduce the existence of a continuous function $s : X^2 \to C(I, X)$ satisfying $s(\xi_i, \mu_i) = x_i$, $i = 1, 2$ and $s(\xi, \mu) \in \mathcal{S}(\xi, \mu)$, $(\xi, \mu) \in X^2$. Let $h : [0, 1] \to X^2$ be a continuous mapping such that $h(0) = (\xi_1, \mu_1)$ and $h(1) = (\xi_2, \mu_2)$. Then the function $s \circ h : [0, 1] \to C(I, X)$ is continuous and verifies

$$s \circ h(0) = x_1, \quad s \circ h(1) = x_2, \quad s \circ h(\tau) \in \mathcal{S}(h(\tau)) \subset \mathcal{S}, \quad \tau \in [0, 1],$$

which completes the proof. □

REFERENCES

[1] A. Baliki, M. Benchohra, and J.R. Graef, *Global existence and stability of second order functional evolution equations with infinite delay*, Electronic J. Qual. Theory Differ. Equations, 2016, no. 23, (2016), pp. 1–10.

[2] A. Baliki, M. Benchohra, and J.J. Nieto, *Qualitative analysis of second-order functional evolution equations*, Dynamic Syst. Appl., 24 (2015), pp. 559–572.

[3] M. Benchohra, and I. Medjadj, *Global existence results for second order neutral functional differential equations with state-dependent delay*, Comment. Math. Univ. Carolin., 57 (2016), pp. 169–183.

[4] M. Benchohra, and N. Rezzoug, *Measure of noncompactness and second-order evolution equations*, Gulf J. Math., 4 (2016), pp. 71–79.

[5] A. Cernea, *On the solution set of some classes of nonconvex nonclosed differential inclusions*, Portugaliae Math., 65 (2008), pp. 485–496.

[6] A. Cernea, *On the solution set of a nonconvex nonclosed Sturm-Liouville type differential inclusion*, Comment. Math., 49 (2009), pp. 139–146.

[7] A. Cernea, *On the solution set of a nonconvex nonclosed hyperbolic differential inclusion of third order*, J. Nonlinear Convex Anal., 17 (2016), pp. 1171–1179.

[8] H.R. Henriquez, *Existence of solutions of nonautonomous second order functional differential equations with infinite delay*, Nonlinear Anal., 74 (2011), pp. 3333–3352.
[9] H.R. Henriquez, V. Poblete, and J.C. Pozo, *Mild solutions of non-autonomous second order problems with nonlocal initial conditions*, J. Math. Anal. Appl., 412 (2014), pp. 1064–1083.
[10] M. Kozak, *A fundamental solution of a second-order differential equation in a Banach space*, Univ. Iagel. Acta. Math., 32 (1995), pp. 275–289.
[11] S. Marano, *Fixed points of multivalued contractions with nonclosed, nonconvex values*, Atti. Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur., 5 (1994), pp. 203–212.
[12] S. Marano, and V. Staicu, *On the set of solutions to a class of nonconvex nonclosed differential inclusions*, Acta Math. Hungar., 76 (1997), pp. 287–301.

Proceedings of EQUADIFF 2017
pp. 29–36

DYNAMICAL MODEL OF VISCOPLASTICITY

KONRAD KISIEL *

Abstract. This paper discusses the existence theory to dynamical model of viscoplasticity and show possibility to obtain existence of solution without assuming weak safe-load condition.

Key words. viscoplasticity, coercive approximation, Yosida approximation, safe-load condition, mixed boundary condition

AMS subject classifications. 35Q74, 35A01, 74C10

1. Introduction.

Systems of equations describing an inelastic deformation of metals, under fundamental assumption of small deformations, consist of linear partial differential equations coupled with nonlinear differential inclusion.

The differential inclusion (inelastic constitutive equation) is experimental and depend on the considered material. Therefore, there are many different inelastic constitutive equations. H.-D. Alber in [1] defines a very large class of constitutive equations (of pre-monotone type) which contains all models proposed in engineering sciences known by author. However, the existence theory for such wide class of constitutive equations is not complete. Therefore, we will focus on a certain subclass of possible constitutive equations called viscoplastic models of gradient type (see Definition 1.1). For models equipped with such constitutive equation it is quite common to assume specific indirect assumption on data called *weak safe-load condition* (see Definition 1.2) as for example in: [3], [5], or [6]. However, in paper [4] authors were able to omit this indirect assumption in the case of dynamical visco-poroplasticity. We observed that similar methods can be used in case of viscoplastic models of gradient type.

1.1. Formulation of the model.

We assume that considered material (with the constant mass density $\rho > 0$) lies within the subset $\Omega \subset \mathbb{R}^3$ with smooth boundary $\partial\Omega$. The system of equations describing the inelastic deformation process can be written in the following form

$$\begin{aligned} \rho u_{tt}(x,t) - \operatorname{div}_x T(x,t) &= F(x,t), \\ T(x,t) &= \mathcal{D}\left(\varepsilon\left(u(x,t)\right) - \varepsilon^p(x,t)\right), \\ \varepsilon^p_t(x,t) &\in M(T(x,t)), \end{aligned} \tag{1.1}$$

where $\varepsilon\left(u(x,t)\right)$ denotes the symmetric part of the gradient of function $u(x,t)$ i.e.

$$\varepsilon\left(u(x,t)\right) = \frac{1}{2}\left(\nabla_x u(x,t) + \nabla_x^T u(x,t)\right).$$

The first equation $(1.1)_1$ is the balance of momentum coupled with the generalized Hooke's law (equation $(1.1)_2$). The given functions are: $F : \Omega \times [0, T_e] \to \mathbb{R}^3$ which describes a density of applied body forces and $\mathcal{D} : \mathcal{S}(3) \to \mathcal{S}(3) = \mathbb{R}^{3\times 3}_{sym}$ which is an elasticity tensor. $\mathcal{D}$ is assumed to be linear, symmetric, positive-definite and

*Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland (K.Kisiel@mini.pw.edu.pl).

constant in time and space. Last equation $(1.1)_3$ is called constitutive equation, where $M : D(M) \subset \mathcal{S}(3) \to \mathcal{P}(\mathcal{S}(3))$ is a given constitutive multifunction.

For any fixed $T_e > 0$ we are interested in finding the following

- the displacement field $u : \Omega \times [0, T_e] \to \mathbb{R}^3$,
- the inelastic deformation tensor $\varepsilon^p : \Omega \times [0, T_e] \to \mathcal{S}(3) = \mathbb{R}^{3\times 3}_{sym}$,
- the Cauchy stress tensor $T : \Omega \times [0, T_e] \to \mathcal{S}(3)$,

Problem (1.1) will be considered with mixed boundary conditions

$$\begin{aligned} u(x,t) &= g_D(x,t), \qquad x \in \Gamma_D,\ t \geqslant 0, \\ T(x,t)n(x) &= g_N(x,t), \qquad x \in \Gamma_N,\ t \geqslant 0, \end{aligned} \tag{1.2}$$

where $n(x)$ is the outward pointing, unit normal vector at point $x \in \partial\Omega$. The sets Γ_D, Γ_N, are open subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$.

Furthermore, we also need to assume initial conditions in the form

$$\text{for } x \in \Omega \qquad u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \qquad \varepsilon^p(x,0) = \varepsilon^p_0(x). \tag{1.3}$$

In this paper we consider viscoplastic models of gradient type. Therefore, we assume that the inelastic constitutive multifunction M is viscoplastic of gradient type which means

Definition 1.1.
We say that constitutive multifunction $M : \mathcal{S}(3) \to \mathcal{P}(\mathcal{S}(3))$ is viscoplastic of gradient type if there exist a convex function $M_0 : \mathcal{S}(3) \to \mathbb{R}$ such that

$$M(T) = \partial M_0(T).$$

1.2. Main results.

In [3] K. Chełmiński introduce the coercive approximation process of model (1.1). Moreover, in [3] author proved that the approximate solutions converge to the solution of the original problem. However, in order to obtain needed estimates author assumed *weak safe-load condition* in the following form

Definition 1.2 (weak safe-load condition).
We say that the functions g_D, g_N satisfy the weak safe-load conditions *if there exist the initial conditions $u_0^*, u_1^* \in H^1(\Omega;\mathbb{R}^3)$ and the function $F^* \in H^1(0,T_e;L^2(\Omega;\mathbb{R}^3))$ such that, there exists a solution (u^*, T^*) of the linear system*

$$\begin{aligned} \rho u^*_{tt}(x,t) - \operatorname{div}_x T^*(x,t) &= F^*(x,t), \\ T^*(x,t) &= \mathcal{D}(\varepsilon(u^*(x,t))), \end{aligned}$$

with the initial–boundary conditions

$$\begin{aligned} u^*(x,0) &= u_0^*(x) && \text{for } x \in \Omega, \\ u^*_t(x,0) &= u_1^*(x) && \text{for } x \in \Omega, \\ u^*(x,t) &= g_D(x,t) && \text{for } x \in \Gamma_D, \quad t \geqslant 0, \\ T^*(x,t)n(x) &= g_N(x,t) && \text{for } x \in \Gamma_N, \quad t \geqslant 0, \end{aligned}$$

and the regularity

$$u^* \in W^{2,\infty}(0,T_e;L^2(\Omega;\mathbb{R}^3)), \quad \varepsilon(u^*) \in W^{1,\infty}(0,T_e;L^2(\Omega;\mathcal{S}^3)),$$
$$T^* \in L^\infty(0,T_e;L^\infty(\Omega;\mathcal{S}^3)).$$

This indirect assumption on data is very difficult to check (especially in the case of mixed boundary conditions). Therefore, natural question arise: If such assumption is needed in case of viscoplasticity?

Now we are able to answer this question. It occurs that in order to prove the existence of solution to viscoplasticity problem the *weak safe-load condition* can be completely omitted. Namely, we are able to proof the following theorem

THEOREM 1.3 (Main result).
Consider dynamical model of viscoplasticity (1.1) *(where constitutive function is viscoplastic of gradient type) with the initial-boundary conditions* (1.2)–(1.3). *Assume that the initial conditions, boundary data and external force satisfy* (2.1)–(2.5) *then, there exists a solution* (u, ε^p, T) *in the sense of Definition 2.2.*

2. General information.
Before we start the main part of the discussion we would like to introduce regularity assumptions then it is important to define the notation of a solution. Finally in the last part of this section we introduce coercive approximation of the problem (1.1)–(1.3) along with the existence result for approximate model.

2.1. Regularity assumption on data.
First of all let us state the regularity assumptions for needed data. To obtain existence of solution to the problem (1.1)–(1.3) we assume the following (it is worth mentioning that in fact we can prove existence under slightly lower assumption on data but, for simplicity, we state them this way).
• Regularities of the external force

$$F \in H^1(0, T_e; L^2(\Omega; \mathbb{R}^3)). \tag{2.1}$$

• Regularities of the boundary conditions

$$\begin{gathered} g_D \in W^{3,\infty}(0, T_e; H^{\frac{3}{2}}(\Gamma_D; \mathbb{R}^3)), \\ g_N \in W^{2,\infty}\left(0, T_e; L^\infty\left(\Gamma_N; \mathbb{R}^3\right)\right) \cap W^{2,\infty}\left(0, T_e; H^{-\frac{1}{2}}\left(\Gamma_N; \mathbb{R}^3\right)\right), \end{gathered} \tag{2.2}$$

• Regularities of the initial conditions

$$u_0 \in H^2(\Omega; \mathbb{R}^3), \quad u_1 \in H^1(\Omega; \mathbb{R}^3), \quad \varepsilon_0^p \in L^2_{\mathrm{div}}(\Omega; \mathcal{S}(3)). \tag{2.3}$$

Moreover, we require compatibility conditions of the form

$$\begin{array}{rcll} u_0(x) & = & g_D(x,0), & x \in \Gamma_D, \\ u_1(x) & = & g_{D,t}(x,0), & x \in \Gamma_D, \\ T_0(x)\, n(x) & = & g_N(x,0), & x \in \Gamma_N, \end{array} \tag{2.4}$$

where $T_0(x) := \mathcal{D}(\varepsilon(u_0(x)) - \varepsilon_0^p(x))$ is an initial stress.

We also need to assume that the initial stress lies in a domain of the constitutive multifunction M, which means

DEFINITION 2.1.
The initial data (u_0, ε_0^p) *are said to be admissible for problem* (1.1) *if*

$$\exists M^* \in L^2(\Omega; \mathcal{S}(3)) \quad \text{such that} \quad M^*(x) \in M(\mathcal{D}(\varepsilon(u_0(x)) - \varepsilon_0^p(x))) \tag{2.5}$$

for almost every $x \in \Omega$.

2.2. Definition of solution.
We were able to obtain solution in the same sense as it is done in [3]. Our solution satisfy problem (1.1) almost everywhere. Namely

Definition 2.2 (Solution).
We say that (u, ε^p, T) *is a solution of the problem* (1.1)–(1.3) *if:*
1. *The following regularities are satisfied*

$$\begin{gathered} u \in W^{2,\infty}\left(0, T_e; L^2\left(\Omega; \mathbb{R}^3\right)\right), \qquad \varepsilon(u) \in W^{1,1}\left(0, T_e; L^1\left(\Omega; \mathcal{S}(3)\right)\right), \\ \varepsilon^p \in W^{1,1}\left(0, T_e; L^1\left(\Omega; \mathcal{S}(3)\right)\right), \\ T \in W^{1,\infty}\left(0, T_e; L^2\left(\Omega; \mathcal{S}(3)\right)\right), \qquad \operatorname{div} T \in L^\infty\left(0, T_e; L^2\left(\Omega; \mathbb{R}^3\right)\right). \end{gathered}$$

2. *For almost every* $(x,t) \in \Omega \times (0, T_e)$ *the following problem is satisfied*

$$\begin{aligned} \rho u_{tt}(x,t) - \operatorname{div} T(x,t) &= F(x,t), \\ T(x,t) &= \mathcal{D}\left(\varepsilon\left(u(x,t)\right) - \varepsilon^p(x,t)\right), \\ \varepsilon^p_t(x,t) &\in M\left(T(x,t)\right). \end{aligned}$$

3. *By* γ *let us denote the trace operator. Then*

$$\begin{aligned} \gamma_{|\Gamma_D \times [0,T_e]}(u) &= g_D, \\ \gamma_{|\Gamma_N \times [0,T_e]}(T\,n) &= g_N. \end{aligned}$$

4. *For almost every* $x \in \Omega$ *initial conditions:*

$$u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \qquad \varepsilon^p(x,0) = \varepsilon^p_0(x)$$

are satisfied.

2.3. Approximation of the model.
Observe that the free energy of (1.1) is given by

$$\rho\psi(\varepsilon, \varepsilon^p)(t) = \frac{1}{2}\mathcal{D}(\varepsilon - \varepsilon^p)(\varepsilon - \varepsilon^p).$$

The energy is only a positive semi–definite quadratic form and therefore our system is *non-coercive* (for details see [1]). The lack of coercivity significantly hinders the analysis. As a remedy we introduce a standard idea of the coercive approximation (see for example [3]) of (1.1) as follows

$$\begin{aligned} \rho\, u^k_{tt}(x,t) - \operatorname{div}_x T^k(x,t) &= F(x,t), \\ T^k(x,t) &= \mathcal{D}\left(\left(1 + \frac{1}{k}\right)\varepsilon(u^k(x,t)) - \varepsilon^{p,k}(x,t)\right), \\ \widehat{T}^k(x,t) &= T^k(x,t) - \frac{1}{k}\mathcal{D}(\varepsilon(u^k(x,t))), \\ \varepsilon^{p,k}_t(x,t) &\in \partial M_0(\widehat{T}^k(x,t)), \end{aligned} \tag{2.6}$$

where $k \geqslant 1$.

Now if we fix k, the free energy of (2.6) is given by

$$\rho\psi^k\left(\varepsilon^k, \varepsilon^{p,k}\right)(t) = \frac{1}{2}\mathcal{D}\left(\varepsilon^k - \varepsilon^{p,k}\right)\left(\varepsilon^k - \varepsilon^{p,k}\right) + \frac{1}{2k}\mathcal{D}\left(\varepsilon^k\right)\varepsilon^k.$$

One can see that now the energy is a positive-definite quadratic form. Models with that type of energy are called *coercive*. The total energy of the discussed model is in the form

$$\mathcal{E}^k(u_t^{k,\lambda},\varepsilon^{k,\lambda},\varepsilon^{p,k,\lambda})(t)=\frac{\rho}{2}\int\limits_\Omega\left|u_t^{k,\lambda}(x,t)\right|^2\mathrm{d}x+\int\limits_\Omega\rho\psi^k\left(\varepsilon^{k,\lambda}(x,t),\varepsilon^{p,k,\lambda}(x,t)\right)\mathrm{d}x.$$

Now we can state the existence result for model (2.6).

THEOREM 2.3 (Existence of solution to problem (2.6)).
Assume that the initial conditions u_0, u_1, ε_0^p, given boundary data g_D, g_N and external forces F, have the regularity (2.1)–(2.3). *Moreover, suppose that initial data are admissible and along with boundary data satisfy the compatibility conditions* (2.4). *Then, for every $k\in\mathbb{N}_+$, there exists a unique solution $(u^k,\varepsilon^{p,k},T^k)$ of* (2.6) *with the initial-boundary conditions* (1.2)–(1.3) *such that*

$$u^k\in W^{2,\infty}\left(0,T_e;L^2(\Omega;\mathbb{R}^3)\right),\quad \varepsilon(u^k)\in W^{1,\infty}\left(0,T_e;L^2(\Omega;\mathcal{S}(3))\right),$$
$$\varepsilon^{p,k}\in W^{1,\infty}\left(0,T_e;L^2(\Omega;\mathcal{S}(3))\right),\quad \mathrm{div}T^k\in L^\infty\left(0,T_e;L^2\left(\Omega;\mathbb{R}^3\right)\right).$$

Proof of Theorem 2.3 is very similar to the proof presented in [5, section 4 and 5] (computation is very similar however it have to be done for plasticity not for poroplasticity model). Main idea of the proof is quite simple. One have to approximate differential inclusion be sequence of differential equations given by

$$\varepsilon_t^{p,k,\lambda}(x,t)=(\partial M_0)^\lambda\left(\widehat{T}^{k,\lambda}(x,t)\right),$$

where $(\partial M_0)^\lambda$ denotes the Yosida approximation of the operator ∂M_0. This approximation is maximal-monotone and globally Lipschitz with Lipschitz constant $1/\lambda$ (for details see [2]). Then, in the case when the right hand side of a constitutive equation is globally Lipschitz vector field, one can prove existence by the same reasoning as in [5, section 4] (Galerkin approximation and fixed point method). Therefore, in order to obtain solution to (2.6) for any fixed k one have to pass to the limit with λ in its Yosida approximation which also can be done due to quite standard reasoning (see for example [5, secton 5] or [6, section 4]).

3. Passing to the limit in coercive approximation.
The main part of the classic existence proof, where *weak safe-load condition* is needed, is proving the energy estimates (see [3, Theorem 3]). In the rest of the proof [3, Theorems 4,5,6] this assumption is not essential. Therefore, here we are going to present only the quick sketch of the proof of energy estimates and the rest of reasoning will be omitted.

THEOREM 3.1 (Energy estimates).
Assume (2.1)–(2.5) *then, for every $t\in[0,T_e]$ the following estimates hold:*

$$\operatorname*{ess\,sup}_{\tau\in(0,t)}\mathcal{E}^k\left(u_t^k,\varepsilon^k,\varepsilon^{p,k}\right)(\tau)+\int\limits_0^t\int\limits_\Omega\varepsilon_t^{p,k}\,\widehat{T}^k\,\mathrm{d}x\mathrm{d}\tau\leqslant C,\tag{3.1}$$

$$\operatorname*{ess\,sup}_{\tau\in(0,t)}\mathcal{E}^k\left(u_{tt}^k,\varepsilon_t^k,\varepsilon_t^{p,k}\right)(\tau)\leqslant C,\tag{3.2}$$

$$\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))} \leqslant C, \tag{3.3}$$

where $(u^k, \varepsilon^{p,k})$ *is a solution of* (2.6) *with the initial-boundary conditions* (1.2)–(1.3). *Constant* $C \geqslant 0$ *is independent of* k *and* t.

Proof.
Firstly, one have to prove the following

$$\operatorname*{ess\,sup}_{\tau\in(0,t)} \mathcal{E}^k\left(u_{tt}^k, \varepsilon_t^k, \varepsilon_t^{p,k}\right)(t) \leqslant C + C\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))} \text{ for a.e. } t \in (0, T_e). \tag{a}$$

To begin let us introduce a special notation for translated in time function, i.e.

$$\left(u_{t,h}^k(t), \varepsilon_h^k(t), \varepsilon_h^{p,k}(t)\right) := \left(u_t^k(t+h), \varepsilon^k(t+h), \varepsilon^{p,k}(t+h)\right),$$

where $h > 0$ is a sufficiently small constant.
Then, computing $\frac{1}{h^2}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^k\left(u_{t,h}^k - u_t^k, \varepsilon_h^k - \varepsilon^k, \varepsilon_h^{p,k} - \varepsilon^{p,k}\right)(t)$ and using similar methods as presented in [4, Theorem 7.1] in order to pass to the limit with h give

$$\begin{aligned}\mathcal{E}^k\left(u_{tt}^k, \varepsilon_t^k, \varepsilon_t^{p,k}\right)(t) \leqslant C \cdot &\left(\left\|u_t^k\right\|_{L^\infty(0,t;L^1(\partial\Omega))} + \left\|T^k n\right\|_{L^\infty\left(0,t;H^{-\frac{1}{2}}(\partial\Omega)\right)}\right) \\ &+ C(\nu) + \nu \cdot \mathcal{E}^k\left(u_{tt}^k, \varepsilon_t^k, \varepsilon_t^{p,k}\right)(t),\end{aligned} \tag{3.4}$$

where ν is an arbitrary positive constant. Using trace theorems and some elementary inequalities allows to obtain:

$$\left\|T^k(t)\, n\right\|_{H^{-\frac{1}{2}}(\partial\Omega)} \leqslant C(\nu) + \nu \cdot \operatorname*{ess\,sup}_{(0,t)} \mathcal{E}^k(u_{tt}^k, \varepsilon_t^k, \varepsilon_t^{p,k})(t). \tag{3.5}$$

$$\left\|u_t^k(t)\right\|_{L^1(\partial\Omega)} \leqslant C(\nu) + C\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1)} + \nu \cdot \operatorname*{ess\,sup}_{(0,t)} \mathcal{E}^k(u_{tt}^k, \varepsilon_t^k, \varepsilon_t^{p,k})(t). \tag{3.6}$$

Using (3.5) and (3.6) in (3.4), taking the supremum over $(0,t)$ and fixing a sufficiently small ν finally give (a).

Secondly, one have to prove that

$$\begin{aligned}\operatorname*{ess\,sup}_{\tau\in(0,t)} \mathcal{E}^k\left(u_t^k, \varepsilon^k, \varepsilon^{p,k}\right)(\tau) + \int_0^t \int_\Omega \varepsilon_t^{p,k}\, \widehat{T}^k \,\mathrm{d}x\mathrm{d}\tau \leqslant &C(\mu_1) + \mu_1\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))} \\ &+ C\left\|\varepsilon_t^{p,k}\right\|_{L^1(0,t;L^1(\Omega))},\end{aligned} \tag{b}$$

where μ_1 is an arbitrary positive constant.

We start by computing $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^k\left(u_t^k, \varepsilon^k, \varepsilon^{p,k}\right)(t)$ then, after using few elementary estimates and integrating over time $(0,t)$ one can obtain

$$\begin{aligned}\mathcal{E}^k\left(u_t^k, \varepsilon^k, \varepsilon^{p,k}\right)(t) + \int_0^t \int_\Omega \varepsilon_t^{p,k}\, \widehat{T}^k \,\mathrm{d}x\mathrm{d}\tau \leqslant &C(\nu) + \nu \cdot \operatorname*{ess\,sup}_{(0,t)} \mathcal{E}^k\left(u_t^k, \varepsilon^k, \varepsilon^{p,k}\right) \\ &+ C\left\|T^k n\right\|_{L^\infty\left(0,t;H^{-\frac{1}{2}}(\partial\Omega)\right)} \\ &+ C\left\|u_t^k\right\|_{L^1(0,t;L^1(\partial\Omega))},\end{aligned} \tag{3.7}$$

where ν is an arbitrary positive constant. By using trace theorems and some elementary inequalities one can prove the following inequalities

$$\begin{aligned}\left\|T^k(t)n\right\|_{H^{-\frac{1}{2}}(\partial\Omega)} \leqslant & C(\nu,\mu_1)+\nu\cdot\operatorname*{ess\,sup}_{(0,t)}\mathcal{E}^k(u_t^k,\varepsilon^k,\varepsilon^{p,k})(t)\\ &+\frac{\mu_1}{2}\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))}.\end{aligned}\tag{3.8}$$

$$\begin{aligned}\left\|u_t^k\right\|_{L^1(0,t;L^1(\partial\Omega))} \leqslant & C\left(\nu,\mu_1\right)+\frac{\mu_1}{2}\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))}+C\left\|\varepsilon_t^{p,k}\right\|_{L^1(0,t;L^1(\Omega))}\\ &+\nu\cdot\operatorname*{ess\,sup}_{(0,t)}\mathcal{E}^k(u_t^k,\varepsilon^k,\varepsilon^{p,k})(t).\end{aligned}\tag{3.9}$$

Hence, by using (3.8) and (3.9) in (3.7), taking the supremum over $(0,t)$ and, fixing sufficiently small $\nu>0$ one can obtain (b).

As a third step one have to prove the following inequality:

$$\left\|\varepsilon_t^{p,k}\right\|_{L^1(0,t;L^1(\Omega))}\leqslant C\left(\mu_2\right)+\mu_2\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))},\tag{c}$$

where μ_2 is an arbitrary positive constant.

Due to the monotonicity of ∂M_0 one can obtain that for any $\delta_0>0$ the following inequality holds

$$\left|\varepsilon_t^{p,k}\right|\leqslant\frac{1}{\delta_0}\varepsilon_t^{p,k}\widehat{T}^k+\frac{1}{\delta_0}\sup_{|\sigma|\leqslant\delta_0}\left|m\left(\partial M_0\left(\sigma\right)\right)\right|\left(\left|\widehat{T}^k\right|+\delta_0\right),\tag{3.10}$$

where $m\left(\partial M_0\left(\sigma\right)\right)$ is the element of $\partial M_0\left(\sigma\right)$ of minimal norm.
Integrating (3.10) over $\Omega\times(0,t)$ for $t\leqslant T_e$, using some elementary inequalities along with (b) and fixing a sufficiently large δ_0 (it is possible due to viscoplasticity assumption) give (c).

In the last step one have to prove that

$$\left\|\varepsilon_t^{p,k}\left(\tau\right)\right\|_{L^\infty(0,t;L^1(\Omega))}\leqslant C.\tag{d}$$

which finally allows to close estimates (c), (b) (a) and therefore ends the proof.

Using inequality (c) in (b) gives

$$\mathcal{E}^k\left(u_t^k,\varepsilon^k,\varepsilon^{p,k}\right)(t)+\int_0^t\int_\Omega\varepsilon_t^{p,k}\,\widehat{T}^k\,\mathrm{d}x\mathrm{d}\tau\leqslant C(\mu)+\mu\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))},\tag{3.11}$$

where $\mu>0$ is an arbitrary constant.
After integrating (3.10) over Ω and using some elementary inequalities along with (3.11) one can obtain for almost every $t\in(0,T_e)$

$$\left\|\varepsilon_t^{p,k}(t)\right\|_{L^1(\Omega)}\leqslant\frac{1}{\delta_0}\int_\Omega\varepsilon_t^{p,k}(t)\,\widehat{T}^k(t)\,\mathrm{d}x+\mu\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))}+C(\mu,\delta_0).\tag{3.12}$$

Computing $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}^k\left(u_t^k,\varepsilon^k,\varepsilon^{p,k}\right)(t)$ and using standard inequalities lead to

$$\begin{aligned}\int_\Omega\varepsilon_t^{p,k}(t)\,\widehat{T}^k(t)\,\mathrm{d}x\leqslant & C-\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{E}^k\left(u_t^k,\varepsilon^k,\varepsilon^{p,k}\right)(t)\right)+\mathcal{E}^k\left(u_t^k,\varepsilon^k,\varepsilon^{p,k}\right)(t)\\ &+C\left\|T^k(t)\,n\right\|_{H^{-\frac{1}{2}}(\partial\Omega)}+C\left\|u_t^k(t)\right\|_{L^1(\partial\Omega)}.\end{aligned}\tag{3.13}$$

Using (3.5) and (3.6) along with (a) leads to

$$\left\|T^k(t)\,n\right\|_{H^{-\frac{1}{2}}(\partial\Omega)} + \left\|u_t^k(t)\right\|_{L^1(\partial\Omega)} \leqslant C + C\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))}. \tag{3.14}$$

Using (3.11), (3.14) and (a) in (3.13) yields for almost every $t \in (0, T_e)$

$$\int_\Omega \varepsilon_t^{p,k}(t)\,\widehat{T}^k(t)\,\mathrm{d}x \leqslant C + C\left\|\varepsilon_t^{p,k}\right\|_{L^\infty(0,t;L^1(\Omega))}. \tag{3.15}$$

After inserting (3.15) into (3.12), taking the essential supremum, fixing sufficiently small $\mu > 0$ and sufficiently large $\delta_0 > 0$ (possible because constitutive function is viscoplastic) one can finally obtain (d), which ends the proof. □

REFERENCES

[1] H.-D. Alber, *Materials with memory*, volume 1682 Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1998.

[2] J.-P. Aubin and A. Cellina, *Differential inclusions*, volume 264 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1984.

[3] K. Chełmiński, *Coercive approximation of viscoplasticity and plasticity*, Asymptotic Anal., 26(2), pp. 105–133, 2001.

[4] K. Kisiel, *Dynamical poroplasticity model – Existence theory for gradient type nonlinearities with Lipschitz perturbations*, J. Math. Anal. Appl., 450(1), pp. 544–577, 2017.

[5] K. Kisiel and K. Kosiba, *Dynamical poroplasticity model with mixed boundary conditions – theory for $\mathcal{LM}$-type nonlinearity*, J. Math. Anal. Appl., 443(1), pp. 187–229, 2016.

[6] S. Owczarek, *Existence of solution to a non-monotone dynamic model in poroplasticity with mixed boundary conditions*, Topol. Methods Nonlinear Anal., 43(2), pp. 297–322, 2014.

Proceedings of EQUADIFF 2017
pp. 37–44

NONLINEAR DIFFUSION EQUATIONS WITH PERTURBATION TERMS ON UNBOUNDED DOMAINS

SHUNSUKE KURIMA*

Abstract. This paper considers the initial-boundary value problem for the nonlinear diffusion equation with the perturbation term

$$u_t + (-\Delta + 1)\beta(u) + G(u) = g \quad \text{in } \Omega \times (0,T)$$

in an unbounded domain $\Omega \subset \mathbb{R}^N$ with smooth bounded boundary, where $N \in \mathbb{N}$, $T > 0$, β is a single-valued maximal monotone function on $\mathbb{R}$, e.g.,

$$\beta(r) = |r|^{q-1}r \ (q > 0, q \neq 1)$$

and G is a function on $\mathbb{R}$ which can be regarded as a Lipschitz continuous operator from $(H^1(\Omega))^*$ to $(H^1(\Omega))^*$. The present work establishes existence and estimates for the above problem.

Key words. porous media equations, fast diffusion equations, subdifferential operators

AMS subject classifications. 35K59, 35K35, 47H05

1. Introduction. In this paper we consider the initial-boundary value problem for the nonlinear diffusion equation with the perturbation term

$$\begin{cases} u_t + (-\Delta + 1)\beta(u) + G(u) = g & \text{in } \Omega \times (0,T), \\ \partial_\nu \beta(u) = 0 & \text{on } \partial\Omega \times (0,T), \\ u(0) = u_0 & \text{in } \Omega, \end{cases} \tag{P}$$

where Ω is an *unbounded* domain in $\mathbb{R}^N$ ($N \in \mathbb{N}$) with smooth bounded boundary $\partial\Omega$ (e.g., $\Omega = \mathbb{R}^N \setminus \overline{B(0,R)}$, where $B(0,R)$ is the open ball with center 0 and radius $R > 0$) or $\Omega = \mathbb{R}^N$ or $\Omega = \mathbb{R}^N_+$, $T > 0$, and ∂_ν denotes the derivative with respect to the outward normal of $\partial\Omega$. Though the precise conditions for β, G, g and u_0 will be given in (A1)-(A4) stated later, we roughly explain that β is a single-valued maximal monotone function, e.g.,

$$\beta(r) = |r|^{q-1}r,$$

where the problem is the porous media equation in the case that $q > 1$ (see e.g., [1, 13, 17, 18]) and is the fast diffusion equation in the case that $0 < q < 1$ (see e.g., [5, 15, 17]); G can be regarded as a Lipschitz continuous operator from $(H^1(\Omega))^*$ to $(H^1(\Omega))^*$; g and u_0 are known functions.

Nonlinear diffusion equations on unbounded domains are not so substantially studied from a viewpoint of the operator theory, whereas in the case that $\Omega = \mathbb{R}^N$ the equations are studied by the method of real analysis (see e.g., [11]). The case of unbounded domains would be important in both mathematics and physics. Also, since compact methods do not work directly on unbounded domains, it would be worth studying the case of unbounded domains mathematically. Also, the perturbation term $G(u)$ makes proving existence for (P) without growth conditions for β be difficult (see

*Department of Mathematics, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan (shunsuke.kurima@gmail.com).

Remark in the end of this section). Although we give an example of G only as $G(u) = u$ in this paper, if we can weaken the condition for G and we can take $G(u) = -\beta(u)$, then we can possibly deal with the "pure" diffusion equation as $u_t - \Delta\beta(u) = g$.

In the case that $G \equiv 0$, in [12] and [9], existence of weak solutions to (P) and their estimates were shown by monotonicity methods.

The new point of this paper is that the perturbation term "$G(u)$" is added to the left-hand side of the equation $u_t + (-\Delta + 1)\beta(u) = g$ studied in [12] and [9]. The purpose of this paper is to show existence of weak solutions to (P) and to obtain their estimates. In particular, we prove existence for (P) by using Brézis's theory which is a monotonicity method for an abstract evolution equation including a subdifferential operator and a perturbation term.

We first give assumptions, notations and definitions used in this paper before introducing main results.

Assume that β, G, g and u_0 satisfy the following conditions:

(A1) The following (A1a), (A1b) and (A1c) hold:

(A1a) $\beta : \mathbb{R} \to \mathbb{R}$ is a single-valued maximal monotone function and

$$\beta(r) = \hat{\beta}\,'(r) = \partial\hat{\beta}(r),$$

where $\hat{\beta}\,'$ and $\partial\hat{\beta}$ respectively denote the differential and subdifferential of a proper differentiable (lower semicontinuous) convex function $\hat{\beta} : \mathbb{R} \to [0, +\infty]$ satisfying $\hat{\beta}(0) = 0$. This entails $\beta(0) = 0$.

(A1b) There exist constants $m > 1$ and $c_0, c_0' > 0$ such that for all $r \in \mathbb{R}$,

$$\hat{\beta}(r) \geq c_0|r|^m$$

and

$$|\beta(r)| \leq c_0'|r|^{m-1}$$

hold.

(A1c) For all $z \in H^1(\Omega)$, if $\hat{\beta}(z) \in L^1(\Omega)$, then $\beta(z) \in L^1_{\mathrm{loc}}(\Omega)$. For all $z \in H^1(\Omega)$ and all $\psi \in C_{\mathrm{c}}^\infty(\Omega)$, if $\hat{\beta}(z) \in L^1(\Omega)$, then $\hat{\beta}(z + \psi) \in L^1(\Omega)$.

(A2) $G : (H^1(\Omega))^* \to (H^1(\Omega))^*$ is a Lipschitz continuous operator.

(A3) $g \in L^2\big(0, T; L^2(\Omega)\big)$.

(A4) $u_0 \in L^2(\Omega)$ and $\hat{\beta}(u_0) \in L^1(\Omega)$.

From (A3) we can fix a solution $f \in L^2\big(0, T; H^2(\Omega)\big)$ of

$$\begin{cases} (-\Delta + 1)f(t) = g(t) & \text{a.e. in } \Omega, \\ \partial_\nu f(t) = 0 & \text{in the sense of traces on } \partial\Omega \end{cases}$$

for a.a. $t \in (0, T)$, that is,

$$\int_\Omega \nabla f(t) \cdot \nabla z + \int_\Omega f(t)z \ = \int_\Omega g(t)z \quad \text{for all } z \in H^1(\Omega).$$

An example of (A2) is given as $G(v^*) = v^*$ for all $v^* \in (H^1(\Omega))^*$.

We define the Hilbert spaces

$$H := L^2(\Omega), \quad V := H^1(\Omega)$$

with inner products $(\cdot,\cdot)_H$ and $(\cdot,\cdot)_V$, respectively. Moreover we put

$$W := \{z \in H^2(\Omega) \mid \partial_\nu z = 0 \quad \text{a.e. on } \partial\Omega\}.$$

The notation V^* denotes the dual space of V with duality pairing $\langle\cdot,\cdot\rangle_{V^*,V}$. Moreover the Riesz representation theorem ensures the existence of a bijective mapping $F : V \to V^*$ satisfying

$$\langle Fv_1, v_2\rangle_{V^*,V} := (v_1, v_2)_V \quad \text{for all } v_1, v_2 \in V$$

and we define the inner product in V^* as

$$(v_1^*, v_2^*)_{V^*} := \langle v_1^*, F^{-1}v_2^*\rangle_{V^*,V} \quad \text{for all } v_1^*, v_2^* \in V^*.$$

We remark that (A3) implies

$$Ff(t) = g(t) \quad \text{for a.a. } t \in (0,T). \tag{1.1}$$

We give the definition of weak solutions to (P).

DEFINITION 1.1. *A pair* (u,μ) *with*

$$u \in H^1(0,T;V^*), \quad \mu \in L^2(0,T;V)$$

is called a weak solution of (P) *if* (u,μ) *satisfies*

$$\langle u'(t) + G(u(t)), z\rangle_{V^*,V} + (\mu(t), z)_V = 0 \quad \textit{for all } z \in V \textit{ and a.a. } t \in (0,T), \tag{1.2}$$

$$\mu(t) = \beta(u(t)) - f(t) \quad \textit{in } V \quad \textit{for a.a. } t \in (0,T), \tag{1.3}$$

$$u(0) = u_0 \quad \textit{a.e. on } \Omega. \tag{1.4}$$

We next state the main result which asserts existence and estimates for (P).

THEOREM 1.2. *Assume* (A1)-(A4). *Then there exists a unique weak solution* (u,μ) *of* (P) *satisfying* $u \in H^1(0,T;V^*), \mu \in L^2(0,T;V)$. *Moreover, if it holds that* $G(v^*) = av^*$ *for* $v^* \in V^*$, *where* $a \in \mathbb{R}$, *then there exists a constant* $M > 0$ *such that for a.a.* $t \in (0,T)$, $u(t) \in H$ *and*

$$|u(t)|_H^2 \le M, \tag{1.5}$$

$$\int_0^t |u'(s)|_{V^*}^2\, ds + a|u(t)|_{V^*}^2 \le M, \tag{1.6}$$

$$\int_0^t |\mu(s)|_V^2\, ds \le M, \tag{1.7}$$

$$\int_0^t |\beta(u(s))|_V^2\, ds \le M. \tag{1.8}$$

In the case that $G \equiv 0$, in [12], existence of weak solutions to (P) was proved by rewriting (P) to

$$u'(t) + \partial\phi(u(t)) = g(t) \quad \text{in } V^*,$$

where ϕ is a proper lower semicontinuous convex function on V^* and $\partial\phi$ is the subdifferential of ϕ, and by applying Brézis's theory ([3, Theorem 3.6]). Also, the m-growth

condition for β was assumed to derive the lower semicontinuity of $\phi : V^* \to \overline{\mathbb{R}}$. The examples are the porous media equation and the fast diffusion equation. Recently, in [9], the approximation

$$u_\varepsilon'(t) + (-\Delta + 1)(\varepsilon(-\Delta + 1)u_\varepsilon(t) + \beta(u_\varepsilon(t)) + \pi_\varepsilon(u_\varepsilon(t))) = g \qquad (\mathrm{P})_\varepsilon$$

was considered and existence of weak solutions to $(\mathrm{P})_\varepsilon$ with their estimates was shown; moreover, existence of weak solutions to (P) with their estimates was obtained without growth conditions for β, and existence of weak solutions to (P), their estimates were obtained without growth conditions for β by passing to the limit in $(\mathrm{P})_\varepsilon$ as $\varepsilon \searrow 0$. In addition to the porous media equation and fast diffusion equation, the examples of (P) include the Stefan problem (see e.g., [2, 4, 6, 7, 8, 10]) which is described by (P) with

$$\beta(r) = \begin{cases} k_s r & \text{if } r < 0, \\ 0 & \text{if } 0 \le r \le L, \\ k_\ell (r - L) & \text{if } r > L \end{cases}$$

for all $r \in \mathbb{R}$, where k_s, k_ℓ, L are positive constants.

The strategy for the proof of Theorem 1.2 is to prove existence for (P) under the m-growth condition for β by setting some proper lower semicontinuous convex function $\phi : V^* \to \overline{\mathbb{R}}$ appropriately as in [12, Section 3], by rewriting (P) to

$$u'(t) + \partial\phi(u(t)) + G(u(t)) = g \quad \text{in } V^*$$

and by applying Brézis's theory to the above abstract evolution equation with the perturbation term.

Remark. At the moment, we do not know whether existence of weak solutions to (P) can be proved in a similar way to [9] or not. Since existence of weak solutions to the approximation of (P)

$$u_\varepsilon'(t) + (-\Delta + 1)\big(\varepsilon(-\Delta + 1)u_\varepsilon(t) + \beta(u_\varepsilon(t)) + \pi_\varepsilon(u_\varepsilon(t))\big) + G(u_\varepsilon(t)) = g \qquad (1.9)$$

can be proved in a similar way to the above strategy for (P), we can prove existence of weak solutions to (P) by passing the limit in (1.9) if we can obtain estimates for (1.9). In this paper we will directly prove existence of weak solutions to (P) under the m-growth condition for β without approximation (1.9). We hope that we can avoid growth conditions for β in a future work.

The plan of this paper is as follows. In Section 2 we prove existence of weak solutions to (P). Section 3 obtains estimates for (P). In Section 4 we present the porous media equation and the fast diffusion equation as examples of (P).

2. Proof of Theorem 1.2 (existence). In this section we will prove existence of a unique weak solution to (P). The following lemma is known in the Brézis's theory for a nonlinear evolution equation with a perturbation term including a subdifferential operator (see e.g., [3, Proposition 3.12]) and plays an important role in this section.

LEMMA 2.1. *Let X be a real Hilbert space, let $\psi : X \to \overline{\mathbb{R}}$ be a proper l.s.c. convex function and let $G : X \to X$ be a Lipschitz continuous operator. If $u_0 \in D(\psi)$ and $\tilde{f} \in L^2(0, T; X)$, then there exists a unique function u such that $u \in H^1(0, T; X)$, $u(t) \in D(\partial\psi)$ for a.a. $t \in (0, T)$ and u solves the following initial value problem:*

$$\begin{cases} u'(t) + \partial\psi(u(t)) + G(u(t)) \ni \tilde{f}(t) & \text{in } X \quad \text{for a.a. } t \in (0, T), \\ u(0) = u_0 & \text{in } X. \end{cases}$$

Proof of Theorem 1.2 (*existence*). Defining $\phi : V^* \to \overline{\mathbb{R}}$ as

$$\phi(z) = \begin{cases} \displaystyle\int_\Omega \hat{\beta}(z(x))\,dx & \text{if } z \in D(\phi) := \{z \in V^* \cap L^m(\Omega) \mid \hat{\beta}(z) \in L^1(\Omega)\}, \\ +\infty & \text{otherwise,} \end{cases}$$

we deduce from [12, Section 3] that this ϕ is proper lower semicontinuous convex on V^* and

$$\beta(z) \in V, \quad \partial\phi(z) = F\beta(z) \tag{2.1}$$

hold for all $z \in D(\partial\phi)$. Hence, from (1.1) and (2.1) we can rewrite (1.2)-(1.4) in Definition 1.1 to

$$\begin{cases} u'(t) + \partial\phi(u(t)) + G(u(t)) = g(t) \quad \text{in } V^* \quad \text{for a.a. } t \in [0,T], \\ u(0) = u_0 \quad \text{in } V^*. \end{cases} \tag{2.2}$$

Invoking Lemma 2.1, we can find a unique solution $u \in H^1(0,T;V^*)$ of (2.2) and $u(t) \in D(\partial\phi)$ for a.a. $t \in (0,T)$. Hence there exists a unique weak solution of (P). □

3. Proof of Theorem 1.2 (estimates). We will obtain the estimates for weak solutions of (P) in this section.

Proof of Theorem 1.2 (*estimates*). In addition to (A2) we assume further that

$$G(v^*) = av^*$$

for all $v^* \in V^*$, where $a \in \mathbb{R}$. For $\lambda > 0$ we put

$$\begin{aligned} &A := -\Delta : D(A) := W \subset H \to H, \\ &J_\lambda := (I + \lambda A)^{-1} : H \to H, \\ &A_\lambda := \lambda^{-1}(I - J_\lambda) : H \to H, \end{aligned}$$

and

$$\begin{aligned} &\tilde{A} := F - I : V \to V^*, \\ &\tilde{J}_\lambda := \left(I + \lambda\tilde{A}\right)^{-1} : V^* \to V. \end{aligned}$$

Let $u \in H^1(0,T;V^*)$ be a unique solution of (2.2). We first show (1.5). Noting that $\tilde{J}_\lambda^{1/2} : V^* \to H$ is defined as a bounded operator (see e.g., [14, Lemma 3.3]) and putting

$$u_\lambda(t) := \tilde{J}_\lambda^{1/2} u(t) \quad \text{for all } t \in (0,T),$$

we derive from [12, Lemma 3.3] that

$$u_\lambda \in H^1(0,T;H)$$

and

$$u_\lambda'(t) + \tilde{J}_\lambda^{1/2} F\beta(u(t)) + \tilde{J}_\lambda^{1/2} G(u(t)) = J_\lambda^{1/2} g(t).$$

Then we obtain

$$\frac{1}{2}\frac{d}{ds}|u_\lambda(s)|_H^2 \le \frac{1}{2}|g(s)|_H^2 + \frac{1}{2}|u_\lambda(s)|_H^2 + \big(\tilde{J}_\lambda^{1/2}G(u(s)), u_\lambda(s)\big)_H \tag{3.1}$$

in a similar way to [12, Section 3]. Here we have

$$\begin{aligned}\big(\tilde{J}_\lambda^{1/2}G(u(s)), u_\lambda(s)\big)_H &= a\big(\tilde{J}_\lambda^{1/2}u(s), u_\lambda(s)\big)_H \\ &= a|u_\lambda(s)|_H^2. \end{aligned}\tag{3.2}$$

Thus combining (3.1) and (3.2) yields

$$\frac{1}{2}\frac{d}{ds}|u_\lambda(s)|_H^2 \le \frac{1}{2}|g(s)|_H^2 + \Big(a+\frac{1}{2}\Big)|u_\lambda(s)|_H^2,$$

and hence the inequality

$$|u_\lambda(t)|_H^2 \le e^{|2a+1|T}|u_0|_H^2 + e^{|2a+1|T}|g|_{L^2(0,T;H)}^2$$

holds for all $t \in (0,T)$. Therefore it follows from a similar way to [12, Section 3, Proof of Theorem 1.1 (continued)] that for a.a. $t \in (0,T)$,

$$u(t) \in H$$

and there exists a positive constant C such that

$$\|u\|_{L^\infty(0,T;H)} \le C,$$

which means that the estimate (1.5) holds.

Next we verify (1.6). The equation in (2.2) yields that

$$\begin{aligned}\big|u'(s)\big|_{V^*}^2 &= -\big(u'(s), \partial\phi(u(s))\big)_{V^*} + \big(u'(s), Ff(s)\big)_{V^*} - a\big(u'(s), u(s)\big)_{V^*} \\ &= -\big(u'(s), \partial\phi(u(s))\big)_{V^*} + \big(u'(s), Ff(s)\big)_{V^*} - \frac{a}{2}\frac{d}{ds}|u(s)|_{V^*}^2. \end{aligned}\tag{3.3}$$

Here we have

$$\big(u'(s), \partial\phi(u(s))\big)_{V^*} = \frac{d}{ds}\phi(u(s))$$

(see e.g., Showalter [16, Lemma IV.4.3]) and it follows from the definition of $(\cdot,\cdot)_{V^*}$ and Young's inequality that

$$\begin{aligned}\big(u'(s), Ff(s)\big)_{V^*} &= \big\langle u'(s), f(s)\big\rangle_{V^*,V} \\ &\le \frac{1}{2}\big|u'(s)\big|_{V^*}^2 + \frac{1}{2}|f(s)|_V^2. \end{aligned}$$

Integrating (3.3) combined with these facts leads to the inequality

$$\frac{1}{2}\int_0^t \big|u'(s)\big|_{V^*}^2\,ds \le -\phi(u(t)) + \phi(u_0) + \frac{1}{2}|f|_{L^2(0,T;V)}^2 - \frac{a}{2}|u(t)|_{V^*}^2 + \frac{a}{2}|u_0|_{V^*}^2,$$

i.e.,

$$\frac{1}{2}\int_0^t \big|u'(s)\big|_{V^*}^2\,ds + \int_\Omega \hat{\beta}(u(t)) + \frac{a}{2}|u(t)|_{V^*}^2 \le \int_\Omega \hat{\beta}(u_0) + \frac{1}{2}|f|_{L^2(0,T;V)}^2 + \frac{a}{2}|u_0|_{V^*}^2.$$

Since (A1a) implies

$$\int_\Omega \hat{\beta}(u(t)) \geq 0,$$

it holds that

$$\int_0^t \big|u'(s)\big|_{V^*}^2 \, ds + a|u(t)|_{V^*}^2 \leq 2\int_\Omega \hat{\beta}(u_0) + |f|_{L^2(0,T;V)}^2 + a|u_0|_{V^*}^2.$$

Next we show (1.7). The fact that $\mu(s) = -F^{-1}\big(u'(s)\big)$ implies

$$\begin{aligned}\int_0^t |\mu(s)|_V^2 \, ds &= \int_0^t \big|F^{-1}\big(u'(s)\big)\big|_V^2 \, ds \\ &= \int_0^t \big|u'(s)\big|_{V^*}^2 \, ds.\end{aligned}$$

Hence (1.7) can be obtained from (1.6).

Next we prove (1.8). We see from (1.3) and the definition of $\mu(\cdot)$ that

$$\begin{aligned}|\beta(u(s))|_V^2 &= \big(-F^{-1}\big(u'(s)\big) - au(s) + f(s), \beta(u(s))\big)_V \\ &\leq \big(\big|F^{-1}\big(u'(s)\big)\big|_V + |a||u(s)|_{V^*} + |f(s)|_V\big)|\beta(u(s))|_V \\ &\leq \big|F^{-1}\big(u'(s)\big)\big|_V^2 + a^2|u(s)|_{V^*}^2 + |f(s)|_V^2 + \frac{3}{4}|\beta(u(s))|_V^2 \\ &= \big|u'(s)\big|_{V^*}^2 + a^2|u(s)|_{V^*}^2 + |f(s)|_V^2 + \frac{3}{4}|\beta(u(s))|_V^2.\end{aligned}$$

Integrating this inequality, we have

$$\int_0^t |\beta(u(s))|_V^2 \, ds \leq 4\int_0^t \big|u'(s)\big|_{V^*}^2 \, ds + 4a^2 \int_0^t |u(s)|_{V^*}^2 \, ds + 4|f|_{L^2(0,T;V)}^2.$$

Therefore there exists a constant $M > 0$ satisfying (1.5), (1.6), (1.7) and (1.8). Moreover, (1.5) means that $u \in L^\infty(0,T;H)$. □

4. Examples. An example of $G : (H^1(\Omega))^* \to (H^1(\Omega))^*$ is given by $G(v^*) = v^*$ for all $v^* \in (H^1(\Omega))^*$. As to β, we give the following two examples.

The porous media equation. We consider

$$\beta(r) = |r|^{q-1}r \quad (q > 1).$$

This β is the function in the porous media equation (see e.g., [1, 13, 17, 18]).

The fast diffusion equation. Consider

$$\beta(r) = |r|^{q-1}r \quad (0 < q < 1).$$

This β is the function in the fast diffusion equation (see e.g., [5, 15, 17]).

In both examples we can show that β satisfies (A1), (A4) (see [12, Section 6]).

Acknowledgments. The author would like to thank the referee for helpful comments and suggestions and to appreciate that Professor Tomomi Yokota encouraged him very kindly and gave comments on the manuscript.

REFERENCES

[1] G. Akagi, G. Schimperna, and A. Segatti, *Fractional Cahn–Hilliard, Allen–Cahn and porous medium equations*, J. Differential Equations, 261 (2016), pp. 2935–2985.
[2] D. Blanchard and A. Porretta, *Stefan problems with nonlinear diffusion and convection*, J. Differential Equations, 210 (2005), pp. 383–428.
[3] H. Brézis, "Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Especes de Hilbert", North-Holland, Amsterdam, 1973.
[4] A. Damlamian, *Some results on the multi-phase Stefan problem*, Comm. Partial Differential Equations, 2 (1977), pp. 1017–1044.
[5] E. DiBenedetto, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp. 83–118.
[6] A. Friedman, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 133 (1968), pp. 51–87.
[7] T. Fukao, N. Kenmochi, and I. Pawłow, *Transmission problems arising in Czochralski process of crystal growth*, Mathematical aspects of modelling structure formation phenomena (Będlewo/Warsaw, 2000), pp. 228–243, GAKUTO Internat. Ser. Math. Sci. Appl., 17, Gakkōtosho, Tokyo, 2001.
[8] T. Fukao, *Convergence of Cahn–Hilliard systems to the Stefan problem with dynamic boundary conditions*, Asymptot. Anal., 99 (2016), pp. 1–21.
[9] T. Fukao, S. Kurima, and T. Yokota, *Nonlinear diffusion equations as asymptotic limits of Cahn–Hilliard systems on unbounded domains via Cauchy's criterion*, preprint.
[10] A. Haraux and N. Kenmochi, *Asymptotic behaviour of solutions to some degenerate parabolic equations*, Funkcial. Ekvac., 34 (1991), pp. 19–38.
[11] C. E. Kenig, "Degenerate Diffusions", Initial value problems and local regularity theory. EMS Tracts in Mathematics, 1. European Mathematical Society (EMS), Zürich, 2007.
[12] S. Kurima and T. Yokota, *Monotonicity methods for nonlinear diffusion equations and their approximations with error estimates*, J. Differential Equations, 263 (2017), pp. 2024–2050.
[13] G. Marinoschi, *Well-posedness of singular diffusion equations in porous media with homogeneous Neumann boundary conditions*, Nonlinear Anal., 72 (2010), pp. 3491–3514.
[14] N. Okazawa, T. Suzuki, and T. Yokota, *Energy methods for abstract nonlinear Schrödinger equations*, Evol. Equ. Control Theory, 1 (2012), pp. 337–354.
[15] A. Rodriguez and J. L. Vázquez, *Obstructions to existence in fast-diffusion equations*, J. Differential Equations, 184 (2002), pp. 348–385.
[16] R. E. Showalter, "Monotone Operators in Banach Space and Nonlinear Partial Differential Equations", Mathematical Surveys and Monographs, 49, American Mathematical Society, Providence, RI, 1997.
[17] J. L. Vázquez, "The Porous Medium Equation", Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, Oxford, 2007.
[18] H. -M. Yin, *On a degenerate parabolic system*, J. Differential Equations, 245 (2008), pp. 722–736.

Proceedings of EQUADIFF 2017
pp. 45–52

ON BEHAVIOR OF SOLUTIONS TO A CHEMOTAXIS SYSTEM WITH A NONLINEAR SENSITIVITY FUNCTION*

TAKASI SENBA† AND KENTAROU FUJIE‡

Abstract. In this paper, we consider solutions to the following chemotaxis system with general sensitivity

$$\begin{cases} \tau u_t = \Delta u - \nabla \cdot (u \nabla \chi(v)) & \text{in } \Omega \times (0, \infty), \\ \eta v_t = \Delta v - v + u & \text{in } \Omega \times (0, \infty), \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega \times (0, \infty). \end{cases}$$

Here, τ and η are positive constants, χ is a smooth function on $(0, \infty)$ satisfying $\chi'(\cdot) > 0$ and Ω is a bounded domain of $\mathbf{R}^n$ $(n \geq 2)$.

It is well known that the chemotaxis system with direct sensitivity ($\chi(v) = \chi_0 v$, $\chi_0 > 0$) has blowup solutions in the case where $n \geq 2$. On the other hand, in the case where $\chi(v) = \chi_0 \log v$ with $0 < \chi_0 \ll 1$, any solution to the system exists globally in time and is bounded.

We present a sufficient condition for the boundedness of solutions to the system and some related systems.

Key words. Chemotaxis system, nonlinear sensitivity, time-global existence

AMS subject classifications. 35B45, 35K45, 35Q92, 92C17

1. Introduction.

We treat this system,

$$(PP) \begin{cases} \tau u_t = \nabla \cdot (\nabla u - u \nabla \chi(v)) & \text{in } \Omega \times (0, T), \\ \eta v_t = \Delta v - v + u & \text{in } \Omega \times (0, T), \\ \dfrac{\partial u}{\partial \nu} = \dfrac{\partial v}{\partial \nu} = 0 & \text{on } \partial\Omega \times (0, T), \\ u(\cdot, 0) = u_0, v(\cdot, 0) = v_0 & \text{in } \Omega. \end{cases}$$

Here, η and τ (time constants) are positive constants, $\Omega \subset \mathbf{R}^n$ $(n \geq 2)$ is a bounded domain with smooth boundary $\partial\Omega$, χ is smooth on $(0, \infty)$ satisfying $\chi'(v) > 0$ $(v > 0)$, $\nu = \nu(x)$ is the outer normal unite vector at $x \in \partial\Omega$ and initial conditions u_0 and v_0 are positive in $\overline{\Omega}$.

This system (PP) is introduced to describe the aggregation of cellular slime molds. Normally the living things move around as individual amoebas, performing a simple random walk. But when the environmental situation worsens, they suddenly change their behavior and aggregate to a single milt-cellular body. During this aggregation process, a chemical signal is secreted by cells to guide the collective movements. Unknown functions u and v in (PP) represent the density of the living things and the chemical concentration, respectively.

The maximal principle guarantees that

$$u > 0 \quad \text{and} \quad v > 0 \quad \text{in } \Omega \times (0, T_{max}).$$

*The first author was partially supported by Grant-in-Aid for Scientific Research (C) (No. 26400172), Japan Society for the Promotion of Science.

†Faculty of Science, Fukuoka University, Fukuoka, 814-0180,JAPAN (senba@fukuoka-u.ac.jp).

‡Faculty of Science Division I, Tokyo University of Science, Tokyo, 162-8601, JAPAN (fujie@rs.tus.ac.jp).

Here, T_{max} is the maximal existence time of the classical solution (u, v). It follows from the boundary condition that

$$\|u(t)\|_{L^1(\Omega)} = \|u_0\|_{L^1(\Omega)} \quad \text{for } t \in [0, T_{max}). \tag{1.1}$$

The function $\chi(v)$ represents the relation between the movement of cells and the chemical concentration. The term $u\chi'(v)\nabla v$ represents the flow due to the stimulus of the chemical substance. This property is so called chemotaxis. Then, the positivity of χ' means that the chemical substance is an attractant. When $\chi(v) = av$ and a is positive constant, we refer to this function as linear sensitivity function. The following functions are used in biological models frequently.

$$\chi(v) = av,\ a\log v,\ \frac{av}{b+v} \quad (a > 0,\ b > 0).$$

Except the linear sensitivity function, they satisfy that

$$\lim_{v\to\infty} \chi'(v) = 0. \tag{1.2}$$

This property represents saturation of the stimulus.

The following are our problem and our landmark.

Our problem

(i) Find a condition of sensitivity functions for the boundedness of solutions.
(ii) Find a condition of sensitivity functions for the existence of blowup solutions.

Our conjecture

(i) All solutions exist globally in time and are bounded, if one of the following two conditions holds:
· $\lim_{v\to\infty} \chi'(v) = 0$ and $n = 2$, or
· $\limsup_{v\to\infty} v\chi'(v) < \frac{n}{n-2}$ and $n \geq 3$.
(ii) There exist blowup solutions, if $\limsup_{v\to\infty} v\chi'(v) > \frac{n}{n-2}$ and $n \geq 3$.

Here, we say that a solution (u, v) to (PP) blows up at a time T, if

$$\limsup_{t\to T} \left(\|u(t)\|_{L^\infty(\Omega)} + \|v(t)\|_{L^\infty(\Omega)}\right) = \infty.$$

2. Known results. In this section, we describe known results.

Firstly, we describe those in the case where $\chi(v)$ is a linear function.

THEOREM 2.1. *Suppose that $\chi(v) = \chi_1 v$, $\chi_1 > 0$, $\eta > 0$ and $\tau > 0$ and that Ω is a bounded domain of $\mathbf{R}^n$ ($n \geq 2$) with smooth boundary. Then, the following hold:*

(i) *Suppose $n = 2$. Then, solutions exist globally in time and are bounded, if one of the following two conditions holds ([10]):*
· $\|u_0\|_{L^1(\Omega)} < 4\pi/\chi_1$, *or*
· *Ω is a bounded disk and u_0 is a radial function satisfying $\|u_0\|_{L^1(\Omega)} < 8\pi/\chi_1$.*

(ii) *If Ω is a bounded disk of $\mathbf{R}^2$ and u_0 is a radial function satisfying $\|u_0\|_{L^1(\Omega)} > 8\pi/\chi_1$, there are blowup solutions ([7]).*

(iii) *If Ω is a bounded ball of $\mathbf{R}^n$ ($n \geq 3$), there are blowup solutions ([16]).*

Then, in the linear sensitivity case, the behavior of solutions depends on the constant χ_1 and the L^1 norm of the solution u if $n = 2$, and there exist blowup solutions for any positive constant χ_1 if $n \geq 3$.

When χ is a nonlinear function satisfying (1.2), classical solutions to (PP) satisfy the following properties.

THEOREM 2.2. *Suppose that Ω is a bounded domain of $\mathbf{R}^n$ ($n \geq 2$) with smooth boundary. Then, the following hold:*

(i) *If $\chi'(v) \le a/(b+v)^p$, $a > 0$ and $p > 1$, solutions to (PP) exist globally in time and are bounded ([14, 5]).*

(ii) *If $\chi(v) = a \log v$ and $a < \sqrt{2/n}$, solutions to (PP) exist globally in time and are bounded ([15, 1]).*

The above sensitivity functions $\chi(v)$ satisfy that $\limsup_{v\to\infty} v\chi'(v) < \sqrt{2/n}$. Then, those conditions for global existence of classical solutions are not critical in the sense of our conjecture.

3. limiting systems. When the sensitivity function is a linear function, the condition for global existence of classical solutions is critical. The condition comes from a Lyapunov function and the Trudinger-Moser inequality ([10, 7, 16]). On the other hand, when the sensitivity function is not linear, it seems that conditions presented at the moment are not critical. In this case, we do not have any tools such as the Lyapunov function. Then, we consider the limiting system of (PP) as τ or $\eta = 0$. Because, those systems are simpler than (PP).

First, we consider the limiting system of (PP) as $\tau = 0$. For simplicity, we assume $\eta = 1$.

$$(PE)\begin{cases} u_t = \nabla\cdot(\nabla u - u\nabla\chi(v)) & \text{in } \Omega\times(0,T),\\ 0 = \Delta v - v + u & \text{in } \Omega\times(0,T),\\ \dfrac{\partial u}{\partial\nu} = \dfrac{\partial v}{\partial\nu} = 0 & \text{on } \partial\Omega\times(0,T),\\ u(\cdot,0) = u_0 & \text{in } \Omega. \end{cases}$$

Classical solutions to this system satisfy the following properties.

THEOREM 3.1. *Suppose that Ω is a bounded domain of $\mathbf{R}^n$ ($n \ge 2$) with smooth boundary and that $\lim_{v\to\infty}\chi'(v) = 0$. Then, the following hold:*

(i) *If $n = 2$, solutions to (PE) exist globally in time and are bounded ([2]).*

(ii) *If $n \ge 3$, Ω is a bounded ball, u_0 is radial, $\chi(v) = a\log v$ and $a < 2/(n-2)$, then solutions to (PE) exist globally in time and are bounded ([11]).*

(iii) *If $n \ge 3$, Ω is a bounded ball, u_0 is radial, $\chi(v) = a\log v$ and $a > 2n/(n-2)$, there are blowup solutions to (PE) ([11]).*

We think that the assumption (1.2) is almost necessary condition in two dimensional case. Because, if Ω is a bounded disk of $\mathbf{R}^2$ and $\inf_{v>0}\chi'(v) > 0$, we can find blowup solutions to (PE) by using an argument similar to the one in [9].

In the case of $n \ge 3$, the conditions for global existence of solutions and existence of blowup solutions are not critical. Because, in our conjecture, the critical number is $n/(n-2)$.

Next, we consider the limiting system of (PP) as $\tau = 0$. For simplicity, we assume $\eta = 1$.

$$(EP)\begin{cases} 0 = \nabla\cdot(\nabla u - u\nabla\chi(v)) & \text{in } \Omega\times(0,T),\\ v_t = \Delta v - v + u & \text{in } \Omega\times(0,T),\\ \dfrac{\partial u}{\partial\nu} = \dfrac{\partial v}{\partial\nu} = 0 & \text{on } \partial\Omega\times(0,T),\\ v(\cdot,0) = v_0 & \text{in } \Omega,\\ \displaystyle\int_\Omega u(x,t)dx = \lambda & \text{in } (0,T), \end{cases}$$

where λ is a given positive constant. The last condition means the conservation of mass. Since solutions to the original system (PP) satisfy (1.1), then we impose this property also for solutions to (EP).

This system (EP) can be transformed into a non-local parabolic equation. In fact, the first equation and the boundary condition of (EP) guarantee that

$$\log u = \chi(v) + C,$$

where C is a constant. This and the last condition of (EP) ensure that

$$u = \frac{\lambda \exp(\chi(v))}{\int_\Omega \exp(\chi(v))dx}.$$

Therefore, the system (EP) is equivalent to the following system,

$$(NLP)\begin{cases} v_t = \Delta v - v + \dfrac{\lambda \exp(\chi(v))}{\int_\Omega \exp(\chi(v))dx} & \text{in } \Omega \times (0,T), \\ u = \dfrac{\lambda \exp(\chi(v))}{\int_\Omega \exp(\chi(v))dx} & \text{in } \Omega \times (0,T), \\ \dfrac{\partial v}{\partial \nu} = 0 & \text{on } \partial\Omega \times (0,T), \\ v(\cdot,0) = v_0 & \text{in } \Omega. \end{cases}$$

Here, λ is a given positive constant.

Classical solutions to (NLP) satisfy the following properties.

THEOREM 3.2 ([12]). *Suppose that Ω is a bounded domain of $\mathbf{R}^n$ ($n \geq 2$) with smooth boundary and that χ satisfies (1.2). Then, the following hold:*

(i) *If $n = 2$, solutions to (NLP) exist globally in time and are bounded.*

(ii) *If $n \geq 3$ and $\limsup_{v\to\infty} v\chi'(v) < n/(n-2)$, solutions to (NLP) exist globally in time and are bounded.*

(iii) *If $n \geq 3$, Ω is a bounded ball of $\mathbf{R}^n$, $\chi(v) = a \log v$ and $a > n/(n-2)$, there are blowup solutions to (NLP).*

In two dimensional case, (1.2) is the sufficient condition for the global existence of solutions to (PE) and (NLP). We expect that (1.2) is also the sufficient condition for (PP). In the case of $n \geq 3$, the threshold number $n/(n-2)$ in Theorem 3.2 is same as the one in our conjecture. Then, we think that this result is an evidence for our conjecture.

4. Our results. Considering results on the limiting systems mentioned in the previous section, we consider also almost limiting systems which are the systems (PP) in the case where τ or η is sufficient small.

In two dimensional case, classical solutions to those almost systems satisfy the following properties.

THEOREM 4.1 ([3, 4]). *Suppose that Ω is a bounded domain of $\mathbf{R}^2$ with smooth boundary and that $\lim_{v\to\infty} \chi'(v) = 0$. Then, the following hold:*

(i) *If Ω is a bounded disk, (u_0, v_0) is radial and η is sufficiently small, then solutions to (PP) exist globally in time and are bounded.*

(ii) *If Ω is convex and τ is sufficiently small, then solutions to (PP) exist globally in time and are bounded.*

REMARK 4.2. *If our conjecture is correct, the smallness of constants η and τ and the symmetry of (u_0, v_0) are not necessary in two dimensional case.*

In high dimensional case, classical solutions to the almost limiting system satisfy the following property.

THEOREM 4.3 ([4]). *If $n \geq 3$, Ω is a bounded and convex domain of $\mathbf{R}^n$, τ is sufficiently small and $\limsup_{v\to\infty} v\chi'(v) < n/(n-2)$, then solutions to (PP) exist globally in time and are bounded.*

REMARK 4.4. *If our conjecture is correct, we expect that the smallness of τ and the convexity of Ω are not necessary. Moreover, the research on blowup solutions is necessary.*

5. Idea of proof of Theorem 4.3. In this section, we describe the idea of the proof of Theorem 4.3. For simplicity, we assume $\eta = 1$.

LEMMA 5.1. *There exist positive constants $T_{min} > 0$ and $L > 0$ satisfying*

$$\|(u, v)\|_{C([0,T_{min}]\times\overline{\Omega})} \le L \quad \text{for } \tau \in (0, 1].$$

LEMMA 5.2. *There exists a positive constant v_{min} satisfying*

$$v \ge v_{min} \quad \text{in } \Omega \times [0, T_{max}) \quad \text{for } \tau \in (0, 1].$$

Lemma 5.1 comes from the standard energy argument and Lemma 5.2 comes from $\min_{\overline{\Omega}} v_0 > 0$ and $\|u\|_{L^1(\Omega)} > 0$.

Let $z = \dfrac{\exp(\chi(v))}{\int_\Omega \exp(\chi(v))dx}$ and $w = \dfrac{u}{z}$. Those functions satisfy the following system,

$$(TPP)\begin{cases} \dfrac{\partial v}{\partial t} = \Delta v - v + w\dfrac{\exp(\chi(v))}{\int_\Omega \exp(\chi(v))\,dx} & \text{in } \Omega\times(0,T), \\ \tau\dfrac{\partial w}{\partial t} = \dfrac{1}{z}\nabla\cdot(z\nabla w) - \dfrac{\tau}{z}\dfrac{\partial z}{\partial t}w & \text{in } \Omega\times(0,T), \\ \dfrac{\partial v}{\partial \nu} = \dfrac{\partial w}{\partial \nu} = 0 & \text{on } \partial\Omega\times(0,T), \\ v(\cdot,0) = v_0, \quad w(\cdot,0) = \displaystyle\int_\Omega \exp(\chi(v_0))dx\frac{u_0}{\exp(\chi(v_0))} & \text{in } \Omega. \end{cases}$$

Let $H = 2\max(\|u_0\|_{L^1(\Omega)}, \|w(0)\|_{L^\infty(\Omega)}, L)$ and let

$$S(\tau) = \sup\{T > 0;\ \sup_{0<t<T} \|w(t)\|_{L^\infty(\Omega)} \le H\},$$

where L is the constant in Lemma 5.1.

LEMMA 5.3. *There exists a constant $\theta \in (0, 1)$ such that*

$$\|v\|_{C^{2+\theta,(2+\theta)/2}(\Omega\times[0,S(\tau)])} < C(H),$$

where here and henceforth we will denote by $C(H)$ a positive generic constant (possibly changing from line to line) depending on H.

Proof. For $q > n/2$, $n/q < 2\beta < 2$, the semi-group property of the Laplacian guarantees that

$$\begin{aligned} \|v(t)\|_{L^\infty(\Omega)} &\le \|v_0\|_{L^\infty(\Omega)} + \int_0^t \|e^{(t-s)(\Delta-1)}w(s)z(s)\|_{L^\infty(\Omega)}ds \\ &\le \|v_0\|_{L^\infty(\Omega)} + C\int_0^t \frac{e^{s-t}}{(t-s)^\beta}\|w(s)\|_{L^\infty(\Omega)}\|z(s)\|_{L^q(\Omega)}ds. \end{aligned}$$

Here and henceforth, we will denote by C a positive generic constant (possibly changing from line to line). We see that

$$\|z(t)\|_{L^q(\Omega)} = \frac{\|\exp(\chi(v)))\|_{L^q(\Omega)}}{\|\exp(\chi(v)))\|_{L^1(\Omega)}}$$

$$\le \frac{\|\exp(\chi(v))\|_{L^1(\Omega)}^{1/q}\|\exp(\chi(v))\|_{L^\infty(\Omega)}^{(q-1)/q}}{\|\exp(\chi(v))\|_{L^1(\Omega)}}$$
$$\le C(H)\frac{\|(v+1)^\mu\|_{L^\infty(\Omega)}^{(q-1)/q}}{(|\Omega|\exp(\chi(v_{min})))^{(q-1)/q}}.$$

Since $\limsup_{v\to\infty} v\chi'(v) < \mu < n/(n-2)$, we can take q and β such that

$$q > n/2, \quad n/q < 2\beta < 1 \quad \text{and} \quad \mu\frac{q-1}{q} < 1.$$

Then, we have that $\|v\|_{L^\infty(\Omega\times[0,S(\tau)])} \le C(H)$. We obtain this lemma from this estimate and the parabolic regularity argument. □

By those and the parabolic regularity argument, we get a unique classical solution (v,w) to (TPP) in $\Omega\times[0,S(\tau)]$.

We will show that $S(\tau)=\infty$ if τ is sufficiently small. Assume to the contrary that $S(\tau)<\infty$ for $\tau\in(0,1]$. For an integer $J\ge 2$ and $j=0,1,2,\cdots,J$, put $T=S(\tau)/J$ and $z_j=z(jT)$. Then, for $j=0,1,2,\cdots,J-1$ and $t\in(jT,(j+1)T]$ we have

$$\tau w_t = \frac{1}{z_j}\nabla\cdot z_j\nabla w + \nabla\log\frac{z}{z_j}\cdot\nabla w - \tau\frac{z_t}{z}w \quad \text{in } \Omega.$$

For $j=0,1,2,\cdots,J-1$, put $\zeta=(t-jT)/\tau$, $W(x,\zeta)=w(x,t)$, $Z(x,\zeta)=z(x,t)$, $Z_0(x)=z_j(x)$ and $Q(x,\zeta)=z_t(x,t)/z(x,t)$. Then, those functions satisfy that

$$\frac{\partial W}{\partial\zeta} = \frac{1}{Z_0}\nabla\cdot Z_0\nabla W + \nabla\log\frac{Z}{Z_0}\cdot\nabla W - \tau QW \quad \text{in } \Omega\times(0,T/\tau).$$

Put $\mathcal{A}=Z_0^{-1}\nabla\cdot Z_0\nabla$ in Ω with $\partial\cdot/\partial\nu=0$ on $\partial\Omega$. The function W satisfies that

$$W(\zeta) = e^{\zeta\mathcal{A}}W(0) + \int_0^\zeta e^{(\zeta-\xi)\mathcal{A}}F(\xi)d\xi \quad \text{for } \zeta\in(0,T/\tau),$$

where

$$F = \nabla\log\frac{Z}{Z_0}\cdot\nabla W - \tau QW.$$

There exists a positive constant Λ depending on $\inf_\Omega Z_0$, $\|Z_0\|_\infty$ and Ω such that

$$\|\nabla e^{\zeta\mathcal{A}}W(0)\|_{L^q(\Omega)} \le Ce^{-\zeta\Lambda}\|\nabla W(0)\|_{L^q(\Omega)} \quad \text{for } \zeta\in(0,T/\tau),$$
$$\int_0^\zeta \|\nabla e^{(\zeta-\xi)\mathcal{A}}\nabla\log\frac{Z(\xi)}{Z_0}\cdot\nabla W(\xi)\|_{L^q(\Omega)}d\xi \le C(H)T^{\theta/2}e^{-\zeta\Lambda}\sup_{\xi\in[0,\zeta]}e^{\xi\Lambda}\|\nabla W(\xi)\|_{L^q(\Omega)}$$
$$\text{for } \zeta\in(0,T/\tau),$$
$$\int_0^\zeta \|\nabla e^{(\zeta-\xi)\mathcal{A}}\tau Q(\xi)W(\xi)\|_{L^q(\Omega)}d\xi \le C(H)\tau^{(q-1)/q} \quad \text{for } \zeta\in(0,T/\tau)$$

and that

$$\sup_{\xi\in[0,T/\tau]} e^{\xi\Lambda}\|\nabla W(\xi)\|_{L^q(\Omega)}$$
$$\le C\|\nabla W(0)\|_{L^q(\Omega)} + C(H)T^{\theta/2}\sup_{\xi\in[0,T/\tau]}e^{\xi\Lambda}\|\nabla W(\xi)\|_{L^q(\Omega)} + C(H)\tau^{(q-1)/q}e^{T\Lambda/\tau}.$$

Here, θ is the constant in Lemma 5.3. Taking $0 < \tau \ll T \ll 1$, we have that

$$\|\nabla w((j+1)T)\|_{L^q(\Omega)} \le Ce^{-T\Lambda/\tau}\|\nabla w(jT)\|_{L^q(\Omega)} + C(H)\tau^{(q-1)/q} \quad \text{for } j = 0, 1, 2, \cdots, J-1.$$

and that

$$\|\nabla w(jT)\|_{L^q(\Omega)} \le Ce^{-jT\Lambda/\tau}\|\nabla w(jT)\|_{L^q(\Omega)} + C(H)\tau^{(q-1)/q} \quad \text{for } j = 1, 2, 3, \cdots, J.$$

Those estimates guarantee that

$$\|\nabla w(t-jT)\|_{L^q(\Omega)} \le Ce^{-(t-jT)\Lambda/\tau}\|\nabla w(jT)\|_{L^q(\Omega)} + C(H)\tau^{(q-1)/q}$$

for $t \in [jT, (j+1)T]$. Take $x(t) \in \overline{\Omega}$ such that $w(x(t), t) = \|w(t)\|_{L^\infty(\Omega)}$. We have that

$$\begin{aligned}\lambda &= \int_\Omega u(t)dx = \int_\Omega w(t)z(t)dx \\ &\ge \int_\Omega w(x(t), t)z(t)dx - \mathrm{diam}(\Omega)\int_\Omega \frac{|w(x,t) - w(x(t),t)|}{|x - x(t)|}z(t)dx,\end{aligned}$$

where $\mathrm{diam}(\Omega) = \sup\{|x-y|; x, y \in \Omega\}$. Then, we obtain that

$$\|w(t)\|_{L^\infty(\Omega)} \le \lambda + C(\Omega, H, q)\|\nabla w(t)\|_{L^q(\Omega)} < H \quad \text{for } t \in [0, S(\tau)],$$

if τ is sufficiently small. This means that $S(\tau) = \infty$, a contradiction. Then, we have that $S(\tau) = \infty$ if τ is sufficiently small. Therefore, we get Theorem 4.3.

REFERENCES

[1] K. Fujie, Boundedness in a fully parabolic chemotaxis system with singular sensitivity, J. Math. Anal. Appl., 424 (2015), pp. 675–684.

[2] K. Fujie and T. Senba, Global existence and boundedness in a parabolic-elliptic Keller-Segel system with general sensitivity, Discrete Contin. Dyn. Syst. Ser. B, 21 (2016), pp. 81–102.

[3] K. Fujie and Senba T., Global existence and boundedness of radial solutions to a two dimensional fully parabolic chemotaxis system with general sensitivity, Nonlinearity, 29 (2016), pp. 2417–2450.

[4] K. Fujie and T. Senba, A sufficient condition of sensitivity functions for boundedness of solutions to a parabolic-parabolic chemotaxis system, Preprint.

[5] K. Fujie and T. Yokota, Boundedness in a fully parabolic chemotaxis system with strongly singular sensitivity, Appl. Math. Lett, 38 (2014), pp. 140–143.

[6] D. Horstmann and G. Wang, Blow-up in a chemotaxis model without symmetry assumptions, European J. Appl. Math., 12 (2001), pp. 159–177.

[7] N. Mizoguchi and M. Winkler, Is finite-time blow-up a generic phenomenon in the two-dimensional Keller-Segel system ?, Preprint.

[8] X. Mora, Semilinear parabolic problems define semiflows on C^k spaces, Trans. Amer. Math. Soc, 278 (1983), pp. 21–55.

[9] T. Nagai, Blow-up of radially symmetric solutions to a chemotaxis system, Adv.Math.Sci. Appl. 5 (1995), pp. 581–601.

[10] T. Nagai, T. Senba and K. Yoshida, Application of the Trudinger-Moser inequality to a parabolic system of chemotaxis, Funkcial. Ekvac., 40 (1997), pp. 411-433.

[11] T. Nagai and T. Senba, Global existence and blow-up of radial solutions to a parabolic-elliptic system of chemotaxis, Adv. Math. Sci. Appl, 8 (1998), PP. 145–156.

[12] P. Quittner and P. Souplet, Superlinear parabolic problems, Birkhäuser advanced text Basler Lehrbücher. Birkhäuser, Berlin, 2007.

[13] C. Stinner and M. Winkler, Global weak solutions in a chemotaxis system with large singular sensitivity, Nonlinear Analysis: Real World Applications, 12 (2011), pp. 3727–3740.

[14] M. Winkler, Aggregation vs. global diffusive behavior in the higher-dimensional Keller-Segel model, J. Differential Equations, 248 (2010), pp. 2889–2905.
[15] W. Winkler, Global solutions in a fully parabolic chemotaxis system with singular sensitivity, Math. Methods Appl. Sci., 34 (2011), pp. 176–190.
[16] M. Winkler, Finite-time blow-up in the higher-dimensional parabolic-parabolic Keller-Segel system, J. Math. Pures Appl., 100 (2013), pp. 748–767.

Proceedings of EQUADIFF 2017
pp. 53–60

VIRAL INFECTION MODEL WITH DIFFUSION AND STATE-DEPENDENT DELAY: A CASE OF LOGISTIC GROWTH

ALEXANDER V. REZOUNENKO*

Abstract. We propose a virus dynamics model with reaction-diffusion and logistic growth terms, intracellular state-dependent delay and a general non-linear infection rate functional response. Classical solutions with Lipschitz in-time initial functions are investigated. This type of solutions is adequate to the discontinuous change of parameters due to, for example, drug administration. The Lyapunov functions approach is used to analyse stability of interior infection equilibria which describe the cases of a chronic disease.

Key words. Reaction-diffusion, evolution equations, Lyapunov stability, state-dependent delay, virus infection model.

AMS subject classifications. 93C23, 34K20,35K57, 97M60

1. Introduction. Our goal is to discuss a wide class of mathematical models of viral diseases. Many viruses (as Ebola virus, Zika virus, HIV, HBV, HCV and others) continue to be a major global public health issues, according to World Health Organization. Particularly, from The Global hepatitis report (WHO, April 2017) [25] we know that "a large number of people - about 325 million worldwide in 2015 - are carriers of hepatitis B or C virus infections, which can remain asymptomatic for decades." and "Viral hepatitis caused 1.34 million deaths in 2015, a number comparable to deaths caused by tuberculosis and higher than those caused by HIV. However, the number of deaths due to viral hepatitis is increasing over time, while mortality caused by tuberculosis and HIV is declining."

In such a situation any steps toward understanding the dynamics of viral diseases are important.

There are variety of models described by systems of ordinary differential equations and/or partial differential equations with or without delays which describe dynamics of different viral infections. Delays could be bounded or unbounded, concentrated or distributed, constant, time-dependent or state-dependent.

The classical models [12, 14] contain ordinary differential equations (without delay) for three variables: susceptible host cells T, infected host cells T^* and free virus particles V. The intracellular delay is an important property of the biological problem, so we start with the delay problem

$$\begin{cases} \dot{T}(t) = \lambda - dT(t) - f(T(t), V(t)), \\ \dot{T}^*(t) = e^{-\omega h} f(T(t-h), V(t-h)) - \delta T^*(t), \\ \dot{V}(t) = N\delta T^*(t) - cV(t). \end{cases} \tag{1.1}$$

In system (1.1), susceptible cells T are produced at a rate λ, die at rate dT, and become infected at rate $f(T, V)$. Properties and examples of incidence function f are discussed below. Infected cells T^* die at rate δT^*, free virions V are produced by infected cells at rate $N\delta T^*$ and are removed at rate $cV(t)$. In (1.1) h denotes the delay

*V.N.Karazin Kharkiv National University, Kharkiv, 61022, Ukraine (rezounenko@gmail.com) *and* Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, P.O. Box 18, 182 08 Praha, CR

between the time a virus particle contacts a target cell and the time the cell becomes actively infected (start producing new *free* virions). It is clear that the constancy of the delay is just an extra assumption which essentially simplifies the study, but has no biological background.

To the best of our knowledge, viral infection models with state-dependent delay (SDD) have been considered for the first time in [20] (see also [21]). It is well known that differential equations with discrete state-dependent delay are always non-linear by its nature (see the review [5] for more details and discussion).

As usual in the study of delay systems with (maximal) delay $h > 0$ [4, 8], for a function $v(t), t \in [a-h, b] \subset \mathbf{R}, b > a$, we denote the history segment $v_t = v_t(\theta) \equiv v(t+\theta), \theta \in [-h, 0], t \in [a, b]$.

Consider a connected bounded domain $\Omega \subset \mathbf{R^n}$ with a smooth boundary $\partial\Omega$. Let $T(t,x), T^*(t,x), V(t,x)$ represent the densities of uninfected cells, infected cells and free virions at position $x \subset \Omega$ at time t.

In [22] the following system with SDD η is investigated

$$\begin{cases} \dot{T}(t,x) = \lambda - dT(t,x) - f(T(t,x), V(t,x)) + d^1 \Delta T(t,x), \\ \dot{T}^*(t,x) = e^{-\omega h} f(T(t-\eta(u_t),x), V(t-\eta(u_t),x)) - \delta T^*(t,x) + d^2 \Delta T^*(t,x), \\ \dot{V}(t,x) = N\delta T^*(t,x) - cV(t,x) + d^3 \Delta V(t,x). \end{cases} \tag{1.2}$$

Here the dot over a function denotes the partial time derivative i.g, $\dot{T}(t,x) = \frac{\partial T(t,x)}{\partial t}$, all the constants $\lambda, d, \delta, N, c, \omega$ are positive while $d^i, i = 1, 2, 3$ (diffusion coefficients) are non negative. In (1.2) (and in (1.3) below), a solution denoted by $u(t) = u(t,\cdot) = (T(t,\cdot), T^*(t,\cdot), V(t,\cdot))$, see the argument of the state-dependent delay η in the second equation. The precise definition of a solution is given below (Def. 2.1).

We consider a general functional response $f(T, V)$ satisfying natural assumptions presented below. In earlier models (with constant or without delay) the study was started in case of bilinear $f(T, V) = \text{const} \cdot TV$ and then extended to more general classes of non-linearities. For more details and discussion see [1, 3, 7, 11, 22].

We mention that the term $e^{-\omega h}$ in front of f (see the second equation (1.2)), in fact, states that only *a part* of the cell population survived during the virus incubation period. Clearly, it should be less than 1. It is an assumption which is not too precise in *nonlinear systems*. It could be regarded as a coefficient (strictly smaller than 1) and could be easily incorporated into the definition of the function f. We keep this coefficient in the form of $e^{-\omega h}$ for the only reason to simplify for the reader the comparison of computations with the constant delay case.

In this note we are interested in the following PDEs system with state-dependent delay η

$$\begin{cases} \dot{T}(t,x) = rT(t,x)\left(1 - \frac{T(t,x)}{T_K}\right) - dT(t,x) - f(T(t,x), V(t,x)) + d^1 \Delta T(t,x), \\ \dot{T}^*(t,x) = e^{-\omega h} f(T(t-\eta(u_t),x), V(t-\eta(u_t),x)) - \delta T^*(t,x) + d^2 \Delta T^*(t,x), \\ \dot{V}(t,x) = N\delta T^*(t,x) - cV(t,x) + d^3 \Delta V(t,x). \end{cases} \tag{1.3}$$

Let us discuss the principal difference in the first equations of (1.2) and (1.3). In system (1.2), uninfected target cells T are produced by the body at a constant rate λ which is relevant, for example, in case of HIV. In contrast, the first term in the first equation of (1.3) is the classical logistic growth term (Pierre Verhulst term) for the population of uninfected cells T. The constant T_K is the so-called carrying capacity for the population T, which has the clear biological meaning. System (1.3) is more

relevant in case of chronic infections with viruses such as, for example, hepatitis B (HBV) and hepatitis C (HCV). Here $T(t,x)$ and $T^*(t,x)$ represent uninfected and infected liver cells (hepatocytes). The carrying capacity could be also considered for the sum of uninfected and infected cells (c.f. [6]), but we decide to use it for uninfected hepatocytes (liver cells) only for the following biological reason. It is well-known that the development of HBV, HCV infections is usually connected with development of fibrosis. The last indicates that the regeneration of healthy hepatocytes is not quick enough to fill all the available (free) space in liver. This available space appears as a result of natural death of both uninfected and infected hepatocytes as well as killing of infected cells by immune system. The above suggests that the presence of infected cells does not make essential restriction on the regeneration of healthy hepatocytes T.

Boundary conditions are of Neumann type for the corresponding unknown if $d^i \neq 0$ i.e. $\frac{\partial T(t,x)}{\partial n}|_{\partial\Omega} = 0$ if $d^1 \neq 0$ and similarly for $T^*(t,x)$ and $V(t,x)$. Here $\frac{\partial}{\partial n}$ is the outward normal derivative on $\partial\Omega$. In case $d^i = 0$, no boundary conditions are needed for the corresponding unknown(s). For more discussion see [22].

Our main goals are to present the existence and uniqueness results for the model (1.3) in the sense of classical solutions, and to study the local asymptotic stability of non-trivial disease equilibria. We apply the Lyapunov approach [9] to the state-dependent delay PDE model (1.3) and allow, but not require, diffusion terms in each state equation. For the Lyapunov approach in context of viral infection models (with constant delay or nondelay cases) see e.g. works by A.Korobeinikov, C.McCluskey [7, 11] and references therein. Our main interest is in discussion of the state-dependent delay.

2. Main results. We use the basic functional framework described in [10] and applied to the system (1.2) in [22].

Define the following linear operator $-\mathcal{A}^0 = diag\,(d^1\Delta, d^2\Delta, d^3\Delta)$ in $C(\overline{\Omega};\mathbf{R^3})$ with $D(\mathcal{A}^0) \equiv D(d^1\Delta) \times D(d^2\Delta) \times D(d^3\Delta)$. Here, for $d^i \neq 0$ we set $D(d^i\Delta) \equiv \{v \in C^2(\overline{\Omega}) : \frac{\partial v(x)}{\partial n}|_{\partial\Omega} = 0\}$ and $D(d^j\Delta) \equiv C(\overline{\Omega})$ for $d^j = 0$. We omit the space coordinate x, for short, for unknown $u(t) = (T(t), T^*(t), V(t)) \in X \equiv [C(\overline{\Omega})]^3 \equiv C(\overline{\Omega};\mathbf{R^3})$. It is well-known that the closure $-\mathcal{A}$ (in X) of the operator $-\mathcal{A}^0$ generates a C_0-semigroup $e^{-\mathcal{A}t}$ on X which is analytic and nonexpansive [10, p.5]. We denote the space of continuous functions by $C \equiv C([-h,0];X)$ equipped with the sup-norm $||\psi||_C \equiv \max_{\theta\in[-h,0]} ||\psi(\theta)||_X$.

We write, the system (1.3) in the following abstract form

$$\frac{d}{dt}u(t) + \mathcal{A}u(t) = F(u_t), \qquad t > 0. \tag{2.1}$$

The non-linear continuous mapping $F : C \to X$ is defined by

$$F(\varphi)(x) = \begin{pmatrix} r\,\varphi^1(0,x)\left(1 - \frac{\varphi^1(0,x)}{T_K}\right) - d\varphi^1(0,x) - f(\varphi^1(0,x), \varphi^3(0,x)) \\ e^{-\omega h} f(\varphi^1(-\eta(\varphi),x), \varphi^3(-\eta(\varphi),x)) - \delta\varphi^2(0,x) \\ N\delta\varphi^2(0,x) - c\varphi^3(0,x) \end{pmatrix}. \tag{2.2}$$

Here $\varphi = (\varphi^1, \varphi^2, \varphi^3) \in C$. Mapping F is *not* Lipschitz on the space C which is typical for a mapping which includes discrete state-dependent delays (see review [5] for ODE case and works [15, 16, 17, 2] for PDEs).

We need initial conditions $u(\theta,x) = \varphi(\theta,x) = (T(\theta,x), T^*(\theta,x), V(\theta,x)), \theta \in [-h,0]$ for the delay problem (2.1) (c.f. (1.3)):

$$\varphi \in Lip([-h,0];X) \equiv \left\{ \psi \in C : \sup_{s\neq t} \frac{||\psi(s)-\psi(t)||_X}{|s-t|} < \infty \right\}, \quad \varphi(0) \in D(\mathcal{A}). \tag{2.3}$$

In our study we use the standard (c.f. [13, Def. 2.3, p.106] and [13, Def. 2.1, p.105])

DEFINITION 2.1. *A function $u \in C([-h,T];X)$ is called a* `mild solution` *on $[-h,T)$ of the initial value problem (2.1), (2.3) if it satisfies (2.3) and $u(t) = e^{-\mathcal{A}t}\varphi(0) + \int_0^t e^{-\mathcal{A}(t-s)}F(u_s)\,ds, \quad t \in [0,T)$.*

A function $u \in C([-h,T);X) \bigcap C^1((0,T);X)$ is called a `classical solution` *on $[-h,T)$ of the initial value problem (2.1), (2.3) if it satisfies (2.3), $u(t) \in D(A)$ for $0 < t < T$ and (2.1) is satisfied on $(0,T)$.*

Assume the non-linear function $f : \mathbf{R^2} \to \mathbf{R}$ is Lipschitz continuous and satisfies

$$(\mathbf{Hf_1}) \quad \text{there exists } \mu > 0 \text{ such that } |f(T,V)| \le \mu|T| \text{ for all } T,V \in \mathbf{R}. \tag{2.4}$$

We have the following result

THEOREM 2.2. *Let nonlinear function f be Lipschitz and satisfy $(\mathbf{Hf_1})$ (see (2.4)), state-dependent delay $\eta : C \to [0,h]$ is locally Lipschitz. Then the initial value problem (2.1), (2.3) has a unique classical solution which is global in time i.e. defined for all $t \ge 0$.*

Proof of Theorem 2.2 follows the line of the proof of [22, Proposition 1].

Define the set (c.f. (2.3)), which is different from the one Ω_{Lip} in [22]:

$$\Omega_{Lip}^{log} \equiv \big\{ \varphi = (\varphi^1,\varphi^2,\varphi^3) \in Lip([-h,0];X)) \subset C,\ \varphi(0) \in D(\mathcal{A}) :$$

$$0 \le \varphi^1(\theta) \le \left(1-\frac{d}{r}\right) T_K, 0 \le \varphi^2(\theta) \le \frac{\mu}{\delta}\left(1-\frac{d}{r}\right) T_K e^{-\omega h},$$

$$0 \le \varphi^3(\theta) \le \frac{N\mu}{c}\left(1-\frac{d}{r}\right) T_K e^{-\omega h}, \quad \theta \in [-h,0] \Big\}, \tag{2.5}$$

where μ is defined in $(\mathbf{Hf_1})$ and all the inequalities hold pointwise w.r.t. $x \in \overline{\Omega}$.

We need further assumptions (which include $(\mathbf{Hf_1})$) on Lipschitz function f :

$$(\mathbf{Hf_1}+) \begin{cases} f(T,0) = f(0,V) = 0, \quad \text{and} \quad f(T,V) > 0 \text{ for all } T > 0, V > 0; \\ f \text{ is strictly increasing in both coordinates for all } T > 0, V > 0; \\ \text{there exists } \mu > 0 \text{ such that } |f(T,V)| \le \mu|T| \text{ for all } T,V \in \mathbf{R}. \end{cases} \tag{2.6}$$

We have the following result

THEOREM 2.3. *Let non-linear Lipschitz function f satisfy $(\mathbf{Hf_1}+)$ (see (2.6)), state-dependent delay $\eta : C \to [0,h]$ is locally Lipschitz. Then Ω_{Lip}^{log} is invariant i.e. for any $\varphi \in \Omega_{Lip}^{log}$ the unique solution to problem (2.1), (2.3) satisfies $u_t \in \Omega_{Lip}^{log}$ for all $t \ge 0$.*

Proof of Theorem 2.3. The existence and uniqueness of solution is proven in theorem 2.2. The proof of the invariance part follows the invariance result of [10] with the use of the *almost Lipschitz property* of nonlinearity F. The estimates (for the subtangential condition) are the same as for the constant delay case, see e.g. [11, Theorem 2.2]. We do not repeat it here. It is important to notice that the solutions are classic for all $t \ge 0$ (but not for $t \ge h$ as could be in the case of merely continuous initial functions $\varphi \in C$). For more details see, e.g. [22]. The proof of Theorem 2.3 is complete.

2.1. Stationary solutions. Let us discuss stationary solutions of (1.3). By such solutions we mean time independent $\widehat{u}$ which, in general, may depend on $x \in \overline{\Omega}$. Consider the system (1.3) with $u(t) = u(t - \eta(u_t)) = \widehat{u}$ and denote the coordinates (a possible triple of coordinates) of a stationary solution by $(\widehat{T}, \widehat{T^*}, \widehat{V}) = \widehat{u} \equiv \widehat{\varphi}(\theta)$, $\theta \in [-h, 0]$. Since stationary solutions of (1.3) do not depend on the type of delay (state-dependent or constant) we have

$$\begin{cases} 0 = r\widehat{T}\left(1 - \frac{\widehat{T}}{T_K}\right) - d\widehat{T} - f(\widehat{T}, \widehat{V}), \qquad 0 = e^{-\omega h} f(\widehat{T}, \widehat{V}) - \delta\widehat{T^*}, \\ 0 = N\delta\widehat{T^*} - c\widehat{V}. \end{cases} \tag{2.7}$$

Equations hold pointwise w.r.t. $x \in \overline{\Omega}$.

It is easy to see that the trivial stationary solution $(\left(1 - \frac{d}{r}\right) T_K, 0, 0)$ always exists. We are interested in nontrivial disease stationary solutions of (1.3). We have from the first and second equations of (2.7) $\widehat{T^*} = \frac{r}{\delta} e^{-\omega h} \cdot \widehat{T}\left(1 - \frac{\widehat{T}}{T_K}\right) - \frac{d}{\delta} e^{-\omega h}\widehat{T}$ and from the third equation $\widehat{V} = \frac{N\delta}{c}\widehat{T^*}$. It gives the condition on the coordinate $\widehat{T}$ which should belong to $(0, \left(1 - \frac{d}{r}\right) T_K]$. Denote (c.f. [11, 20])

$$h_f^{log}(s) \equiv f\left(s\,, \frac{Nr}{c} e^{-\omega h} \cdot s\left(1 - \frac{s}{T_K}\right) - \frac{Nd}{c} e^{-\omega h} \cdot s\right)$$

$$-r \cdot s\left(1 - \frac{s}{T_K}\right) + d \cdot s. \tag{2.8}$$

Assume f satisfies

($\mathbf{Hf_2^{log}}$) $h_f^{log}(s) = 0$ has at least one and at most a finite number of roots on $(0, \left(1 - \frac{d}{r}\right) T_K]$.

We denote an arbitrary root of $h_f^{log}(s) = 0$ by $\widehat{T}$ and define the corresponding $\widehat{T^*} \equiv \frac{r}{\delta} e^{-\omega h} \cdot \widehat{T}\left(1 - \frac{\widehat{T}}{T_K}\right) - \frac{d}{\delta} e^{-\omega h}\widehat{T}$ and $\widehat{V} \equiv \frac{N\delta}{c}\widehat{T^*} = \frac{Nr}{c} e^{-\omega h} \cdot \widehat{T}\left(1 - \frac{\widehat{T}}{T_K}\right) - \frac{Nd}{c} e^{-\omega h}\widehat{T}$. The point $(\widehat{T}, \widehat{T^*}, \widehat{V})$ satisfies (2.7), so it is a a disease stationary solution of (1.3). We notice that in [22] the corresponding equation was written for coordinate $\widehat{T^*}$, while (2.8) is designed for $s = \widehat{T}$.

Remark (c.f. [22]). *We notice that the finiteness of roots (which are obviously isolated) does not allow the existence of equilibria which depend on spatial coordinate $x \in \Omega$. We remind that Ω is a connected set, so a function $v \in C(\overline{\Omega})$ may take either one or continuum values. Assumption* ($\mathbf{Hf_2^{log}}$) *implies $\widehat{T^*}(x) \equiv \widehat{T^*} \in \mathbf{R}$, so $(\widehat{T}, \widehat{T^*}, \widehat{V})$ is independent of $x \in \overline{\Omega}$.*

Remark. *It is important to mention that usually in study of stability properties of stationary solutions (for viral dynamics problems) one uses conditions on the so-called reproduction numbers. These conditions are used to separate the case of a unique stationary solution. Then the global stability of the equilibrium is investigated. In our study, taking into account the state-dependence of the delay, we discuss the local stability. As a consequence, it allows the co-existence of multiple equilibria. We believe this framework provides a way to model more complicated situations with rich dynamics (in contrast to a globally stable equilibrium). The conditions on the reproduction numbers do not appear explicitly here, but could be seen as particular sufficient conditions for* ($\mathbf{Hf_2^{log}}$).

2.2. Stability of disease stationary solutions. In this section we use the following *local* assumptions on f in a small neighborhood of a disease equilibrium (given by $(\mathbf{Hf_2^{log}})$).

$$(\mathbf{Hf_3}) \qquad \left(\frac{V}{\widehat{V}} - \frac{f(T,V)}{f(T,\widehat{V})}\right)\cdot\left(\frac{f(T,V)}{f(T,\widehat{V})} - 1\right) > 0. \tag{2.9}$$

One can check that the DeAngelis-Beddington functional response [1, 3] of the form $f(T,V) = \frac{kTV}{1+k_1T+k_2V}$, with $k, k_1 \geq 0, k_2 > 0$ satisfies $(\mathbf{Hf_3})$ globally. We also mention that the DeAngelis-Beddington functional response includes as a special case ($k_1 = 0$) the *saturated incidence* rate $f(T,V) = \frac{kTV}{1+k_2V}$.

We will also use the assumption

$(\mathbf{Hf_4})$ Function f is differentiable in a neighborhood of $(\widehat{T},\widehat{V})$.

The main result is the following

THEOREM 2.4. *Let the nonlinear Lipschitz function f satisfy $(\mathbf{Hf_1}+), (\mathbf{Hf_2^{log}})$, $(\mathbf{Hf_3}), (\mathbf{Hf_4})$ (see (2.6), (2.9)), a root $\widehat{T}$ of $h_f^{log}(s) = 0$ (see (2.8) and $(\mathbf{Hf_2^{log}})$) satisfy $\widehat{T} > \frac{1}{2}(1-\frac{d}{r})T_K$. Let state-dependent delay $\eta : C \to [0,h]$ be locally Lipschitz in C and continuously differentiable in a neighbourhood of equilibrium $\widehat{\varphi} \equiv (\widehat{T},\widehat{T^*},\widehat{V})$. Then the stationary solution $\widehat{\varphi}$ is locally asymptotically stable.*

In the proof we use the following Lyapunov functional with *state-dependent delay* along a solution of (1.3)

$$U^{\mathrm{sdd}}(t) \equiv \int_\Omega \left\{ \left(T(t,x) - \widehat{T} - \int_{\widehat{T}}^{T(t,x)} \frac{f(\widehat{T},\widehat{V})}{f(\theta,\widehat{V})}\, d\theta \right) e^{-\omega h} + \widehat{T^*}\cdot v\left(\frac{T^*(t,x)}{\widehat{T^*}}\right) \right.$$

$$\left. + \frac{\widehat{V}}{N}\cdot v\left(\frac{V(t,x)}{\widehat{V}}\right) + \delta\widehat{T^*}\int_{t-\eta(u_t)}^{t} v\left(\frac{f(T(\theta,x),V(\theta,x))}{f(\widehat{T},\widehat{V})}\right) d\theta \right\} dx. \tag{2.10}$$

In (2.10) the Volterra function $v(s) = s - 1 - \ln s : (0,+\infty) \to \mathbf{R}_+$ (c.f. [7, 11]) is used. The form of the functional is standard except the low limit of the last integral in (2.10) which is state-dependent. This state-dependence was first considered in [20] (see also [21]). For PDE (1.2) with constant delay case and $d^1 = d^2 = 0$, see e.g. [11] and for PDE with state-dependent delay (1.2) see [22]. We do not repeat here detailed calculations of the time derivative of $U^{\mathrm{sdd}}(t)$ along a solution of (1.3). They are similar to the ones of [22] and differ in the parts where the connection between coordinates of the stationary solution $\widehat{\varphi} = (\widehat{T},\widehat{T^*},\widehat{V})$ is used. The logistic growth term also makes difference to the study presented in [20, 22].

Acknowledgments. The author is thankful to A.Korobeinikov for useful discussions. This work was supported in part by GA CR under project 16-06678S.

REFERENCES

[1] J. R. Beddington, *Mutual interference between parasites or predators and its effect on searching efficiency*, Journal of Animal Ecology, **44** (1975), 331-340. pp.

[2] I.D. Chueshov and A.V. Rezounenko, *Finite-dimensional global attractors for parabolic nonlinear equations with state-dependent delay*, Communications on Pure and Applied Analysis, 14/5 (2015), pp.1685-1704.

[3] D. L. DeAngelis, R. A. Goldstein and R. V. O'Neill, *A model for tropic interaction*, Ecology, 56 (1975), pp.881–892.

[4] J. K. Hale, *Theory of Functional Differential Equations*, Springer, Berlin- Heidelberg- New York, 1977.

[5] F. Hartung, T. Krisztin, H.-O. Walther and J. Wu, *Functional differential equations with state-dependent delays: Theory and applications*, In: *Canada, A., Drabek., P. and A. Fonda (Eds.) Handbook of Differential Equations, Ordinary Differential Equations,* Elsevier Science B.V., North Holland, 3 (2006), pp.435–545.

[6] S. Hews, S. Eikenberry, J.D. Nagy et al., *Rich dynamics of a hepatitis B viral infection model with logistic hepatocyte growth*, Journal of Mathematical Biology, Volume 60, Issue 4, (2010), pp.573-590.

[7] A. Korobeinikov, *Global properties of infectious disease models with nonlinear incidence*, *Bull. Math. Biol.*, 69 (2007), pp.1871-1886.

[8] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Mathematics in Science and Engineering, 191. Academic Press, Inc., Boston, MA, 1993.

[9] A. M. Lyapunov, *The General Problem of the Stability of Motion*, Kharkov Mathematical Society, Kharkov, 1892, 251p.

[10] R.H. Martin, Jr., H.L. Smith, *Abstract functional-differential equations and reaction-diffusion systems*, Trans. Amer. Math. Soc., 321 (1990), pp.1-44.

[11] C. McCluskey, Yu.Yang, *Global stability of a diffusive virus dynamics model with general incidence function and time delay*, Nonlinear Anal. Real World Appl, 25 (2015), pp.64-78.

[12] M. Nowak and C. Bangham, *Population dynamics of immune response to persistent viruses*, Science, 272 (1996), pp.74-79.

[13] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Applied Mathematical Sciences, 44. Springer-Verlag, New York, 1983. viii+279 pp.

[14] A. Perelson, A. Neumann, M. Markowitz, J. Leonard and D. Ho, *HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time*, Science, 271 (1996), pp.1582-1586.

[15] A. V. Rezounenko, *Partial differential equations with discrete and distributed state-dependent delays*, Journal of Mathematical Analysis and Applications, 326 (2007), pp.1031-1045.

[16] A. V. Rezounenko, *Differential equations with discrete state-dependent delay: Uniqueness and well-posedness in the space of continuous functions*, Nonlinear Analysis: Theory, Methods and Applications, 70 (2009), pp.3978-3986.

[17] A. V. Rezounenko, *Non-linear partial differential equations with discrete state-dependent delays in a metric space*, Nonlinear Analysis: Theory, Methods and Applications, 73 (2010), pp.1707-1714.

[18] A. V. Rezounenko, *A condition on delay for differential equations with discrete state-dependent delay*, Journal of Mathematical Analysis and Applications, 385 (2012), pp.506-516.

[19] A.V. Rezounenko, P. Zagalak, *Non-local PDEs with discrete state-dependent delays: well-posedness in a metric space*, Discrete and Continuous Dynamical Systems - Series A, 33:2 (2013), pp.819-835.

[20] A. V. Rezounenko, *Stability of a viral infection model with state-dependent delay, CTL and antibody immune responses*, Discrete and Continuous Dynamical Systems - Series B, Vol. 22 (2017), pp.1547-1563; Preprint arXiv:1603.06281v1 [math.DS], 20 March 2016, arxiv.org/abs/1603.06281v1.

[21] A. V. Rezounenko, *Continuous solutions to a viral infection model with general incidence rate, discrete state-dependent delay, CTL and antibody immune responses*, Electron. J. Qual. Theory Differ. Equ., 79 (2016), pp.1-15.

[22] A. V. Rezounenko, *Viral infection model with diffusion and state-dependent delay: stability of classical solutions*, Discrete and Continuous Dynamical Systems - Series B, Vol. 23, N. 3, May 2018, to appear; Preprint arXiv:1706.08620 [math.DS], 26 Jun 2017, arxiv.org/abs/1706.08620.

[23] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, Mathematical Surveys and Monographs, 41. American Mathematical

Society, Providence, RI, 1995.
[24] H. Smith, *An Introduction to Delay Differential Equations with Sciences Applications to the Life*, Texts in Applied Mathematics, vol. 57, Springer, New York, Dordrecht, Heidelberg, London, 2011.
[25] World Health Organization, *Global hepatitis report-2017, April 2017,* ISBN: 978-92-4-156545-5; apps.who.int/iris/bitstream/10665/255016/1/9789241565455-eng.pdf?ua=1

Proceedings of EQUADIFF 2017
pp. 61–68

BOUNDEDNESS IN A FULLY PARABOLIC CHEMOTAXIS SYSTEM WITH SIGNAL-DEPENDENT SENSITIVITY AND LOGISTIC TERM*

MASAAKI MIZUKAMI†

Abstract. This paper deals with the chemotaxis system with signal-dependent sensitivity and logistic term

$$u_t = \Delta u - \nabla \cdot (u\chi(v)\nabla v) + \mu u(1-u),$$
$$v_t = \Delta v + u - v$$

in $\Omega \times (0,\infty)$, where Ω is a bounded domain in $\mathbb{R}^n$ ($n \geq 2$) with smooth boundary, $\mu > 0$ is a constant and χ is a function generalizing

$$\chi(s) = \frac{K}{(1+s)^2} \quad (K > 0,\ s > 0).$$

In the case that $\mu = 0$ global existence and boundedness were established under some conditions ([14]); however, conditions for global existence and boundedness in the above system have not been studied. The purpose of this paper is to construct conditions for global existence and boundedness in the above system.

Key words. chemotaxis; signal-dependent sensitivity; logistic term; global existence.

AMS subject classifications. Primary: 35K51; Secondary: 35A01, 92C17.

1. Introduction. Chemotaxis is the property such that species move towards higher concentration of a chemical substance when they plunge into hunger. The following problem which describes the movement of species with chemotaxis

$$u_t = \Delta u - \nabla \cdot (u\chi(v)\nabla v) + \mu u(1-u), \quad v_t = \Delta v + u - v,$$

where χ is a function and $\mu \geq 0$ is a constant, is called a *Keller–Segel system* or a *chemotaxis system*, and is studied intensively. The function χ appearing in the above problem is called *signal-dependent sensitivity*, and examples of this function χ are as follows: $\chi(s) = K$ (constant), $\chi(s) = \frac{K}{s}$ (singular), $\chi(s) = \frac{K}{(1+s)^2}$ (regular) for $s > 0$ with some constant $K > 0$. Previous works which deal with the constant sensitivity can be found in [2, 7, 8, 15, 18, 19]; the singular sensitivity is treated in [3, 5, 6, 9, 10]; we can find works related to the regular sensitivity in [5, 6, 11, 13, 14, 16, 17, 20]; variation of chemotaxis systems are in [1]. Here we focus on the case that χ is a function generalizing the regular sensitivity:

$$\chi(s) \leq \frac{K}{(a+s)^k} \quad (s > 0) \tag{1.1}$$

with some constants $a \geq 0$, $k > 1$ and $K > 0$. In a mathematical view, one of difficulties caused by the sensitivity function χ is to deal with the additional term $u\chi'(v)|\nabla v|^2$ which does not appear in the case that χ is a constant. In the case that

*This work was supported by JSPS Research Fellowships for Young Scientists (No. 17J00101).

†Department of Mathematics, Tokyo University of Science, 1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan (masaaki.mizukami.math@gmail.com).

$\mu = 0$, by using an energy estimate to overcome the difficulties of the sensitivity function, under the condition that χ fulfils (1.1) with some constants $a \geq 0$, $k > 1$ and $K > 0$ satisfying

$$K < k(a+\eta)^{k-1}\sqrt{\frac{2}{n}}, \tag{1.2}$$

where η is a constant defined as

$$\eta := \sup_{\tau>0}\left(\min\left\{e^{-2\tau}\min_{x\in\overline{\Omega}} v_0(x),\ c_0\|u_0\|_{L^1(\Omega)}(1-e^{-\tau})\right\}\right) \geq 0$$

(see [4, 14]), global existence and boundedness were established ([14]). Recently, Fujie–Senba [5, 6] established conditions for global existence and boundedness in a problem generalizing the chemotaxis system with $\mu = 0$. More related works which deal with a two-species chemotaxis system with competitive kinetics can be found in [11, 12, 13, 16, 17, 20]; global existence and boundedness are in [11, 13, 16, 17, 20]; asymptotic behavior is shown in [11, 12].

In summary, the conditions (1.1)–(1.2) lead to global existence and boundedness in the chemotaxis system with $\mu = 0$. However, the case that $\mu > 0$ has not been studied. The purpose of this work is to derive conditions for global existence and boundedness in the chemotaxis system.

In this paper we consider the chemotaxis system with signal-dependent sensitivity and logistic term

$$\begin{cases} u_t = \Delta u - \nabla\cdot(u\chi(v)\nabla v) + \mu u(1-u), & x\in\Omega,\ t>0,\\ v_t = \Delta v + u - v, & x\in\Omega,\ t>0,\\ \nabla u\cdot\nu = \nabla v\cdot\nu = 0, & x\in\partial\Omega,\ t>0,\\ u(x,0) = u_0(x),\ v(x,0) = v_0(x), & x\in\Omega, \end{cases} \tag{1.3}$$

where Ω is a bounded domain in $\mathbb{R}^n$ $(n \geq 2)$ with smooth boundary $\partial\Omega$ and ν is the outward normal vector to $\partial\Omega$; $\mu > 0$ is a constant; the initial data u_0 and v_0 are assumed to be nonnegative functions. The unknown function $u(x,t)$ represents the population density of species and $v(x,t)$ shows the concentration of the substance at place x and time t. As to the sensitivity function χ, we are interested in functions generalizing

$$\chi(s) = \frac{K}{(1+s)^2} \quad (s>0),$$

where $K > 0$ is a constant.

In order to achieve our purpose we shall suppose that χ satisfies that

$$\chi \in C^{1+\lambda}((0,\infty)) \quad \text{and} \quad 0 \leq \chi(s) \leq \frac{K}{(a+s)^k} \quad (s>0) \tag{1.4}$$

with some $\lambda > 0$, $k > 1$, $a > 0$ and $K > 0$ fulfiling

$$K < ka^{k-1}\sqrt{\frac{2}{n}}. \tag{1.5}$$

Now the main result reads as follows.

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^n$ $(n \geq 2)$ be a bounded domain with smooth boundary and let $\mu > 0$. Assume that χ satisfies* (1.4) *with some $\lambda > 0$, $k > 1$, $a > 0$, $K > 0$ fulfiling* (1.5). *Then for any u_0, v_0 satisfying*

$$0 \leq u_0 \in C(\overline{\Omega}) \setminus \{0\} \quad \textit{and} \quad 0 \leq v_0 \in W^{1.q}(\Omega) \setminus \{0\} \tag{1.6}$$

with some $q > n$, there exists an exactly one pair (u, v) of positive functions

$$u,\ v \in C(\overline{\Omega} \times [0, \infty)) \cap C^{2,1}(\overline{\Omega} \times (0, \infty))$$

which solves (1.3). *Moreover, the solution (u, v) is uniformly bounded, i.e., there exists a constant $C > 0$ such that*

$$\|u(\cdot, t)\|_{L^\infty(\Omega)} + \|v(\cdot, t)\|_{W^{1,q}(\Omega)} \leq C$$

for all $t > 0$.

Here we give one remark: The condition (1.5) is more restricted condition than (1.2) except the case that $\eta = 0$ (which is the case that $\min_{x \in \Omega} v_0(x) = 0$). The reason is that it is difficult to see the uniform-in-time lower estimate for v because of lacking information about the lower estimate for u. Moreover, the condition (1.5) is independent of $\mu > 0$: The question "can the logistic term relax conditions for global existence and boundedness?" is still open problem in (1.3).

The strategy for the proof of Theorem 1.1 is to construct the L^p-estimate for u with some $p > \frac{n}{2}$. One of keys for this strategy is to derive the inequality

$$\frac{d}{dt} \int_\Omega u^p \varphi(v) \leq c \int_\Omega u^p \varphi(v) - \mu p \int_\Omega u^{p+1} \varphi(v)$$

for some constant $c > 0$, where

$$\varphi(s) := \exp\left\{ -r \int_0^s \frac{1}{(a + \tau)^k}\, d\tau \right\} \quad (s \geq 0)$$

with some $r > 0$. Thanks to this strategy, we obtain

$$\int_\Omega u^p \varphi(v) \leq C$$

with some $C > 0$, which together with the lower estimate for φ implies the L^p-estimate for u. Thus in light of the well-known semigroup estimates, we can attain the L^∞-estimate for u.

2. Proof of the main result. In this section we will prove Theorem 1.1. We first recall the well-known result about local existence of solutions to (1.3) (see e.g., [1, Lemma 3.1]).

LEMMA 2.1. *Assume that χ satisfies* (1.4) *with some $\lambda > 0$, $k > 1$, $a > 0$, $K > 0$ and the initial data u_0, v_0 fulfil* (1.6) *for some $q > n$. Then there exist $T_{\max} \in (0, \infty]$ and exactly one pair (u, v) of positive functions*

$$u \in C(\overline{\Omega} \times [0, T_{\max})) \cap C^{2,1}(\overline{\Omega} \times (0, T_{\max})),$$
$$v \in C(\overline{\Omega} \times [0, T_{\max})) \cap C^{2,1}(\overline{\Omega} \times (0, T_{\max})) \cap L^{\infty}_{\mathrm{loc}}([0, T_{\max}); W^{1,q}(\Omega))$$

which solves (1.3) *in the classical sense. Moreover, if $T_{\max} < \infty$, then*

$$\lim_{t \nearrow T_{\max}} (\|u(\cdot, t)\|_{L^{\infty}(\Omega)} + \|v(\cdot, t)\|_{W^{1,q}(\Omega)}) = \infty.$$

In the following, we let (u, v) be the solution of (1.3) on $[0, T_{\max})$ as in Lemma 2.1. For the proof of Theorem 1.1 we will recall a useful fact to derive the L^{∞}-estimate for u.

LEMMA 2.2. *Assume that the solution (u, v) of* (1.3) *satisfies*

$$\|u(\cdot, t)\|_{L^p(\Omega)} \le C(p) \tag{2.1}$$

for all $t \in (0, T_{\max})$ with some $p > \frac{n}{2}$ and $C(p) > 0$. Then there exists a constant $C' > 0$ such that

$$\|u(\cdot, t)\|_{L^{\infty}(\Omega)} + \|v(\cdot, t)\|_{W^{1,q}(\Omega)} \le C'$$

for all $t \in (0, T_{\max})$.

Proof. The same argument as in the proof of [1, Lemma 3.2] yields this result. □

Thanks to Lemmas 2.1 and 2.2 we will only make sure that the L^p-estimate for u holds with some $p > \frac{n}{2}$ to show global existence and boundedness of solutions to (1.3). To establish (2.1) we introduce the functions g and φ by

$$g(s) := -r \int_0^s \frac{1}{(a + \tau)^k}\, d\tau, \quad \varphi(s) := \exp\{g(s)\} \quad (s \ge 0), \tag{2.2}$$

where $r > 0$ is a constant fixed later. Here we note from straightforward calculations that

$$\varphi(s) = C_{\varphi} \exp\left\{ \frac{r}{(k-1)(a+s)^{k-1}} \right\}$$

with $C_{\varphi} = \exp\{-r(k-1)^{-1} a^{-k+1}\} > 0$. Now we shall prove the following inequality by using the test function $\varphi(v)$.

LEMMA 2.3. *Assume that χ satisfies* (1.4) *with some $\lambda > 0$, $k > 1$, $a > 0$, $K > 0$. Then there exists $c > 0$ such that*

$$\frac{d}{dt} \int_{\Omega} u^p \varphi(v) \le \int_{\Omega} u^p H_r(v) \varphi(v) |\nabla v|^2 + c \int_{\Omega} u^p \varphi(v) - \mu p \int_{\Omega} u^{p+1} \varphi(v), \tag{2.3}$$

where H_r is the function defined by

$$H_r(s) := -\frac{kr}{(a+s)^{k+1}} + \left(\frac{p(p-1)K^2}{4} + \frac{r^2}{p-1} \right) \frac{1}{(a+s)^{2k}} \tag{2.4}$$

for $s \geq 0$.

Proof. Let $p \geq 1$. From (1.3) we have

$$\frac{d}{dt}\int_\Omega u^p\varphi(v) = p\int_\Omega u^{p-1}\varphi(v)\nabla\cdot(\nabla u - u\chi(v)\nabla v) + \mu p\int_\Omega u^p\varphi(v)(1-u) + \int_\Omega u^p\varphi'(v)(\Delta v - v + u). \tag{2.5}$$

Then integration by parts derives

$$\begin{aligned} &p\int_\Omega u^{p-1}\varphi(v)\nabla\cdot(\nabla u - u\chi(v)\nabla v) + \int_\Omega u^p\varphi'(v)\Delta v \\ &= -p\int_\Omega \nabla(u^{p-1}\varphi(v))\cdot(\nabla u - u\chi(v)\nabla v) - \int_\Omega \nabla(u^p\varphi'(v))\cdot\nabla v \\ &= -p(p-1)\int_\Omega u^{p-2}\varphi(v)|\nabla u|^2 + \int_\Omega u^{p-1}\left(p(p-1)\varphi(v)\chi(v) - 2p\varphi'(v)\right)\nabla u\cdot\nabla v \\ &\quad + \int_\Omega u^p(-\varphi''(v) + p\varphi'(v)\chi(v))|\nabla v|^2. \end{aligned} \tag{2.6}$$

Due to the Young inequality, we infer that

$$\begin{aligned} &\int_\Omega u^{p-1}\left(p(p-1)\varphi(v)\chi(v) - 2p\varphi'(v)\right)\nabla u\cdot\nabla v \\ &\leq p(p-1)\int_\Omega u^{p-2}\varphi(v)|\nabla u|^2 + \int_\Omega u^p\frac{(p(p-1)\varphi(v)\chi(v) - 2p\varphi'(v))^2}{4p(p-1)\varphi(v)}|\nabla v|^2. \end{aligned} \tag{2.7}$$

Thus a combination of (2.5), (2.6) and (2.7) yields that

$$\frac{d}{dt}\int_\Omega u^p\varphi(v) \leq \int_\Omega u^p F_\varphi(v)|\nabla v|^2 + \mu p\int_\Omega u^p\varphi(v)(1-u) + \int_\Omega u^p\varphi'(v)(-v+u), \tag{2.8}$$

where

$$F_\varphi(s) := -\varphi''(s) + \frac{p(p-1)}{4}\chi(s)^2\varphi(s) + \frac{p\varphi'(s)^2}{(p-1)\varphi(s)} \quad (s \geq 0).$$

Noting that

$$\varphi'(s) = g'(s)\varphi(s) \quad \text{and} \quad \varphi''(s) = g''(s)\varphi(s) + g'(s)^2\varphi(s) \quad (s \geq 0),$$

we can rewrite the function $F_\varphi(s)$ as

$$F_\varphi(s) = \left(-g''(s) + \frac{p(p-1)}{4}\chi(s)^2 + \frac{g'(s)^2}{p-1}\right)\varphi(s) \quad (s \geq 0).$$

Recalling by (2.2) that

$$g'(s) = \frac{-r}{(a+s)^k} \quad \text{and} \quad g''(s) = \frac{rk}{(a+s)^{k+1}} \quad (s \geq 0),$$

we obtain from (1.4) that

$$F_\varphi(s) \leq H_r(s)\varphi(s) \quad \text{for all } s \geq 0, \tag{2.9}$$

where H_r is defined as (2.4). Therefore we see from (2.8) together with (2.9) that

$$\frac{d}{dt}\int_\Omega u^p\varphi(v) \le \int_\Omega u^p H_r(v)\varphi(v)|\nabla v|^2 + \mu p\int_\Omega u^p\varphi(v)(1-u) - r\int_\Omega u^p\varphi(v)\frac{(-v+u)}{(a+v)^k}.$$

We finally verify from the boundedness of the function $s \mapsto \frac{s}{(a+s)^k}$ on $[0,\infty)$ $(k>1)$ and the positivity of u, v and φ that there is a constant $c_1 > 0$ satisfying

$$-r\int_\Omega u^p\varphi(v)\frac{(-v+u)}{(a+v)^k} \le c_1\int_\Omega u^p\varphi(v),$$

and thus we obtain (2.3). □

Now we shall confirm the following inequality which enables us to see the L^p-boundedness of u.

LEMMA 2.4. *Assume that* (1.4) *and* (1.5) *are satisfied with some* $\lambda > 0$, $k > 1$, $a > 0$ *and* $K > 0$. *Then there exist* $p > \frac{n}{2}$ *and* $r > 0$ *such that*

$$H_r(s) \le 0 \quad \textit{for all } s \ge 0, \tag{2.10}$$

where H_r *is defined as* (2.4)*, which implies that*

$$\frac{d}{dt}\int_\Omega u^p\varphi(v) \le c\int_\Omega u^p\varphi(v) - \mu p\int_\Omega u^{p+1}\varphi(v) \tag{2.11}$$

holds.

Proof. The same argument as in the proof of [14, Lemma 4.1] with $\varepsilon = 0$ leads to (2.10). Moreover, from a combination of Lemma 2.3 and (2.10) we obtain (2.11). □

Now we are ready to show the L^p-estimate for u. By using an argument similar to that in the proof of [13, Lemma 3.2] we can verify the following lemma.

LEMMA 2.5. *Assume that* (1.4) *and* (1.5) *are satisfied with some* $\lambda > 0$, $k > 1$, $a > 0$ *and* $K > 0$. *Then there exist* $p > \frac{n}{2}$ *and* $C > 0$ *such that*

$$\|u(\cdot,t)\|_{L^p(\Omega)} \le C$$

for all $t \in (0, T_{\max})$.

Proof. From Lemma 2.4 we obtain (2.11) with some $p > \frac{n}{2}$ and $r > 0$. We shall show the L^p-estimate for u by using (2.11). We first note from the definition of φ (see (2.2)) that

$$C_\varphi \le \varphi(s) \le 1 \quad (s \ge 0). \tag{2.12}$$

Noticing from the Hölder inequality and (2.12) that

$$\int_\Omega u^p\varphi(v) \le \left(\int_\Omega \varphi(v)\right)^{\frac{1}{p+1}}\left(\int_\Omega u^{p+1}\varphi(v)\right)^{\frac{p}{p+1}} \le |\Omega|^{\frac{1}{p+1}}\left(\int_\Omega u^{p+1}\varphi(v)\right)^{\frac{p}{p+1}},$$

we infer from (2.11) that

$$\frac{d}{dt}\int_\Omega u^p\varphi(v) \le c\int_\Omega u^p\varphi(v) - \mu p|\Omega|^{-\frac{1}{p+1}}\left(\int_\Omega u^p\varphi(v)\right)^{\frac{p+1}{p}},$$

which implies that there exists $C > 0$ satisfying

$$\int_\Omega u^p \varphi(v) \le C.$$

Therefore we obtain from (2.12) that

$$\int_\Omega u^p \le C C_\varphi^{-1},$$

which entails this lemma. □

Proof of Theorem 1.1. Lemmas 2.2 and 2.5 directly lead to the conclusion of Theorem 1.1. □

REFERENCES

[1] N. Bellomo, A. Bellouquid, Y. Tao, and M. Winkler. Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues. *Math. Models Methods Appl. Sci.*, 25:1663–1763, 2015.

[2] X. Cao. Global bounded solutions of the higher-dimensional Keller–Segel system under smallness conditions in optimal spaces. *Discrete Contin. Dyn. Syst.*, 35:1891–1904, 2015.

[3] K. Fujie. Boundedness in a fully parabolic chemotaxis system with singular sensitivity. *J. Math. Anal. Appl.*, 424:675–684, 2015.

[4] K. Fujie. *Study of reaction-diffusion systems modeling chemotaxis*. PhD thesis, Tokyo University of Science, 2016.

[5] K. Fujie and T. Senba. Global existence and boundedness of radial solutions to a two dimensional fully parabolic chemotaxis system with general sensitivity. *Nonlinearity*, 29:2417–2450, 2016.

[6] K. Fujie and T. Senba. A sufficient condition of sensitivity functions for boundedness of solutions to a parabolic-parabolic chemotaxis system. preprint.

[7] X. He and S. Zheng. Convergence rate estimates of solutions in a higher dimensional chemotaxis system with logistic source. *J. Math. Anal. Appl.*, 436:970–982, 2016.

[8] D. Horstmann and G. Wang. Blow-up in a chemotaxis model without symmetry assumptions. *Eur. J. Appl. Math.*, 12:159–177, 2001.

[9] J. Lankeit. A new approach toward boundedness in a two-dimensional parabolic chemotaxis system with singular sensitivity. *Math. Methods Appl. Sci.*, 39:394–404, 2016.

[10] J. Lankeit and M. Winkler. A generalized solution concept for the Keller–Segel system with logarithmic sensitivity: global solvability for large nonradial data. *NoDEA Nonlinear Differential Equations Appl.*, 24:24:49, 2017.

[11] M. Mizukami. Boundedness and asymptotic stability in a two-species chemotaxis-competition model with signal-dependent sensitivity. *Discrete Contin. Dyn. Syst. Ser. B*, 22:2301–2319, 2017.

[12] M. Mizukami. Improvement of conditions for asymptotic stability in a two-species chemotaxis-competition model with signal-dependent sensitivity. submitted, arXiv:1706.04774 [math.AP].

[13] M. Mizukami and T. Yokota. Global existence and asymptotic stability of solutions to a two-species chemotaxis system with any chemical diffusion. *J. Differential Equations*, 261:2650–2669, 2016.

[14] M. Mizukami and T. Yokota. A unified method for boundedness in fully parabolic chemotaxis systems with signal-dependent sensitivity. *Math. Nachr.*, to appear.

[15] T. Nagai, T. Senba, and K. Yoshida. Application of the Trudinger-Moser inequality to a parabolic system of chemotaxis. *Funkcial. Ekvac.*, 40:411–433, 1997.

[16] M. Negreanu and J. I. Tello. On a two species chemotaxis model with slow chemical diffusion. *SIAM J. Math. Anal.*, 46:3761–3781, 2014.

[17] M. Negreanu and J. I. Tello. Asymptotic stability of a two species chemotaxis system with non-diffusive chemoattractant. *J. Differential Equations*, 258:1592–1617, 2015.

[18] M. Winkler. Aggregation vs. global diffusive behavior in the higher-dimensional Keller–Segel model. *J. Differential Equations*, 248:2889–2905, 2010.

[19] M. Winkler. Global asymptotic stability of constant equilibria in a fully parabolic chemotaxis system with strong logistic dampening. *J. Differential Equations*, 257:1056–1077, 2014.
[20] Q. Zhang and X. Li. Global existence and asymptotic properties of the solution to a two-species chemotaxis system. *J. Math. Anal. Appl.*, 418:47–63, 2014.

Proceedings of EQUADIFF 2017
pp. 69–78

KOLMOGOROV'S ε-ENTROPY OF THE ATTRACTOR OF THE STRONGLY DAMPED WAVE EQUATION IN LOCALLY UNIFORM SPACES*

JAKUB SLAVÍK†

Abstract. We establish an upper bound on the Kolmogorov's entropy of the locally compact attractor for strongly damped wave equation posed in locally uniform spaces in subcritical case using the method of trajectories.

Key words. Strongly damped wave equation, unbounded domains, locally compact attractor, Kolmogorovs entropy.

AMS subject classifications. 37L30, 35B41, 35L05.

1. Introduction. We are interested in the asymptotic properties of the strongly damped wave equation

$$u_{tt} + \beta u_t - \alpha \Delta u_t - \Delta u + f(u) = g, \qquad t > 0, \quad x \in \mathbb{R}^d, \tag{1.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear function specified later and α, $\beta > 0$, supplemented by the initial datum

$$u(0) = u_0 \in W^{1,2}_b(\mathbb{R}^d), \qquad u_t(0) = u_1 \in L^2_b(\mathbb{R}^d).$$

The strongly damped wave equation has a number of relevant physical applications, see e.g. [5].

Asymptotic properties of the equation (1.1) in bounded domains have been thoroughly studied. Let us mention some of the results briefly. In [2] the authors establish the existence of global attactor for the critical case. Exponential attractors in the subcritical and critical case have been studied in [11] and [16]. The existence of global attractor for critical and supercritical exponents has been shown for a strongly damped wave equation with memory in [5]. The finite dimensionality of the attractor has been shown in [6]. The situation in supercritical case is studied in detail in [8].

In the non-autonomous case when $g = g(t)$, the resulting uniform attractor might have infinite fractal dimension induced by the time-dependence of the external forces. To measure the complexity of the attractor one can employ Kolmogorov's ε-entropy instead of fractal dimension. In [9] the authors establish an upper bound on Kolmogorov's ε-entropy of the attractor of equation similar to (1.1) in bounded domain and show that if the time-dependent right-hand side is finite-dimensional in the appropriate sense, the resulting attractor is finite dimensional.

In unbounded domains the results are more scarce. In [1] and [4] the authors study the equation (1.1) posed in the classical space $W^{1,2}(\mathbb{R}^d) \times L^2(\mathbb{R}^d)$ and show the existence of a connected universal attractor in the subcritical and critical case. In the context of locally uniform spaces, the non-autonomous wave equation with weak linear damping, i.e. with $\alpha = 0$, has been studied in detail in [17] including an upper bound

*This research was supported by the Charles University, project GA UK No. 200716.

†Department of Mathematical Analysis, Charles University, Sokolovská 83, Prague 186 75, Czech Republic (slavikj@karlin.mff.cuni.cz).

on Kolmogorov's ε-entropy of an attractor reflecting the non-compactness induced both by time-dependent external forces and the unboundedness of the spatial domain. The strongly damped wave equation has been studied in [3], where the well-posedness of the equation in a more regular subspace of locally uniform space $W^{2,p}_b(\mathbb{R}^d)\times L^p_b(\mathbb{R}^d)$, $p > d/2$, $p \geq 2$, and the existence of a locally compact attractor have been shown for the critical case. In [15] the authors generalized these results to the space of locally uniform functions $W^{1,2}_b(\mathbb{R}^d) \times L^2_b(\mathbb{R}^d)$ and obtained a result on the asymptotic regularity of the solutions, cf. the end of this section. In [14] the author studies a variant of the strongly damped wave equation with fractional damping and shows the existence of a locally compact attractor in the critical case together with space-time regularity of the solutions.

The aim of this paper is to establish an upper bound on the Kolmogorov's ε-entropy of the attractor of the equation (1.1) in the subcritical case. To this end we use the method of trajectories and a technique similar to the ones used for a wave equation with nonlinear damping in [12] for bounded domains, resp. in [10] for unbounded domains. However, compared to [10] or [14], solutions of (1.1) do not possess neither a finite speed of propagation nor a smoothing property, and thus the argument must be adapted.

Let ϕ be a weight function, $\bar{x} \in \mathbb{R}^d$ and $\varepsilon > 0$. We denote

$$\begin{gathered}\Phi_{\bar{x},\varepsilon} = W^{1,2}_{\bar{x},\varepsilon}(\mathbb{R}^d) \times L^2_{\bar{x},\varepsilon}(\mathbb{R}^d), \quad W_{\bar{x},\varepsilon} = W^{1,2}_{\bar{x},\varepsilon}(\mathbb{R}^d) \times W^{1,2}_{\bar{x},\varepsilon}(\mathbb{R}^d),\\ \Phi_{b,\phi} = W^{1,2}_{b,\phi}(\mathbb{R}^d) \times L^2_{b,\phi}(\mathbb{R}^d), \quad W_{b,\phi} = W^{1,2}_{b,\phi}(\mathbb{R}^d) \times W^{1,2}_{b,\phi}(\mathbb{R}^d),\\ W_{\mathrm{loc}} = W^{1,2}_{\mathrm{loc}}(\mathbb{R}^d) \times W^{1,2}_{\mathrm{loc}}(\mathbb{R}^d),\end{gathered}$$

with the convention that we omit the subscript ϕ if $\phi \equiv 1$ and write for example Φ_b instead of $\Phi_{b,1}$. For definitions of weight functions and weighted and locally uniform spaces see Section 2.

For simplicity let us choose $\alpha = \beta = 1$. The nonlinear term $f \in C^1(\mathbb{R},\mathbb{R})$ satisfies the following conditions:

- (*growth condition*) there exist $C > 0$ and $0 \leq q \leq 4/(d-2)$ such that

$$|f(r) - f(s)| \leq C|r-s|\left(1 + |r|^q + |s|^q\right), \qquad \forall r, s \in \mathbb{R}. \tag{1.2}$$

  The nonlinearity is *critical* if $q = 4/(d-2)$ and *subcritical* if $q < 4/(d-2)$.
- (*dissipation condition*) there exist $k \geq 1$ and $\mu_0 > 0$ such that for every $\mu \in (0, \mu_0]$ there exist $C_\mu, C_0 \in \mathbb{R}$ such that

$$kF(s) + \mu s^2 - C_\mu \leq sf(s), \quad -C_0 \leq F(s) \qquad \forall s \in \mathbb{R},$$

  where $F(s) = \int_0^s f(r)\,dr$.

These conditions are the same as in [3] and [15].

The *weak solution* of (1.1) is defined in the sense of distributions on $(0,\infty)\times\mathbb{R}^d$ and has the regularity

$$(u, u_t) \in C([0,T];\Phi_{\bar{x},\varepsilon}), \qquad \|u\|^2_{W^{1,2}_b} + \|u_t\|^2_{L^2_b} \in L^\infty((0,T)),$$

for every $T > 0$, $\bar{x} \in \mathbb{R}^d$ and $\varepsilon > 0$. Using a standard density argument it can be shown that the equation can be tested by functions

$$\varphi \in L^2(0,T;W^{1,2}_{\bar{x},\varepsilon}(\mathbb{R}^d)) \cap W^{1,2}(0,T;L^2_{\bar{x},\varepsilon}(\mathbb{R}^d))$$

for arbitrary $T > 0$, $\bar{x} \in \mathbb{R}^d$, $\varepsilon > 0$.

The existence and uniqueness of weak solutions has been shown in [15, Section 3] using semigroup theory in the subspace of more regular initial data continuous with respect to spatial to translations. We also have the following dissipative estimates: there exist t_0, $C > 0$ such that for every $t > t_0$ we have

$$\|u\|_{W_b^{1,2}} + \|u_t\|_{W_b^{1,2}} + \|u_{tt}\|_{L_b^2} \le C. \tag{1.3}$$

For proofs see [15, Section 4]. Let us denote the absorbing set by $\mathcal{B}$ and assume that $\mathcal{B}$ is closed and positively invariant.

In [15], the authors also show the existence of a locally compact attractor in the critical case, namely the existence an invariant set $\mathcal{A} \subseteq \Phi_b$ bounded and closed in $W_b^{2,2}(\mathbb{R}^d) \times W_b^{1,2}(\mathbb{R}^d)$ and compact in W_{loc}, which attracts the bounded sets of Φ_b in the W_{loc}-norm, and the asymptotic regularity, namely the existence of a closed and bounded set $\mathcal{B}_1 \subseteq W_b^{2,2}(\mathbb{R}^d) \times W_b^{1,2}(\mathbb{R}^d)$, a constant $\nu > 0$, and a positive monotonically increasing function $Q(\cdot)$ such that for every bounded $B \subseteq \Phi_b$ we have

$$\operatorname{dist}_{\Phi_b}(S(t)B, \mathcal{B}_1) \le Q(\|B\|_{\Phi_b}) e^{-\nu t} \qquad \forall t > 0.$$

For proofs see [15, Theorem 1.1 and 1.2]. It is worth noting that the technique presented in this paper do not rely on the asymptotic regularity of the attractor.

This paper is organized as follows: in Section 2 we review the basic definitions of function spaces used in the rest of the paper. In Section 3 we define the trajectory spaces and the trajectory semigroup and show that the trajectory semigroup has a squeezing property which is then used in Section 4 to establish an upper estimate on the locally compact attractor of the equation (1.1).

2. Function spaces. A function $\phi : \mathbb{R}^d \to (0, \infty)$ is called a *weight function* of growth $\mu \ge 0$ if

$$C_\phi^{-1} e^{-\mu|x-y|} \le \phi(x)/\phi(y) \le C_\phi e^{\mu|x-y|}, \ |\nabla \phi| \le \tilde{C}_\phi \mu \phi, \quad \text{for a.e. } x, y \in \mathbb{R}^d, \tag{2.1}$$

for some $C_\phi \ge 1$ and some $\tilde{C}_\phi > 0$. For $\bar{x} \in \mathbb{R}^d$ and $\varepsilon > 0$ we denote

$$\phi_{\bar{x},\varepsilon}(x) = \exp(-\varepsilon|x - y|).$$

Clearly $\phi_{\bar{x},\varepsilon}$ is a weight function of growth ε.

For $p \in [1, \infty)$, $\bar{x} \in \mathbb{R}^d$ and $\varepsilon > 0$ we define the *weighted Lebesgue space* $L^p_{\bar{x},\varepsilon}(\mathbb{R}^d)$ by

$$L^p_{\bar{x},\varepsilon}(\mathbb{R}^d) = \{u \in L^p_{\mathrm{loc}}(\mathbb{R}^d); \|u\|^p_{L^p_{\bar{x},\varepsilon}} = \int_{\mathbb{R}^d} |u(x)|^p \phi_{\bar{x},\varepsilon}(x)\, dx < \infty\}.$$

In the case $p = 2$ we use the notation $\|\cdot\|_{L^2_{\bar{x},\varepsilon}} \equiv \|\cdot\|_{\bar{x},\varepsilon}$ and denote the scalar product in $L^2_{\bar{x},\varepsilon}(\mathbb{R}^d)$ by $(\cdot,\cdot)_{\bar{x},\varepsilon}$. The weighted Sobolev spaces are defined in an obvious manner.

Observe that the space $W^{k,p}_{\bar{x},\varepsilon}(\mathbb{R}^d)$ cannot be embedded into $L^q_{\bar{x},\varepsilon}(\mathbb{R}^d)$ for any $q > p$. However, assuming that k, $l \in \mathbb{N}_0$ and p, $q \in [1, \infty)$ satisfy $k \ge l$, $q \ge p$ and $W^{k,p}(\mathbb{R}^d) \hookrightarrow W^{l,q}(\mathbb{R}^d)$, then for $\tilde{\varepsilon} = \varepsilon q/p$ we have the continuous embedding $W^{k,p}_{\bar{x},\varepsilon}(\mathbb{R}^d) \hookrightarrow W^{l,q}_{\bar{x},\tilde{\varepsilon}}(\mathbb{R}^d)$. Moreover, if the embedding $W^{k,p}(B) \hookrightarrow\hookrightarrow W^{l,q}(B)$ is compact, where $B = B(0,1) \subseteq \mathbb{R}^d$ then for $\tilde{\varepsilon} > \varepsilon q/p$ the embedding $W^{k,p}_{\bar{x},\varepsilon}(\mathbb{R}^d) \hookrightarrow\hookrightarrow W^{l,q}_{\bar{x},\tilde{\varepsilon}}(\mathbb{R}^d)$ is compact as well.

Let ϕ be a weight function and $p \in [1, \infty)$. We define the *weighted locally uniform space* $L^p_{b,\phi}(\mathbb{R}^d)$ by

$$L^p_{b,\phi}(\mathbb{R}^d) = \{u \in L^p_{\mathrm{loc}}(\mathbb{R}^d);\, \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x})^{1/p} \|u\|_{L^p(C^1_{\bar{x}})} < \infty\},$$

where C^R_x denotes the cube in $\mathbb{R}^d$ of side $R > 0$ and centred at $x \in \mathbb{R}^d$. We equip the space with a norm equivalent to $\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x})^{1/p} \|u\|_{L^p(C^1_{\bar{x}})}$ defined by

$$\|u\|_{L^p_b} = \sup_{k\in\mathbb{Z}^d} \phi(k)^{1/p} \|u\|_{L^p(C^1_k)}. \tag{2.2}$$

Also one can see that if we take any bounded neighbourhood of $\bar{x}$ in (2.2) instead of C^1_k, we again obtain an equivalent norm.

THEOREM 2.1 (see e.g. [7, Theorem 2.1]). *Let $k \in \mathbb{N}_0$, $p \in [1, \infty)$ and $\varepsilon > 0$. Let ϕ be a weight function of growth rate $0 \le \mu < \varepsilon$ and $u \in W^{k,p}_{\mathrm{loc}}(\mathbb{R}^d)$. Then $u \in W^{k,p}_{b,\phi}(\mathbb{R}^d)$ if and only if $u \in W^{k,p}_{\bar{x},\varepsilon}(\mathbb{R}^d)$ for every $\bar{x} \in \mathbb{R}^d$ and*

$$\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x})^{1/p} \|u\|_{W^{k,p}_{\bar{x},\varepsilon}} < \infty. \tag{2.3}$$

Moreover, the left-hand side of (2.3) *defines a norm equivalent to the $W^{k,p}_{b,\phi}(\mathbb{R}^d)$-norm.*

For $\mathcal{O} \subseteq \mathbb{R}^d$ denote $\mathbb{I}(\mathcal{O}) = \{k \in \mathbb{Z}^d; C^1_k \cap \mathcal{O} \neq \emptyset\}$ and we define the $W^{k,p}_{b,\phi}(\mathcal{O})$-seminorm by

$$\|u\|_{W^{k,p}_{b,\phi}(\mathcal{O})} = \sup_{l\in\mathbb{I}(\mathcal{O})} \phi(l)^{1/p} \|u\|_{W^{k,p}(C^1_l)}. \tag{2.4}$$

LEMMA 2.2 ([17, Proposition 1.2]). *For $1 \le p < \infty$ and $\varepsilon > 0$ fixed there exist C_1, $C_2 > 0$ such that for $\bar{x} \in \mathbb{R}^d$ and $u \in L^p_{\bar{x},\varepsilon}(\mathbb{R}^d)$ with we have*

$$C_1 \|u\|^p_{L^p_{\bar{x},\varepsilon}} \le \int_{\mathbb{R}^d} \phi_{\bar{x},\varepsilon}(x) \|u\|^p_{L^p(B(x,1)}\, dx \le C_2 \|u\|^p_{L^p_{\bar{x},\varepsilon}}.$$

Let $\ell > 0$ and let ϕ be a weight function. We define the parabolic locally uniform spaces $L^2_{b,\phi}(0,\ell; L^2(\mathbb{R}^d))$, $L^2_{b,\phi}(0,\ell; W^{1,2}(\mathbb{R}^d)) \subseteq L^2_{\mathrm{loc}}((0,\ell)\times\mathbb{R}^d)$ by

$$L^2_{b,\phi}(0,\ell;L^2) = \{u; \|u\|^2_{L^2_{b,\phi}(0,\ell;L^2)} = \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \|u\|^2_{L^2(0,\ell;L^2(C^1_{\bar{x}}))} < \infty\},$$
$$L^2_{b,\phi}(0,\ell;W^{1,2}) = \{u; \|u\|^2_{L^2_{b,\phi}(0,\ell;W^{1,2})} = \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \|u\|^2_{L^2(0,\ell;W^{1,2}(C^1_{\bar{x}}))} < \infty\}.$$

LEMMA 2.3 ([7, Theorem 2.4]). *Let $\varepsilon > 0$ be fixed and let ϕ be a weight function of growth rate $\mu \in [0, \varepsilon)$. Then*

$$\|u\|^2_{L^2_{b,\phi}(0,\ell;L^2)} \approx \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{\mathbb{R}^d} |u(x,t)|^2 \phi_{\bar{x},\varepsilon}(x)\, dx\, dt,$$
$$\|u\|^2_{L^2_{b,\phi}(0,\ell;W^{1,2})} \approx \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{\mathbb{R}^d} \left(|u(x,t)|^2 + |\nabla u(x,t)|^2\right) \phi_{\bar{x},\varepsilon}(x)\, dx\, dt.$$

In particular the previous lemma implies that for a weight function ϕ of growth rate $\mu \in [0, \min\{\varepsilon_1, \varepsilon_2\})$ for some $\varepsilon_1, \varepsilon_2 > 0$ one has

$$\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{\mathbb{R}^d} |u(x)|^2 \phi_{\bar{x},\varepsilon_2}(x)\,dx\,dt \approx \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{\mathbb{R}^d} |u(x)|^2 \phi_{\bar{x},\varepsilon_1}(x)\,dx\,dt$$

and similarly in the case of $L^2_{b,\phi}(0,\ell;W^{1,2})$. For $\mathcal{O} \subseteq \mathbb{R}^d$ we can define the seminorms $L^2_{b,\phi}(0,\ell;L^2(\mathcal{O}))$ and $L^2_{b,\phi}(0,\ell;W^{1,2}(\mathcal{O}))$ similarly as in (2.4).

LEMMA 2.4 (Ehrling's lemma in weighted spaces, see e.g. [13, Lemma 7.6]). *Let $p,\ q \geq 1$ and $\varepsilon,\ \tilde{\varepsilon} > 0$ be such that the embedding $W^{1,p}_{\bar{x},\varepsilon}(\mathbb{R}^d) \hookrightarrow\hookrightarrow L^q_{\bar{x},\tilde{\varepsilon}}(\mathbb{R}^d)$ holds. Then for every $\theta > 0$ and $1 \leq \alpha < q$ there exist $C,\ R > 0$ such that for every $u : (0,\ell)\times\mathbb{R}^d \to \mathbb{R}$ one has*

$$\int_0^\ell \|u(t)\|^\alpha_{L^q_{\bar{x},\tilde{\varepsilon}}}\,dt \leq \theta \int_0^\ell \|u(t)\|^\alpha_{W^{1,p}_{\bar{x},\varepsilon}}\,dt + C\int_0^\ell \int_{B(\bar{x},R)} |u(t,x)|^\alpha\,dxdt. \tag{2.5}$$

3. Squeezing property. We define the energy functional by

$$E[u](t,x) = \frac{1}{2}\left(|u_t(t,x)|^2 + |u(t,x)|^2 + |\nabla u(t,x)|^2\right).$$

Let us define the space of trajectories

$$\mathcal{X} = \{(\chi,\chi_t);\chi \in L^2_{\mathrm{loc}}((0,\ell)\times\mathbb{R}^d) \text{ solves (1.1) on } (0,\ell) \text{ with } (\chi(0),\chi_t(0)) \in \mathcal{B}\}.$$

Let $\ell > 0$ be fixed. The trajectory semigroup $L(t) : \mathcal{X} \to \mathcal{X}$ and the end-point operator $e : \mathcal{X} \to \Phi_b$ are given by

$$(L(t)(\chi,\chi_t))(s) = (S(t)\chi(s), \partial_t S(t)\chi),\ s \in (0,\ell), \qquad e(\chi) = (\chi(\ell),\chi_t(\ell)).$$

Let us also denote $L \equiv L(\ell)$. For a weight function ϕ we also define

$$\Phi^\ell_{b,\phi} = L^2_{b,\phi}(0,\ell;W^{1,2}(\mathbb{R}^d)) \times L^2_{b,\phi}(0,\ell;L^2(\mathbb{R}^d)),$$
$$W^\ell_{b,\phi} = L^2_{b,\phi}(0,\ell;W^{1,2}(\mathbb{R}^d)) \times L^2_{b,\phi}(0,\ell;W^{1,2}(\mathbb{R}^d))$$

and define respective seminorms similarly as in (2.4) for the parabolic spaces.

LEMMA 3.1. *There exists $\mu_0 > 0$ such that for all weight functions of growth $\mu \in [0,\mu_0)$ and all $\ell > 0$ the operators $L : \Phi^\ell_{b,\phi} \to W^\ell_{b,\phi}$ and $e : \Phi^\ell_{b,\phi} \to W_{b,\phi}$ are Lipschitz continuous on $\mathcal{X}$.*

In the next section we will use a weaker version of Lemma 3.1, more precisely the Lipschitz continuities $L : W^\ell_{b,\phi} \to W^\ell_{b,\phi}$ and $e : W^\ell_{b,\phi} \to W_{b,\phi}$, both of which follow from the proof by adding $\|\nabla w_t(s)\|^2_{\bar{x},\varepsilon}$ to the right-hand side of (3.1). A similar remark also applies to Lemma 4.1.

Proof. Let $\chi_1,\ \chi_2 \in \mathcal{X}$, let u_1 and u_2 be the respective solutions and denote $w = u_1 - u_2$. By Lemma [15, Lemma 9.2] the semigroup $S(t) : \Phi_{\bar{x},\varepsilon} \to W_{\bar{x},\varepsilon}$ is Lipschitz continuous on $\mathcal{B}$ uniformly w.r.t. $t \in [0,T]$, i.e.

$$\|w(t)\|^2_{\bar{x},\varepsilon} + \|\nabla w(t)\|^2_{\bar{x},\varepsilon} + \|w_t(t)\|^2_{\bar{x},\varepsilon} + \|\nabla w_t(t)\|^2_{\bar{x},\varepsilon}$$
$$\leq C_{t,s}\left(\|w(s)\|^2_{\bar{x},\varepsilon} + \|\nabla w(s)\|^2_{\bar{x},\varepsilon} + \|w_t(s)\|^2_{\bar{x},\varepsilon}\right) \tag{3.1}$$

for $0 < s < t$ and $\varepsilon > 0$ sufficiently small. The Lipschitz continuity of L then follows by integration over $s \in (0, \ell)$, $t \in (\ell, 2\ell)$, multiplication by $\phi(\bar{x})$, applying supremum over $\bar{x} \in \mathbb{R}^d$ to both sides of the estimate and using the equivalence of norms from Lemma 2.3. The Lipschitz continuity of e follows in a similar manner. □

DEFINITION 3.2. *The mapping $L : \mathcal{X} \to \mathcal{X}$ has a* squeezing property *for weight function ϕ if there exists $\varepsilon > 0$ such that for every $\gamma > 0$ we may find ℓ, κ, $R > 0$ so that for every χ_1, $\chi_2 \in \mathcal{X}$ and their respective solutions u_1 and u_2 we have*

$$\begin{aligned}\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{\mathbb{R}^d} \left(E[w] + |\nabla w_t|^2\right) \phi_{\bar{x},\varepsilon}\, dx\, dt \le \gamma \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^{\ell} \int_{\mathbb{R}^d} E[w] \phi_{\bar{x},\varepsilon}\, dx\, dt \\ + \kappa \left(\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^{\ell} \int_{B(\bar{x},R)} |w|^2\, dx\, dt + \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^{\ell} \int_{B(\bar{x},R)} |w_t|^2\, dx\, dt \right) \quad (3.2) \\ + \kappa \left(\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w|^2\, dx\, dt + \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w_t|^2\, dx\, dt \right),\end{aligned}$$

where $w = u_1 - u_2$.

LEMMA 3.3. *Let the nonlinear term f be subcritical, i.e. let $0 \le q < 4/(d-2)$. Then for every weight function ϕ of sufficiently small growth the operator L has the squeezing property.*

Proof. The proof is similar to [12, Lemma 3.1]. Let χ_1, $\chi_2 \in \mathcal{X}$ and let u_1, u_2 be the respective solutions. Let $0 < \tau < \ell$ and denote $w = u_1 - u_2$. We test both the equations for u_1 and u_2 by $w_t + w/2$ to get

$$\begin{aligned}\frac{1}{2}\left(\|w_t(2\ell) + \frac{1}{2}w(2\ell)\|_{\bar{x},\varepsilon}^2 + \frac{1}{8}\|w(2\ell)\|_{\bar{x},\varepsilon}^2 + \frac{3}{4}\|\nabla w(2\ell)\|_{\bar{x},\varepsilon}^2 \right) + \frac{1}{2}\int_\tau^{2\ell} \|w_t\|_{\bar{x},\varepsilon}^2\, dt \\ + \int_\tau^{2\ell} \|\nabla w_t\|_{\bar{x},\varepsilon}^2 + \frac{1}{2}\|\nabla w\|_{\bar{x},\varepsilon}^2\, dt + \int_\tau^{2\ell} \left(f(u_1) - f(u_2), w_t + \frac{1}{2}w \right)_{\bar{x},\varepsilon} dt \\ + \int_\tau^{2\ell} \left(\nabla w_t, (w_t + \frac{1}{2}w)\nabla\phi_{\bar{x},\varepsilon} \right) + \left(\nabla w, (w_t + \frac{1}{2}w)\nabla\phi_{\bar{x},\varepsilon} \right) dt \\ = \frac{1}{2}\left(\|w_t(\tau) + \frac{1}{2}w(\tau)\|_{\bar{x},\varepsilon}^2 + \frac{1}{8}\|w(\tau)\|_{\bar{x},\varepsilon}^2 + \frac{3}{4}\|\nabla w(\tau)\|_{\bar{x},\varepsilon}^2 \right). \quad (3.3)\end{aligned}$$

Relying on a standard but a rather tedious argument comprised of using Lemma 2.2, Hölder's and Young's inequalities, subcritical growth estimates (1.2) on the nonlinearity f and compact Sobolev embedding on bounded domains together with dissipation estimates (1.3) we obtain

$$\left| \int_{\mathbb{R}^d} (f(u_1) - f(u_2))(w_t + w)\phi_{\bar{x},\varepsilon}\, dx \right| \le \eta(\|w\|_{\bar{x},\varepsilon}^2 + \|\nabla w\|_{\bar{x},\varepsilon}^2) + C\|w_t\|_{L^p_{\bar{x},\varepsilon}}^2$$

for $\eta > 0$ determined later and $1 \le p < 2d/(d-2)$. Putting the previous estimates into (3.3) and employing (2.1) and Young's inequality we get

$$\begin{aligned}C\left(\|w_t(2\ell) + \frac{1}{2}w(2\ell)\|_{\bar{x},\varepsilon}^2 + \|w(2\ell)\|_{\bar{x},\varepsilon}^2 + \|\nabla w(2\ell)\|_{\bar{x},\varepsilon}^2 \right) \\ + \zeta \int_\ell^{2\ell} \|w_t\|_{\bar{x},\varepsilon}^2 + \|\nabla w_t\|_{\bar{x},\varepsilon}^2 + \|\nabla w\|_{\bar{x},\varepsilon}^2 + \|w\|_{\bar{x},\varepsilon}^2\, dt \\ \le \int_{\mathbb{R}^d} E[w](\tau)\phi_{\bar{x},\varepsilon}\, dx + C\int_0^{2\ell} \|w\|_{\bar{x},\varepsilon}^2 + \|w_t\|_{L^p_{\bar{x},\varepsilon}}^2 + \|w\|_{L^p_{\bar{x},\varepsilon}}^2\, dt\end{aligned}$$

for some $\zeta > 0$. We note that from now on the value of ε will not change. We integrate over $\tau \in (0, \ell)$ and apply the weighted version of Ehrling's lemma (Lemma 2.4) to the functions $w(t)$ and $w_t(t)$ both on the time intervals $(0, \ell)$ and $(\ell, 2\ell)$ to obtain

$$\begin{aligned}\zeta\ell \int_\ell^{2\ell} \int_{\mathbb{R}^d} \left(E[w] + |\nabla w_t|^2\right) \phi_{\bar{x},\varepsilon}\, dx\, dt &\le \int_0^\ell \int_{\mathbb{R}^d} E[w]\phi_{\bar{x},\varepsilon}\, dx\, dt + C\ell \int_0^{2\ell} \|w\|^2_{\bar{x},\varepsilon}\, dt \\ &+ C\ell\theta \left(\int_0^\ell \|w\|^2_{W^{1,2}_{\bar{x},\tilde{\varepsilon}}}\, dt + \int_0^\ell \|w_t\|^2_{W^{1,2}_{\bar{x},\tilde{\varepsilon}}}\, dt + \int_\ell^{2\ell} \|w\|^2_{W^{1,2}_{\bar{x},\tilde{\varepsilon}}}\, dt + \int_\ell^{2\ell} \|w_t\|^2_{W^{1,2}_{\bar{x},\tilde{\varepsilon}}}\, dt \right) \\ &+ C\ell \left(\int_0^\ell \int_{B(\bar{x},R)} |w|^2 + |w_t|^2\, dx\, dt + \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w|^2 + |w_t|^2\, dx\, dt \right)\end{aligned}$$

for some $R > 0$ fixed, $\theta > 0$ determined later and some $\tilde{\varepsilon} > 0$ such that $W^{1,2}_{\bar{x},\tilde{\varepsilon}}(\mathbb{R}^d) \hookrightarrow\hookrightarrow L^q_{\bar{x},\varepsilon}(\mathbb{R}^d)$, i.e. $2\varepsilon/q > \tilde{\varepsilon}$. If we restrict ourselves to weight functions ϕ of growth $\mu \in [0, \min\{\varepsilon, \tilde{\varepsilon}\})$, multiply by $\phi(\bar{x})$ and apply supremum over $\bar{x} \in \mathbb{R}^d$, then by Lemma 2.3 and by choosing θ sufficiently small we obtain

$$\begin{aligned}\tilde{\zeta}\ell \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{\mathbb{R}^d} \left(E[w] + |\nabla w_t|^2\right) \phi_{\bar{x},\varepsilon}\, dx\, dt &\le C \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{\mathbb{R}^d} E[w]\phi_{\bar{x},\varepsilon}\, dx\, dt \\ &+ C\ell \left(\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{B(\bar{x},R)} |w|^2\, dx\, dt + \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_0^\ell \int_{B(\bar{x},R)} |w_t|^2\, dx\, dt \right) \\ &+ C\ell \left(\sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w|^2\, dx\, dt + \sup_{\bar{x}\in\mathbb{R}^d} \phi(\bar{x}) \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w_t|^2\, dx\, dt \right).\end{aligned}$$

for some $0 < \tilde{\zeta} < \zeta$. The conclusion follows by dividing by $\tilde{\zeta}\ell$ and choosing ℓ sufficiently large. □

4. Entropy estimate. Let X be a metric space and let $K \subseteq X$ be precompact. We define the *Kolmogorov's ε-entropy* by

$$H_\varepsilon(K, X) = \ln N_\varepsilon(K, X),$$

where $N_\varepsilon(K, X)$ is the smallest number of ε-balls in X with centres in K that cover the set K.

LEMMA 4.1. *Let $\mathcal{O} \subseteq \mathbb{R}^d$ be bounded and let*

$$\mathbb{I}(\mathcal{O}) \le C_0 \operatorname{vol}(\mathcal{O}) \tag{4.1}$$

for some $C_0 > 0$. Let $\varepsilon > 0$ and $\theta \in (0, 1)$. Let $(u_0, u_1) \in \mathcal{B}$ and let $(\chi_0, (\chi_0)_t)$ be the trajectory starting from (u_0, u_1). Let ϕ be a weight function such that the operator L has the squeezing property for ϕ and denote $B = B_\varepsilon((\chi_0, (\chi_0)_t); \Phi^\ell_{b,\phi}) \cap \mathcal{X}$. Then there exist C_1, $\ell > 0$ such that

$$H_{\theta\varepsilon}\left((LB)|_{\mathcal{O}}, W^\ell_{b,\phi}(\mathcal{O})\right) \le C_1 \operatorname{vol}(\mathcal{O}),$$

where the constant C_1 depends only on C_0 and θ and is independent of (u_0, u_1), ε, ϕ and $\mathcal{O}$ as long as (4.1) holds and the constants in (2.1) remain the same.

Proof. The proof combines the technique of [12, Lemma 4.1] and [7, Lemma 2.6] and adapts these to the squeezing property at hand. We will prove the assertion for

$\phi \equiv 1$. The general case then follows by the same argument as in [7, Lemma 2.6], namely by showing that $\|\chi\|_{L^2_{b,\phi}(0,\ell;W^{1,2}(\mathcal{O}))} \approx \|F\chi\|_{L^2_{b,1}(0,\ell;W^{1,2}(\mathcal{O}))}$ with $F : \chi \to \phi^{1/2}\chi$.

First fix $0 < \gamma < \theta^2$ and using Lemma 3.3 find κ, $\ell > 0$ such that L has the squeezing property for the weight function ϕ and γ. Let $\delta > 0$ be such that $\gamma + 4\kappa\delta^2 < \theta^2$. For $x_1, x_2, x_3, x_4 \in \mathbb{R}^d$ fixed we denote

$$P_{x_1,x_2,x_3,x_4}\left((\chi, \partial_t\chi)\right) = \left(\chi|_{B(x_1,R)}, \partial_t\chi|_{B(x_2,R)}, L\chi|_{B(x_3,R)}, \partial_t L\chi|_{B(x_4,R)}\right),$$

where $R > 0$ comes from the squeezing property (3.2). Employing the standard Aubin-Lions lemma and the Lipschitz continuity of L we observe that the set

$$X(x_1, x_2, x_3, x_4) = \left\{P_{x_1,x_2,x_3,x_4}\left((\chi, \partial_t\chi)\right); (\chi, \partial_t\chi) \in B\right\}$$

equipped with the product topology $\prod_{i=1}^4 L^2(0, \ell; L^2(B(x_i, R)))$ can be covered by N balls of diameter $\delta\varepsilon$ with N independent of ε and x_i.

Let now $\chi_1, \chi_2 \in B$, let u_1, u_2 be their respective solutions and set $w = u_1 - u_2$. Then we find $x_i^M \in \mathbb{R}^d$ such that

$$\begin{aligned}
&\sup_{\bar{x}\in\mathbb{R}^d} \int_0^\ell \int_{B(\bar{x},R)} |w|^2 \, dx \, dt + \sup_{\bar{x}\in\mathbb{R}^d} \int_0^\ell \int_{B(\bar{x},R)} |w_t|^2 \, dx \, dt \\
&\qquad + \sup_{\bar{x}\in\mathbb{R}^d} \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w|^2 \, dx \, dt + \sup_{\bar{x}\in\mathbb{R}^d} \int_\ell^{2\ell} \int_{B(\bar{x},R)} |w_t|^2 \, dx \, dt \\
&\qquad\qquad \le \int_0^\ell \int_{B(x_1^M,R)} |w|^2 \, dx \, dt + \int_0^\ell \int_{B(x_2^M,R)} |w_t|^2 \, dx \, dt \\
&\qquad\qquad\qquad + \int_\ell^{2\ell} \int_{B(x_3^M,R)} |w|^2 \, dx \, dt + \int_\ell^{2\ell} \int_{B(x_4^M,R)} |w_t|^2 \, dx \, dt + \frac{1}{M}
\end{aligned}$$

with $M \in \mathbb{N}$ large enough to have $\gamma\varepsilon^2 + 4\kappa\delta^2\varepsilon^2 + \kappa/M \le \theta^2\varepsilon^2$. By the previous observation we may cover the set $X(x_1^M, x_2^M, x_3^M, x_4^M)$ by $\delta\varepsilon$-balls centered at $P_{x_1^M,x_2^M,x_3^M,x_4^M}\left(\left(\chi^i, \partial_t\chi^i\right)\right)$ for some $(\chi^i, \partial_t\chi^i) \in B$, $i = 1, \dots, N$. For arbitrary $(\chi, \partial_t\chi) \in B$ we may now find $(\chi^i, \partial_t\chi^i) \in B$ such that

$$\|P_{x^M}\left((\chi, \partial_t\chi)\right) - P_{x^M}\left(\left(\chi^i, \partial_t\chi^i\right)\right)\|_{X(x_1^M,x_2^M,x_3^M,x_4^M)} < \delta\varepsilon.$$

The squeezing property now leads to the estimate

$$\sup_{\bar{x}\in\mathbb{R}^d} \int_\ell^{2\ell} \int_{\mathbb{R}^d} \left(E[w] + |\nabla w_t|^2\right) dx \, dt \le \gamma\varepsilon^2 + 4\kappa\delta^2\varepsilon^2 + \frac{\kappa}{M} \le \theta^2\varepsilon^2,$$

which finishes the proof. □

We will use the following auxiliary function in the spirit of [17]: let $\bar{x} \in \mathbb{R}^d$, $R > 0$ and $\nu > 0$. Define

$$\psi(\bar{x}, R) = \psi(\bar{x}, R)(x) = \begin{cases} 1, & |x - \bar{x}| \le R + \sqrt{d}, \\ \exp\left(\nu\left(R + \sqrt{d} - |x - \bar{x}|\right)\right), & \text{otherwise.} \end{cases}$$

The function $\psi(\bar{x}, R)$ is clearly a weight function of growth ν with, in the notation of (2.1), $C_{\psi(\bar{x},R)} = 1$ for every $\bar{x} \in \mathbb{R}^d$ and $R > 0$. Also we have

$$H_\varepsilon\left(B|_{B(\bar{x},R)}, W_b(B(\bar{x}, R))\right) \le H_\varepsilon\left(B, W_{b,\psi(\bar{x},R)}\right), \tag{4.2}$$

where $W_b(B(\bar{x},R))$ is a seminorm defined similarly as in (2.4) and $B \subseteq W^\ell_b$.

LEMMA 4.2 ([7, Lemma 5.4]). *For every $\varepsilon_0 > 0$ we there exists $R' > 0$ such that for every $\bar{x} \in \mathbb{R}^d$, $R \geq 1$, $\varepsilon \in (0,\varepsilon_0)$ and $\chi_1, \chi_2 \in W^\ell_{b,\psi(\bar{x},R)}$ one has*

$$\|\chi_1 - \chi_2\|_{W^\ell_{b,\psi(\bar{x},R)}} \leq \max\left\{\varepsilon, \|\chi_1 - \chi_2\|_{W^\ell_{b,\psi(\bar{x},R)}(B(\bar{x},R+R'\ln(\varepsilon_0/\varepsilon)))}\right\}.$$

Recall that $\mathcal{A} \subseteq W^{2,2}_b(\mathbb{R}^d) \times W^{1,2}_b(\mathbb{R}^d)$ is the locally compact attractor of the set (1.1) defined in Section 1.

THEOREM 4.3. *There exist constants C_0, C_1, $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0,\varepsilon_0)$, $\bar{x} \in \mathbb{R}^d$ and $R \geq 1$ one has the estimate*

$$H_\varepsilon\left(\mathcal{A}|_{B(\bar{x},R)}, W_b(B(\bar{x},R))\right) \leq C_0\left(R + C_1 \ln\frac{\varepsilon_0}{\varepsilon}\right)^d \ln\frac{\varepsilon_0}{\varepsilon}.$$

Proof. The proof is standard and runs in almost the same way as in [10, Theorem 6.5] and [7, Theorem 5.1] with only minor differences.

Let $\bar{x} \in \mathbb{R}^d$, $R \geq 1$ and let $\psi(\bar{x},R)$ be of sufficiently small growth such that L has the squeezing property for $\psi(\bar{x},R)$ and let $\ell > 0$ be such that Lemma 4.1 holds with $\theta = 1/2\,\mathrm{Lip}(L) < 1$, where $\mathrm{Lip}(L)$ denotes the Lipschitz constant of L from Lemma 3.1. The smallness of growth of $\psi(\bar{x},R)$ can always be achieved by choosing ν small in the definition of $\psi(\bar{x},R)$. By the Lipschitz continuity of e and the property of the weight function $\psi(\bar{x},R)$ (4.2) we get

$$H_\varepsilon\left(\mathcal{A}|_{B(\bar{x},R)}, W_b(B(\bar{x},R))\right) \leq H_\varepsilon\left(\mathcal{A}, W_{b,\psi(\bar{x},R)}\right) \leq H_{\varepsilon/\mathrm{Lip}(e)}\left(\mathcal{A}_\ell, W^\ell_{b,\psi(\bar{x},R)}\right),$$

where $\mathcal{A}_\ell = \{(\chi,\chi_t) \in \Phi^\ell_b; (\chi(0),\chi_t(0) \in \mathcal{A}\}$. By the dissipation estimates (1.3) and the invariance of $\mathcal{A}$ we observe that actually $\mathcal{A}_\ell \subseteq W^\ell_b$ and $\mathcal{A}_\ell$ is invariant w.r.t. $L(t)$. Also the dissipation estimates (1.3) imply that for some $\chi \in \mathcal{A}_\ell$ and $\varepsilon_0 > 0$ sufficiently large we have

$$H_{\varepsilon_0/\mathrm{Lip}(e)}\left(\mathcal{A}_\ell, W^\ell_{b,\psi(\bar{x},R)}\right) = 0.$$

The key part of the proof is to show that for $k \in \mathbb{N} \cup \{0\}$ one has

$$H_{\varepsilon_0 2^{-k}/\mathrm{Lip}(e)}\left(\mathcal{A}_\ell, W^\ell_{b,\psi(\bar{x},R)}\right) \leq C\left(R + C'\ln 2^k\right)^d k \tag{4.3}$$

for some $C' > 0$. Indeed, once we have established (4.3) for given $\varepsilon \in (0,\varepsilon_0)$ we may find $k \in \mathbb{N}$ such that $2^{-k}\varepsilon_0 \leq \varepsilon < 2^{-k+1}\varepsilon_0$ and the desired entropy bound follows.

The estimate (4.3) clearly holds for $k = 0$. Assume that (4.3) holds for $k \geq 0$, i.e.

$$\mathcal{A}_\ell \subseteq \bigcup_{i=1}^{N_k} B_{\varepsilon_0 2^{-k}/\mathrm{Lip}(e)}\left((\chi^i,\chi^i_t); W^\ell_{b,\psi(\bar{x},R)}\right) \tag{4.4}$$

for some $N_k \in \mathbb{N}$ such that $\ln N_k \leq C(R+C'\ln 2^k)^d k$ and $(\chi^i,\chi^i_t) \in \mathcal{A}_\ell$ for $1 \leq i \leq N_k$. Applying L to (4.4) and recalling the invariance of $\mathcal{A}_\ell$ under L and the Lipschitz continuity of L, we get

$$\mathcal{A}_\ell = L(\mathcal{A}_\ell) \subseteq \bigcup_{i=1}^{N} B_{\mathrm{Lip}(L)\varepsilon_0 2^{-k}/\mathrm{Lip}(e)}\left((L\chi^i, \partial_t L\chi^i); W^\ell_{b,\psi(\bar{x},R)}\right) \tag{4.5}$$

By Lemma 4.1 with $\theta = 1/2\,\mathrm{Lip}(L)$ each of the balls on the right-hand side of (4.5) localized to the spatial domain $B(\bar{x}, R + R' \ln 2^{k+1})$ can be covered by $\varepsilon_0 2^{-(k+1)}$-balls in the space $W^\ell_{b,\psi(\bar{x},R)}$ in such a way that

$$\begin{aligned} H_{\varepsilon_0 2^{-(k+1)}/\,\mathrm{Lip}(e)} &\left(\mathcal{A}_\ell|_{B(\bar{x},R+R'\ln 2^{k+1})}, W^\ell_{b,\psi(\bar{x},R)}(B(\bar{x}, R + R' \ln 2^{k+1}))\right) \\ &\le H_{\varepsilon_0 2^{-k}/\,\mathrm{Lip}(e)} \left(\mathcal{A}_\ell, W^\ell_{\bar{x},\psi(\bar{x},\varepsilon)}\right) + C\left(R + R' \ln 2^{k+1}\right)^d \\ &\le C\left(R + R' \ln 2^{k+1}\right)^d (k+1). \end{aligned}$$

The proof is finished since by Lemma 4.2 every $\varepsilon_0 2^{-(k+1)}/\,\mathrm{Lip}(e)$-covering in the space $W^\ell_{b,\psi(\bar{x},R)}(B(\bar{x}, R(\varepsilon_0 2^{-(k+1)})))$ is also an $\varepsilon_0 2^{-(k+1)}/\,\mathrm{Lip}(e)$-covering in $W^\ell_{b,\psi(\bar{x},R)}$. □

Acknowledgements. The author would like to thank D. Pražák for discussions leading to the results of this paper.

REFERENCES

[1] V. Belleri and V. Pata, *Attractors for semilinear strongly damped wave equations on* $\mathbb{R}^3$, Discrete Contin. Dynam. Systems, 7 (2001), pp. 719–735.
[2] A. N. Carvalho and J. W. Cholewa, *Attractors for strongly damped wave equations with critical nonlinearities*, Pacific J. Math., 207 (2002), pp. 287–310.
[3] J. W. Cholewa and T. Dlotko, *Strongly damped wave equation in uniform spaces*, Nonlinear Anal., 64 (2006), pp. 174–187.
[4] M. Conti, V. Pata, and M. Squassina, *Strongly damped wave equations on* $\mathbb{R}^3$ *with critical nonlinearities*, Commun. Appl. Anal., 9 (2005), pp. 161–176.
[5] F. Di Plinio, V. Pata, and S. Zelik, *On the strongly damped wave equation with memory*, Indiana Univ. Math. J., 57 (2008), pp. 757–780.
[6] J.-M. Ghidaglia and A. Marzocchi, *Longtime behaviour of strongly damped wave equations, global attractors and their dimension*, SIAM J. Math. Anal., 22 (1991), pp. 879–895.
[7] M. Grasselli, D. Pražák, and G. Schimperna, *Attractors for nonlinear reaction-diffusion systems in unbounded domains via the method of short trajectories*, J. Differential Equations, 249 (2010), pp. 2287–2315.
[8] V. Kalantarov and S. Zelik, *Finite-dimensional attractors for the quasi-linear strongly-damped wave equation*, J. Differential Equations, 247 (2009), pp. 1120–1155.
[9] H. Li and S. Zhou, *Kolmogorov* ε*-entropy of attractor for a non-autonomous strongly damped wave equation*, Commun. Nonlinear Sci. Numer. Simul., 17 (2012), pp. 3579–3586.
[10] M. Michálek, D. Pražák, and J. Slavík, *Semilinear damped wave equation in locally uniform spaces*, Commun. Pure Appl. Anal., 16 (2017), pp. 1673–1695.
[11] V. Pata and M. Squassina, *On the strongly damped wave equation*, Comm. Math. Phys., 253 (2005), pp. 511–533.
[12] D. Pražák, *On finite fractal dimension of the global attractor for the wave equation with nonlinear damping*, J. Dynam. Differential Equations, 14 (2002), pp. 763–776.
[13] T. Roubíček, *Nonlinear partial differential equations with applications*, Birkhäuser/Springer Basel AG, Basel, International Series of Numerical Mathematics 153 (2013).
[14] A. Savostianov, *Infinite energy solutions for critical wave equation with fractional damping in unbounded domains*, Nonlinear Anal., 136 (2016), pp. 136–167.
[15] M. Yang and C. Sun, *Dynamics of strongly damped wave equations in locally uniform spaces: attractors and asymptotic regularity*, Trans. Amer. Math. Soc., 361 (2009), pp. 1069–1101.
[16] M. Yang and C. Sun, *Exponential attractors for the strongly damped wave equations*, Nonlinear Anal. Real World Appl., 11 (2010), pp. 913–919.
[17] S. V. Zelik, *The attractor for a nonlinear hyperbolic equation in the unbounded domain*, Discrete Contin. Dynam. Systems, 7 (2001), pp. 593–641.

Proceedings of EQUADIFF 2017
pp. 79–88

THE TREE-GRID METHOD WITH CONTROL-INDEPENDENT STENCIL

IGOR KOSSACZKÝ , MATTHIAS EHRHARDT , AND MICHAEL GÜNTHER *

Abstract. The Tree-Grid method is a novel explicit convergent scheme for solving stochastic control problems or Hamilton-Jacobi-Bellman equations with one space dimension. One of the characteristics of the scheme is that the stencil size is dependent on space, control and possibly also on time. Because of the dependence on the control variable, it is not trivial to solve the optimization problem inside the method. Recently, this optimization part was solved by brute-force testing of all permitted controls. In this paper, we present a simple modification of the Tree-Grid scheme leading to a control-independent stencil. Under such modification an optimal control can be found analytically or with the Fibonacci search algorithm.

Key words. Tree-Grid Method, Hamilton-Jacobi-Bellman equation, Stochastic control problem, Fibonacci algorithm

AMS subject classifications. 65M75, 65C40

1. Introduction. Stochastic control problems (SCP) arise in many fields where some stochastic process is controlled in order to maximize (or minimize) an expected value of an uncertain outcome. An effective approach to solve such problems presents the Hamilton-Jacobi-Bellman (HJB) equation. As the analytical solutions are in most cases not feasible, the development of numerical methods dealing either with HJB equation or directly with the SCP is essential. A large class of methods is based on approximating the stochastic process by a Markov chain [5]. Another way presented e.g. in [2] is to solve the HJB equation with an implicit finite-difference method (FDM). A method based on Ricatti transformation of the HJB equation was proposed in [3].

Recently a new method having similarities with Markov chain approximations as well as with the explicit FDMs was presented in [4]. The advantage of this method is its independence on the space-stepping of the grid, as well as its unconditional convergence. However, as well as in FDMs and Markov chain methods, an optimization problem needs to be solved in each step.

In this paper, we want to present a modification of the Tree-Grid method, that will allow us to solve the optimization problem more effectively.

2. Problem formulation. The Tree-Grid method is a numerical scheme for searching the value function $V(s,t)$ of the following *general stochastic control problem*:

*University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany
igor.vyr@gmail.com, ehrhardt@math.uni-wuppertal.de, guenther@math.uni-wuppertal.de

$$V(s,t) = \max_{\theta(s,t)\in\bar{\Theta}} \mathbb{E}\left(\int_t^T \exp\left(\int_t^k r(S_l,l,\theta(S_l,l))dl\right) f(S_k,k,\theta(S_k,k))dk \right.$$
$$\left. + \exp\left(\int_t^T r(S_k,k,\theta(S_k,k))dk\right) V_T(S_T)\Big| S_t = s\right), \tag{2.1}$$
$$dS_t = \mu(S_t,t,\theta(S_t,t))dt + \sigma(S_t,t,\theta(S_t,t))dW_t, \tag{2.2}$$
$$0 < t < T, \quad s \in \mathbb{R},$$

where s is the state variable and t denotes time. Here, $\bar{\Theta}$ is the space of all *suitable* control functions from $\mathbb{R}\times[0,T]$ to a set Θ. In the original Tree-Grid method [4], Θ is supposed to be discrete. If this is not the case, the set Θ should be discretized. Another option arising from this paper would be to search for an optimum analytically, that will be discussed later. Now following Bellman's principle, the following *dynamic programming equation* holds:

$$V(s,t_j) = \max_{\theta(s,t)\in\bar{\Theta}_{t_j}} \mathbb{E}\left(\int_{t_j}^{t_{j+1}} \exp\left(\int_{t_j}^k r(S_l,l,\theta(S_l,l))dl\right) f(S_k,k,\theta(S_k,k))dk \right.$$
$$\left. + \exp\left(\int_{t_j}^{t_{j+1}} r(S_k,k,\theta(S_k,k))dk\right) V(S_{t_{j+1}},t_{j+1})\Big| S_{t_j} = s\right), \tag{2.3}$$

where $0 \le t_j < t_{j+1} \le T$ are some time-points and $\bar{\Theta}_{t_j}$ is a set of control functions from $\bar{\Theta}$ restricted to the $\mathbb{R}\times[t_j,t_{j+1})$ domain. Using this equation (2.3), it can be shown [7], that solving the SCP (2.1),(2.2) is equivalent to solving the so-called Hamilton-Jacobi-Bellman (HJB) equation:

$$\frac{\partial V}{\partial t} + \max_{\theta\in\Theta}\left(\frac{\sigma(\cdot)^2}{2}\frac{\partial^2 V}{\partial s^2} + \mu(\cdot)\frac{\partial V}{\partial s} + r(\cdot)V + f(\cdot)\right) = 0, \tag{2.4}$$
$$V(s,T) = V_T(s), \tag{2.5}$$
$$0 < t < T, \quad s \in \mathbb{R},$$

where $\sigma(\cdot)$, $\mu(\cdot)$, $r(\cdot)$, $f(\cdot)$ are functions of s,t,θ. This HJB formulation was used to prove the convergence of the scheme [4].

We should note that the maximum operator in (2.1) and (2.4) can be replaced by a minimum, (supremum, infimum) operator and the whole following analysis will hold analogously.

3. The Tree-Grid Method. The main idea of the Tree-Grid method is approximating the continuous stochastic process (2.2) with a discrete one, attaining only values from the grid inside the computational domain, or values outside the computational domain, that are assumed to be predefined. Then, a discretized version of (2.3) is used to compute the approximation of the value function in each node of the grid. The underlying discretized stochastic process can be easily represented by a scenario tree. However, such a tree is "growing" from every time-space node of an (arbitrarily chosen) grid, what explains the name of the method. We illustrate this structure in Figure 3.1. Alternatively, the method can be also interpreted in terms of finite differences which is discussed concisely in [4]. We will use this alternative representation also in the Sections 5, 6.

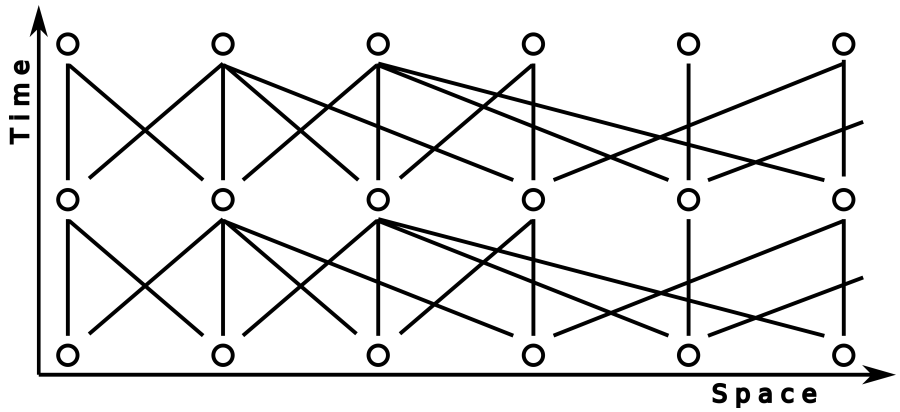


FIG. 3.1. *Illustration of the Tree-Grid structure. From each grid node in current time layer three branches are growing (bottom-to-top), determining which values from grid in later time layer influence the value in the current node.*

Now we will quickly recapitulate the Tree-Grid method algorithm. We compute the approximation of the solution on a rectangular domain $[s_L, s_R] \times [0, T]$ with some grid as in usual finite difference schemes for PDEs. The grid-points are denoted as $[s_i, t_j]$, $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, M\}$, $k < l \Rightarrow s_k < s_l, t_k < t_l$, $t_1 = 0, t_M = T$, $s_1 = s_L, s_N = s_R$. The grid is possibly non-equidistant in space with space-steps $\Delta_i s = s_{i+1} - s_i$ and $\Delta s = \max_i \Delta_i s$. We will use an equidistant discretization in time with a time-step Δt. A generalization to non-equidistant time-stepping is straightforward, however the implementation is less effective in means of computational time in that case. The numerical approximation of $V(s_i, t_j)$ will be denoted by v_i^j.

The whole scheme is then defined by the discrete approximation of the dynamic programming equation (2.3)

$$v_i^j = \max_{\theta \in \Theta} \Big(f_i^j(\theta)\Delta t + (1 + r_i^j(\theta)\Delta t) \cdot \Big(p_{(i-,\theta)} v_{(i-,\theta)}^{j+1} + p_{(i,\theta)} v_i^{j+1} + p_{(i+,\theta)} v_{(i+,\theta)}^{j+1} \Big) \Big). \tag{3.1}$$

for $i = 2, 3, \ldots, N-1$ and

$$v_1^j = BC_L(s_1, t_j), \quad v_N^j = BC_R(s_N, t_j). \tag{3.2}$$

Here, $f_i^j(\theta) = f(s_i, t_j, \theta)$, $r_i^j(\theta) = r(s_i, t_j, \theta)$ and

$$v_{(i*,\theta)}^{j+1} = \begin{cases} v_k^{j+1} \quad \text{so that } s_k = s_{(i*,\theta)} & \text{if } s_{(i*,\theta)} \in \{s_1, s_2, \ldots, s_N\} \\ BC_L(s_{(i*,\theta)}, t_{j+1}) & \text{if } s_{(i*,\theta)} < s_1 \\ BC_R(s_{(i*,\theta)}, t_{j+1}) & \text{if } s_{(i*,\theta)} > s_N \end{cases}$$

for the $* \in \{-, +\}$. Here $BC_L(s, t)$ and $BC_R(s, t)$ are functions defining an approximation of the value function behind the boundaries and $s_{(i-,\theta)}$, s_i, $s_{(i+,\theta)}$ are states that the discretized process may attain with the probabilities $p_{(i-,\theta)}$, p_i, $p_{(i+,\theta)}$

under the control θ after the time-step Δt if the previous state was s_i. It holds $s_{(i-,\theta)} < s_i < s_{(i+,\theta)}$. In order to match the moments of this discretized process with the original time-continuous process (2.2) the probabilities are chosen in the following manner:

$$p_{(i-,\theta)} = \frac{-\mu\Delta t(\Delta_+ s - \mu\Delta t) + Var}{\Delta_- s(\Delta_- s + \Delta_+ s)}, \tag{3.3}$$

$$p_{(i,\theta)} = \frac{(-\Delta_- s - \mu\Delta t)(\Delta_+ s - \mu\Delta t) + Var}{-\Delta_- s\Delta_+ s}, \tag{3.4}$$

$$p_{(i+,\theta)} = \frac{(-\Delta_- s - \mu\Delta t)(-\mu\Delta t) + Var}{(\Delta_+ s + \Delta_- s)\Delta_+ s}. \tag{3.5}$$

Here, $\Delta_+ s = s_{(i+,\theta)} - s_i$, $\Delta_- s = s_i - s_{(i-,\theta)}$, $\mu := \mu(s_i, t_j, \theta)$ and $Var := Var(s_i, t_j, \theta)$ is chosen in such manner, that $Var/\Delta t$ is equal or at least converges to $\sigma^2(s_i, t_j, \theta)$ with $\Delta t, \Delta s \to 0$. As explained in [4], these probabilities sum up to one. However, we need to choose states $s_{(i-,\theta)}$, $s_{(i+,\theta)}$ such that all probabilities are positive. Let us assume that the drift μ is positive. Then $p_{(i+,\theta)}$ is positive, and $p_{(i-,\theta)}$, $p_{(i,\theta)}$ are positive if the following condition holds:

$$\Delta_- s\Delta_+ s + \mu\Delta t(\Delta_+ s - \Delta_- s) \geq (\mu\Delta t)^2 + Var \geq \mu\Delta t\Delta_+ s \tag{3.6}$$

We choose

$$s_{(i-,\theta)} = \left\lfloor s_i - \sqrt{(\mu(s_i, t_j, \theta)\Delta t)^2 + Var(s_i, t_j, \theta)} \right\rfloor_s, \tag{3.7}$$

$$s_{(i+,\theta)} = \left\lceil s_i + \sqrt{(\mu(s_i, t_j, \theta)\Delta t)^2 + Var(s_i, t_j, \theta)} \right\rceil_s, \tag{3.8}$$

where $\lceil\rceil_s$ denotes rounding to the nearest greater element from $s_1, s_2, \ldots, s_N$, and $\lfloor\rfloor_s$ denotes rounding to the nearest smaller element from $s_1, s_2, \ldots, s_N$. If such element does not exist, $\lceil x\rceil_s$ and $\lfloor x\rfloor_s$ will return just x. This corresponds to the boundary cases where $x < s_1$ or $x > s_N$. Now it holds

$$\sqrt{(\mu\Delta t)^2 + Var} \leq \Delta_- s, \Delta_+ s \leq \sqrt{(\mu\Delta t)^2 + Var} + \Delta s \tag{3.9}$$

and the first inequality in (3.6) holds. For the second inequality in (3.6) it is sufficient if

$$(\mu\Delta t)^2 + Var \geq \left(\sqrt{(\mu\Delta t)^2 + Var} + \Delta s\right)\mu\Delta t \tag{3.10}$$

For $Var = A(s_i, t_j, \theta)$ with

$$A(s_i, t_j, \theta) = 1/2\left(-(\mu\Delta t)^2 + 2|\mu|\Delta t\Delta s + |\mu|\Delta t\sqrt{(\mu\Delta t)^2 + 4|\mu|\Delta t\Delta s}\right) \tag{3.11}$$

condition (3.10) is fulfilled as equality, for larger Var as inequality. Therefore we set

$$Var = \max\left(\sigma^2(s_i, t_j, \theta)\Delta t, A(s_i, t_j, \theta)\right) \tag{3.12}$$

and compute $s_{(i-,\theta)}$, $s_{(i+,\theta)}$ according to (3.7), (3.8) using this value. We should note, that in (3.11) we replaced μ with $|\mu|$ to cover also the analogous case of a negative drift μ. Now, also the second part of the inequality (3.6), is fulfilled. It holds $Var/\Delta t \to \sigma^2(s_i, t_j, \theta)$ with $\Delta t, \Delta s \to 0$ and it is easy to check that the difference $|Var - \sigma^2(s_i, t_j, \theta)\Delta t|$ is smaller or equal than in the original paper [4]. Following [4], the scheme is then consistent and formula (3.12) is even better then the original version [4], as potentially less artificial diffusion is added.

4. Modification: control-independent stencil. The dependence of the possible states $s_{(i-,\theta)}$, $s_{(i+,\theta)}$ on the control variable θ implies also a dependence of $v^{j+1}_{(i-,\theta)}$, $v^{j+1}_{(i+,\theta)}$ on θ and makes the optimization problem in (3.1) harder to solve. Therefore, our goal now is to find a θ-independent choice of possible states s_{i-}, s_{i+}, while preserving condition (3.6) (and its analogue for negative drift). We will assume a positive drift $\mu(s_i, t_j, \theta)$, the case of negative drift is treated analogously.

Let us define

$$\begin{aligned} W_M &= \max_{\theta\in\Theta} \left(\sigma^2(s_i, t_j, \theta)\Delta t + (\mu(s_i, t_j, \theta)\Delta t)^2\right) \\ &= \sigma^2(s_i, t_j, \theta_M)\Delta t + (\mu(s_i, t_j, \theta_M)\Delta t)^2, \qquad (4.1) \\ E &= \max_{\theta\in\Theta} |\mu(s_i, t_j, \theta)\Delta t|, \qquad (4.2) \\ W_E &= 1/2\left(E^2 + 2\Delta s E + E\sqrt{E^2 + 4\Delta s E}\right). \qquad (4.3) \end{aligned}$$

It holds $W_E = E(\sqrt{W_E} + \Delta s)$ and for all $W \geq W_E : W > E(\sqrt{W} + \Delta s)$. Finally, let us define

$$W = \max(W_E, W_M) \qquad (4.4)$$

and

$$s_{i-} = \left\lfloor s_i - \sqrt{W} \right\rfloor_s \geq s_i - (\sqrt{W} + \Delta s), \qquad (4.5)$$

$$s_{i+} = \left\lceil s_i + \sqrt{W} \right\rceil_s \leq s_i + (\sqrt{W} + \Delta s). \qquad (4.6)$$

Moreover, we redefine also the variance $Var(s_i, t_j, \theta)$:

$$Var = \max\left(\sigma^2\Delta t, \quad |\mu\Delta t|\,(\sqrt{W} + \Delta s) - (\mu\Delta t)^2\right), \qquad (4.7)$$

where $\sigma = \sigma(s_i, t_j, \theta)$, $\mu = \mu(s_i, t_j, \theta)$. It is easy to check that $Var/\Delta t \to \sigma^2$ as $\Delta t, \Delta s \to 0$ and therefore the consistency is preserved. Now it holds

$$\Delta_- s, \;\; \Delta_+ s \geq \sqrt{W} \geq \sqrt{W_M} = \sqrt{\sigma^2(s_i, t_j, \theta_M)\Delta t + (\mu(s_i, t_j, \theta_M)\Delta t)^2}.$$

Therefore it also holds

$$\begin{aligned} \Delta_- s\Delta_+ s + \mu\Delta t(\Delta_- s - \Delta_+ s) &\geq \sigma^2(s_i, t_j, \theta_M)\Delta t + (\mu(s_i, t_j, \theta_M)\Delta t)^2 \\ &\geq \sigma^2(s_i, t_j, \theta)\Delta t + (\mu(s_i, t_j, \theta)\Delta t)^2. \qquad (4.8) \end{aligned}$$

It also holds

$$\Delta_- s\Delta_+ s + \mu\Delta t(\Delta_- s - \Delta_+ s) \geq W \geq E(\sqrt{W} + \Delta s) \geq |\mu\Delta t|(\sqrt{W} + \Delta s). \qquad (4.9)$$

From (4.8) and (4.9) the first inequality of (3.6) holds. The second inequality of (3.6) holds, because

$$Var + (\mu(s_i, t_j, \theta)\Delta t)^2 \geq \mu \Delta t \Delta_+ s. \tag{4.10}$$

Equation (4.10) also holds if we replace $\mu\Delta t\Delta_+ s$ with $|\mu\Delta t|\Delta_- s$ which is important for the case of a negative drift. Now substituting $s_{(i-,\theta)}, s_{(i+,\theta)}$ with s_{i-}, s_{i+} for all values of θ, we get also θ-independent values $v^{j+1}_{(i-,\theta)}, v^{j+1}_{(i+,\theta)}$ (that can be written as v^{j+1}_{i-}, v^{j+1}_{i+}, and the scheme (3.1) still remains consistent and monotone ($p_{(i-,\theta)}, p_i, p_{(i-,\theta)} \geq 0$). In the next section, we employ this *"modified scheme"* to effectively solve the control problem arising in each node in equation (2.3).

5. Analytical solution of the control problem in the modified scheme. According to [4] where also relationship of the Tree-Grid method with the FDMs is discussed, the numerical scheme (3.1) can be written as

$$\begin{aligned} v_i^j =& \max_{\theta\in\Theta} \Big(f_i^j(\theta)\Delta t + (1 + r_i^j(\theta)\Delta t) \\ & \cdot \left(v_i^{j+1} + \mu_i^j(\theta)\Delta_j t D_1 v_i^{j+1} + 1/2 \left(Var_i^j(\theta) + (\mu_i^j(\theta)\Delta_j t)^2 \right) D_2 v_i^{j+1} \right) \Big) \\ :=& \max_{\theta\in\Theta} F_i^j(\theta), \end{aligned} \tag{5.1}$$

where $\mu_i^j(\theta) = \mu(s_i, t_j, \theta)$, $Var_i^j(\theta) = Var(s_i, t_j, \theta)$ and D_1, D_2 are standard finite difference approximations of the first and second derivative on nonuniform grids:

$$D_1 v_i^{j+1} = \left(\frac{s_{i+} - s_i}{s_{i+} - s_{i-}} \right) \frac{v_i^{j+1} - v_{i-}^{j+1}}{s_i - s_{i-}} + \left(\frac{s_i - s_{i-}}{s_{i+} - s_{i-}} \right) \frac{v_{i+}^{j+1} - v_i^{j+1}}{s_{i+} - s_{i-}}, \tag{5.2}$$

$$D_2 v_i^{j+1} = \left(\frac{v_{i+}^{j+1} - v_i^{j+1}}{s_{i+} - s_{i-}} - \frac{v_i^{j+1} - v_{i-}^{j+1}}{s_i - s_{i-}} \right) \Big/ \left(\frac{s_{i+} - s_{i-}}{2} \right). \tag{5.3}$$

Now, under the modification presented in the previous section, s_{i+} and s_{i-} are control-independent and hence also $D_1 v_i^{j+1}$ and $D_2 v_i^{j+1}$ are control independent. Then, for a fixed node (s_i, t_j) the function $F_i^j(\theta)$ is some combination of the functions $f_i^j(\theta)$, $r_i^j(\theta)$, $\mu_i^j(\theta)$ and $Var_i^j(\theta)$. As these functions are typically in closed form, it should be possible to search for the $\max_{\theta\in\Theta} F_i^j(\theta)$ analytically, and it is not necessary to discretize Θ (if it is for example an interval).

However, $Var_i^j(\theta)$ is defined as the maximum of two different functions in (4.7) and therefore may switch its form in several points of the interval Θ. This can make the analytical computation of $\max_{\theta\in\Theta} F_i^j(\theta)$ quite difficult. This problem is not present, if we can assure $Var_i^j(\theta) = \sigma(s_i, t_j, \theta)^2\Delta t$. That condition is typically fulfilled for a relatively large diffusion coefficient σ compared to the drift coefficient μ.

6. Fibonacci algorithm for finding the optimal control. Because of the possible complications arising by the search for the analytical solution of the control problem $\max_{\theta\in\Theta} F_i^j(\theta)$ presented in the previous section, our aim is now to present another, more straightforward approach.

Let us suppose:

1. Θ is a one-dimensional interval.
2. Discount rate $r_i^j(\theta)$ is constant in θ.
3. Increment rate $f_i^j(\theta)$ and drift $\mu_i^j(\theta)$ are linear in θ.
4. Volatility $\sigma^2(s_i, t_j, \theta)$ is convex in θ.

These conditions are fulfilled in many applications. Under these conditions, it is easy to verify, that also $1/2(Var_i^j(\theta) + (\mu_i^j(\theta)\Delta_j t)^2)$ is convex. Then, $F_i^j(\theta)$ is convex or concave and therefore has at most one local (and global) extreme inside the interval Θ and has at least one extreme on the boundary. This makes the problem $\max_{\theta\in\Theta} F_i^j(\theta)$ suitable for the Fibonacci algorithm for maximum search [1]:

Discretize the interval Θ into Φ_n points $\theta_1, \theta_2, \ldots \theta_{\Phi_n}$ where Φ_n is the n-th Fibonacci number.
Set $a = 1$, $b = \Phi_n$, $c_1 = \Phi_{n-2}$, $c_2 = \Phi_{n-1}$
for $j = n-1, n-2, \ldots, 3$ **do**
  if $F_i^j(\theta_{c_1}) > F_i^j(\theta_{c_1})$ **then**
    $b := c_2$;
    $c_2 := c_1$;
    $c_1 := a - 1 + \Phi_{j-2}$;
  else
    $a := c_1$;
    $c_1 := c_2$;
    $c_2 := a - 1 + \Phi_{j-1}$;
  end
end
$\max_{\theta\in\Theta} F_i^j(\theta) \approx \max(F_i^j(\theta_a), F_i^j(\theta_{c_1}), F_i^j(\theta_{c_2}), F_i^j(\theta_b), F_i^j(\theta_1), F_i^j(\theta_{\Phi_n}))$

Algorithm 1: Fibonacci algorithm for finding the optimal control

In the last step of the algorithm we included for testing also values $F_i^j(\theta_1), F_i^j(\theta_{\Phi_n})$ for the case that the function $F_i^j(\theta)$ is convex and the maximum is on the boundary. The computational time of the Fibonacci algorithm is $\mathcal{O}(n) = \mathcal{O}(\log(\Phi_n))$ which is much better than the computational time of the brute-force search approach [4] that is $\mathcal{O}(\Phi_n)$ for Φ_n controls.

7. Numerical experiment. We will test this modified Tree-Grid method with control-independent stencil, and the Fibonacci algorithm for control search on a Passport option pricing problem. This problem is solved with implicit FDM in [6]. In [4], a "capped payoff" is used as terminal condition, and the performance of the implicit FDM and of the Tree-Grid method is compared. Here, we will use the same parameters, terminal and boundary conditions as in [4]. For convenience we repeat here briefly the problem formulation. Passport options are contracts that allow the buyer to run a trading account for a certain amount of time. After the maturity, the buyer

of this contract can keep the profit, or some part of it, however the potential loss will be covered by the seller. The HJB equation for the price of a passport option is

$$\frac{\partial V}{\partial t} + \max_{|\theta|\leq 1}\Big(\frac{\sigma^2}{2}(x-\theta)^2\frac{\partial^2 V}{\partial x^2} + \Big((r-r_c-\gamma)\theta - (r-r_t-\gamma)x\Big)\frac{\partial V}{\partial x} - \gamma V\Big) = 0 \tag{7.1}$$

Here, t is time, V is the option price divided by asset price S and $x = W/S$, where W is wealth accumulated on the trading account. By r, we denote the risk-free interest rate, γ is the dividend rate, r_c is the cost of carry rate, r_t is the interest rate for the trading account and σ is the volatility. The number of shares that the investor holds (control variable) is denoted by θ, and it does not have to be an integer. In this case the seller of the option requires the constraint $|\theta| \leq 1$. We used the same parameter values as in [6]: $r = 0.08, \gamma = 0.03, r_c = 0.12, r_t = 0.05, \sigma = 0.2$.

Computational domain: The maturity of the option will be one year ($T = 1$), the spatial domain will be restricted to $[-3, 4]$. The grid will be uniformly spaced in time, and non-uniformly in space. On the coarsest grid, the time-step size is 0.01. At each refinement, a four-times smaller time-step is taken. Basis for the space grid is vector of nodes:

$$\begin{aligned} S_0 = [&-3, -2, -1.5, -1, -0.75, -0.5, -0.375, -0.25, -0.1875, -0.125, \\ &-0.0625, 0, 0.0625, 0.125, 0.1875, 0.25, 0.375, 0.5, 0.75, 1, 1.5, 2, 3, 4] \end{aligned} \tag{7.2}$$

On the coarsest grid, 15 another nodes are equidistantly inserted between each two neighbouring nodes of S_0. Moreover, at each refinement, a new space-node is inserted between each two neighbouring space-nodes.

Terminal and boundary conditions: As terminal condition we use the "capped" payoff:

$$V(T,x) = V_T(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases},$$

and the Dirichlet boundary conditions:

$$\begin{aligned} &V(x_{min}, t) = BC_L(x_{min}) = 0, \quad V(x_{max}, t) = BC_R(x_{max}) = 1, \\ &x_{min} = -3, \quad x_{max} = 4. \end{aligned}$$

Results: In the Figure 7.1, we illustrate results of numerical simulations. The left figure presents a comparison of error and computational time of the original Tree-Grid method [4] with the modified Tree-Grid method with control-independent stencil for different discretizations. To compute the error, we used as a benchmark solution a solution computed on a very fine grid (having twice as much space and time nodes as the grid at the last refinement level) with an implicit FDM from [6]. In both cases, the control interval was discretized into 9 different controls, and we used brute-force search for the optimal control. We see that the modified Tree-Grid

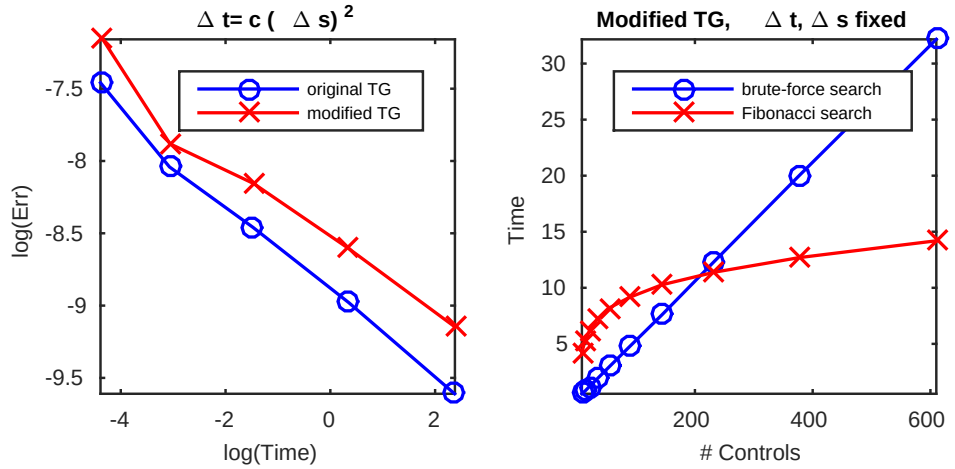


FIG. 7.1. *Left: Comparison of the natural logarithm of estimated absolute error of numerical solution against natural logarithm of computational time (in seconds) for the original Tree-Grid (TG) method and the modified Tree-Grid method with control independent stencil. Brute-force search for optimal control is done in both cases. Right: Computational time (in seconds) of the modified Tree-Grid method with control independent stencil for different number of controls in cases of brute-force search and Fibonacci search for optimal control.*

(TG) method converges, however the original method performs better. This may be of course compensated for finer discretizations of the control interval, if the optimal control is searched analytically or with a Fibonacci search algorithm in the modified scheme.

This illustrates the right figure. Here we used a coarse grid with 24 space-nodes defined by (7.2), 100 (equidistant) time-steps and a varying number of controls. As number of controls (on the x-axis), we used the Fibonacci numbers from the fifth (8) to the 14th (610). We compared the computational time of the modified Tree-Grid method with a brute-force search for control and with a Fibonacci search for control. We observe that for a large number of controls the Fibonacci search performs better due to its logarithmic time-complexity (in contrast to the linear time complexity of brute-force search). We should note that the actual values presented here in the figure are strongly implementation dependent, but they are sufficient in illustrating the proof of concept.

8. Conclusion. In this paper we presented modification of the Tree-Grid method [4] leading to a control independent "stencil" (control independent possible future states s_-^{j+1}, s_+^{j+1}). Due to this modification, it is possible to solve the optimization problem arising in each step analytically. As this approach may be still quite complicated in some cases, we proposed solving the control problem with a Fibonacci search algorithm, if certain conditions on the problem parameters are fulfilled. We analyzed the performance of the original and the modified method using an example of HJB equation from finance, and illustrated the logarithmic time-complexity of the Fibonacci search algorithm that can be applied in the modified scheme. In Section 3, we also improved the strategy of adding artificial diffusion from [4].

REFERENCES

[1] D.E. Ferguson, Fibonaccian searching. *Communications of the ACM*, 3(12):648, 1960.
[2] P. A. Forsyth, and G. Labahn, Numerical methods for controlled Hamilton-Jacobi-Bellman PDEs in finance. *Journal of Computational Finance*, 11(2):1, 2007.
[3] S. Kilianová, and D. Ševčovič, A transformation method for solving the Hamilton-Jacobi-Bellman equation for a constrained dynamic stochastic optimal allocation problem. *The ANZIAM Journal*, 55(01):14–38, 2013.
[4] I. Kossaczký, M. Ehrhardt, and M. Günther, A new convergent explicit Tree-Grid method for HJB equations in one space dimension. Preprint 17/06, University of Wuppertal, to appear in Numerical Mathematics: Theory, Methods and Applications, 2017.
[5] H. Kushner, and P.G. Dupuis, *Numerical methods for stochastic control problems in continuous time*, volume 24. Springer Science & Business Media, 2013.
[6] J. Wang, and P.A. Forsyth, Maximal use of central differencing for Hamilton-Jacobi-Bellman PDEs in finance. *SIAM Journal on Numerical Analysis*, 46(3):1580–1601, 2008.
[7] Jiongmin Yong, and Xun Yu Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.

Proceedings of EQUADIFF 2017
pp. 89–96

MULTIPLE POSITIVE SOLUTIONS FOR A p-LAPLACE CRITICAL PROBLEM (p > 1), VIA MORSE THEORY

GIUSEPPINA VANNELLA*

Abstract. We consider the quasilinear elliptic problem

$$(P_\lambda)\quad \begin{cases} -\Delta_p u = \lambda u^{q-1} + u^{p^*-1} & \text{in } \Omega \\ u > 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where Ω is bounded in $\mathbf{R}^N$, $N \geq p^2$, $1 < p \leq q < p^*$, $p^* = \frac{Np}{N-p}$, $\lambda > 0$ is a parameter.
Denoting by $\mathcal{P}_1(\Omega)$ the Poincaré polynomial of Ω, we state that, for any $p > 1$, there exists $\lambda^* > 0$ such that, for any $\lambda \in (0, \lambda^*)$, either (P_λ) has at least $\mathcal{P}_1(\Omega)$ distinct solutions or, if not, (P_λ) can be approached by a sequence of problems $(P_n)_{n\in\mathbf{N}}$, each having at least $\mathcal{P}_1(\Omega)$ distinct solutions. These results have been proved in [12] only as regards the case $p \geq 2$, while they will be completely proved in the forthcoming work [13] in the case $p \in (1, 2)$.
Note that, when $p \in (1, 2)$, the Euler functional associated to (P_λ) is never C^2, so the approach already used for $p \geq 2$ fails. This problem will be faced exploiting recent results given in [7] and [8].

Key words. Morse theory in Banach spaces, p-laplace equations, critical exponent, critical groups, multiplicity, perturbation results, functionals with lack of smoothness, generalized Morse index

AMS subject classifications. 58E05, 35J92, 35B33, 35B20

1. Introduction.

Let us consider the quasilinear elliptic problem

$$(P_\lambda)\quad \begin{cases} -\Delta_p\, u = \lambda u^{q-1} + u^{p^*-1} & \text{in } \Omega \\ u > 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where Ω is a bounded domain in $\mathbf{R}^N$ with smooth boundary, $N \geq p^2$, $1 < p \leq q < p^*$, $p^* = Np/(N-p)$, $\lambda > 0$ is a parameter.

This problem was introduced by Brezis and Nirenberg in the famous paper [3], in the semilinear case in which $p = q = 2$. Their result was later extended to the quasilinear case $p = q \neq 2$ by Azorero and Peral [2], and Guedda and Veron [14]. Alves and Ding in [1] achieved a multiplicity result for the quasilinear problem (P_λ), under the hypothesis $p \geq 2$. More precisely, they proved that, if $N \geq p^2$ and $2 \leq p \leq q < p^*$, then (P_λ) has at least $cat(\Omega)$ solutions, where $cat(\Omega)$ denotes the Ljusternik-Schnirelmann category of Ω in itself.

Our goal is to exploit Morse theory in order to improve the previous result and extend it to the case $p > 1$.

Solutions to (P_λ) are critical points of the energy functional $I_\lambda : W_0^{1,\,p}(\Omega) \to \mathbf{R}$ defined by

$$I_\lambda(u) = \frac{1}{p}\int_\Omega |\nabla u|^p\, dx - \frac{\lambda}{q}\int_\Omega (u^+)^q\, dx - \frac{1}{p^*}\int_\Omega (u^+)^{p^*}\, dx.$$

*Department of Mechanics, Mathematics and Management (DMMM), Polytechnic University of Bari, Campus Universitario, Via Orabona 4, 70126 Bari, Italy (giuseppina.vannella@poliba.it).

When $p \neq 2$, $W_0^{1,p}(\Omega)$ is a Banach space, but not a Hilbert one, and this brings a lot of problems when trying to apply Morse theory.

Furthermore, when $p \in (1,2)$, I_λ is just a C^1 functional, not C^2.

2. Morse theory: recalls and considerations.

We need to recall some notions about this topic (cf. [5, 6]).
In the sequel, let $\mathbf{K}$ be a field.

DEFINITION 2.1. *For any $B \subset A \subset \mathbf{R}^n$, we denote $\mathcal{P}_t(A,B)$ the Poincaré polynomial of the topological pair (A,B), defined by*

$$\mathcal{P}_t(A,B) = \sum_{k=0}^{+\infty} \dim H^k(A,B)\, t^k.$$

where $H^k(A,B)$ stands for the k-th Alexander-Spanier relative cohomology group of (A,B), with coefficient in $\mathbf{K}$; we also set $H^k(A) = H^k(A,\emptyset)$ so that

$$\mathcal{P}_t(A) = \mathcal{P}_t(A,\emptyset) \tag{2.1}$$

is the Poincaré polynomial of A.

DEFINITION 2.2. *Let Y be a Banach space and J a C^1 functional on Y. Let C be a closed subset of Y. A sequence (u_n) in C is a Palais-Smale sequence for J if $\|J(u_n)\| \leq M$ uniformly in n, while $J'(u_n) \to 0$ as $n \to +\infty$.*
We say that J satisfies (P.S.) on C, if any Palais-Smale sequence in C has a strongly convergent subsequence.
Let $c \in \mathbf{R}$. We say that J satisfies $(P.S.)_c$ if any sequence (u_n) in Y, such that $J(u_n) \to c$ and $J'(u_n) \to 0$ as $n \to +\infty$, has a strongly convergent subsequence.

DEFINITION 2.3. *Let Y be a Banach space, $J \in C^2(Y,\mathbf{R})$ and z a critical point of J. The Morse index of J in z is the supremum of the dimensions of the subspaces of Y on which $J''(z)$ is negative definite. It is denoted by $m(J,z)$. The large Morse index of J in z is the supremum of the dimensions of the subspaces of Y on which $J''(z)$ is negative semidefinite. It is denoted by $m^*(J,z)$.*

DEFINITION 2.4. *Let Y be a Banach space, $J \in C^1(Y,\mathbf{R})$ and z a critical point of J. We call*

$$C_q(J,z) = H^q(J^c, J^c \setminus \{z\})$$

the q-th critical group of J at z, where $c = J(z)$, $q \in \mathbf{N}$ and $H^q(A,B)$ stands for the q-th Alexander-Spanier cohomology group of the pair (A,B) with coefficients in $\mathbf{K}$.
We call multiplicity of z the number

$$\sum_{q=0}^{+\infty} \dim C_q(J,z). \tag{2.2}$$

In the context of a C^1 functional defined on a Banach space, a topological version of Morse relation holds.

THEOREM 2.5. *Let Y be a Banach space, $J \in C^1(Y, \mathbf{R})$ and z a critical point of J. Let $a, b \in \mathbf{R}$ be two regular values for J, with $a < b$. If J satisfies $(P.S.)_c$ condition for any $c \in (a, b)$, and $z_1, \dots, z_l$ are the critical points of J in $J^{-1}(a, b)$, then*

$$\sum_{q=0}^{+\infty} \left(\sum_{j=1}^{l} \dim C_q(J, z_j) \right) t^q = \mathcal{P}_t(J^b, J^a) + (1+t)Q(t)$$

where $Q(t)$ is a formal series with coefficients in $\mathbf{N} \cup \{+\infty\}$.

By the previous theorem, building a suitable barycenter map, in [12] we proved that there exists $\lambda^* > 0$ such that, for any $\lambda \in (0, \lambda^*)$, (P_λ) has at least $\mathcal{P}_1(\Omega)$ solutions, possibly counted with their multiplicities, (see (2.1) and (2.2)).

This is an improvement on the result previously obtained in [1] via Ljusternik-Schnirelmann theory, at least when Ω is a topologically rich domain. In fact, for example, if Ω is obtained by cutting off k holes from an open contractible domain, then $\mathcal{P}_1(\Omega) = k + 1$, while $cat(\Omega) = 2$.

At the same time, we do not know what is the minimum number of *distinct* solutions to (P_λ), as we have no information about the multiplicity of each solution. The situation would have been different if the energy functional had been defined on a Hilbert space.

In fact, if H is a Hilbert space, $J \in C^2(H, \mathbf{R})$, and z is a nondegenerate critical point of J, namely if $J''(z) : H \to H^*$ is invertible, then, using Morse splitting Lemma, a crucial relation between differential and topological information about z holds.

THEOREM 2.6. *If z is a nondegenerate critical point of J, then*

$$C_q(J, z) \cong \mathbf{K} \quad \text{if } q = m(J, z),$$
$$C_q(J, z) = \{0\} \quad \text{if } q \neq m(J, z)$$

being $m(J, z)$ the Morse index of J in z.

Consequently, in a Hilbert space, the multiplicity of any nondegenerate critical point is 1.

Moreover, nondegeneracy assumption holds quite often, as proved by the remarkable result [15] due to Marino and Prodi, in which it is showed that, if the second derivative is a Fredholm operator, an isolated degenerate critical point can be "solved" in a finite number of nondegenerate critical points by a small local perturbation of J.

When we pass to consider a functional $J \in C^2(Y, \mathbf{R})$ defined on a Banach space (not Hilbert) Y, a lot of difficulties arise, in fact:

- it is not clear what can be a reasonable definition of nondegenerate critical point, as it makes no sense to require that the second derivative of J in a critical point is invertible, since a Banach space, in general (and $W_0^{1,p}(\Omega)$, in particular), is not isomorphic to its dual space;
- moreover in [16] it has been proved that the existence of a nondegenerate critical point having finite Morse index, implies the existence of an equivalent Hilbert structure on Y;
- Morse Lemma does not hold;

- $J''(z)$ is not a even Fredholm operator from Y to Y^*, because if $J''(z)$ is a Fredholm operator, then Y is isomorphic to its dual space;
- extensions of Morse lemma of Gromoll-Meyer type can not be applied, and no perturbation argument of Marino-Prodi type can be applied.

In this context, is it possible to relate critical groups to differential features of critical points?

Various experts addressed the issue. We just quote, among them, Uhlenbeck [18], Tromba [17] and Chang [4].

First of all, they tried to give a suitable definition of nondegenerate critical points in Banach spaces, but these definitions were quite involved and not easy to verify. For example, let us see the following one given in [4].

DEFINITION 2.7. *Let X be a Banach space and $f : X \to \mathbf{R}$ a C^2 function. A critical point x_0 of f is said to be s-nondegenerate if*

- *x_0 is isolated;*
- *there exists a neighborhood U of x_0 and an hyperbolic operator $L_{x_0} \equiv L : X \to X$ such that*

  $\langle f''(x_0)Lx, y\rangle = \langle f''(x_0)x, Ly\rangle \quad \forall x, y \in X;$
  $\langle f''(x_0)Lx, x\rangle > 0 \quad \forall\, x \in X \setminus \{0\};$
  $\langle f'(x), L(x - x_0)\rangle > 0 \quad \forall\, x \in f^c \cap (U \setminus \{x_0\})$, *where* $c = f(x_0)$.

Our approach is different.

3. A new approach: I) case $p \geq 2$.

Let us return to consider problem (P_λ). In the case $p \geq 2$, the energy functional I_λ is of class C^2 on $W_0^{1,\,p}(\Omega)$, which is not a Hilbert space when $p > 2$. In [11] we consider a class of functionals including

$$J_{\alpha,f}(u) = \frac{1}{p}\int_\Omega \left(\left(\alpha^2 + |\nabla u|^2\right)^{\frac{p}{2}}\right) dx - \frac{\lambda}{q}\int_\Omega (u^+)^q dx - \frac{1}{p^*}\int_\Omega (u^+)^{p^*} dx - \int f(x)u(x)dx$$

where $\alpha > 0$ and $f \in C^1(\overline{\Omega})$. We give the following new definition of nondegenerate critical point, introduced for the first time in [10].

DEFINITION 3.1. *A critical point u of $J_{\alpha,f}$ is nondegenerate if*

$$J''_{\alpha,f}(u) : W_0^{1,p}(\Omega) \to W^{-1,p'}(\Omega) \quad \textit{is injective.}$$

In [11], using this new notion, we obtain critical groups estimates in the spirit of differential Morse relation (cf. Theorem 2.6). More precisely:

THEOREM 3.2. *Let $p > 2$, $\lambda > 0$, $\alpha > 0$, $f \in C^1(\overline{\Omega})$ and u be a nondegenerate critical point of $J_{\alpha,f}$. Then u is isolated, the Morse index $m = m(J_{\alpha,f}, u)$ is finite and*

$$C_q(J_{\alpha,f}, u) \cong \delta_{q,m}\mathbf{K}.$$

So, in particular, if u is nondegenerate, then its multiplicity is 1.
Moreover, if u is nondegenerate, then u isolated. Hence, from this new definition, we can infer something that Definition 2.7 needed to assume.

We remark that in 1969 Smale, as reported by Uhlenbeck in [18], conjectured that mere injectivity could be enough for developing Morse theory in some Banach settings. So the previous result proves that, as regards $J_{\alpha,f}$, Smale's conjecture is true.

In order to give an interpretation of *multiplicity* for a solution to (P_λ), we need a deep insight into this notion. We do it taking advantage of the following abstract result, proved by Cingolani, Lazzo and Vannella in [9].

THEOREM 3.3. *Let A be an open subset of a Banach space Y. Let I be a C^1 functional on A and $z \in A$ be an isolated critical point of I. Assume that there exists an open neighborhood U of z such that $\overline{U} \subset A$, z is the only critical point of I in $\overline{U}$ and I satisfies the Palais–Smale condition in $\overline{U}$.*
Then, there exists $\mu > 0$ such that, for any $J \in C^1(A, \mathbf{R})$ such that

- $\|I - J\|_{C^1(A)} < \mu$,
- *J satisfies the Palais–Smale condition in $\overline{U}$,*
- *J has a finite number $\{u_1, u_2, \dots, u_m\}$ of critical points in U,*

we have

$$\sum_{j=1}^{m} \mathcal{P}_t(J, u_j) = \mathcal{P}_t(I, z) + (1 + t)Q(t),$$

where $Q(t)$ is a formal series with coefficients in $\mathbf{N} \cup \{+\infty\}$.

So, in particular,

$$\sum_{j=1}^{m} \textit{multiplicity } (J, u_j) \geq \textit{multiplicity } (I, z).$$

In what follows, we say that $\partial\Omega$ satisfies the interior sphere condition if for each $x_0 \in \partial\Omega$ there exists a ball $B_R(x_1) \subset \Omega$ such that $\overline{B_R(x_1)} \cap \partial\Omega = \{x_0\}$.

Due to the previous abstract theorem, considering also that, if B is bounded in $W_0^{1,\,p}(\Omega)$, then

$$\lim_{\alpha \to 0^+} \|J_{\alpha,0} - I_\lambda\|_{C^1(B)} = 0 \quad \text{and} \quad \lim_{\|f\|_{C^1(\overline{\Omega})} \to 0} \|J_{\alpha,f} - J_{\alpha,0}\|_{C^1(B)} = 0,$$

in [12] we proved the following result.

THEOREM 3.4. *Assume that $\partial\Omega$ satisfies the interior sphere condition and that $N \geq p^2$, $2 < p \leq q < p^*$, $p^* = Np/(N-p)$. There exists $\lambda^* > 0$ such that, for any $\lambda \in (0, \lambda^*)$, either (P_λ) has at least $\mathcal{P}_1(\Omega)$ distinct solutions or, if not, for any sequence (α_n), with $\alpha_n > 0$, $\alpha_n \to 0$, there exists a sequence (f_n) with $f_n \in C^1(\overline{\Omega})$, $\|f_n\|_{C^1} \to 0$ such that problem*

$$(P_n) \quad \begin{cases} -div\big((|\nabla u|^2 + \alpha_n)^{(p-2)/2}\nabla u\big) = \lambda u^{q-1} + u^{p^*-1} + f_n & \text{in } \Omega \\ u > 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

has at least $\mathcal{P}_1(\Omega)$ distinct solutions, for n large enough. Moreover, if $p = 2$, the statement holds also if $\alpha_n \geq 0$.

The previous theorem is a sharp interpretation of the multiplicity of a critical point of (P_λ) in terms of approximating elliptic problems. We remark that this approach is new also for the case $p = 2$. Indeed the perturbation results by Marino and Prodi furnish an interpretation of the multiplicity in terms of C^1 locally approximating functionals, which can not be, in general, the Euler functionals of some semilinear problems.

4. A new approach: II) case $1 < p < 2$.

When we consider (P_λ) in the case $p \in (1, 2)$, an additional difficulty arises, as I_λ is just in $C^1(W_0^{1,p}(\Omega), \mathbf{R})$ and not in $C^2(W_0^{1,p}(\Omega), \mathbf{R})$. So it seems even not possible to give a notion of Morse index, and, more in general, the approach used in [12], where $p \geq 2$, fails.

In order to face this further challenge, we used suitable approximations of (P_λ), suggested by recent results given in [7] and [8]. In these papers we considered a class of functionals including

$$J_{\alpha,f}(u) = \frac{1}{p}\int_\Omega \Big(\big(\alpha^2 + |\nabla u|^2\big)^{\frac{p}{2}}\Big)dx - \frac{\lambda}{q}\int_\Omega \big(\alpha + (u^+)^s\big)^{\frac{q}{s}}dx - \frac{1}{p^*}\int_\Omega \big(\alpha + (u^+)^s\big)^{\frac{p^*}{s}}dx - \int f(x)u(x)dx,$$

where $\alpha > 0$, $s > 2$ and $f \in C^1(\overline{\Omega})$. Functionals $J_{\alpha,f}$ are still just in $C^1(W_0^{1,p}(\Omega), \mathbf{R})$ and not in $C^2(W_0^{1,p}(\Omega), \mathbf{R})$.

However, if u_0 is a critical point of $J_{\alpha,f}$, it can be proved that $u_0 \in C^1(\overline{\Omega})$, so we introduce a suitable quadratic form Q_{u_0} defined on $W_0^{1,2}(\Omega)$ (which is embedded in $W_0^{1,p}(\Omega)$, as $p < 2$) by

$$Q_{u_0}(z) = \int_\Omega (\alpha + |\nabla u_0|^2)^{\frac{p-2}{2}}|\nabla z|^2 + (p-2)\int_\Omega (\alpha + |\nabla u_0|^2)^{\frac{p-4}{2}}(\nabla u_0/\nabla z)^2 - \lambda\int_\Omega \left(\big((s-1)\alpha(u_0^+)^{s-2} + (q-1)(u_0^+)^{2s-2}\big)\big(\alpha + (u_0^+)^s\big)^{\frac{q-2s}{s}}\right)z^2 - \int_\Omega \left(\big((s-1)\alpha(u_0^+)^{s-2} + (p^*-1)(u_0^+)^{2s-2}\big)\big(\alpha + (u_0^+)^s\big)^{\frac{p^*-2s}{s}}\right)z^2.$$

Through this quadratic form, we give the following generalized definition of Morse indices.

DEFINITION 4.1. *We denote by $m(J_{\alpha,f}, u_0)$, the (generalized) Morse index of $J_{\alpha,f}$ at u_0, defined as the supremum of the dimensions of the linear subspaces of $W_0^{1,2}(\Omega)$ where Q_{u_0} is negative definite.*

In a similar way, we denote by $m^*(J_{\alpha,f}, u_0)$ the (generalized) large Morse index of $J_{\alpha,f}$ at u_0, defined as the supremum of the dimensions of the linear subspaces of $W_0^{1,2}(\Omega)$ where Q_{u_0} is negative semidefinite.

We remark that

$$m(J_{\alpha,f}, u_0) \leq m^*(J_{\alpha,f}, u_0) < +\infty.$$

Moreover these generalized Morse indices coincide with the usual ones when $p \geq 2$.

In [8] we proved that even in this case a critical group estimate result holds.

THEOREM 4.2. *Let $p \in (1,2)$, $q \in [p, p^*)$, $\lambda > 0$, $f \in C^1(\overline{\Omega})$ and $\alpha > 0$. If u_0 is a critical point of $J_{\alpha,f}$ and*

$$m(J_{\alpha,f}, u_0) = m^*(J_{\alpha,f}, u_0) = m$$

then u_0 is an isolated critical point of $J_{\alpha,f}$ and

$$C_q(J_{\alpha,f}, u_0) \cong \delta_{q,m}\, \mathbf{K}$$

In particular, if $m(J_{\alpha,f}, u_0) = m^*(J_{\alpha,f}, u_0)$, then multiplicity of u_0 is 1.

Exploiting in a suitable way Theorem 3.3, in [13] we prove the following result.

THEOREM 4.3.
Assume that $\partial\Omega$ satisfies the interior sphere condition and that $N \geq p^2$, $p \in (1,2)$, $p \leq q < p^*$, $p^* = Np/(N-p)$. There exists $\lambda^* > 0$ such that, for any $\lambda \in (0, \lambda^*)$, either (P_λ) has at least $\mathcal{P}_1(\Omega)$ distinct solutions or, if not, there exists $s > 2$ such that, for any sequence (α_n), with $\alpha_n > 0$, $\alpha_n \to 0$, there exists a sequence (f_n) with $f_n \in C^1(\overline{\Omega})$, $\|f_n\|_{C^1} \to 0$, such that problem

$$(P_n) \quad \begin{cases} -div\big((|\nabla u|^2 + \alpha_n)^{(p-2)/2}\nabla u\big) & \\ = \lambda u^{s-1}(\alpha_n + u^s)^{\frac{q-s}{s}} + u^{s-1}(\alpha_n + u^s)^{\frac{p^*-s}{s}} + f_n & \text{in } \Omega \\ u > 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

has at least $\mathcal{P}_1(\Omega)$ distinct *solutions, for n large enough.*

REFERENCES

[1] C.O. ALVES AND Y.H. DING, *Multiplicity of positive solutions to a p-Laplacian equation involving critical nonlinearity*, J. Math. Anal. Appl., 279 (2003), pp. 508–521.

[2] J.G. AZORERO AND I. PERAL, *Existence and nonuniqueness for the p-Laplacian: Nonlinear eigenvalues*, Comm. Partial Differential Equations, 12 (1987), pp. 1389–1430.

[3] H. BREZIS, L. NIRENBERG, *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents*, Comm. Pure Appl. Math., 36 (1983), pp. 437–477.

[4] K. C. CHANG, *Morse theory on Banach space and its applications to partial differential equations*, Chinese Ann. Math. Ser. B, 4 (1983), pp. 381–399.

[5] K. C. CHANG *Infinite dimensional Morse theory and multiple solution problems*, Birkhäuser, Boston, MA, 1993.

[6] K. C. CHANG *Morse theory in nonlinear analysis*, in Nonlinear Functional Analysis and Applications to Differential Equations, A. Ambrosetti, K. C. Chang and I. Ekeland, eds., World Scientific Singapore, River Edge, NJ, 1998, pp. 60–101.

[7] S. CINGOLANI, M. DEGIOVANNI, AND G. VANNELLA, *On the critical polynomial of functionals related to p-area $(1 < p < \infty)$ and p-Laplace $(1 < p \leq 2)$ type operators*, Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl., 26 (2015), pp. 49–56.

[8] S. Cingolani, M. Degiovanni, and G. Vannella, *Amann-Zehnder type results for p-Laplace problems*, Ann. Mat. Pura Appl., to appear. doi:10.1007/s10231-017-0694-8.
[9] S. Cingolani, M. Lazzo, G. Vannella, *Multiplicity results for a quasilinear elliptic system via Morse theory*, Commun. Contemp. Math., 7 (2005), pp. 227–249.
[10] S. Cingolani and G. Vannella, *Critical groups computations on a class of Sobolev Banach spaces via Morse index*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 271–292.
[11] S. Cingolani and G. Vannella, *Morse index and critical groups for p-Laplace equations with critical exponents*, Mediterr. J. Math., 3 (2006), pp. 495–512.
[12] S. Cingolani and G. Vannella, *Multiple positive solutions for a critical quasilinear equation via Morse theory*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 26 (2009), pp. 397–413.
[13] S. Cingolani and G. Vannella, *The Brezis-Nirenberg type problem for the p-laplacian ($1 < p < 2$): multiple positive solutions*, in preparation.
[14] M. Guedda and L. Veron, *Quasilinear elliptic equations involving critical Sobolev exponents*, Nonlinear Anal., 13 (1989), pp. 879–902.
[15] A. Marino and G. Prodi, *Metodi perturbativi nella teoria di Morse*, Boll. Un. Mat. Ital., 11 (1975), pp. 1–32.
[16] F. Mercuri and G. Palmieri, *Problems in extending Morse theory to Banach spaces*, Boll. Un. Mat. Ital., 12 (1975), pp. 397–401.
[17] A. J. Tromba, *A general approach to Morse theory*, J. Differential Geometry, 12 (1977), pp. 47–85.
[18] K. Uhlenbeck, *Morse theory on Banach manifolds*, J. Functional Analysis, 10 (1972), pp. 430–445.

Proceedings of EQUADIFF 2017
pp. 97–106

PROPAGATION OF ERRORS IN DYNAMIC ITERATIVE SCHEMES

BARBARA ZUBIK-KOWAL*

Abstract. We consider iterative schemes applied to systems of linear ordinary differential equations and investigate their convergence in terms of magnitudes of the coefficients given in the systems. We address the question of whether the reordering of equations in a given system improves the convergence of an iterative scheme.

Key words. Dynamic iterations, waveform relaxation, Gauss-Seidel schemes, convergence, error bounds

AMS subject classifications. 65L04, 65L20, 65L70

1. Introduction. We investigate convergence of dynamic iteration schemes, see e.g. [2], [4], whose successive iterates are vector functions of the time variable t rather than just a collection of scalars (as in static iterations). The schemes are also called waveform relaxation techniques and their advantages are described e.g. in [3]. The references [3], [2], [4] provide a broad overview on the dynamic iteration schemes (designed for time-dependent initial value problems) versus static iteration schemes (designed for linear algebraic systems). Convergence analyses for dynamic iteration schemes are provided in [3], [2], [4] and the references therein. However, the comparison of different choices of dynamic iteration schemes obtained through a change in the order of the differential equations in a given system is not considered in these references.

In this paper, we show that the choice of the components to be computed using previous iterates and the components to be computed using present iterates affects the efficiency of resulting iterative schemes. To illustrate this, we consider dynamic iterative schemes for the following system

$$\begin{cases} \dfrac{d}{dt}x_1(t) = a_{11}x_1(t) + a_{12}x_2(t) + g_1(t), \\ \dfrac{d}{dt}x_2(t) = a_{21}x_1(t) + a_{22}x_2(t) + g_2(t), \quad t > 0. \end{cases} \tag{1.1}$$

supplemented by the initial conditions

$$x_1(0) = x_{1,0}, \quad x_2(0) = x_{2,0}, \tag{1.2}$$

where $a_{11} \leq 0$, $a_{22} \leq 0$, a_{12}, a_{21}, $x_{1,0}$, $x_{2,0}$ are given real numbers and $g_i(t)$ are given real valued functions.

For (1.1)–(1.2), we consider the following alternative iterative schemes

$$\begin{cases} \dfrac{d}{dt}x_1^{k+1}(t) = a_{11}x_1^{k+1}(t) + a_{12}x_2^{k}(t) + g_1(t), \\ \dfrac{d}{dt}x_2^{k+1}(t) = a_{21}x_1^{k+1}(t) + a_{22}x_2^{k+1}(t) + g_2(t), \quad t > 0. \end{cases} \tag{1.3}$$

*Department of Mathematics, Boise State University, 1910 University Drive, Boise, Idaho 83725, USA, (zubik@math.boisestate.edu).

and

$$\begin{cases} \dfrac{d}{dt}y_2^{k+1}(t) = a_{22}y_2^{k+1}(t) + a_{21}y_1^k(t) + g_2(t), \\ \dfrac{d}{dt}y_1^{k+1}(t) = a_{12}y_2^{k+1}(t) + a_{11}y_1^{k+1}(t) + g_1(t), \quad t > 0. \end{cases} \tag{1.4}$$

supplemented by the initial conditions

$$x_1^k(0) = y_1^k(0) = x_{1,0}, \quad x_2^k(0) = y_2^k(0) = x_{2,0}. \tag{1.5}$$

Scheme (1.3) is initiated from an arbitrary function $x_2^0(t)$ and (1.4) is initiated from another arbitrary function $y_1^0(t)$. Schemes (1.3) and (1.4) are called Gauss-Seidel waveform relaxation schemes see, e.g., [2], [4].

Note that (1.4) is obtained from (1.1) by switching the equations in (1.1). Moreover, schemes (1.3) and (1.4) differ through the fact that scheme (1.3) is slowed down by the previous iterate $x_2^k(t)$ that is multiplied by the coefficient a_{12} while scheme (1.4) is slowed down by the previous iterate $y_1^k(t)$ multiplied by a_{21}.

Suppose that both k^{th} iterates $x_2^k(t)$ and $y_1^k(t)$ give rise to the same error

$$\mathcal{E}^k(t) = x_2^k(t) - x_2(t) = y_1^k(t) - y_1(t).$$

Then, in scheme (1.3), the error $\mathcal{E}^k(t)$ is multiplied by the coefficient a_{12} while, in scheme (1.4), $\mathcal{E}^k(t)$ is multiplied by a_{21}. Let us additionally suppose that a_{12} is much greater than a_{21}, for example, $a_{12} = 10^6$ and $a_{21} = 10^{-6}$. Then, in scheme (1.3), the error $\mathcal{E}^k(t)$ is multiplied by 10^6 (that is, it is significantly enlarged) while, in scheme (1.4), the error $\mathcal{E}^k(t)$ is multiplied by 10^{-6} (so, it is significantly reduced). Therefore, a natural question arises. Which of the schemes (1.3) or (1.4) is faster? In other words, which of the sequences

$$\Big\{\big(x_1^k(t), x_2^k(t)\big)\Big\}_{k=0}^{\infty}, \qquad \Big\{\big(y_1^k(t), y_2^k(t)\big)\Big\}_{k=0}^{\infty} \tag{1.6}$$

converges to $\big(x_1(t), x_2(t)\big)$ faster?

This brings about further questions. Is it better to reorder the differential equations in system (1.1) before the Gauss-Seidel waveform relaxation scheme is applied to get faster convergence of the resulting successive iterates? The goal of the paper is to address the above questions.

2. Convergence analysis involving the spectral radius of a linear integral operator. In this section, we follow [3] and define the linear integral operator

$$\mathcal{K}x(t) = \int_0^t \exp\big((t-s)A\big)Bx(s)ds,$$

where A and B are complex square matrices of the same size. Then system (1.3) is written in the form

$$x^{k+1}(t) = \mathcal{K}x^k(t) + \int_0^t \exp\big((t-s)A\big)g(s)ds + \exp\big((t-s)A\big)x_0$$

with

$$A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & a_{12} \\ 0 & 0 \end{bmatrix}, \quad g(t) = \begin{bmatrix} g_1(t) \\ g_2(t) \end{bmatrix}, \quad x_0 = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}$$

and the spectral radius of $\mathcal{K}$ is written in the form

$$\rho(\mathcal{K}) = \Big|\frac{a_{12}a_{21}}{a_{11}a_{22}}\Big|,$$

see [3]. If

$$\tilde{A} = \begin{bmatrix} a_{22} & 0 \\ a_{12} & a_{11} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 & a_{21} \\ 0 & 0 \end{bmatrix}, \quad \tilde{g}(t) = \begin{bmatrix} g_2(t) \\ g_1(t) \end{bmatrix}, \quad \tilde{x}_0 = \begin{bmatrix} x_{2,0} \\ x_{1,0} \end{bmatrix}$$

then (1.4) is written in the form

$$y^{k+1}(t) = \tilde{\mathcal{K}}y^k(t) + \int_0^t \exp\big((t-s)\tilde{A}\big)\tilde{g}(s)ds + \exp\big((t-s)\tilde{A}\big)\tilde{x}_0,$$

where

$$\tilde{\mathcal{K}}x(t) = \int_0^t \exp\big((t-s)\tilde{A}\big)\tilde{B}x(s)ds,$$

and

$$\rho(\tilde{\mathcal{K}}) = \Big|\frac{a_{12}a_{21}}{a_{11}a_{22}}\Big|.$$

Note that the spectral radius for (1.4) is the same as for (1.3). Therefore, the spectral radius does not give rise to any answer to the question of which of the schemes (1.3) or (1.4) converge faster, though numerical experiments presented in Section 5 illustrate that both schemes converge at different rates, showing that one is more efficient than the other.

3. Explicit formulas for errors and conclusions for improving convergence of iterative schemes. The roles of the parameters in the propagation of errors can be traced more precisely from exact formulas of the errors than from error bounds. In this section, we investigate the roles of the parameters a_{11}, a_{12}, a_{21}, a_{22} in the propagation of errors arising during computations of the sequences of vector functions (1.6) from the alternative numerical schemes (1.3) or (1.4) and address the question of which of the schemes converges faster.

To realize this goal, we investigate exact formulas for the errors

$$e_i^k(t) = x_i(t) - x_i^k(t), \quad i = 1, 2,\ k = 0, 1, \ldots \tag{3.1}$$

and

$$E_i^k(t) = x_i(t) - y_i^k(t), \quad i = 1, 2,\ k = 0, 1, \ldots, \tag{3.2}$$

which are provided through the following theorem.

THEOREM 3.1. *Let*

$$w(\xi) = \sum_{k=1}^{\infty} \frac{\xi^k}{k!} \sum_{i=0}^{k-1} a_{11}^{k-1-i} a_{22}^i. \tag{3.3}$$

Then the errors (3.1) are given by the formulas

$$e_1^k(t_{k+1}) = a_{12}^k a_{21}^{k-1} \int_0^{t_{k+1}} \int_0^{t_k} \dots \int_0^{t_2} e^{a_{11}(t_{k+1}-t_k)} \prod_{j=1}^{k-1} w(t_{j+1}-t_j) \tag{3.4}$$
$$e_2^0(t_1)dt_1dt_2\dots dt_k,$$
$$e_2^k(t_{k+1}) = a_{12}^k a_{21}^k \int_0^{t_{k+1}} \int_0^{t_k} \dots \int_0^{t_2} \prod_{j=1}^{k} w(t_{j+1}-t_j) \tag{3.5}$$
$$e_2^0(t_1)dt_1dt_2\dots dt_k,$$

where $0 < t_1 < t_2 < \dots < t_{k+1}$ *and* $k = 1, 2, \dots$

Proof. From (1.1)–(3.1), we have

$$\begin{cases} \dfrac{d}{dt}e_1^{k+1}(t) = a_{11}e_1^{k+1}(t) + a_{12}e_2^k(t), \\ \dfrac{d}{dt}e_2^{k+1}(t) = a_{21}e_1^{k+1}(t) + a_{22}e_2^{k+1}(t), \end{cases} \tag{3.6}$$

and

$$e_1^k(0) = e_2^k(0) = 0.$$

Therefore, the error $e^k(t) = (e_1^k(t), e_2^k(t))^T$ is given recursively by

$$e^{k+1}(t) = \int_0^t \exp\left((t-s)\begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & a_{12} \\ 0 & 0 \end{bmatrix} e^k(s)ds, \tag{3.7}$$

for $k = 0, 1, 2\dots$. It can be proved by induction that

$$\begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}^k = \begin{bmatrix} a_{11}^k & 0 \\ a_{21}\sum_{j=0}^{k-1} a_{11}^{k-1-j}a_{22}^j & a_{22}^k \end{bmatrix},$$

for $k = 1, 2, \dots$ This leads to

$$\exp\left((t-s)\begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{(t-s)^1}{1!}\begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix} + \dots +$$
$$\frac{(t-s)^i}{i!}\begin{bmatrix} a_{11}^i & 0 \\ a_{21}\sum_{j=0}^{i-1} a_{11}^{i-1-j}a_{22}^j & a_{22}^i \end{bmatrix} + \dots$$
$$= \begin{bmatrix} \sum_{i=0}^{\infty} \frac{a_{11}^i(t-s)^i}{i!} & 0 \\ a_{21}\sum_{i=1}^{\infty}\frac{(t-s)^i}{i!}\sum_{j=0}^{i-1} a_{11}^{i-1-j}a_{22}^j & \sum_{i=0}^{\infty}\frac{a_{22}^i(t-s)^i}{i!} \end{bmatrix}$$
$$= \begin{bmatrix} e^{a_{11}(t-s)} & 0 \\ a_{21}w(t-s) & e^{a_{22}(t-s)} \end{bmatrix},$$

which gives

$$\exp\left((t-s)\begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & a_{12} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & a_{12}e^{a_{11}(t-s)} \\ 0 & a_{12}a_{21}w(t-s) \end{bmatrix}.$$

From this and from (3.7) we have

$$e_1^{k+1}(t) = a_{12}\int_0^t \exp\Big(a_{11}(t-s)\Big)e_2^k(s)ds, \tag{3.8}$$

$$e_2^{k+1}(t) = a_{12}a_{21}\int_0^t w(t-s)e_2^k(s)ds, \tag{3.9}$$

for $k = 0, 1, \ldots$. We now use (3.9) to prove (3.5). It is easy to check that (3.9) for $k = 0$ implies (3.5) for $k = 1$, (here, $t_2 = t$ and $t_1 = s$). Assuming (3.5) holds for some k, we will prove it for $k+1$. From (3.9) we have

$$\begin{aligned} e_2^{k+1}(t_{k+2}) &= a_{12}a_{21}\int_0^{t_{k+2}} w(t_{k+2}-t_{k+1})e_2^k(t_{k+1})dt_{k+1} \\ &= a_{12}^{k+1}a_{21}^{k+1}\int_0^{t_{k+2}} w(t_{k+2}-t_{k+1})\int_0^{t_{k+1}}\int_0^{t_k}\cdots\int_0^{t_2}\prod_{j=1}^{k} w(t_{j+1}-t_j)\times \\ &\quad e_2^0(t_1)dt_1dt_2\ldots dt_kdt_{k+1} \\ &= a_{12}^{k+1}a_{21}^{k+1}\int_0^{t_{k+2}}\int_0^{t_{k+1}}\int_0^{t_k}\cdots\int_0^{t_2}\prod_{j=1}^{k+1} w(t_{j+1}-t_j)e_2^0(t_1)dt_1dt_2\ldots dt_kdt_{k+1}, \end{aligned}$$

which proves (3.5). We now use (3.5) and (3.8) to prove (3.4). From (3.5) and (3.8) we have

$$\begin{aligned} e_1^{k+1}(t_{k+2}) &= a_{12}\int_0^{t_{k+2}} \exp\Big(a_{11}(t_{k+2}-t_{k+1})\Big)e_2^k(t_{k+1})dt_{k+1} \\ &= a_{12}^{k+1}a_{21}^{k}\int_0^{t_{k+2}} \exp\Big(a_{11}(t_{k+2}-t_{k+1})\Big)\int_0^{t_{k+1}}\int_0^{t_k}\cdots\int_0^{t_2}\prod_{j=1}^{k} w(t_{j+1}-t_j)\times \\ &\quad e_2^0(t_1)dt_1dt_2\ldots dt_kdt_{k+1}, \end{aligned}$$

which finishes the proof of the theorem. □

We now apply Theorem 3.1 to (1.4) and compare the errors arising in both schemes, (1.3) and (1.4). Since

$$\sum_{i=0}^{k-1} a_{11}^{k-1-i}a_{22}^i = \sum_{i=0}^{k-1} a_{22}^{k-1-i}a_{11}^i,$$

from (3.4) and (3.5), we have

$$E_2^k(t_{k+1}) = a_{21}^k a_{12}^{k-1} \int_0^{t_{k+1}} \int_0^{t_k} \dots \int_0^{t_2} e^{a_{22}(t_{k+1}-t_k)} \times \prod_{j=1}^{k-1} w(t_{j+1}-t_j) E_1^0(t_1) dt_1 dt_2 \dots dt_k, \tag{3.10}$$

$$E_1^k(t_{k+1}) = a_{21}^k a_{12}^k \int_0^{t_{k+1}} \int_0^{t_k} \dots \int_0^{t_2} \prod_{j=1}^{k} w(t_{j+1}-t_j) \times E_1^0(t_1) dt_1 dt_2 \dots dt_k, \tag{3.11}$$

for $k = 1, 2, \dots$ and $t_{k+1} > 0$.

REMARK. *Note that the starting function $x_1^0(t)$ has no influence on the convergence of the scheme (1.3) and the starting function $y_2^0(t)$ has no influence on the convergence of the scheme (1.4).*

The formulas (3.4)–(3.5), for the scheme (1.3), and the formulas (3.10)–(3.11), for the scheme (1.4), show how the starting error $e_2^{(0)} = x_2 - x_2^{(0)}$ propagates in (1.3) and how the starting error $e_1^{(0)} = x_1 - y_1^{(0)}$ propagates in (1.4).

To choose the faster scheme, we compare (3.4)–(3.5) with (3.10)–(3.11) in the following Corollary.

COROLLARY 3.2. *If*

$$e_2^0 \equiv E_1^0. \tag{3.12}$$

and

$$a_{11} < a_{22} \quad \text{and} \quad |a_{12}| < |a_{21}|, \tag{3.13}$$

then scheme (1.3) converges faster than scheme (1.4). If (3.12) holds and the inequalities in (3.13) are reversed then scheme (1.4) converges faster than scheme (1.3).

Corollary 3.2 shows that if (3.13) holds, then even though (1.3) and (1.4) are initiated with the same error, it propagates differently in both schemes.

Results for higher-dimensional systems are developed in [5].

4. Using parameters in the derivation of error bounds. Applying the variation of constants formula it is easy to obtain the following classical error bound

$$\|e^k(t)\| \le \frac{1}{k!}\Big(\exp(t\|L+D\|)\|U\|\Big)^k \max\{\|e^0(s)\| : 0 \le s \le t\},$$

see [2]. However, sharper error estimation can be obtained by using the exact formulas (3.4) and (3.5).

THEOREM 4.1. *Let*

$$S_k = \frac{1}{k!}\Big(\frac{|a_{12}a_{21}|}{|a_{11}|+|a_{22}|}\Big)^k \int_0^t s^k \exp\Big(s(|a_{11}|+|a_{22}|)\Big) ds \max_{s\in[0,t]} |e_2^0(s)|, \tag{4.1}$$

for $k = 0, 1, \dots$. Then

$$|e_1^k(t)| < |a_{12}| S_{k-1}, \tag{4.2}$$

$$|e_2^k(t)| < \frac{|a_{12}a_{21}|}{|a_{11}|+|a_{22}|} S_{k-1}, \tag{4.3}$$

for $k = 1, 2, \ldots$. *Moreover*

$$\lim_{k\to\infty} S_k = 0. \tag{4.4}$$

Proof. Let w be defined as in Theorem 3.1 and $\alpha = |a_{11}| + |a_{22}|$. Since

$$0 < t_1 < t_2 < \cdots < t_k < t_{k+1}$$

in (3.4) and (3.5), then from the definition (3.3) we have

$$\begin{aligned}
\Big|w(t_{j+1} - t_j)\Big| &\le \sum_{k=1}^{\infty} \frac{(t_{j+1} - t_j)^k}{k!} \sum_{i=0}^{k-1} |a_{11}|^{k-1-i} |a_{22}|^i \\
&\le \sum_{k=1}^{\infty} \frac{(t_{j+1} - t_j)^k}{k!} \sum_{i=0}^{k-1} \binom{k-1}{i} |a_{11}|^{k-1-i} |a_{22}|^i \\
&= \sum_{k=1}^{\infty} \frac{(t_{j+1} - t_j)^k}{k!} \alpha^{k-1} < \frac{1}{\alpha} \exp\Big(\alpha(t_{j+1} - t_j)\Big),
\end{aligned}$$

and

$$\Big|\prod_{j=1}^{k-1} w(t_{j+1} - t_j)\Big| = \prod_{j=1}^{k-1} \Big|w(t_{j+1} - t_j)\Big| < \prod_{j=1}^{k-1} \frac{1}{\alpha} \exp\Big(\alpha(t_{j+1} - t_j)\Big) = \frac{1}{\alpha^{k-1}} \exp\Big(\alpha(t_k - t_1)\Big).$$

This, together with (3.4), implies that

$$\begin{aligned}
|e_1^k(t_{k+1})| \le |a_{12}|^k |a_{21}|^{k-1} &\int_0^{t_{k+1}} \int_0^{t_k} \cdots \int_0^{t_2} \exp\Big(a_{11}(t_{k+1} - t_k)\Big) \\
&\frac{1}{\alpha^{k-1}} \exp\Big(\alpha(t_k - t_1)\Big) |e_2^0(t_1)| dt_1 dt_2 \ldots dt_k \\
\le |a_{12}|^k |a_{21}|^{k-1} \alpha^{1-k} &\max_{0\le\tau\le t_{k+1}} |e_2^0(\tau)| \int_0^{t_{k+1}} \int_0^{t_k} \cdots \int_0^{t_2} \exp\Big(\alpha(t_{k+1} - t_1)\Big) dt_1 dt_2 \ldots dt_k.
\end{aligned}$$

We now show that

$$\int_0^{t_{k+1}} \int_0^{t_k} \cdots \int_0^{t_2} \exp\Big(\alpha(t_{k+1} - t_1)\Big) dt_1 dt_2 \ldots dt_k = \frac{1}{(k-1)!} \int_0^{t_{k+1}} s^{k-1} e^{\alpha s} ds. \tag{4.5}$$

Since

$$\frac{1}{k-1} \int_0^t s^{k-1} e^{\alpha s} ds = \int_0^t e^{\alpha(t-z)} \int_0^z s^{k-2} e^{\alpha s} ds dz,$$

the right-hand side of (4.5) is

$$\frac{1}{(k-1)!} \int_0^{t_{k+1}} t_k^{k-1} e^{\alpha t_k} dt_k = \frac{1}{(k-2)!} \int_0^{t_{k+1}} e^{\alpha(t_{k+1} - t_k)} \int_0^{t_k} t_{k-1}^{k-2} e^{\alpha t_{k-1}} dt_{k-1} dt_k =$$

$$\frac{1}{(k-3)!} \int_0^{t_{k+1}} e^{\alpha(t_{k+1} - t_k)} \int_0^{t_k} e^{\alpha(t_k - t_{k-1})} \int_0^{t_{k-1}} t_{k-2}^{k-3} e^{\alpha t_{k-2}} dt_{k-2} dt_{k-1} dt_k = \ldots$$

$$\frac{1}{1!} \int_0^{t_{k+1}} e^{\alpha(t_{k+1} - t_k)} \int_0^{t_k} e^{\alpha(t_k - t_{k-1})} \int_0^{t_{k-1}} e^{\alpha(t_{k-1} - t_{k-2})} \ldots \int_0^{t_3} t_2 e^{\alpha t_2} dt_2 \ldots dt_{k-2} dt_{k-1} dt_k =$$

$$\int_0^{t_{k+1}} \int_0^{t_k} \int_0^{t_{k-1}} \ldots \int_0^{t_3} t_2 e^{\alpha(t_{k+1} - t_3 + t_2)} dt_2 \ldots dt_{k-2} dt_{k-1} dt_k =$$

$$\int_0^{t_{k+1}} \int_0^{t_k} \int_0^{t_{k-1}} \ldots \int_0^{t_3} (t_3 - t_2) e^{\alpha(t_{k+1} - t_2)} dt_2 \ldots dt_{k-2} dt_{k-1} dt_k.$$

This, together with

$$\int_0^{t_3}(t_3-t_2)e^{\alpha(t_{k+1}-t_2)}dt_2=\int_0^{t_3}(t_3-t_2)\Big(\frac{d}{dt_2}\int_0^{t_2}e^{\alpha(t_{k+1}-t_1)}dt_1\Big)dt_2=$$
$$\Big[(t_3-t_2)\int_0^{t_2}e^{\alpha(t_{k+1}-t_1)}dt_1\Big]_{t_2=0}^{t_2=t_3}+\int_0^{t_3}\int_0^{t_2}e^{\alpha(t_{k+1}-t_1)}dt_1dt_2,$$

implies (4.5) and the proof of (4.2) is finished. The proof of (4.3) is similar. We now show (4.4). Since

$$0\le\frac{S_k}{S_{k-1}}\le\frac{t}{k}\frac{|a_{12}a_{21}|}{|a_{11}|+|a_{22}|},$$

it follows that

$$\lim_{k\to\infty}\frac{S_k}{S_{k-1}}=0,$$

which proves (4.4) and finishes the proof of the theorem. □

5. Numerical experiments. In this section, we present results of numerical experiments for (1.1). We apply the alternative schemes (1.3) and (1.4) to (1.1) and compare their corresponding errors. To integrate (1.3) and (1.4) in time, we apply BDF3 with the step size $h=10^{-3}$. Time integration gives rise to the numerical approximations

$$x_{1,n}^k\approx x_1(t_n),\quad x_{2,n}^k\approx x_2(t_n),$$

for (1.3) and

$$y_{1,n}^k\approx x_1(t_n),\quad y_{2,n}^k\approx x_2(t_n),$$

for (1.4), at the grid-points $t_n=nh$, $n=0,1,\dots$. We measure the errors

$$\max_{i=1,2}\big\{|x_i(t_n)-x_{i,n}^k|\big\},\tag{5.1}$$

$$\max_{i=1,2}\big\{|x_i(t_n)-y_{i,n}^k|\big\},\tag{5.2}$$

and observe the convergence of the schemes (1.3) and (1.4) by plotting (5.1) and (5.2) as functions of $k=0,1,\dots$ for a fixed n.

The errors (5.1) and (5.2) resulting from the different schemes ((5.1) corresponds to (1.3) and (5.2) corresponds to (1.4)) are plotted in Figures 5.1 and 5.2 for $n=1000$. In both figures, the dotted line presents the error (5.1) and the solid line presents the error (5.2).

Figure 5.1 presents the errors for problem (1.1)–(1.2) with $g_1\equiv g_2\equiv 0$ and the initial values $x_{1,0}=0$ and $x_{2,0}=0$. Figure 5.2 presents the errors for problem (1.1)–(1.2) with the initial values $x_{1,0}=1$ and $x_{2,0}=0$ and the inhomogeneous functions $g_1(t)$ and $g_2(t)$ defined in such a way that the exact solution to this problem is $x_1(t)=\cos t$, $x_2(t)=\sin t$, cp. [1, Sec. 203].

Figures 5.1 and 5.2 illustrate that scheme (1.3) converges faster than scheme (1.4). Note that condition (3.13) is satisfied by the scheme whose error is presented by the

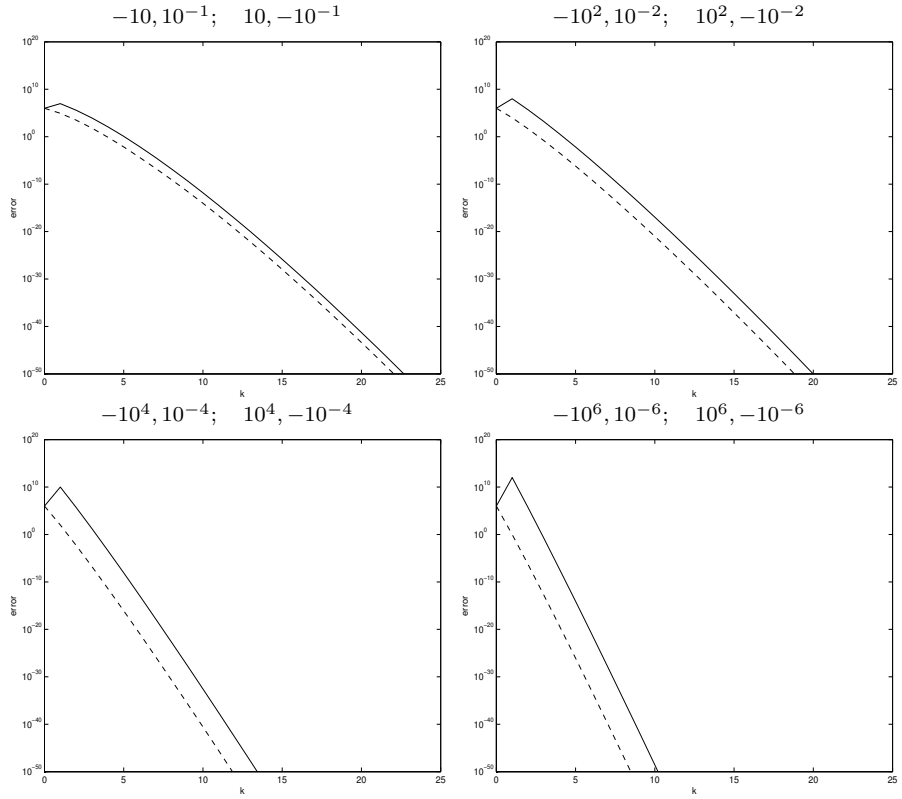


FIG. 5.1. *Numerical errors (5.1) using (1.3) (dotted) and numerical errors (5.2) using (1.4) (solid) for (1.1)–(1.2) in the homogeneous case with* $g_1 \equiv g_2 \equiv 0$.

dotted line and is not satisfied by the scheme whose error is presented by the solid line. This illustrates the conclusion derived in Corollary 3.2 in both homogeneous and non-homogeneous cases.

The errors presented in Figures 5.1 and 5.2 were obtained by running numerical experiments with different coefficients, which we list above each subfigure in the order a_{11}, a_{12}, a_{21}, a_{22}. Note that all these lists of coefficients satisfy condition (3.13) and, therefore, Corollary 3.2 implies that for all these problems (each problem with a different list of a_{ij}) scheme (1.3) convergerges faster than scheme (1.4).

Note that the error (5.1) (presented by the dotted lines), that is,

$$\Big(x_i(t_n) - x_i^k(t_n)\Big) + \Big(x_i^k(t_n) - x_{i,n}^k\Big),$$

is composed of two components: the error $x_i(t_n) - x_i^k(t_n)$ of the iteration and the error $x_i^k(t_n) - x_{i,n}^k$ of the ODE solver. Since integration in t is exact for the problem considered in Figure 5.1, the only non-zero component of (5.1) is the error $e_i^k(t_n)$ of the iteration presented in Figure 5.1. The same conclusion can be derived for the error (1.4) presented by the solid lines.

In Figure 5.2, the error (5.1) (dotted lines) has two non-zero components: the iteration error $e_i^k(t_n)$, which tends to zero as $k \to \infty$, and the time integration error $x_i^k(t_n) - x_{i,n}^k$ which is illustrated by the persistent horizontal lines in Figure 5.2. The same conclusion can be derived for the error (1.4) (solid lines).

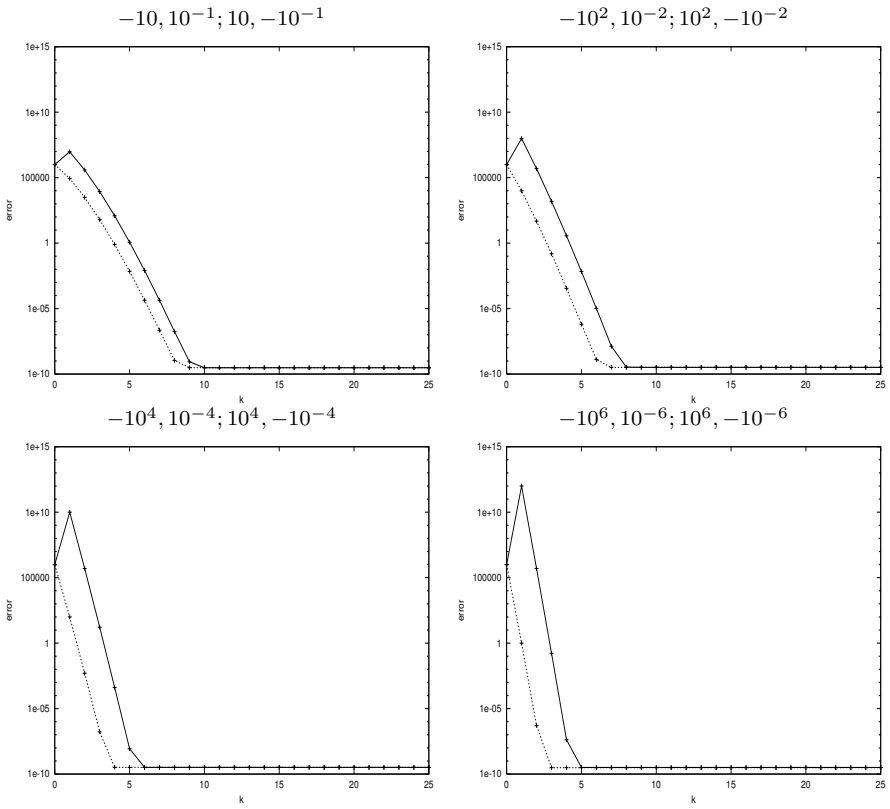


FIG. 5.2. *Numerical errors (5.1) using (1.3) (dotted) and numerical errors (5.2) using (1.4) (solid) for (1.1)–(1.2) in the non-homogeneous case with non-zero source functions $g_1(t)$ and $g_2(t)$.*

6. Concluding remarks and future work. In this paper, we addressed the question of whether the convergence of dynamic iterations depends on the magnitudes of the coefficients multiplied by present and previous iterates. From Sections 3, 4, and 5, we conclude that the order of the differential equations given in a larger dimensional system may slow down or speed up the convergence of the dynamic iterations applied to it. Therefore, we conclude that the order of the equations should be thoughtfully optimized before dynamic iterations are used. The conclusions derived from Sections 3, 4, and 5 give suggestions for choices of present and previous iterates in larger dimensional systems. Our future work [5] addresses the questions raised in this paper in the case of higher-dimensional systems.

REFERENCES

[1] J. C. BUTCHER, *Numerical Methods for Ordinary Differential Equations*, Second edition, John Wiley & Sons, Ltd., Chichester, 2008.
[2] K. BURRAGE, *Parallel and Sequential Methods for Ordinary Differential Equations*, Oxford University Press, Oxford, 1995.
[3] U. MIEKKALA AND O. NEVANLINNA, *Convergence of dynamic iteration methods for initial value problems* SIAM J. Sci. Stat. Comput. 8 (1987), pp. 459–482.
[4] U. MIEKKALA AND O. NEVANLINNA, *Iterative solution of systems of linear differential equations*, Acta Numerica (1996), pp. 259–307.
[5] B. ZUBIK-KOWAL, *Improving the convergence of iterative schemes*, in preparation.

Proceedings of EQUADIFF 2017
pp. 107–116

ON LYAPUNOV STABILITY IN HYPOPLASTICITY*

VICTOR A. KOVTUNENKO†, PAVEL KREJČÍ‡, ERICH BAUER§, LENKA SIVÁKOVÁ¶, AND ANNA V. ZUBKOVA‖

Abstract. We investigate the Lyapunov stability implying asymptotic behavior of a nonlinear ODE system describing stress paths for a particular hypoplastic constitutive model of the Kolymbas type under proportional, arbitrarily large monotonic coaxial deformations. The attractive stress path is found analytically, and the asymptotic convergence to the attractor depending on the direction of proportional strain paths and material parameters of the model is proved rigorously with the help of a Lyapunov function.

Key words. Nonlinear ODE, rate-independent problem, asymptotic behavior, attractor, Lyapunov function, proportional loading, hypoplasticity, granular media

AMS subject classifications. 37B25, 34D20, 74C15

1. Introduction. A rate-independent nonlinear ODE system describing the constitutive stress–strain relation for hypoplastic granular materials like cohesionless soil or broken rock is investigated here. The hypoplastic constitutive equation is of the rate type, incrementally non-linear and based on the hypoplastic concept proposed by Kolymbas [8]. Various physical aspects of hypoplastic models are discussed in engineering literature, e. g., [3, 5, 6, 11, 12, 13]. For mathematical approaches to granular and multiphase media within the variational theory, we refer to [1, 7, 9]. An important feature of the hypoplastic concept is the asymptotic behavior under monotonic proportional loading paths accompanied with a sweeping out of the memory on the initial state. This is a general property also observed in experiments with granular materials. Although for particular monotonic strain paths some numerical simulations and analytical investigations indicate the existence of asymptotic states pointed out, e. g., in [10, Chapter 3.4], a rigorous mathematical proof is missing so far. The main difficulty of developing proper mathematical tools suitable for hypoplastic models is a strongly nonlinear behavior of the corresponding ODE.

For a particular simplified version of a hypoplastic model by Bauer [2] we identify the domain of physical parameters of the model which guarantee that proportional strain paths are stable in the sense of Lyapunov. Our proof of asymptotic stability for unrestricted monotonic deformations is inspired by the rate-independent technique developed in [4].

*This work was supported by the OeAD Scientific & Technological Cooperation (WTZ CZ 01/2016) financed by the Austrian Federal Ministry of Science, Research and Economy (BMWFW) and by the Czech Ministry of Education, Youth and Sports (MŠMT).

†Institute for Mathematics and Scientific Computing, University of Graz, NAWI Graz, Heinrichstr.36, 8010 Graz, Austria, and Lavrent'ev Institute of Hydrodynamics, Siberian Division of the Russian Academy of Sciences, 630090 Novosibirsk, Russia (victor.kovtunenko@uni-graz.at).

‡Institute of Mathematics, Czech Academy of Sciences, Žitná 25, 115 67 Praha 1, Czech Republic (krejci@math.cas.cz).

§Institute of Applied Mechanics, Graz University of Technology, Technikerstr.4, 8010 Graz, Austria (erich.bauer@tugraz.at).

¶Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Praha 6, Czech Republic, (lenka.sivak@gmail.com).

‖Institute for Mathematics and Scientific Computing, University of Graz, NAWI Graz, Heinrichstr.36, 8010 Graz, Austria (anna.zubkova@uni-graz.at).

2. Problem of Lyapunov stability. Consider the general form of a hypoplastic constitutive equation of the Kolymbas type [8] in which the objective stress rate can be stated as follows:

$$\overset{\circ}{\boldsymbol{\sigma}} = \mathbf{L}(\boldsymbol{\sigma}) : \dot{\boldsymbol{\varepsilon}} + \mathbf{N}(\boldsymbol{\sigma})\|\dot{\boldsymbol{\varepsilon}}\|, \tag{2.1}$$

where $\mathbf{L}(\boldsymbol{\sigma})$ is a fourth order tensor and $\mathbf{N}(\boldsymbol{\sigma})$ is a second order tensor depending on the stress $\boldsymbol{\sigma}$. The current Cauchy stress tensor $\boldsymbol{\sigma}$ and the strain rate tensor $\dot{\boldsymbol{\varepsilon}}$ are assumed to be symmetric and of second order. The right-hand side of (2.1) is positively homogeneous of degree one in $\dot{\boldsymbol{\varepsilon}}$. With respect to the Frobenius norm $\|\dot{\boldsymbol{\varepsilon}}\| = \sqrt{\dot{\boldsymbol{\varepsilon}} : \dot{\boldsymbol{\varepsilon}}}$ the constitutive equation is incrementally nonlinear.

Here we consider the particular version of (2.1) by Bauer [2] in a simplified manner:

$$\overset{\circ}{\boldsymbol{\sigma}} = c\big\{a^2\mathrm{tr}(\boldsymbol{\sigma})\dot{\boldsymbol{\varepsilon}} + \frac{1}{\mathrm{tr}(\boldsymbol{\sigma})}(\boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}})\boldsymbol{\sigma} + a(2\boldsymbol{\sigma} - \frac{1}{3}\mathrm{tr}(\boldsymbol{\sigma})I)\|\dot{\boldsymbol{\varepsilon}}\|\big\}, \tag{2.2}$$

with the constitutive constants $c < 0$ and $a > 0$. We emphasize that the second term in the right-hand side of (2.2) is nonlinear in $\boldsymbol{\sigma}$.

For the following investigations we consider cartesian coordinates and we assume coaxial deformations such that $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0$ and $\dot{\varepsilon}_{12} = \dot{\varepsilon}_{13} = \dot{\varepsilon}_{23} = 0$. Then the objective stress rate $\overset{\circ}{\boldsymbol{\sigma}}$ equals to the material time derivative, i.e. the rate $\dot{\boldsymbol{\sigma}}$. For the constitutive equation (2.2) only negative principal stresses are relevant. In this case, using the Voigt notation of the time-dependent 3-vector-valued functions

$$t \mapsto \sigma : \ \mathbb{R}_+ \mapsto \mathbb{R}^3_-, \quad t \mapsto \dot{\varepsilon} : \ \mathbb{R}_+ \mapsto \mathbb{R}^3,$$

the stress components σ_i and strain rate components $\dot{\varepsilon}_i$ can be combined in

$$\sigma = (\sigma_1, \sigma_2, \sigma_3)^\top := (\sigma_{11}, \sigma_{22}, \sigma_{33})^\top,$$
$$\dot{\varepsilon} = (\dot{\varepsilon}_1, \dot{\varepsilon}_2, \dot{\varepsilon}_3)^\top := (\dot{\varepsilon}_{11}, \dot{\varepsilon}_{22}, \dot{\varepsilon}_{33})^\top.$$

Here $^\top$ swaps between rows and columns. We use respective vector notation for the inner product and the associated Euclidean norm:

$$\sigma \cdot \dot{\varepsilon} := \sum_{i=1}^{3} \sigma_i \dot{\varepsilon}_i, \quad \|\dot{\varepsilon}\| := \sqrt{\dot{\varepsilon} \cdot \dot{\varepsilon}}, \quad \mathrm{tr}(\sigma) := \sigma_1 + \sigma_2 + \sigma_3.$$

In this case, $\mathrm{tr}(\boldsymbol{\sigma}) = \mathrm{tr}(\sigma)$ and $\boldsymbol{\sigma} : \dot{\boldsymbol{\varepsilon}} = \sigma \cdot \dot{\varepsilon}$, hence from (2.2) we derive the corresponding matrix equation

$$\dot{\sigma} = c\big(L(\sigma)\dot{\varepsilon} + N(\sigma)\|\dot{\varepsilon}\|\big), \tag{2.3a}$$

with the corresponding 3-by-3 symmetric matrix L depending on σ:

$$L(\sigma) = a^2\mathrm{tr}(\sigma)I + \frac{1}{\mathrm{tr}(\sigma)} \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 \\ \sigma_1\sigma_2 & \sigma_2^2 & \sigma_2\sigma_3 \\ \sigma_1\sigma_3 & \sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, \quad I := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.3b}$$

and the 3-vector

$$N(\sigma) = 2a\sigma - \frac{a}{3}\mathrm{tr}(\sigma)\mathbf{1}, \quad \mathbf{1} := (1, 1, 1)^\top, \tag{2.3c}$$

where we have employed the usual matrix product rule, e.g.:

$$\{L(\sigma)\dot{\varepsilon}\}_{i=1}^{3} = \Big\{\sum_{j=1}^{3} L(\sigma)_{ij}\dot{\varepsilon}_j\Big\}_{i=1}^{3}.$$

We consider here strain paths pointing in one fixed direction. Since the dynamical system (2.3) is rate-independent, without loss of generality we can assume that the loading speed is constant and consider the strain in the form

$$\varepsilon(t) = tU, \quad \|U\| = 1, \quad t \geq 0, \tag{2.4a}$$

along a prescribed unit vector $U = (U_1, U_2, U_3)^\top \in \mathbb{R}^3$. Here t is to be interpreted as a dimensionless monotonically increasing time-like loading parameter. Physically, $\mathrm{tr}(U) < 0$ corresponds to proportional compression and $\mathrm{tr}(U) > 0$ to extension. After inserting (2.4a) in (2.3a), due to $d\,\varepsilon/dt = U$ we get the equivalent system

$$\frac{d}{dt}\sigma = c\{L(\sigma)U + N(\sigma)\}. \tag{2.4b}$$

The ODE (2.4b) for the unknown vector $\sigma(t)$ is considered for $t > 0$, with a prescribed initial condition

$$\sigma(0) = \sigma^0, \tag{2.4c}$$

where $\sigma^0 = (\sigma_1^0, \sigma_2^0, \sigma_3^0)^\top \in \mathbb{R}^3_-$ is a fixed vector. In the next sections we study the asymptotic behavior as $t \nearrow \infty$ of solutions to the Cauchy problem (2.4).

3. Isotropic proportional loading. The strain path (2.4a) is said to be *isotropic* if its direction is parallel to the vector $\mathbf{1}$. In the following we consider two proportional strain paths, i.e. isotropic compression and isotropic extension.

3.1. Isotropic compression. According to (2.4a), the case of the monotonic isotropic compression $\dot{\varepsilon}_1 = \dot{\varepsilon}_2 = \dot{\varepsilon}_3 < 0$ implies that

$$\varepsilon(t) = tU, \quad U = -\frac{1}{\sqrt{3}}\mathbf{1}. \tag{3.1a}$$

In this particular case, due to $\sigma \cdot \dot{\varepsilon} = \sigma \cdot U = -\frac{1}{\sqrt{3}}\mathrm{tr}(\sigma)$, we have

$$L(\sigma)\dot{\varepsilon} = -\frac{1}{\sqrt{3}}\big(a^2\mathrm{tr}(\sigma)\mathbf{1} + \sigma\big),$$

and the system (2.4b) turns out to be linear

$$\frac{d}{dt}\sigma = A\sigma, \quad t > 0, \tag{3.1b}$$

with the 3-by-3 system matrix

$$A = b\mathbb{1} + dI, \quad b = -c\big(\frac{a^2}{\sqrt{3}} + \frac{a}{3}\big), \quad d = c\big(2a - \frac{1}{\sqrt{3}}\big), \tag{3.1c}$$

where $\mathbb{1}$ stands for the 3-by-3 matrix of ones: $\mathbb{1} := \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$.

The characteristic equation for (3.1) is calculated as

$$\det(A-\lambda I)=(d-\lambda)^2(d+3b-\lambda)=0, \tag{3.2a}$$

it has one double and one single roots:

$$\lambda_1=\lambda_2=d,\quad \lambda_3=d+3b=-c\big(\sqrt{3}a^2-a+\frac{1}{\sqrt{3}}\big)=-c\frac{3a^3+\frac{1}{\sqrt{3}}}{\sqrt{3}a+1}. \tag{3.2b}$$

Recalling that $c<0$, we have

$$\lambda_3>0,\quad \lambda_1=\lambda_2<0\quad \text{for } a>\frac{1}{2\sqrt{3}}. \tag{3.2c}$$

From a physical point of view, the lower bound for the constitutive parameter a in (3.2c) implies a restriction of the granular friction angle as discussed in Section 5. For the isotropic case, this condition is necessary and sufficient for the Lyapunov stability as stated in Theorem 3.1.

Let vectors $V^1,V^2,V^3\in\mathbb{R}^3$ form an orthonormal eigenbasis for the eigenvalues from (3.2b) such that $(A-\lambda_i I)V^i=0$, i.e.

$$\big(b\mathbb{1}+(d-\lambda_i)I\big)V^i=0,\quad i=1,2,3. \tag{3.2d}$$

We note that V^1 and V^2 with the corresponding negative eigenvalues λ_1 and λ_2 lie in the deviatoric stress plane due to $\mathrm{tr}(V^i)\mathbf{1}=\mathbb{1}V^i=0$ for $i=1,2$ in (3.2d), thus

$$V^1=\frac{(p,q,-p-q)^\top}{\sqrt{2(p^2+q^2+pq)}},\quad V^2=\frac{(2p+q,-p-2q,-p+q)^\top}{\sqrt{6(p^2+q^2+pq)}},\quad p,q\in\mathbb{R},$$

for example, $V^1=\frac{1}{\sqrt{6}}(1,1,-2)^\top$ and $V^2=\frac{1}{\sqrt{2}}(1,-1,0)^\top$. For the positive eigenvalue λ_3, we normalize the eigenvector perpendicular to the deviatoric stress plan as follows

$$V^3=-\frac{1}{\sqrt{3}}\mathbf{1}, \tag{3.2e}$$

which coincides with U in the isotropic case.

The following exponential stability theorem is a straightforward consequence of the formulas (3.2).

THEOREM 3.1. *(Isotropic compression)*

The solution of the linear problem (3.1) *with initial condition* (2.4c) *for given* $\sigma^0\in\mathbb{R}^3_-$ *is expressed by the explicit formula*

$$\sigma(t)=\sum_{i=1}^{3}(\sigma^0\cdot V^i)V^i e^{\lambda_i t} \tag{3.3a}$$

in terms of the orthonormal eigenbasis (V^1,V^2,V^3) *corresponding to the eigenvalues* $\lambda_1=\lambda_2$ *and* λ_3 *from* (3.2).

If $a>a_\star=\frac{1}{2\sqrt{3}}$ *and* $c<0$, *then the dynamic system* (3.1) *is exponentially stable as* $t\nearrow\infty$ *in the sense of Lyapunov:*

$$\sigma(t)-\sigma_{V^3}(t)=\big(\sigma(0)-\sigma_{V^3}(0)\big)e^{2c\big(a-\frac{1}{2\sqrt{3}}\big)t} \tag{3.3b}$$

with respect to the attractive trajectory along the V^3-axis:

$$\sigma_{V^3}(t) = (\sigma^0 \cdot V^3)V^3 e^{\lambda_3 t}. \tag{3.3c}$$

Conversely, if $a < a_\star$, then $\sigma(t) - \sigma_{V^3}(t)$ diverges according to (3.3b).

A typical configuration is illustrated in Figure 3.1. In the left plot (a), the strain

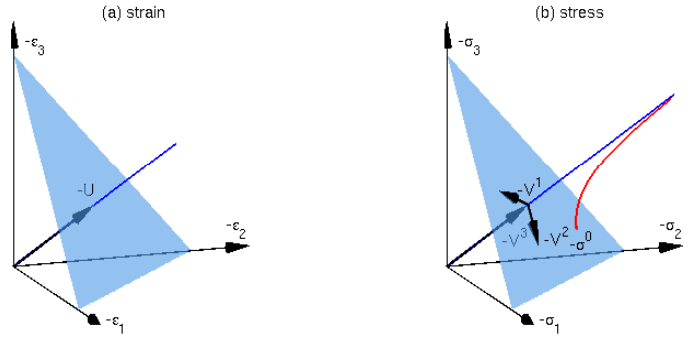


FIG. 3.1. *(a) strain space ; (b) stress space*

path in the direction of -U is depicted in the first octant of the $(-\varepsilon_1, -\varepsilon_2, -\varepsilon_3)$-coordinates. In the right plot (b), in the first octant of the $(-\sigma_1, -\sigma_2, -\sigma_3)$-coordinate system there are presented the stress path attracting the axis along $-V^3$ vector, and the eigenbasis vectors $-V^1$ and $-V^2$ lying in the deviatoric stress plane.

If the initial stress in (2.4c) is isotropic such that $\sigma^0 = sV^3$ with some $s \in \mathbb{R}_+$, then $\sigma(t) = sV^3 e^{\lambda_3 t}$ uniquely solves the system (3.1) under the initial condition (2.4c). This case is the direct consequence of the formula of the solution (3.3a) given in Theorem 3.1. Such $\sigma(t)$ remains isotropic and propagates along the V^3-axis as $t \nearrow \infty$. In the general case when $\sigma^0 \neq sV^3$, an asymptotic stress path attracting the V^3-axis is illustrated in plot (b) of Figure 3.1.

3.2. Isotropic extension. In the case of monotonic isotropic extension, we have $U = \frac{1}{\sqrt{3}}\mathbf{1}$ in (3.1a). It follows that $b = c(\frac{a^2}{\sqrt{3}} - \frac{a}{3})$ and $d = c(2a + \frac{1}{\sqrt{3}})$ in (3.1c). Calculated from (3.2b), the corresponding eigenvalues $\lambda_1 = \lambda_2 = c(2a + \frac{1}{\sqrt{3}})$ and $\lambda_3 = c(\sqrt{3}a^2 + a + \frac{1}{\sqrt{3}})$ are negative since $c < 0$. Therefore, due to the representation formula (3.3a), starting at arbitrary initial stress $\sigma^0 \in \mathbb{R}^3_-$, the stress $\sigma(t)$ decays exponentially to zero as $t \nearrow \infty$ under isotropic extension.

In the next section we investigate the stress path under a non-isotropic loading.

4. Non-isotropic proportional strain paths. For the case of non-isotropic proportional loading, the strain is expressed by formula (2.4a) with an arbitrary unit vector $U \in \mathbb{R}^3$. As mentioned above, this can describe both loading, i.e. compression, and unloading, i.e. extension, tests according to the sign of the trace of U.

The constitutive equation (2.4b) with L and N from (2.3b) and (2.3c) takes the

specific form depending on U as a parameter:

$$\frac{d}{dt}\sigma = c\big\{a\big(aU - \frac{1}{3}\mathbf{1}\big)\mathrm{tr}(\sigma) + \big(2a + \frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big)\sigma\big\}. \tag{4.1a}$$

The right-hand side of (4.1a) is a nonlinear vector function of σ and represents the principal difficulty in the analysis.

We start with the following two consequences of formula (4.1a) which will be used in the sequel. First, after scalar multiplication of (4.1a) with $\mathbf{1}$ using the fact that $\sigma\cdot\mathbf{1} = \mathrm{tr}(\sigma)$, it follows that

$$\frac{d}{dt}\mathrm{tr}(\sigma) = c\big\{a(a\mathrm{tr}(U) + 1)\mathrm{tr}(\sigma) + (\sigma\cdot U)\big\}. \tag{4.1b}$$

Second, multiplying (4.1a) with $-U$ we get

$$\frac{d}{dt}(-\sigma\cdot U) = c\big\{a\big(-a + \frac{1}{3}\mathrm{tr}(U)\big)\mathrm{tr}(\sigma) - \big(2a + \frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big)(\sigma\cdot U)\big\}. \tag{4.1c}$$

Analogously with (3.3c) we look for a linear attractive trajectory of (4.1a) such that $\dfrac{d}{dt}\sigma = \lambda_3\sigma$ which can be expressed in the form

$$\sigma(t) = (\sigma^0\cdot V^3)V^3 e^{\lambda_3 t} \tag{4.2a}$$

with unknown parameters $\lambda_3\in\mathbb{R}$ and nonzero $V^3\in\mathbb{R}^3$ such that $\mathrm{tr}(V^3)\neq 0$. When $\sigma^0\cdot V^3 = 0$, this special case describes the attractive point 0.

Inserting (4.2a) in (4.1b), since $\dfrac{d}{dt}\mathrm{tr}(\sigma) = \lambda_3\mathrm{tr}(\sigma)$ and $\dfrac{\sigma\cdot U}{\mathrm{tr}(\sigma)} = \dfrac{V^3\cdot U}{\mathrm{tr}(V^3)}$ we get

$$\lambda_3 = c\big(a^2\mathrm{tr}(U) + a + \frac{V^3\cdot U}{\mathrm{tr}(V^3)}\big).$$

Substituting this expression together with (4.2a) in (4.1a) such that

$$(a^2\mathrm{tr}(U) - a)V^3 = (a^2U - \frac{1}{3}a\mathbf{1})\mathrm{tr}(V^3),$$

we find a vector $V^3 = a^2U - \dfrac{1}{3}a\mathbf{1}$ with the trace $\mathrm{tr}(V^3) = a^2\mathrm{tr}(U) - a$ satisfying this equality, then $\dfrac{V^3\cdot U}{\mathrm{tr}(V^3)} = \dfrac{-a + \frac{1}{3}\mathrm{tr}(U)}{-a\,\mathrm{tr}(U) + 1}$, and, consequently, after normalization we arrive at

$$\lambda_3 = c\,\frac{\mathrm{tr}(U)\big(-a^3\mathrm{tr}(U) + \frac{1}{3}\big)}{-a\,\mathrm{tr}(U) + 1}, \quad V^3 = \frac{aU - \frac{1}{3}\mathbf{1}}{\sqrt{a^2 - \frac{2}{3}a\,\mathrm{tr}(U) + \frac{1}{3}}}. \tag{4.2b}$$

The above formula is meaningless if $U = \dfrac{1}{3a}\mathbf{1}$, that is, $a = \dfrac{1}{\sqrt{3}}$ and $U = \dfrac{1}{\sqrt{3}}\mathbf{1}$. According to (4.1a), this corresponds to the special case of fully isotropic extension along every stress direction. As well the case $a\,\mathrm{tr}(U) - 1 = 0$ implying $\mathrm{tr}(V^3) = 0$ should be excluded from the consideration.

If $\lambda_3 > 0$ in (4.2b), then $\sigma(t)$ from (4.2a) propagates as $t \nearrow \infty$ exponentially along the V^3-direction. This behavior corresponds to the sketch in Figure 3.1. Otherwise, if $\lambda_3 < 0$, then $\sigma(t) \searrow 0$ which implies unloading.

We note that $-V^3$ in (4.2b) will be directed strictly inside the first octant $\mathbb{R}^3_+$, if a and the direction U of loading in (2.4a) are such that $-aU + \frac{1}{3}\mathbf{1} > 0$, hence $-a\,\mathrm{tr}(U) + 1 > 0$. In this case, for $\sigma(t) \in \mathbb{R}^3_-$ it holds $\sigma^0 \cdot V^3 > 0$.

In particular, for the isotropic compression with $U = -\frac{1}{\sqrt{3}}\mathbf{1}$, from (4.2b) it follows formulas (3.2b) of λ_3 and (3.2e) of V^3.

Next we look for the orthogonal projection of any solution σ of (4.1a) on the V^3-axis, that is

$$\sigma_{V^3}(t) := (\sigma(t) \cdot V^3)V^3. \tag{4.3a}$$

The equivalent form of (4.1a) reads

$$\frac{d}{dt}\sigma = c\big\{a\sqrt{(a^2 - \frac{2}{3}a\,\mathrm{tr}(U) + \frac{1}{3})}\; V^3\mathrm{tr}(\sigma) + \big(2a + \frac{\sigma \cdot U}{\mathrm{tr}(\sigma)}\big)\sigma\big\}, \tag{4.3b}$$

after multiplication (4.3b) with V^3 we derive the following equation

$$\frac{d}{dt}(\sigma_{V^3}) = c\big\{a\sqrt{(a^2 - \frac{2}{3}a\,\mathrm{tr}(U) + \frac{1}{3})}\; V^3\mathrm{tr}(\sigma) + \big(2a + \frac{\sigma \cdot U}{\mathrm{tr}(\sigma)}\big)\sigma_{V^3}\big\} \tag{4.3c}$$

for σ_{V^3} from (4.3a). The subtraction of (4.3c) from (4.3b) provides formula for the difference

$$\frac{d}{dt}(\sigma - \sigma_{V^3}) = c\big(2a + \frac{\sigma \cdot U}{\mathrm{tr}(\sigma)}\big)(\sigma - \sigma_{V^3}). \tag{4.3d}$$

Now we introduce the Lyapunov function $\Lambda : \mathbb{R}_+ \mapsto \mathbb{R}_+$ by

$$\Lambda(t) := \frac{1}{2}\|\sigma(t) - \sigma_{V^3}(t)\|^2, \tag{4.4a}$$

which expresses the distance between the trajectories $\sigma(t)$ and $\sigma_{V^3}(t)$. Differentiating (4.4a) with respect to time and using (4.3d) we get the differential equation for Λ:

$$\frac{d}{dt}\Lambda(t) = 2c\big(2a + \frac{\sigma(t) \cdot U}{\mathrm{tr}(\sigma(t))}\big)\Lambda(t), \quad t > 0. \tag{4.4b}$$

Either negative or positive sign of the factor $2c\big(2a + \frac{\sigma(t) \cdot U}{\mathrm{tr}(\sigma(t))}\big)$ in (4.4b) provides, respectively, either Lyapunov stability or instability of the system. This is the key issue of the following theorem.

THEOREM 4.1. *(Non-isotropic proportional loading) Let $\delta > 0$ be arbitrary fixed, and let $U \in \mathbb{R}^3$ and $a > 0$ be such that $U \neq \frac{1}{3a}\mathbf{1}$ and the following inequalities hold*

$$-a\,\mathrm{tr}(U) + 1 > 0, \quad \big(-2a^2 + \frac{1}{3}\big)\mathrm{tr}(U) + a > 0. \tag{4.5a}$$

For $c<0$ and arbitrary initial data $\sigma^0\in\mathcal{C}_\delta$ lying in the cone

$$\mathcal{C}_\delta := \big\{\sigma\in\mathbb{R}^3_- :\ (-\sigma)\cdot\big(U+(2a+\frac{\delta}{2c})\mathbf{1}\big)>0\big\}, \tag{4.5b}$$

any solution $\sigma(t)$ of the nonlinear problem (4.1a) *endowed with the initial condition* (2.4c) *satisfies the inequality*

$$-2a-\frac{\sigma(t)\cdot U}{\mathrm{tr}(\sigma(t))}<\frac{\delta}{2c},\quad t\geq 0. \tag{4.5c}$$

Moreover, if $\sigma(t)\in\mathbb{R}^3_-$, then $\sigma(t)\in\mathcal{C}_\delta$ for all $t\geq 0$.

In particular, by virtue of (4.5c)*, the dynamical system* (4.1a) *is exponentially stable as $t\nearrow\infty$ in the sense of Lyapunov:*

$$\|\sigma(t)-\sigma_{V^3}(t)\|\leq\|\sigma(0)-\sigma_{V^3}(0)\|e^{-\frac{1}{2}\delta t} \tag{4.5d}$$

with respect to the orthogonal projection $\sigma_{V^3}(t)=(\sigma(t)\cdot V^3)V^3$ on the V^3-axis, where V^3 is determined in formula (4.2b)*.*

Proof. The main challenge is to prove the uniform bound in (4.5c). To do so, we subtract the equation (4.1b), multiplied with $-\dfrac{\sigma\cdot U}{\mathrm{tr}^2(\sigma)}$, from the equation (4.1c), divided by $\mathrm{tr}(\sigma)$, to calculate that

$$\frac{d}{dt}\big(-\frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big)=c\big\{a\big(-a+\frac{1}{3}\mathrm{tr}(U)\big)-a(-a\mathrm{tr}(U)+1)\frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big\}.$$

By adding and subtracting the term $2a^2(-a\mathrm{tr}(U)+1)$ here, this yields

$$\begin{aligned}\frac{d}{dt}\big(-2a-\frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big)&=c\big\{a\big[\big(-2a^2+\frac{1}{3}\big)\mathrm{tr}(U)+a\big]+a(-a\mathrm{tr}(U)+1)\big(-2a-\frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big)\big\}\\&<ca(-a\mathrm{tr}(U)+1)\big(-2a-\frac{\sigma\cdot U}{\mathrm{tr}(\sigma)}\big),\end{aligned}$$

where we have used the second inequality in (4.5a) and $c<0$ for the estimation. The integration of this inequality with respect to t and employing the initial condition (2.4c) results in the following upper bounds

$$-2a-\frac{\sigma(t)\cdot U}{\mathrm{tr}(\sigma(t))}<\big(-2a-\frac{\sigma^0\cdot U}{\mathrm{tr}(\sigma^0)}\big)e^{ca(-a\mathrm{tr}(U)+1)t}\leq -2a-\frac{\sigma^0\cdot U}{\mathrm{tr}(\sigma^0)},$$

when the first inequality in (4.5a) holds. This proves the inequality (4.5c) for the initial data σ^0 chosen such that

$$-2a-\frac{\sigma^0\cdot U}{\mathrm{tr}(\sigma^0)}<\frac{\delta}{2c}.$$

Since $-\sigma^0$ is chosen in the first octant, multiplying the latter inequality with $\mathrm{tr}(\sigma^0)<0$ we obtain the equivalent inequality

$$-\sigma^0\cdot U-\mathrm{tr}(\sigma^0)\big(2a+\frac{\delta}{2c}\big)=(-\sigma^0)\cdot\big(U+(2a+\frac{\delta}{2c})\mathbf{1}\big)>0,$$

which determines the cone in (4.5b).

If (4.5c) holds, then the integration of (4.4) leads immediately to the inequality (4.5d) and completes the proof. □

4.1. Analytic expression of the normalized stress. As a corollary, we consider the normalized stress $\hat{\sigma}$ defined as

$$\hat{\sigma} = \frac{\sigma}{\operatorname{tr}(\sigma)}. \tag{4.6a}$$

Similarly to (4.1) we derive the linear equation

$$\frac{d}{dt}\hat{\sigma} = \frac{1}{\operatorname{tr}(\sigma)}\frac{d}{dt}\sigma - \frac{\hat{\sigma}}{\operatorname{tr}(\sigma)}\frac{d}{dt}(\operatorname{tr}(\sigma)) = ca\big\{(-a\operatorname{tr}(U)+1)\hat{\sigma} + aU - \frac{1}{3}\mathbf{1}\big\}, \tag{4.6b}$$

which can be solved analytically:

$$\hat{\sigma}(t) = \frac{-aU+\frac{1}{3}\mathbf{1}}{-a\operatorname{tr}(U)+1} + \Big(\hat{\sigma}(0) - \frac{-aU+\frac{1}{3}\mathbf{1}}{-a\operatorname{tr}(U)+1}\Big)e^{ca(-a\operatorname{tr}(U)+1)t}. \tag{4.6c}$$

This analytical formula entails directly the next result.

THEOREM 4.2. *(Normalized stress) If $U \in \mathbb{R}^3$ is such that $-a\operatorname{tr}(U)+1 > 0$, then $\hat{\sigma}(t) \to \dfrac{-aU+\frac{1}{3}\mathbf{1}}{-a\operatorname{tr}(U)+1}$ exponentially as $t \to \infty$ according to* (4.6).

From Theorem 4.2 we also conclude that no restriction is imposed on a for proportional loading with $\operatorname{tr}(U) < 0$.

5. Discussion. Let us make a few comments on Theorem 4.1.

Remark 1. According to (4.5d) we can establish that the maximal cone $\mathcal{C}_\delta$ is not less than $\mathcal{C}_0$ when passing $\delta \searrow 0^+$.

Remark 2. Conditions (4.5a) are sufficient for the Lyapunov stability.

Remark 3. If $\operatorname{tr}(U) \le 0$, in particular, when $U \in \mathbb{R}^3_-$ and the vector $-U$ lies in the first octant, then the first inequality in (4.5a) always holds.

Remark 4. If $\operatorname{tr}(U) \le 0$ and $a > \dfrac{1}{2\sqrt{3}}$, then we calculate

$$\big(-2a^2+\frac{1}{3}\big)\operatorname{tr}(U) + a > \frac{1}{6}\big(\operatorname{tr}(U)+\sqrt{3}\big) \ge 0$$

since $|\operatorname{tr}(U)| \le \sqrt{3}\|U\| = \sqrt{3}$ in (2.4a). This suffices the second inequality in (4.5a).

Remark 5. In particular, for $U = -\dfrac{1}{\sqrt{3}}\mathbf{1}$ under isotropic compression, the second inequality in (4.5a) implies that $2\sqrt{3}(a+\frac{1}{\sqrt{3}})(a-\frac{1}{2\sqrt{3}}) > 0$ which holds for $a > \dfrac{1}{2\sqrt{3}}$. The inequality $\dfrac{1}{\sqrt{3}} - 2a < \dfrac{\delta}{2c}$ in (4.5c) determines the cone $\mathcal{C}_\delta$ for σ such that $-\sigma > 0$ and $-\operatorname{tr}(\sigma)(2a - \frac{1}{\sqrt{3}} + \frac{\delta}{2c}) > 0$ in (4.5b). In this particular case, the maximal cone $\mathcal{C}_0$ implies $\sigma < 0$ component-wisely and $-\operatorname{tr}(\sigma)(2a - \frac{1}{\sqrt{3}}) > 0$, that is the first octant when $a > \dfrac{1}{2\sqrt{3}}$. This fact is in accordance with Theorem 3.1.

Remark 6. For the granular friction angle ϕ such that $a = \dfrac{2\sqrt{2}\sin\phi}{\sqrt{3}(3-\sin\phi)}$, from $a > a_\star = \dfrac{1}{2\sqrt{3}} \approx 0.2887$, we have $\phi > \phi_\star$ and calculate the critical value $\sin\phi_\star = \dfrac{3}{1+4\sqrt{2}}$ and $\phi_\star \approx 26.78°$.

The analytical result for the minimum value of parameter a to achieve asymptotic behavior under isotropic straining can also be confirmed with numerical simulation, i.e. by numerical integration of the constitutive equation we could get the same results. For smaller values of a the stress path diverges.

Acknowledgments. V.A.K. and A.V.Z. are supported by the Austrian Science Fund (FWF) project P26147-N26 (PION), V.A.K. thanks the Austrian Academy of Sciences (OeAW) and A.V.Z. thanks IGDK1754 for the support.
P.K. has been supported by the GAČR Grant GA15-12227S and RVO: 67985840.
The authors thank N. Krenn and G.A. Rogacheva for the help in organizing the WTZ Austria - Czech Republic cooperation meetings.

REFERENCES

[1] B. D. Annin, V. A. Kovtunenko and V. M. Sadovskii, *Variational and hemivariational inequalities in mechanics of elastoplastic, granular media, and quasibrittle cracks*, in Analysis, Modelling, Optimization, and Numerical Techniques, G. O. Tost, O. Vasilieva, eds., Springer Proc. Math. Stat., 121 (2015), 49–56.
[2] E. Bauer, *Modelling limit states within the framework of hypoplasticity*, AIP Conf. Proc., 1227, J. Goddard, P. Giovine and J. T. Jenkin, eds., AIP, 2010, pp. 290–305.
[3] E. Bauer and W. Wu, *A hypoplastic model for granular soils under cyclic loading*, Proc. Int. Workshop Modern Approaches to Plasticity, D. Kolymbas, ed., Elsevier, 2010, pp. 247–258.
[4] M. Brokate and P. Krejčí, *Wellposedness of kinematic hardening models in elastoplasticity*, RAIRO Modél. Math. Anal. Numér., (1998), 177–209.
[5] G. Gudehus, *Physical Soil Mechanics*, Springer, Berlin, Heidelberg, 2011.
[6] W. Huang and E. Bauer, *Numerical investigations of shear localization in a micro-polar hypoplastic material*, Int. J. Numer. Anal. Meth. Geomech., 27 (2003), pp. 325–352.
[7] A.M. Khludnev and V.A. Kovtunenko, *Analysis of Cracks in Solids*, WIT-Press, Southampton, Boston, 2000.
[8] D. Kolymbas, *An outline of hypoplasticity*, Arch. Appl. Mech., 61 (1991), pp. 143–151.
[9] V.A. Kovtunenko and A.V. Zubkova, *Mathematical modeling of a discontinuous solution of the generalized Poisson–Nernst–Planck problem in a two-phase medium*, Kinet. Relat. Mod., 11 (2018), pp.119–135.
[10] A. Niemunis, *Extended Hypoplastic Models for Soils*, Habilitation thesis, Ruhr University, Bochum, 2002.
[11] A. Niemunis and I. Herle, *Hypoplastic model for cohesionless soils with elastic strain range*, Mech. Cohes.-Frict. Mat., 2 (1997), pp. 279–299.
[12] B. Svendsen, K. Hutter and L. Laloui, *Constitutive models for granular materials including quasi-static frictional behaviour: toward a thermodynamic theory of plasticity*, Continuum Mech. Therm., 4 (1999), pp. 263–275.
[13] W. Wu, E. Bauer and D. Kolymbas, *Hypoplastic constitutive model with critical state for granular materials*, Mech. Mater., 23 (1996), pp. 45–69.

Proceedings of EQUADIFF 2017
pp. 117–126

EXPONENTIAL CONVERGENCE TO THE STATIONARY MEASURE AND HYPERBOLICITY OF THE MINIMISERS FOR RANDOM LAGRANGIAN SYSTEMS. *

ALEXANDRE BORITCHEV †

Abstract. We consider a class of 1d Lagrangian systems with random forcing in the space-periodic setting:

$$\phi_t + \phi_x^2/2 = F^\omega,\ x \in S^1 = \mathbf{R}/\mathbf{Z}.$$

These systems have been studied since the 1990s by Khanin, Sinai and their collaborators [7, 9, 11, 12, 15]. Here we give an overview of their results and then we expose our recent proof of the exponential convergence to the stationary measure [6]. This is the first such result in a classical setting, i.e. in the dual-Lipschitz metric with respect to the Lebesgue space L_p for finite p, partially answering the conjecture formulated in [11]. In the multidimensional setting, a more technically involved proof has been recently given by Iturriaga, Khanin and Zhang [13].

Key words. Lagrangian dynamics, Random dynamical systems, Invariant measure, Hyperbolicity

AMS subject classifications. 35Q53, 35R60, 35Q35, 37H10, 76M35.

1. Introduction and setting.

We are concerned with 1d random Lagrangian systems of the mechanical type, i.e. of the form:

$$L^\omega(x, v, t) = v^2/2 + F^\omega(x, t),\ x \in S^1 = \mathbf{R}/\mathbf{Z},$$

where $F^\omega(x, t)$ is a smooth function in x and a stationary random process in t (of the kick or white force type: see Section 1.1). The Legendre-Fenchel transform gives us the corresponding Hamiltonian $H^\omega(x, p, t) = p^2/2 - F^\omega(x, t)$, and the Hamilton-Jacobi equation:

$$\phi_t + \phi_x^2/2 = F^\omega. \tag{1.1}$$

Here, we consider only 1-periodic solutions ϕ. In this case the function $u = \phi_x$ satisfies the randomly forced inviscid Burgers equation:

$$u_t + uu_x = (F^\omega)_x,\ x \in S^1 = \mathbf{R}/\mathbf{Z}. \tag{1.2}$$

Note that it is equivalent to consider a solution of (1.2) and a solution of (1.1) defined up to an additive constant. Under the assumptions which are specified below, both of these equations are well-posed and their solutions define Markov processes. The existence and uniqueness of a corresponding stationary measure has been proved by E, Khanin, Mazel and Sinai in the white force case in the seminal work [9]. For more general (multi-d) results, see papers by Khanin and his collaborators [7, 11, 12, 15]. Note that in these papers, there are no explicit estimates on the speed of convergence

*This work was supported by the grants ANR WKBHJ and ANR ISDEEC.

† University of Lyon, CNRS UMR 5208, University Claude Bernard Lyon 1, Institut Camille Jordan, 43 Blvd. du 11 novembre 1918 69622 VILLEURBANNE CEDEX FRANCE (boritchev@math.univ-lyon1.fr).

to the stationary measure; nevertheless, an exponential bound locally in space away from the shocks has been obtained by Bec, Frisch and Khanin in [1]. All these papers use Lagrangian techniques; except in [11] the authors do not consider the equation (1.2) with an additional viscous term νu_{xx}. Note that for $\nu > 0$ there is exponential convergence to the stationary measure, but the speed of convergence is not a priori uniform in ν [16].

In [6], we prove an exponential bound for the speed of convergence to the stationary measure for solutions of (1.2) for $\nu = 0$ in the natural dual-Lipschitz metric with respect to L_p, $p \in [1, \infty)$. This gives a partial answer in the 1d case to the conjecture stated in [11, Section 4]. This bound is the natural SPDE analogue to the results on the exponential convergence of the minimising action curves [7, 9]. The part of the conjecture in [11] which remains open is proving that if we add a positive viscosity coefficient ν, this exponential bound still holds, uniformly in ν.

It is very likely that the estimate we obtain is sharp since it coincides with the optimal one obtained in the generic nonrandom case by Iturriaga and Sanchez-Morgado [14]. Note that the metrics are also optimal since it is impossible to obtain such an estimate in the Lipschitz-dual space corresponding to L_∞. Indeed, solutions of (1.2) are discontinuous with a positive probability.

Finally, we would like to emphasize that our work is part of a series of papers giving a stochastic version of the weak KAM theory developed by Fathi and Mather [10]. In particular, there is a striking correspondence between the scheme of our proof and the one in [14], which follows a general rule: the results which hold in the random case under fairly weak assumptions are similar to the results which hold in the nonrandom case under more stringent genericity assumptions. For more on this subject and the link with the Aubry-Mather theory, see [12].

REMARK 1.1. *Our results extend to the case where ϕ, instead of being periodic in space, satisfies $\phi(x+1) = \phi(x) + b$, $x \in \mathbf{R}$. Indeed, we use the results of [7, 9], which hold for all values of b. Moreover, our results extend to a class of non-mechanical convex in p Hamiltonians of the type $H(p) + F^\omega(t, x)$ with F^ω as above, under assumptions of the Tonelli type [10].*

REMARK 1.2. *After the manuscript [6] has been submitted, Iturriaga, Khanin and Zhang published a preprint containing more general results including also the multidimensional case [13]. Their methods are more technically involved.*

1.1. Random setting. We consider the mechanical Hamilton-Jacobi equation with two different types of additive forcing in the right-hand side and a continuous initial condition ϕ^0. We begin by formulating the assumptions on potentials, which are (except 1.1 (i) where we add an additional assumption for moments of the random variable) the same as in the paper [7]:

ASSUMPTION 1.1. *In the kicked case, we assume that:*

(i) The kicks at integer times j are of the form $F^\omega(j)(x) = \sum_{k=1}^K c_k^\omega(j) F^k(x)$, where F^k are C^∞-smooth potentials on $S^1 = \mathbf{R}/\mathbf{Z}$. The vectors $(c_k^\omega(j))_{1 \le k \le K}$ are independent identically distributed $\mathbf{R}^K$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Their distribution on $\mathbf{R}^K$, denoted by λ, is assumed to be absolutely continuous with respect to the Lebesgue measure, and all of its moments are assumed to be finite.

(ii) The potential 0 *belongs to the support of λ.*

(iii) The mapping from S^1 to $\mathbf{R}^K$ defined by $x \mapsto (F^1(x), ..., F^K(x))$ is an embedding.

ASSUMPTION 1.2. *In the case of the white force potential, we assume that:*

(i) The forcing has the form $F^\omega(x,t) = \sum_{k=1}^K (W_k^\omega)_t(t)F^k(x)$, where F^k are C^∞-smooth potentials on S^1, and $(W_k^\omega)_t$ are independent white noises defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, i.e. time derivatives of independent Wiener processes $W_k^\omega(t)$.

(ii) The mapping from S^1 to $\mathbf{R}^K$ defined by $x \mapsto (F^1(x), ..., F^K(x))$ is an embedding.

REMARK 1.3. *For both types of forcing, our results extend to the case of infinite-dimensional noise, as long as it remains smooth in space (for example independent white noises on each Fourier mode with the amplitude of the noise decreasing exponentially with the wavenumber).*

1.2. Functional spaces and Sobolev norms. Consider an integrable function v on S^1. For $p \in [1, \infty]$, we denote its L_p norm by $|v|_p$. The L_2 norm is denoted by $|v|$, and $\langle \cdot, \cdot \rangle$ stands for the L_2 scalar product. Subindices t and x, which can be repeated, denote partial differentiation with respect to the corresponding variables. We denote by $v^{(m)}$ the m-th derivative of v in the variable x. For brevity, the function $v(t, \cdot)$ is denoted by $v(t)$.

For a nonnegative integer m and $p \in [1, \infty]$, $W^{m,p}$ stands for the Sobolev space of zero mean value functions v on S^1 with finite homogeneous norm $|v|_{m,p} = |v^{(m)}|_p$. In particular, $W^{0,p} = L_p$ for $p \in [1, \infty]$. We will never use Sobolev norms with $m \geq 1$ for non-zero mean functions: in particular, for solutions of (1.1) we will only consider the Lebesgue norms. On the other hand, C^0 (resp. C^∞) will denote the space of C^0-smooth (resp. C^∞-smooth) (not necessarily zero mean value!) functions on S^1.

Since the length of S^1 is 1, we have:

$$|v|_1 \leq |v|_\infty \leq |v|_{1,1} \leq |v|_{1,\infty} \leq \cdots \leq |v|_{m,1} \leq |v|_{m,\infty} \leq \dots$$

We denote by $L_\infty/\mathbf{R}$ the space of functions in L_∞ defined modulo an additive constant endowed with the norm:

$$|u|_{L_\infty/\mathbf{R}} = \inf_{c \in \mathbf{R}} |u - c|_\infty$$

The quantities denoted by K, M or M' are positive constants which only depend on the general features of the system (i.e. the statistical distribution of the forcing): they are nonrandom and do not depend on the initial condition. Moreover the constants $K(p)$ depend on the Lebesgue exponent $p \in [1, \infty)$.

There are two quantities, denoted respectively by $\tilde{C}$ and $\tilde{C}_p$, which are time-independent random variables with all moments finite, which do not depend on the initial condition, but only "pathwise" on the forcing; moreover the quantity $\tilde{C}_p$ depends on the parameter p.

Quantities denoted by C are time-dependent random variables, which also have finite moments and do not depend on the initial condition, but only "pathwise" on the forcing ω. Moreover, these random variables are stationary in the sense that $C(s, \omega)$ coincides with $C(s + t, \theta^t\omega)$ for every t, where θ^t denotes the time shift [9].

We will always denote by $\phi(t, x)$ a solution of (1.1) and by $u(t, x)$ its derivative, which solves (1.2), respectively for initial conditions ϕ^0 and $u^0 = \phi^0_x$. We will denote accordingly the solutions for two initial conditions ϕ^0, $\overline{\phi^0}$.

2. Dynamical objects and stationary measure. Here we introduce the Lagrangian dynamical objects. Note that the results in Sections 2.2 hold under much more general assumptions; nevertheless these hypotheses will be extremely important for the results which will be given in Section 2.3. For more details see [11, 12].

2.1. Lagrangian formulation and minimisers. DEFINITION 2.1. *For a time interval $[s,t]$ and $x,y \in S^1$, we say that a curve $\gamma_{s,t}^{y,x}(\tau)$ is a* ***minimiser*** *if it minimises the action*

$$A(\gamma) = \frac{1}{2}\int_s^t \gamma_t(\tau)^2 d\tau + \sum_{n\in(s,t]} \Big(F(n)(\gamma(n))\Big)$$

in the "kicked" case and the action

$$\begin{aligned} A(\gamma) =& \frac{1}{2}\int_s^t \gamma_t(\tau)^2 d\tau + \int_s^t \left(\gamma_t(\tau)\Big(\frac{\partial G}{\partial x}(\gamma(\tau),s) - \frac{\partial G}{\partial x}(\gamma(\tau),\tau)\Big)\right) d\tau \\ &+ \Big(G(\gamma(t),t) - G(\gamma(t),s)\Big) \end{aligned}$$

in the white force case, respectively, over all absolutely continuous curves γ such that $\gamma(t) = x$ and $\gamma(s) = y$. Here G denotes a primitive in space of F. Note that in the kicked case, minimising curves are linear on intervals $[n, n+1]$ for integer values of n.

DEFINITION 2.2. *For a time interval $[s,t]$, $x \in S^1$ and a continuous function $\phi : S^1 \to \mathbf{R}$, we say that a curve $\gamma_{s,t,\phi}^x(\tau) : [s,t] \to S^1$ is a* ***ϕ-minimiser*** *if it minimises $A(\gamma) + \phi(\gamma(s))$ over all absolutely continuous curves on $[s,t]$ such that $\gamma(t) = x$. In particular, all ϕ-minimisers are minimisers.*

Now we can define the (pathwise) solution to (1.1) for a given $\omega \in \Omega$ and a given continuous initial condition.

DEFINITION 2.3. *For a time interval $[s,t]$ and a continuous initial condition $\phi(s) : S^1 \to \mathbf{R}$, for every ω by definition the (pathwise) solution $\phi : [s,t] \times S^1 \to \mathbf{R}$ of (1.1) is defined using the ω-dependent action A by the Hopf-Lax formula:*

$$\phi(\tau, x) = A(\gamma) + \phi(s, \gamma(s)),\ \tau \in [s,t],$$

where $\gamma = \gamma_{s,\tau,\phi(s)}^x$ is an ω-dependent $\phi(s)$-minimiser defined on $[s,\tau]$ satisfying $\gamma(\tau) = x$.

REMARK 2.4. *It is easy to check that the solution ϕ verifies the semigroup property: in other words, one can define a solution operator*

$$\Sigma_{t_1}^{t_2} :\ \phi(t_1) \mapsto \phi(t_2),\ s \le t_1 \le t_2 \le t,$$

such that for $t_1 \le t_2 \le t_3$, $\Sigma_{t_2}^{t_3} \circ \Sigma_{t_1}^{t_2} = \Sigma_{t_1}^{t_3}$. In particular, for any $\tau \in (s,t)$, the restriction of any $\phi(s)$-minimiser defined on $[s,t]$ to the time interval $[\tau,t]$ is a

$\Sigma_s^\tau \phi(s)$-*minimiser.*

REMARK 2.5. *Note that the solution ϕ is the limit in C^0 of the strong solutions to the equation obtained if we add a viscous term $\nu\phi_{xx}$ to (1.1) and then make ν tend to 0 (see [11]).*

DEFINITION 2.6. *For a time t and a point $x \in S^1$, we say that a curve $\gamma_t^{x,+}(\tau) : [t, +\infty) \mapsto S^1$ is a forward one-sided minimiser if it minimises $A(\gamma)$ over all absolutely continuous curves such that $\gamma(t) = x$ for compact in time perturbations.*

Namely, we require that if for a curve $\tilde{\gamma}$ such that $\tilde{\gamma}(t) = x$ there exists T such that $\tilde{\gamma}(s) \equiv \gamma(s)$ for $s \geq T$, then $A(\gamma) - A(\tilde{\gamma}) \leq 0$ (this difference is well-defined since it is equal to the difference of the actions on the finite interval $[t, T]$).

2.2. Stationary measure and related issues. Here we give a few results which hold under weak assumptions and are sufficient to ensure that the stationary measure corresponding to (1.2) exists and is unique. Up to some natural modifications due to the fact that the forcing is now discrete in time, the convergence estimates can be generalised to the kick force case in 1d [2] and to the multidimensional setting [5].

The flow corresponding to (1.2) induces a Markov process, and then we can define the corresponding semigroup denoted by S_t^*, acting on Borel measures on any L_p, $1 \leq p < \infty$. A *stationary measure* for (1.2) is a Borel probability measure defined on L_p, invariant with respect to S_t^* for every t. A *stationary solution* of (1.2) is a random process v defined for $(t, \omega) \in [0, +\infty) \times \Omega$, satisfying (1.2) and taking values in L_p, such that the distribution of $v(t)$ does not depend on t. This distribution is automatically a stationary measure. Existence of a stationary measure for (1.2) is obtained using uniform bounds for solutions in BV, which is compactly injected into L_p, $p \in [1, \infty)$, and the Bogolyubov-Krylov argument. It is more difficult to obtain uniqueness of a stationary measure, which implies uniqueness for the distribution of a stationary solution.

REMARK 2.7. *The most natural space for our model would be the space $L_\infty/\mathbf{R}$ (for the solutions to the equation (1.1)). Moreover, this is the space in which exponential convergence to the unique stationary solution is proved in the deterministic generic setting in [14]. However, this space is not separable, which makes dealing with the stationary measure a delicate issue.*

DEFINITION 2.8. *Fix $p \in [1, \infty)$. For a continuous function $g : L_p \to \mathbf{R}$, we define its Lipschitz norm as*

$$|g|_{L(p)} := |g|_{Lip} + \sup_{L_p} |g|,$$

where $|g|_{Lip}$ is the Lipschitz constant of g. The set of continuous functions with finite Lipschitz norm will be denoted by $L(p)$.

DEFINITION 2.9. *For two Borel probability measures μ_1, μ_2 on L_p, we denote by $\|\mu_1 - \mu_2\|_{L(p)}^*$ the Lipschitz-dual distance:*

$$\|\mu_1 - \mu_2\|_{L(p)}^* := \sup_{g \in L(p),\ |g|_{L(p)} \leq 1} \Big| \int_{S^1} g d\mu_1 - \int_{S^1} g d\mu_2 \Big|.$$

The following result proved in [2, 3, 5] is, as far as we are aware, the first explicit estimate for the speed of convergence to the stationary measure of the equation (1.2) with an additional viscous term νu_{xx} which is uniform with respect to the viscosity coefficient ν and is formulated in terms of Lebesgue spaces only.

THEOREM 2.10. *There exists $\delta > 0$ such that for every $p \in [1, \infty)$, we have:*

$$\|S_t^*\mu_1 - S_t^*\mu_2\|_{L(p)}^* \leq K(p)t^{-\delta/p}, \qquad t \geq 1,$$

for any probability measures μ_1, μ_2 on L_p.

2.3. Main results and scheme of the proof. Now we are ready to state the main result of the paper.

THEOREM 2.11. *For every $p \in [1, \infty)$, we have:*

$$\|S_t^*\mu_1 - S_t^*\mu_2\|_{L(p)}^* \leq K(p)exp(-M't/p), \qquad t \geq 0, \tag{2.1}$$

for any probability measures μ_1, μ_2 on L_p.

The proof is, in the spirit, similar to the proof of [14, Theorem 1]. In that paper the authors use the objects of the weak KAM theory which do not have any directly available counterparts in our setting. However, there is a straightforward dynamical interpretation of their method.

Namely, consider a mechanical Lagrangian $v^2/2 - V(x)$ such that the *deterministic* potential V is smooth and generic (i.e. it has a unique nongenerate maximum at a unique point y_0). An action-minimising curve on $[0, T]$ remains in a small neighbourhood of y_0 on $[C, T-C]$. We obtain by linearising the Euler-Lagrange equation that at the time $T/2$, all minimisers (independently of the initial condition) are $C\exp(-CT)$-close to y_0, and then we conclude that for any initial conditions ϕ^0, $\overline{\phi^0}$, the solutions of (1.1) at time T are $C\exp(-CT)$-close up to an additive constant.

There are two main ingredients in the proof. On one hand for a given initial condition ϕ^0, the ϕ^0-minimisers corresponding to different final points concentrate exponentially. On the other hand, one-sided minimisers, which are the limits of ϕ_T^0-minimisers on $[0, T]$ as $T \to +\infty$ for *any* set of initial conditions $\left\{\phi_T^0\right\}$, also concentrate exponentially.

Now we introduce some definitions.

The diameter of a closed set Z can be thought of as the minimal length of a closed interval on S^1 containing Z.

DEFINITION 2.12. *Consider a closed subset Z of S^1. Let $a(Z)$ denote the maximal length of a connected component of $S^1 - Z$. We define the diameter of Z as $d(Z) = 1 - a(Z)$.*

DEFINITION 2.13. *For $-\infty < r < s \leq t < +\infty$ and for a fixed function $\phi^0 : S^1 \to \mathbf{R}$, let Ω_{r,s,t,ϕ^0} be the set of points reached, at the time s, by ϕ^0-minimisers on*

$[r,t]$:

$$\Omega_{r,s,t,\phi^0} = \{\gamma^x_{r,t,\phi^0}(s),\ x \in S^1\}.$$

Now we give two key estimates. The first one is - up to notation - [7, Corollary 2.1]. The second one is a forward-in-time version of [9, Lemma 5.6 (a)].

LEMMA 2.14. *We have the inequality:*

$$\sup_{\phi^0 \in C^0} d(\Omega_{0,s,s+s',\phi^0}) \leq C(s') \exp(-Ks').$$

LEMMA 2.15. *We have:*

$$\sup_{\tilde{\gamma}_1,\tilde{\gamma}_2 \in \Gamma} |\tilde{\gamma}_1(t) - \tilde{\gamma}_2(t)| \leq \tilde{C} \exp(-Kt),\ t \geq 0. \tag{2.1}$$

where Γ is the set of all forward one-sided minimisers defined on the time interval $[0,+\infty)$.

COROLLARY 2.16. *Consider an initial condition ϕ^0 and a time $t > 0$. Then for any ϕ^0-minimiser $\gamma : [0,2t] \to S^1$ and any forward one-sided minimiser $\delta :\ [0,+\infty) \to S^1$ we have:*

$$|\gamma(t) - \delta(t)| \leq C(t) \exp(-Kt). \tag{2.2}$$

Proof of Corollary 2.16: Extracting a subsequence of minimisers (for example ϕ^0-minimisers) on $[0,s]$ and taking the limit while letting s go to $+\infty$ (which is possible because of the bounds on the velocity of the minimisers: see Lemma 3.1), one gets a forward one-sided minimiser. In particular, for every ϵ there exists $s(\epsilon) \geq 2t$, a ϕ^0-minimiser $\tilde{\gamma}$ defined on $[0,s]$ and a forward one-sided minimiser $\tilde{\delta}$ on $[0,+\infty)$ such that:

$$|\tilde{\gamma}(t) - \tilde{\delta}(t)| \leq \epsilon.$$

By Lemma 2.15 we have:

$$|\delta(t) - \tilde{\delta}(t)| \leq \tilde{C} \exp(-Kt),$$

and by Lemma 2.14, since the restriction $\tilde{\gamma}|_{[0,2t]}$ is a ϕ^0-minimiser, we have:

$$|\gamma(t) - \tilde{\gamma}(t)| \leq C \exp(-Kt).$$

Combining these three inequalities and then letting ϵ go to 0, we get (2.2).

3. Proof of Theorem 2.11. First we state some useful estimates. For the proof of the first lemma, see [12, Lemma 6].

LEMMA 3.1. *For $t \geq 1$, we have:*

$$\sup_{\phi^0 \in C^0} |\phi_x(t)|_{1,1} \leq C(t); \quad \sup_{s \in [t,t+1], \gamma \in \Gamma} |\gamma_t(s)| \leq C(t),$$

where Γ is the set of minimisers defined on $[0, t+1]$.

LEMMA 3.2. *Consider two minimisers γ_1, γ_2, both defined on $[t, T]$, $T \geq t+1$, and satisfying $\gamma_1(T) = \gamma_2(T)$. If for $\epsilon > 0$ we have $|\gamma_1(t) - \gamma_2(t)| \leq \epsilon$, then we have the following inequality for the actions of the minimisers:*

$$|A(\gamma_1) - A(\gamma_2)| \leq C(t)(\epsilon + \epsilon^2).$$

Proof: By symmetry, it suffices to prove that:

$$A(\gamma_2) \leq A(\gamma_1) + C(\epsilon + \epsilon^2). \tag{3.1}$$

We consider the curve $\tilde{\gamma}_1 : [t, T] \to S^1$ defined by:

$$\tilde{\gamma}_1(s) = \gamma_1(s) + (t+1-s)(\gamma_2(t) - \gamma_1(t)),\ s \in [t, t+1].$$
$$\tilde{\gamma}_1(s) = \gamma_1(s),\ s \in [t+1, T].$$

Using Definition 2.1 and Lemma 3.1, we get:

$$A(\tilde{\gamma}_1) \leq A(\gamma_1) + C(\epsilon + \epsilon^2).$$

On the other hand, since $\tilde{\gamma}_1$ has the same endpoints as the minimiser γ_2, we get $A(\gamma_2) \leq A(\tilde{\gamma}_1)$. Combining these two inequalities yields (3.1).

The proof of the following lemma follows the lines of [14].

LEMMA 3.3. *Consider two solutions ϕ and $\overline{\phi}$ of (1.1) defined on the time interval $[0, +\infty)$. Then we have:*

$$|\phi(t) - \overline{\phi}(t)|_{L_\infty/\mathbf{R}} \leq C(t) \exp(-Mt),\ t \geq 0.$$

Proof of Lemma 3.3: Consider two solutions ϕ and $\overline{\phi}$ to (1.1) corresponding to the same forcing and different initial conditions at time 0. Using Definition 2.3, we get for any $t \geq 1$ and $x \in S^1$:

$$\phi(2t, x) - \overline{\phi}(2t, x) = \phi(t, \gamma_1(t)) + A(\gamma_1|_{[t,2t]}) - \overline{\phi}(t, \gamma_2(t)) - A(\gamma_2|_{[t,2t]}), \tag{3.2}$$

where γ_1 and γ_2 are respectively a ϕ^0- and a $\overline{\phi^0}$-minimiser on $[0, 2t]$ ending at x. By Corollary 2.16, we have:

$$|\gamma_i(t) - y| \leq C \exp(-Kt),\ i = 1, 2, \tag{3.3}$$

where we fix *any* point y such that $y = \delta(t)$ for a one-sided minimiser δ defined on $[0, \infty)$. By Lemma 3.1, this inequality yields that:

$$\begin{aligned} |\phi(t, \gamma_1(t)) - \overline{\phi}(t, \gamma_2(t)) - R| &\leq (|\phi_x(t)|_\infty |\gamma_1(t) - y| + |\overline{\phi}_x(t)|_\infty |\gamma_2(t) - y|) \\ &\leq 2C \exp(-Kt), \end{aligned}$$

where $R = \phi(t, y) - \overline{\phi}(t, y)$. On the other hand, using (3.3), by Lemma 3.2 we get that:

$$|A(\gamma_1|_{[t,2t]}) - A(\gamma_2|_{[t,2t]})| \leq C \exp(-Kt).$$

Therefore, by (3.2), we get:

$$|\phi(2t) - \overline{\phi}(2t)|_{L_\infty/\mathbf{R}} \leq \sup_{x\in S^1} |\phi(2t,x) - \overline{\phi}(2t,x) - R| \leq C\exp(-Kt).$$

This proves the lemma's statement.

COROLLARY 3.4. *Consider two solutions u and $\overline{u}$ of (1.2) defined on the time interval $[0,+\infty)$. Then for any $p>0$ we have:*

$$|u(t) - \overline{u}(t)|_p \leq \tilde{C}_p \exp(-Mt/2p),\ t\geq 0.$$

Proof: This result follows from Lemma 3.3 using the Gagliardo-Nirenberg inequality [8] and Lemma 3.1.

Proof of Theorem 2.11: By the Fubini theorem, it suffices to prove this result in the case when the measures μ_1 and μ_2 are two Dirac measures concentrated at the initial conditions $u^0, \overline{u^0} \in L_p$.

It follows from Corollary 3.4 that if we denote by B the event

$$B = \{\omega\in\Omega \mid |u(t) - \overline{u}(t)|_{L(p)} \geq \exp(-Mt/4p)\},$$

then we have:

$$\mathbf{P}(B) \leq \exp(-Mt/4p)\ \mathbf{E}\ \tilde{C}_p,\ t\geq 0.$$

Now consider a function g defined on L_p which satisfies $|g|_L \leq 1$. We have for $t\geq 0$:

$$\begin{aligned}
&\mathbf{E}\ (|g(u(t)) - g(\overline{u}(t))|_p)\\
&\leq \mathbf{P}(B)\ \mathbf{E}\ (|g(u(t)) - g(\overline{u}(t))|_p \mid B) + \mathbf{P}(\Omega - B)\ \mathbf{E}\ (|g(u(t)) - g(\overline{u}(t))|_p \mid \Omega - B)\\
&\leq 2\mathbf{P}(B) + \mathbf{P}(\Omega - B)\exp(-Mt/4p) \leq (2\mathbf{E}\ \tilde{C}_p + 1)\exp(-Mt/4p).
\end{aligned}$$

REMARK 3.5. *The estimate in Lemma 3.3 is uniform with respect to the initial conditions: in other words, we have*

$$\mathbf{E} \sup_{\phi^0,\overline{\phi^0}\in C^0} |\phi(t) - \overline{\phi}(t)|_{L_\infty/\mathbf{R}} \leq K\exp(-Mt),\ t\geq 0.$$

A similar statement holds for the estimate in Corollary 3.4.

Acknowledgments. I would like to thank P. Bernard, A. Davini, R. Iturriaga, K. Khanin and K. Zhang for helpful discussions.

REFERENCES

[1] J. BEC, U. FRISCH, AND K.KHANIN, *Kicked Burgers turbulence*, Journal of Fluid Mechanics, 416(8) (2000), pp. 239–267.

[2] A. BORITCHEV, *Estimates for solutions of a low-viscosity kick-forced generalised Burgers equation*, Proceedings of the Royal Society of Edinburgh A, 143(2) (2013), pp. 253–268.

[3] A. BORITCHEV, *Sharp estimates for turbulence in white-forced generalised Burgers equation*, Geometric and Functional Analysis, 23(6) (2013), pp. 1730–1771.

[4] A. BORITCHEV, *Erratum to: Multidimensional Potential Burgers Turbulence*, Communications in Mathematical Physics, 344(1) (2016), pp. 369–370, see [5].

[5] A. Boritchev, *Multidimensional Potential Burgers Turbulence*, Communications in Mathematical Physics, 342 (2016), pp. 441–489, with erratum: see [4].
[6] A. Boritchev, *Exponential convergence to the stationary measure for a class of 1D Lagrangian systems with random forcing*, accepted to Stochastic and Partial Differential Equations: Analysis and Computations.
[7] A. Boritchev and K. Khanin, *On the hyperbolicity of minimizers for 1D random Lagrangian systems*, Nonlinearity, 26(1) (2013), pp. 65–80.
[8] C. Doering and J. D. Gibbon, *Applied analysis of the Navier-Stokes equations*, Cambridge Texts in Applied Mathematics, Cambridge University Press, 1995.
[9] Weinan E, K. Khanin, A. Mazel, and Ya. Sinai. , *Invariant measures for Burgers equation with stochastic forcing*, Annals of Mathematics, 151 (2000), pp. 877–960.
[10] A. Fathi, *Weak KAM Theorem in Lagrangian Dynamics, preliminary version*, 2005.
[11] D. Gomes, R. Iturriaga, K. Khanin, and P. Padilla, *Viscosity limit of stationary distributions for the random forced Burgers equation*, Moscow Mathematical Journal, 5 (2005), pp. 613–631.
[12] R. Iturriaga and K. Khanin, *Burgers turbulence and random Lagrangian systems*, Communications in Mathematical Physics, 232:3 (2003), pp. 377–428.
[13] R. Iturriaga, K. Khanin, and K. Zhang, *Exponential convergence of solutions for random Hamilton-Jacobi equation*, Preprint, arxiv: 1703.10218, 2017.
[14] R. Iturriaga and H. Sanchez-Morgado, *Hyperbolicity and exponential convergence of the Lax-Oleinik semigroup*, Journal of Differential Equations, 246(5) (2009), pp. 1744 – 1753.
[15] K. Khanin and K. Zhang, *Hyperbolicity of minimizers and regularity of viscosity solutions for random Hamilton-Jacobi equations*, Communications in Mathematical Physics, 355 (2017), pp. 803.
[16] Y. Sinai, *Two results concerning asymptotic behavior of solutions of the Burgers equation with force*, Journal of Statistical Physics, 64, 1991, pp. 1–12,.

Proceedings of EQUADIFF 2017
pp. 127–136

ANALYSIS OF THE FEM AND DGM FOR AN ELLIPTIC PROBLEM WITH A NONLINEAR NEWTON BOUNDARY CONDITION *

MILOSLAV FEISTAUER †, ONDŘEJ BARTOŠ , FILIP ROSKOVEC , AND ANNA-MARGARETE SÄNDIG‡

Abstract. The paper is concerned with the numerical analysis of an elliptic equation in a polygon with a nonlinear Newton boundary condition, discretized by the finite element or discontinuous Galerkin methods. Using the monotone operator theory, it is possible to prove the existence and uniqueness of the exact weak solution and the approximate solution. The main attention is paid to the study of error estimates. To this end, the regularity of the weak solution is investigated and it is shown that due to the boundary corner points, the solution looses regularity in a vicinity of these points. It comes out that the error estimation depends essentially on the opening angle of the corner points and on the parameter defining the nonlinear behaviour of the Newton boundary condition. Theoretical results are compared with numerical experiments confirming a nonstandard behaviour of error estimates.

Key words. elliptic equation, nonlinear Newton boundary condition, monotone operator method, finite element method, discontinuous Galerkin method, regularity and singular behaviour of the solution, error estimation

AMS subject classifications. 65N15, 65N30

1. Introduction. Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain with boundary $\partial\Omega$. We consider a boundary value problem with a non-linear Newton boundary condition: find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.1}$$

$$\frac{\partial u}{\partial n} + \kappa |u|^{\alpha} u = \varphi \quad \text{on } \partial\Omega, \tag{1.2}$$

with given functions $f : \Omega \to \mathbb{R}$, $\varphi : \partial\Omega \to \mathbb{R}$ and constants $\kappa > 0$, $\alpha \geq 0$.

Such boundary value problems have applications in science and engineering. We can mention modelling of electrolysis of aluminium with the aid of the stream function ([11]), radiation heat transfer problem ([9], [10]) or nonlinear elasticity ([6], [7]). For example, by [2] our problem describes deformation of a flat plate with a nonlinear elastic support on the boundary.

In this paper we are concerned with the application of the finite element method (FEM) and the discontinuous Galerkin method (DGM) applied to the numerical solution of problem (1.1)-(1.2). Main attention is paid to a survey of error estimation. Detailed results are contained in the thesis [3] and the forthcoming paper [5].

2. Weak solution. In what follows we use the standard notation $L^p(\omega)$, $W^{k,p}(\omega)$, $H^k(\omega)$ for the Lebesgue and Sobolev spaces over a set ω. See, e.g., [12].

*This work was supported by Grant No. 17-01747S of the Czech Science Foundation.

†Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Praha 8, Czech Republic (feist@karlin.mff.cuni.cz, Ondra.Bartosh@seznam.cz, roskovec@gmail.com).

‡IANS, University Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany (Anna.Saendig@mathematik.uni-stuttgart.de)

Suppose that $f \in L^2(\Omega)$, $\varphi \in L^2(\partial\Omega)$. We introduce the following forms for u, $v \in H^1(\Omega)$:

$$\begin{aligned} &b(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad d(u,v) = \kappa \int_{\partial\Omega} |u|^\alpha uv \,\mathrm{d}S, \quad L^\Omega(v) = \int_\Omega fv \,\mathrm{d}x, \\ &L^{\partial\Omega}(v) = \int_{\partial\Omega} \varphi v \,\mathrm{d}S, \quad L(v) = L^\Omega(v) + L^{\partial\Omega}(v), \quad A(u,v) = b(u,v) + d(u,v). \end{aligned} \tag{2.1}$$

DEFINITION 2.1. *We say that a function $u : \Omega \to \mathbb{R}$ is a weak solution of problem* (1.1)-(1.2)*, if*

$$u \in H^1(\Omega), \quad A(u,v) = L(v) \quad \forall v \in H^1(\Omega). \tag{2.2}$$

Let us note that for $u, v \in H^1(\Omega)$

$$A(u, u-v) - A(v, u-v) = \int_\Omega |\nabla u - \nabla v|^2 \mathrm{d}x + \kappa \int_{\partial\Omega} (|u|^\alpha u - |v|^\alpha v)(u-v) \,\mathrm{d}S. \tag{2.3}$$

The next section will be devoted to to the analysis of the numerical solution of problem (2.2) by the finite element method and the discontinuous Galerkin method. In the analysis of error estimation, the regularity of the weak solution plays an important role. In [5], the following result is proven.

THEOREM 2.2. *Let $u \in H^1(\Omega)$ be a weak solution of* (2.2) *in a polygonal domain Ω. By ω_0 we denote the largest interior angle at corners on the boundary. Let $f \in L^q(\Omega)$, $\varphi \in W^{1-1/q,q}(\partial\Omega)$, where*

$$\begin{aligned} &q = 1 + \frac{\pi}{2\omega_0 - \pi} - \varepsilon < 2 \quad &&for\ \omega_0 > \pi, \\ &q = 1 + \frac{\pi}{2\omega_0 - \pi} - \varepsilon > 2 \quad &&for\ \frac{\pi}{2} < \omega_0 < \pi, \\ &q \geq 1 \ is\ arbitrary &&for\ \omega_0 \leq \frac{\pi}{2}, \end{aligned} \tag{2.4}$$

and $\varepsilon > 0$ is arbitrarily small. Then $u \in W^{2,q}(\Omega)$.

It is obvious that $4/3 < q < \infty$.

3. Discretization. In what follows we are concerned with the discretization of problem (2.2) by the finite element method and the discontinuous Galerkin method. To this end, in Ω we construct a system of triangulations $\mathcal{T}_h$, $h \in (0, \overline{h})$, with $\overline{h} > 0$, consisting of a finite number of closed triangles T with standard properties, see [4]. If $T \in \mathcal{T}_h$, then by h_T and ρ_T we denote the diameter of T and the radius of the largest circle inscribed into T. We assume that this system of triangulations $\mathcal{T}_h$ is shape regular:

$$\frac{h_T}{\rho_T} \leq C_R \quad \forall T \in \mathcal{T}_h \ \forall h \in (0, \overline{h}). \tag{3.1}$$

The approximate solution is sought in the space

$$S_h^r = \{v_h \in C(\overline{\Omega});\ v_h|_T \in P_r(T),\ T \in \mathcal{T}_h\}, \tag{3.2}$$

in the case of the FEM discretization and in

$$S_h^r = \{v_h \in L^2(\Omega);\ v_h|_T \in P_r(T),\ T \in \mathcal{T}_h\}, \tag{3.3}$$

in the case of the DGM. Here $r \geq 1$ is an integer and $P_r(T)$ denotes the space of piecewise polynomial functions on T of degree $\leq r$.

Because of the DGM discretization we denote the set of all faces of all elements $T \in \mathcal{T}_h$ by $\mathcal{F}_h$ and we further distinguish between the set of all boundary faces $\mathcal{F}_h^B = \{\Gamma \in \mathcal{F}_h;\, \Gamma \subset \partial\Omega\}$, and the set of all inner faces $\mathcal{F}_h^I = \mathcal{F}_h \setminus \mathcal{F}_h^B$. For an integer $k \geq 1$, a number $q \geq 1$ and a triangulation $\mathcal{T}_h$ we define the broken Sobolev space

$$W^{k,q}(\Omega, \mathcal{T}_h) = \{v \in L^2(\Omega);\, v|_T \in W^{k,q}(T),\ T \in \mathcal{T}_h\} \tag{3.4}$$

and put $H^k(\Omega, \mathcal{T}_h) = W^{k,2}(\Omega, \mathcal{T}_h)$. For functions $v \in W^{k,p}(\Omega, \mathcal{T}_h)$ and inner faces $\Gamma \in \mathcal{F}_h^I$, we introduce the notation

$$\begin{aligned} &v|_\Gamma^{(L)} = \text{trace of } v|_{T_\Gamma^{(L)}} \text{ on } \Gamma, \quad v|_\Gamma^{(R)} = \text{trace of } v|_{T_\Gamma^{(R)}} \text{ on } \Gamma, \\ &\langle v \rangle_\Gamma = \frac{1}{2}(v|_\Gamma^{(L)} + v|_\Gamma^{(R)}), \quad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}. \end{aligned} \tag{3.5}$$

Here $T_\Gamma^{(L)}$ and $T_\Gamma^{(R)}$ are elements adjacent to Γ. By n_Γ we denote the outer unit normal vector to $T_\Gamma^{(L)}$ on Γ.

In the FEM we use the forms defined by (2.1). In the case of the DGM for $u, v \in H^2(\Omega, \mathcal{T}_h)$ we introduce their analogies. Namely, we set

$$b_h(u,v) = \sum_{T \in \mathcal{T}_h} \int_T \nabla u \cdot \nabla v \,\mathrm{d}x - \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma (n_\Gamma \cdot \langle \nabla u \rangle [v] + \theta n_\Gamma \cdot \langle \nabla v \rangle [u]) \,\mathrm{d}S. \tag{3.6}$$

The parameter θ can be chosen as $1, 0, -1$, which leads to symmetric, incomplete and non-symmetric versions of the diffusion forms denoted by SIPG, IIPG, NIPG, respectively. Further, we introduce the interior penalty form

$$J_h(u,v) = \sum_{\Gamma \in \mathcal{F}_h^I} \frac{C_W}{h_\Gamma} \int_\Gamma [u][v] \,\mathrm{d}S \tag{3.7}$$

with a paramenter C_W. The form d is again defined by (2.1). Finally, we set

$$a_h(u,v) = b_h(u,v) + J_h(u,v), \tag{3.8}$$

$$A_h(u,v) = a_h(u,v) + d(u,v). \tag{3.9}$$

Definition 3.1. *We say that a function u_h is a FEM approximate solution of problem* (2.2)*, if*

$$u_h \in H_h^r, \quad A(u_h, v_h) = L(v_h) \quad \forall v_h \in H_h^r. \tag{3.10}$$

The function U_h is a DGM approximate solution, if

$$U_h \in S_h^r, \quad A_h(U_h, v_h) = L(v_h) \quad \forall v_h \in S_h^r. \tag{3.11}$$

The error of the FEM will be estimated in the standard norm $\|\cdot\|_{1,2,\Omega}$ and seminorm $|\cdot|_{1,2,\Omega}$ of the Sobolev space $H^1(\Omega)$. For the analysis of the DGM we introduce the seminorm

$$|v|_h = \left(\sum_{T \in \mathcal{T}_h} \int_T |\nabla v|^2 \,\mathrm{d}x + J_h(v,v) \right)^{\frac{1}{2}}, \tag{3.12}$$

and the norm

$$|||v||| = \left(|v|_h^2 + \|v\|_{0,2,\Omega}^2 \right)^{\frac{1}{2}}. \tag{3.13}$$

By $\|\cdot\|_{0,2,\Omega}$ we denote the norm in $L^2(\Omega)$.

4. Properties of the forms A and A_h. In what follows, by the symbols $C_0, C_1, C_2, \ldots$, we denote constants independent of the exact and approximate solutions and of h. Proofs of the following results are rather technical. We refer to [5].

LEMMA 4.1. *There exists a constant $C_0 > 0$ independent of $u, v \in H^1(\Omega)$, $u_h, v_h \in S_h^r$ and $h \in (0, \overline{h})$ such that*

$$A(u, u-v) - A(v, u-v) \geq |u-v|_{1,2,\Omega}^2 + C_0 \|u-v\|_{0,\alpha+2,\partial\Omega}^{\alpha+2} \quad \forall u, v \in H^1(\Omega). \tag{4.1}$$

Moreover, if the constant C_W from the definition of the penalty form J_h satisfies the conditions

$$C_W > 0, \text{ for } \theta = -1 \text{ (NIPG)}, \tag{4.2}$$

$$C_W > 4C_M(1 + C_I), \text{ for } \theta = 1 \text{ (SIPG)}, \tag{4.3}$$

$$C_W > C_M(1 + C_I), \text{ for } \theta = 0 \text{ (IIPG)}, \tag{4.4}$$

then $\forall u_h, v_h \in S_h^r, \forall h \in (0, \overline{h})$

$$A_h(u_h, u_h - v_h) - A_h(v_h, u_h - v_h) \geq \frac{1}{2}|u_h - v_h|_h^2 + C_0 \|u - v\|_{0,\alpha+2,\partial\Omega}^{\alpha+2}. \tag{4.5}$$

Similarly as in [5] we can prove the monotonicity and continuity of the forms A and A_h.

THEOREM 4.2. *The following results hold:*

a) The forms A and A_h are uniformly monotone. Namely, we have

$$A(u, u-v) - A(v, u-v) \geq \varrho(\|u-v\|_{1,2,\Omega}) \quad \forall u, v \in H^1(\Omega), \tag{4.6}$$

where

$$\varrho(t) = \begin{cases} C_1 t^{\alpha+2} & \text{for} \quad 0 \leq t \leq 1, \\ C_1 t^2 & \text{for} \quad t \geq 1, \end{cases} \tag{4.7}$$

with the constant C_1 depending on C_0, κ and α. If C_W satisfies (4.2)-(4.4), then

$$A_h(u_h, u_h - v_h) - A_h(v_h, u_h - v_h) \geq \varrho(|||u_h - v_h|||) \quad \forall u_h, v_h \in S_h^r \quad \forall h \in (0, \overline{h}), \tag{4.8}$$

where the function ϱ is again defined by (4.7).

b) The forms A and A_h are continuous: There exists a constant $C_2 > 0$ such that $\forall u, v, w \in H^1(\Omega)$

$$|A(u, v) - A(w, v)| \leq C_2 \left(1 + \|u\|_{1,2,\Omega}^{\alpha} + \|w\|_{1,2,\Omega}^{\alpha}\right) \|u - w\|_{1,2,\Omega} \|v\|_{1,2,\Omega}. \tag{4.9}$$

Further, if C_W satisfies (4.2)-(4.4), then

$$\begin{aligned} |A_h(u, w) - A_h(v, w)| \leq C_2 \Big\{ & |||u - v||| + R_h(u - v, q) \\ & + G_h(u - v) \left(\|u\|_{1,2,\Omega}^{\alpha} + |||v|||^{\alpha} \right) \Big\} |||w|||, \end{aligned} \tag{4.10}$$

holds for all $u \in W^{2,q}(\Omega)$, $v, w \in S_h^r$, $h \in (0, \overline{h})$, *where*

$$R_h(\phi, q) = \left(C_M \sum_{T \in \mathcal{T}_h} h_T \left|\phi\right|_{1,q',T} \left|\phi\right|_{2,q,T} \right)^{1/2}, \tag{4.11}$$

for $\phi \in W^{2,q}(\Omega, \mathcal{T}_h)$, $q \in \left(\frac{4}{3}, 2\right)$, $\frac{1}{q} + \frac{1}{q'} = 1$ *and*

$$R_h(\phi, q) = \left(C_M \sum_{T \in \mathcal{T}_h} h_T \left|\phi\right|_{1,2,T} \left|\phi\right|_{2,2,T} \right)^{1/2}, \tag{4.12}$$

for $\phi \in W^{2,q}(\Omega, \mathcal{T}_h)$, $q \geq 2$. *If* $s \geq 3$, $q > 1$ *and* $u \in W^{s,q}(\Omega)$, *then* R_h *is defined by* (4.12). *Moreover,*

$$G_h(\phi) = \left(C_M \sum_{T \in \mathcal{T}_h} \left(\left\|\phi\right\|_{0,2,T}^2 h_T^{-1} + \left|\phi\right|_{1,2,T} \left\|\phi\right\|_{0,2,T} \right) \right)^{1/2}. \tag{4.13}$$

5. Error estimates. The basis for the error estimation is an abstract error estimate. Using the results formulated in Theorem 4.2, using approach from [3] and [5], it is possible to prove the following result:

THEOREM 5.1. *Let* $u \in H^1(\Omega)$ *be a weak solution of* (2.2). *There exists a constant* $C_3 > 0$ *such that if* $u_h \in H_h^r$ *is the FEM approximate solution defined by* (3.10), *then*

$$\left\|u - u_h\right\|_{1,2,\Omega} \leq \varrho_1^{-1} \left(C_3 \left\|u - v_h\right\|_{1,2,\Omega}\right) \quad \forall v_h \in H_h^r \quad \forall h \in (0, \overline{h}), \tag{5.1}$$

where

$$\varrho_1(t) = \varrho(t)/t, \tag{5.2}$$

and ϱ_1^{-1} *is its inverse. In the case of the DGM we have*

$$\begin{aligned} \left|\!\left|\!\left| u - U_h \right|\!\right|\!\right| \leq \rho_1^{-1} \left(C_3 \left(\left|\!\left|\!\left| u - v_h \right|\!\right|\!\right| + R_h(u - v_h; q) + G_h(u - v_h) \left(\left\|u\right\|_{1,2,\Omega}^{\alpha} + \left|\!\left|\!\left| v_h \right|\!\right|\!\right|^{\alpha} \right) \right) \right) \\ + \left|\!\left|\!\left| u - v_h \right|\!\right|\!\right|, \quad \forall\, v_h \in S_h^r, \quad \forall\, h \in (0, \overline{h}), \end{aligned} \tag{5.3}$$

where U_h *is the approximate solution satisfying* (3.11). *The function* $\varrho_1(t)$ *is again defined by* (5.2).

Now we can derive error estimates in terms of the size h of triangulations $\mathcal{T}_h$. To this end, it is necessary to introduce suitable H_h^r- and S_h^r-interpolations. Here we apply the Lagrangian intepolation denoted by π_h defined elementwise (cf. e.g. [4]). From the interpolation theory in [4] we get the following result:

LEMMA 5.2. *Let us assume that* $s, m \geq 0$ *be integers and* $p, q \geq 1$, *the piecewise Lagrange interpolation* π_h *preserve polynomials of degree at most* r, *the triangulation* $\mathcal{T}_h$ *be shape regular according to* (3.1) *and the following embeddings hold:*

$$W^{\mu,q}(T) \hookrightarrow C(T), \quad W^{\mu,q}(T) \hookrightarrow W^{m,p}(T),$$

where $\mu = \min(r + 1, s)$. *Then there exists a constant* $C_4 = C_4(\pi, C_R) > 0$ *such that for all* $T \in \mathcal{T}_h$ *and* $h \in (0, \overline{h})$ *we have*

$$\left|u - \pi_h u\right|_{m,p,T} \leq C_4 \left|u\right|_{\mu,q,T} h_T^{\mu - m + \frac{2}{p} - \frac{2}{q}} \quad \forall u \in W^{s,q}(T). \tag{5.4}$$

The application of Theorem 5.1 and Lemma 5.2 combined with Jensen's inequality (Theorem 3.3 in [13]) yields the sought error estimates.

THEOREM 5.3. *Let the solution of* (2.2) *be* $u \in W^{s,q}(\Omega)$, $\mu = \min(r+1, s)$ *and* $W^{\mu,q}(\Omega) \hookrightarrow H^1(\Omega)$. *Then for the FEM approximate solution* u_h *defined by* (3.10) *the error estimate*

$$\|u - u_h\|_{1,2,\Omega} \le \begin{cases} \varrho_1^{-1}\left(C_5|u|_{\mu,q,\Omega} h^{\mu-\frac{2}{q}}\right), & q \in [1,2), \\ \varrho_1^{-1}\left(C_5|u|_{\mu,q,\Omega} h^{\mu-1}\right), & q \in [2,\infty). \end{cases} \tag{5.5}$$

holds for all $h \in (0, \overline{h})$.

In the case of the DGM we obtain the following results. (See [5].)

THEOREM 5.4. *Let* $u \in W^{s,q}(\Omega)$, *where* $q > \frac{4}{3}$ *for* $s = 2$ *and* $q > 1$ *for* $s \ge 3$ *be the weak solution given by* (2.2), *let* U_h *be the discontinuous Galerkin approximation of degree* r *given by* (3.11) *and let* C_W *satisfy* (4.2)-(4.4). *Let us set* $\mu = \min(r+1, s)$. *Then*

$$|||u - U_h||| \le \rho_1^{-1}\left(C_6(\|u\|_{1,2,\Omega})\, h^{\mu-2/q}|u|_{\mu,q,\Omega}\right) + C_7\, h^{\mu-2/q}|u|_{\mu,q,\Omega},\ h \in (0,\overline{h}), \tag{5.6}$$

for $q \in (1,2)$. *If* $q \ge 2$, *then*

$$|||u - U_h||| \le \rho_1^{-1}\left(C_6(\|u\|_{1,2,\Omega})\, h^{\mu-1}|u|_{\mu,q,\Omega}\right) + C_7\, h^{\mu-1}|u|_{\mu,q,\Omega},\ h \in (0,\overline{h}). \tag{5.7}$$

6. Numerical experiments. In order to verify the obtained theoretical results, some numerical experiments are presented. They were realized with the aid of the FEniCS software [1]. We explore the reduction of the order of convergence caused by the nonlinearity and find out how it affects different norms. In both experiments we discretize the problem by the FEM and by the SIPG variant of the DGM. We use uniform triangular meshes with element diameters $h_l = \frac{h_0}{2^l}, l = 0, 1, \ldots, 5$. The amount of degrees of freedom (DOF) is therefore expected to increase about four times with each refinement. Denoting the error of the discrete solution by $e_h = u - u_h$, we compute the experimental order of convergence (EOC) by

$$EOC = \frac{\log \|e_{h_{l-1}}\| - \log \|e_{h_l}\|}{\log h_{l-1} - \log h_l}, \qquad l = 1, 2, \ldots, 5. \tag{6.1}$$

We evaluate the experimental order of convergence separately for the H^1-seminorm and L^2-norm for the FEM, and $|\cdot|_h$-seminorm and L^2-norm for the SIPG variant of DG method. The discrete problems (3.10) and (3.11) represent nonlinear systems for $\alpha > 0$. They are solved by a dampened Newton method with tolerance on the residual 10^{-9}.

6.1. Example 1 - solution is zero on the boundary. In the first experiment we consider the problem (1.1)-(1.2) on the unit square domain $\Omega = (0,1)^2$. The data f and φ are chosen so that the exact solution has the form

$$u(x_1, x_2) = x_1(1-x_1)x_2(1-x_2)\left(x_1^2 + x_2^2\right)^{1/4}. \tag{6.2}$$

This function belongs to $W^{4,q}(\Omega)$, $q \in \left(1, \frac{4}{3}\right)$. As $W^{4,q}(\Omega) \hookrightarrow H^3(\Omega)$ and $4 - 2/\frac{4}{3} = 2.5$, it follows from Theorems 5.3 and 5.4 that the EOC should be in the norms $\|\cdot\|_{1,2,\Omega}$ and $|||\cdot|||$ (at least) $\frac{\min(2.5,r)}{\alpha+1}$.

Table 6.1: Example 1 - number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1$, 2, 3, 4 and $\alpha = 0.5$, 1.0, 1.5, 2.0 in FEM.

$\alpha = 1.5$, $r = 1$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	49	4	9.3448e-02	–	7.9119e-02	–	1.2244e-01	–
0.188	161	6	4.8018e-02	0.96	4.0634e-02	0.96	6.2904e-02	0.96
0.094	577	6	2.7109e-02	0.82	2.0042e-02	1.02	3.3713e-02	0.90
0.047	2177	6	1.5600e-02	0.80	9.8458e-03	1.03	1.8447e-02	0.87
0.023	8449	6	8.8992e-03	0.81	4.8780e-03	1.01	1.0148e-02	0.86
0.012	33281	6	5.0395e-03	0.82	2.4321e-03	1.00	5.5957e-03	0.86
$\alpha = 1.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	161	3	2.6724e-02	–	8.6570e-03	–	2.8091e-02	–
0.188	577	6	1.2058e-02	1.15	2.2618e-03	1.94	1.2268e-02	1.20
0.094	2177	6	5.9243e-03	1.03	5.7373e-04	1.98	5.9520e-03	1.04
0.047	8449	6	2.9464e-03	1.01	1.4479e-04	1.99	2.9499e-03	1.01
0.023	33281	6	1.4700e-03	1.00	3.6421e-05	1.99	1.4704e-03	1.00
0.012	132097	6	7.3425e-04	1.00	9.1384e-06	1.99	7.3430e-04	1.00
$\alpha = 1.5$, $r = 3$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	337	3	1.2840e-02	–	8.3916e-04	–	1.2867e-02	–
0.188	1249	6	4.9724e-03	1.37	1.2809e-04	2.71	4.9741e-03	1.37
0.094	4801	5	3.3908e-03	0.55	1.5021e-05	3.09	3.3908e-03	0.55
0.047	18817	6	1.6746e-03	1.02	2.0634e-06	2.86	1.6746e-03	1.02
0.023	74497	6	8.3301e-04	1.01	2.9962e-07	2.78	8.3301e-04	1.01
0.012	296449	3	4.1014e-04	1.02	4.7016e-08	2.67	4.1014e-04	1.02
$\alpha = 1.5$, $r = 4$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	577	3	9.6870e-03	–	1.4266e-04	–	9.6880e-03	–
0.188	2177	6	5.0551e-03	0.94	1.4161e-05	3.33	5.0551e-03	0.94
0.094	8449	6	2.5318e-03	1.00	2.3612e-06	2.58	2.5318e-03	1.00
0.047	33281	6	1.2653e-03	1.00	4.3600e-07	2.44	1.2653e-03	1.00
0.023	132097	6	6.3245e-04	1.00	8.1398e-08	2.42	6.3245e-04	1.00
0.012	526337	4	2.9917e-04	1.08	1.5154e-08	2.43	2.9917e-04	1.08
$\alpha = 0.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	161	4	2.3779e-03	–	8.6544e-03	–	8.9752e-03	–
0.188	577	5	6.3232e-04	1.91	2.2617e-03	1.94	2.3485e-03	1.93
0.094	2177	4	1.9356e-04	1.71	5.7372e-04	1.98	6.0550e-04	1.96
0.047	8449	3	6.0476e-05	1.68	1.4479e-04	1.99	1.5691e-04	1.95
0.023	33281	3	1.8977e-05	1.67	3.6421e-05	1.99	4.1069e-05	1.93
0.012	132097	3	6.0396e-06	1.65	9.1384e-06	1.99	1.0954e-05	1.91
$\alpha = 1.0$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	161	4	1.0793e-02	–	8.6566e-03	–	1.3835e-02	–
0.188	577	6	3.9942e-03	1.43	2.2618e-03	1.94	4.5901e-03	1.59
0.094	2177	6	1.6433e-03	1.28	5.7373e-04	1.98	1.7406e-03	1.40
0.047	8449	5	6.8640e-04	1.26	1.4479e-04	1.99	7.0150e-04	1.31
0.023	33281	4	2.8784e-04	1.25	3.6421e-05	1.99	2.9014e-04	1.27
0.012	132097	3	1.1988e-04	1.26	9.1384e-06	1.99	1.2023e-04	1.27
$\alpha = 2.0$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	161	3	4.8888e-02	–	8.6572e-03	–	4.9648e-02	–
0.188	577	6	2.5182e-02	0.96	2.2618e-03	1.94	2.5284e-02	0.97
0.094	2177	6	1.3928e-02	0.85	5.7373e-04	1.98	1.3940e-02	0.86
0.047	8449	6	7.7818e-03	0.84	1.4479e-04	1.99	7.7831e-03	0.84
0.023	33281	6	4.3594e-03	0.84	3.6421e-05	1.99	4.3595e-03	0.84
0.012	132097	6	2.4446e-03	0.83	9.1384e-06	1.99	2.4446e-03	0.83

We discretized the problem with FEM and SIPG variant of the DG method. For polynomials of degree $r = 2$ we tested different values of the nonlinearity parameter $\alpha = 0.5, 1.0, 1.5, 2.0$, and for parameter $\alpha = 1.5$ we tested FEM with polynomials of degrees $r = 1, 2, 3, 4$. The results shown in Table 6.1 and Table 6.2 also include the mesh element size $h = \max_{T\in\mathcal{T}_h} h_T$, the number of degrees of freedom and the number of Newton iterations.

The EOC in H^1-seminorm and $|\cdot|_h$-seminorm are $\min(2.5, r)$, i.e. the error seems to be unaffected by the nonlinearity. The most significant part of the error measured in H^1-norm (or $|||\cdot|||$-norm) was its L^2-norm. Our estimates for the L^2-norm give us an order of convergence $\frac{\min(2.5,r)}{\alpha+1}$, which would be $\frac{1}{\alpha+1}, \frac{2}{\alpha+1}, \frac{2.5}{\alpha+1}, \frac{2.5}{\alpha+1}$ for $r = 1, 2, 3, 4$, respectively. The EOC, however, suggests $\frac{2}{\alpha+1}, \frac{2.5}{\alpha+1}, \frac{2.5}{\alpha+1}, \frac{2.5}{\alpha+1}$ for $r = 1, 2, 3, 4$, respectively. The theoretical error estimate is therefore suboptimal for $r = 1, 2$.

6.2. Example 2 - solution not identically zero on the boundary. In the second experiment, we again consider the problem (1.1)-(1.2) on the unit square

Table 6.2: Example 1 - number of DOF and Newton iterations, discretization errors and convergence rates for $r = 2$ and $\alpha = 0.5$, 1.0, 1.5, 2.0 in SIPG variant of DG method.

$\alpha = 0.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	384	4	2.3711e-03	–	7.7517e-03	–	8.1062e-03	–
0.188	1536	5	6.3176e-04	1.91	2.0084e-03	1.95	2.1054e-03	1.94
0.094	6144	4	1.9354e-04	1.71	5.0545e-04	1.99	5.4124e-04	1.96
0.047	24576	3	6.0472e-05	1.68	1.2673e-04	2.00	1.4042e-04	1.95
0.023	98304	3	1.8994e-05	1.67	3.1764e-05	2.00	3.7009e-05	1.92
0.012	393216	3	5.9364e-06	1.68	7.9534e-06	2.00	9.9246e-06	1.90
$\alpha = 1.0$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	384	4	1.0791e-02	–	7.7532e-03	–	1.3288e-02	–
0.188	1536	6	3.9941e-03	1.43	2.0084e-03	1.95	4.4706e-03	1.57
0.094	6144	6	1.6433e-03	1.28	5.0545e-04	1.99	1.7193e-03	1.38
0.047	24576	5	6.8640e-04	1.26	1.2673e-04	2.00	6.9800e-04	1.30
0.023	98304	4	2.8785e-04	1.25	3.1764e-05	2.00	2.8960e-04	1.27
0.012	393216	3	1.1989e-04	1.26	7.9534e-06	2.00	1.2015e-04	1.27
$\alpha = 1.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	384	4	2.6723e-02	–	7.7536e-03	–	2.7825e-02	–
0.188	1536	6	1.2058e-02	1.15	2.0084e-03	1.95	1.2224e-02	1.19
0.094	6144	6	5.9243e-03	1.03	5.0545e-04	1.99	5.9459e-03	1.04
0.047	24576	6	2.9464e-03	1.01	1.2673e-04	2.00	2.9491e-03	1.01
0.023	98304	6	1.4700e-03	1.00	3.1764e-05	2.00	1.4703e-03	1.00
0.012	393216	6	7.3425e-04	1.00	7.9534e-06	2.00	7.3429e-04	1.00
$\alpha = 2.0$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	384	3	4.8888e-02	–	7.7537e-03	–	4.9499e-02	–
0.188	1536	6	2.5182e-02	0.96	2.0084e-03	1.95	2.5262e-02	0.97
0.094	6144	6	1.3928e-02	0.85	5.0545e-04	1.99	1.3937e-02	0.86
0.047	24576	6	7.7818e-03	0.84	1.2673e-04	2.00	7.7828e-03	0.84
0.023	98304	6	4.3594e-03	0.84	3.1764e-05	2.00	4.3595e-03	0.84
0.012	393216	6	2.4446e-03	0.83	7.9534e-06	2.00	2.4446e-03	0.83

domain $\Omega = (0,1)^2$. We prescribe the data f and φ in such a way that the exact solution is $u(x_1, x_2) = \frac{1}{4}(1 + x_1)^2 \sin(2\pi x_1 x_2)$. This function was used in [8]. It is smooth, zero on boundary segments going through the points $[0,1]$, $[0,0]$, $[1,0]$ and nonzero on segments going through the points $[1,0]$, $[1,1]$, $[0,1]$.

In this example we choose $\alpha = 1.5$ and polynomial degrees $r = 1, 2, 3$ for both the FEM and the SIPG variant of the DGM. For the FEM, we have also tried $r = 4$, and $\alpha = 0.5$. The EOC is not affected by the boundary nonlinearity parameter α. The H^1-seminorm and $|\cdot|_h$-seminorm converge with the order of convergence r, and the L^2-norm converges faster with order $r + 1$. The error estimates in Theorems 5.3 and 5.4 are again suboptimal, but in this case, the error is dominated by the H^1-seminorm or the $|\cdot|_h$-seminorm.

7. Additional estimates. On the basis of the numerical experiments we come to the conclusion that the error estimates can be influenced by the behaviour of the exact solution on the boundary $\partial\Omega$, namely, if the exact solution u vanishes on the whole boundary and, on the other hand, if it is nonzero on a sufficiently large subset of the boundary. We present here some theoretical results derived for the FEM.

THEOREM 7.1. *Let the weak solution $u \in W^{s,q}(\Omega)$ given by (2.2) be zero on $\partial\Omega$. Let us set $\mu = \min(r + 1, s)$, where r is the degree of used polynomials. Then*

$$|u - u_h|_{1,2,\Omega} \le \begin{cases} C_8 |u|_{k+1,q,\Omega} h^{\mu - \frac{2}{q}}, & q \in [1,2), \\ C_8 |u|_{k+1,q,\Omega} h^{\mu - 1}, & q \in [2,\infty). \end{cases} \tag{7.1}$$

Proof. Neglecting the last term on the right-hand side of (4.1) gives us $|u - u_h|^2_{1,2,\Omega} \le A(u, u - u_h) - A(u_h, u - u_h)$, using Galerkin orthogonality following from (2.2), (3.10) and $H_h^r \subset H^1(\Omega)$ for a piecewise Lagrange interpolation yields $A(u, u - u_h) - A(u_h, u - u_h) = A(u, u - \pi_h u) - A(u_h, u - \pi_h u)$. The fact that $\pi_h u$ is also zero

Table 6.3: Example 2 - number of DOF and Newton iterations, discretization errors and convergence rates for $r = 1$, 2, 3, 4 and $\alpha = 1.5$, 0.5 in FEM.

$\alpha = 1.5$, $r = 1$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	49	6	2.5883e-01	–	9.5881e-01	–	9.9314e-01	–
0.188	161	5	6.1723e-02	2.07	5.3381e-01	0.84	5.3736e-01	0.89
0.094	577	4	1.5381e-02	2.00	2.8145e-01	0.92	2.8187e-01	0.93
0.047	2177	4	3.9289e-03	1.97	1.4421e-01	0.96	1.4426e-01	0.97
0.023	8449	3	9.9584e-04	1.98	7.2704e-02	0.99	7.2711e-02	0.99
0.012	33281	3	2.4986e-04	1.99	3.6390e-02	1.00	3.6391e-02	1.00
$\alpha = 1.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	161	6	1.4730e-02	–	2.3514e-01	–	2.3560e-01	–
0.188	577	4	1.2493e-03	3.56	5.8813e-02	2.00	5.8826e-02	2.00
0.094	2177	3	1.3819e-04	3.18	1.5173e-02	1.95	1.5173e-02	1.95
0.047	8449	3	1.6986e-05	3.02	3.8676e-03	1.97	3.8676e-03	1.97
0.023	33281	2	2.1254e-06	3.00	9.7489e-04	1.99	9.7489e-04	1.99
0.012	132097	2	2.6587e-07	3.00	2.4425e-04	2.00	2.4425e-04	2.00
$\alpha = 1.5$, $r = 3$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	337	6	4.5914e-03	–	2.3116e-02	–	2.3568e-02	–
0.188	1249	3	2.4182e-04	4.25	3.4931e-03	2.73	3.5015e-03	2.75
0.094	4801	3	1.3800e-05	4.13	4.7873e-04	2.87	4.7893e-04	2.87
0.047	18817	2	8.5542e-07	4.01	6.2363e-05	2.94	6.2369e-05	2.94
0.023	74497	2	5.4140e-08	3.98	7.9229e-06	2.98	7.9231e-06	2.98
0.012	296449	2	3.4211e-09	3.98	9.9474e-07	2.99	9.9474e-07	2.99
$\alpha = 1.5$, $r = 4$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	577	6	8.4789e-05	–	4.2824e-03	–	4.2832e-03	–
0.188	2177	3	3.2227e-06	4.72	3.2812e-04	3.71	3.2813e-04	3.71
0.094	8449	2	1.0740e-07	4.91	2.2035e-05	3.90	2.2036e-05	3.90
0.047	33281	2	3.4969e-09	4.94	1.4299e-06	3.95	1.4299e-06	3.95
0.023	132097	2	1.1140e-10	4.97	9.0809e-08	3.98	9.0809e-08	3.98
0.012	526337	2	3.5005e-12	4.99	5.6988e-09	3.99	5.6988e-09	3.99
$\alpha = 0.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_{1,2,\Omega}$	EOC	$\|\|e\|\|_{1,2,\Omega}$	EOC
0.375	161	6	1.4072e-02	–	2.3527e-01	–	2.3569e-01	–
0.188	577	4	1.2379e-03	3.51	5.8815e-02	2.00	5.8828e-02	2.00
0.094	2177	4	1.3806e-04	3.16	1.5173e-02	1.95	1.5173e-02	1.95
0.047	8449	3	1.6989e-05	3.02	3.8676e-03	1.97	3.8676e-03	1.97
0.023	33281	3	2.1256e-06	3.00	9.7489e-04	1.99	9.7489e-04	1.99
0.012	132097	2	2.6588e-07	3.00	2.4425e-04	2.00	2.4425e-04	2.00

Table 6.4: Example 2 - number of DOF and Newton iterations, discretization errors and convergence rates for $\alpha = 1.5$ and $r = 1$, 2, 3 in SIPG variant of DG method.

$\alpha = 1.5$, $r = 1$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	192	6	2.5073e-01	–	8.7620e-01	–	9.1137e-01	–
0.188	768	5	6.1030e-02	2.04	4.7862e-01	0.87	4.8249e-01	0.92
0.094	3072	4	1.5377e-02	1.99	2.4855e-01	0.95	2.4902e-01	0.95
0.047	12288	4	3.9457e-03	1.96	1.2692e-01	0.97	1.2698e-01	0.97
0.023	49152	3	1.0016e-03	1.98	6.3982e-02	0.99	6.3990e-02	0.99
0.012	196608	3	2.5142e-04	1.99	3.2043e-02	1.00	3.2044e-02	1.00
$\alpha = 1.5$, $r = 2$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	384	6	1.3432e-02	–	2.2029e-01	–	2.2069e-01	–
0.188	1536	4	9.8475e-04	3.77	5.4667e-02	2.01	5.4676e-02	2.01
0.094	6144	3	9.5957e-05	3.36	1.3884e-02	1.98	1.3884e-02	1.98
0.047	24576	3	1.1194e-05	3.10	3.5122e-03	1.98	3.5122e-03	1.98
0.023	98304	2	1.3773e-06	3.02	8.8228e-04	1.99	8.8229e-04	1.99
0.012	393216	2	1.7139e-07	3.01	2.2075e-04	2.00	2.2075e-04	2.00
$\alpha = 1.5$, $r = 3$								
h	DOF	$iter$	$\|\|e\|\|_{0,2,\Omega}$	EOC	$\|e\|_h$	EOC	$\|\|\|e\|\|\|$	EOC
0.375	640	6	4.5720e-03	–	2.7526e-02	–	2.7903e-02	–
0.188	2560	3	2.4012e-04	4.25	4.2359e-03	2.70	4.2427e-03	2.72
0.094	10240	3	1.3676e-05	4.13	5.7642e-04	2.88	5.7658e-04	2.88
0.047	40960	2	8.4847e-07	4.01	8.1035e-05	2.83	8.1039e-05	2.83
0.023	163840	2	5.3738e-08	3.98	1.0459e-05	2.95	1.0460e-05	2.95
0.012	655360	2	3.3983e-09	3.98	1.3431e-06	2.96	1.3431e-06	2.96

on $\partial\Omega$ and the Hölder inequality implies that $A(u, u - \pi_h u) - A(u_h, u - \pi_h u) = \int_\Omega \nabla(u - u_h) \cdot \nabla(u - \pi_h u) \mathrm{d}x \leq |u - u_h|_{1,2,\Omega} |u - \pi_h u|_{1,2,\Omega}$ Dividing by $|u - u_h|_{1,2,\Omega}$ leads to the estimate $|u - u_h|_{1,2,\Omega} \leq |u - \pi_h u|_{1,2,\Omega}$. Now Theorem 5.2 for $H^1(T)$-seminorm gives us the sought estimate. □

Further, we can improve estimates in Theorem 5.3 in such a way that $\rho_1(t) = C_8 t$

for all $t \geq 0$ in the case that the exact solution satisfies the following condition:

$$G \subset \partial\Omega, \quad |G| > 0, \quad |u| \geq \varepsilon > 0 \quad \text{on } G. \tag{7.2}$$

Then the improved error estimate is a consequence of the strong monotonicity of the form A:

THEOREM 7.2. *Let $u \in H^1(\Omega)$ and let the conditions* (7.2) *hold. Then there exists a constant $C_9 = C_9(\Omega, G, \varepsilon) > 0$ such that*

$$A(u, u-v) - A(v, u-v) \geq C_9 \|u-v\|^2_{1,2,\Omega} \quad \forall v \in H^1(\Omega). \tag{7.3}$$

Proof. Since $|u|^\alpha - |v|^\alpha$ and $u^2 - v^2$ have the same sign, it follows that $(|u|^\alpha - |v|^\alpha)(u^2 - v^2) \geq 0$, or equivalently $|u|^\alpha u^2 + |v|^\alpha v^2 \geq |u|^\alpha v^2 + |v|^\alpha u^2$. Thus, we can write

$$\begin{aligned} &2(|u|^\alpha u - |v|^\alpha v)(u-v) = |u|^\alpha(2u^2 - 2uv) + |v|^\alpha(2v^2 - 2uv) \\ &\quad \geq |u|^\alpha(u^2 - 2uv + v^2) + |v|^\alpha(v^2 - 2uv + u^2) = (|u|^\alpha + |v|^\alpha)(u-v)^2. \end{aligned} \tag{7.4}$$

Now (7.4) and (2.3) imply that $A(u, u-v) - A(v, u-v) \geq |u-v|^2_{1,2,\Omega} + \frac{1}{2}\kappa\varepsilon^\alpha \|u - v\|^2_{0,2,G}$. The existence of a constant C_9 from the statement of this theorem follows from Poincaré's inequality $\|u\|_{1,2,\Omega} \leq c_P(|u|_{1,2,\Omega} + \|u\|_{0,2,G})$. □

For the DGM we get analogical results with the norm $|||\cdot|||$ replacing $\|\cdot\|_{1,2,\Omega}$ and the seminorm $|\cdot|_h$ replacing $|\cdot|_{1,2,\Omega}$. An interesting problem is the analysis of the FEM or DGM combined with the use of numerical integration.

REFERENCES

[1] Alnæs M.M., Blechta J., Hake J., Johansson A., Kehlet B. Logg A., Richardson C., Ring J.,Rognes M.E., Wells G.N.: *The FEniCS Project Version 1.5.* Archive of Numerical Software, 2015.

[2] Babuška I.: Private communication, Austin 2017.

[3] O. Bartoš: Discontinuous Galerkin method for the solution of boundary-value problems in non-smooth domains. Master Thesis, Faculty of Mathematics and Physics, Charles University, Praha 2017.

[4] Ciarlet P. G.: *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam, 1978.

[5] Feistauer M., Roskovec F., Sändig A.-M.: Discontinuous Galerkin method for an elliptic problem with nonlinear Newton boundary conditions in a polygon. IMA J. Numer. Anal. (to appear).

[6] Ganesh M., Steinbach O.: Boundary element methods for potential problems with nonlinear boundary conditions. *Mathematics of Computation* 70 (2000), 1031–1042.

[7] Ganesh M., Steinbach O.: Nonlinear boundary integral equations for harmonic problems. *Journal of Integral Equations and Applications* 11 (1999), 437–459.

[8] Harriman K., Houston P., Senior B., Süli E.: *hp-Version Discontinuous Galerkin Methods with Interior Penalty for Partial Differential Equations with Nonnegative Characteristic Form*, Contemporary Mathematics Vol. 330, pp. 89-119, AMS, 2003.

[9] Křížek, M., Liu L., Neittaanmäki P.: Finite element analysis of a nonlinear elliptic problem with a pure radiation condition. In: Proc. Conf. devoted to the 70th birthday of Prof. J. Nečas, Lisbon, 1999.

[10] Liu L., Křížek, M.: Finite element analysis of a radiation heat transfer problem. *J. Comput. Math.* 16 (1998), 327–336.

[11] Moreau R., Ewans J. W.: An analysis of the hydrodynamics of alluminium reduction cells. *J. Electrochem. Soc.* 31 (1984), 2251–2259.

[12] Pick, L., Kufner, A., John, O., Fučík, S.: Function Spaces. De Gruyter Series in Nonlinear Analysis and Applications 14, Berlin, 2013.

[13] Rudin W.: *Real and comples analysis*, McGraw-Hill, 1987

Proceedings of EQUADIFF 2017
pp. 137–146

NUMERICAL HOMOGENIZATION FOR INDEFINITE H(CURL)-PROBLEMS

BARBARA VERFÜRTH*

Abstract. In this paper, we present a numerical homogenization scheme for indefinite, time-harmonic Maxwell's equations involving potentially rough (rapidly oscillating) coefficients. The method involves an $\mathbf{H}(\mathrm{curl})$-stable, quasi-local operator, which allows for a correction of coarse finite element functions such that order optimal (w.r.t. the mesh size) error estimates are obtained. To that end, we extend the procedure of [D. Gallistl, P. Henning, B. Verfürth, *Numerical homogenization for H(curl)-problems*, arXiv:1706.02966, 2017] to the case of indefinite problems. In particular, this requires a careful analysis of the well-posedness of the corrector problems as well as the numerical homogenization scheme.

Key words. multiscale method, wave propagation, Maxwell's equations, finite element method

AMS subject classifications. 65N30, 65N15, 65N12, 35Q61, 78M10

1. Introduction. Time-harmonic Maxwell's equations, which model electromagnetic wave propagation, play an essential role in many physical applications. If the coefficients are rapidly oscillating on a fine scale as in the context of photonic crystals or metamaterials, standard discretizations suffer from bad convergence rates and a large pre-asymptotic range due to the multiscale nature, the low regularity, and the indefiniteness of the problem.

In this paper, we consider a numerical homogenization scheme to cope with the multiscale nature and the resolution condition, which couples the maximal mesh size to the frequency and is typical for indefinite wave propagation problems, see [2]. Analytical homogenization for locally periodic $\mathbf{H}(\mathrm{curl})$-problems shows that the solution can be decomposed into a macroscopic contribution (without rapid oscillations) and a fine-scale corrector, see [3, 9, 10, 17]. In [5], this was extended beyond the periodic case and without assuming scale separation. Using a suitable interpolation operator, the exact solution is decomposed into a coarse part, which is a good approximation in $\mathbf{H}(\mathrm{div})'$, and a corrector contribution, which then gives a good approximation in $\mathbf{H}(\mathrm{curl})$. Furthermore, the corrector can be quasi-localized, allowing for an efficient computation. Analytical homogenization and other multiscale methods can also be applied to indefinite problems [3, 9] so that it is natural to examine this extension also for the numerical homogenization of [5].

The technique of numerical homogenization presented there is known as Localized Orthogonal Decomposition (LOD) and was originally proposed in [12]. Among many applications, we point to elliptic boundary value problems [8], the wave equation [1], mixed elements [7], and in particular Helmholtz problems [6, 16, 15]. The works on the Helmholtz equation reveal that the LOD can also reduce the so-called pollution effect. Only a natural and reasonable resolution condition of a few degrees of freedom per wavelength is needed in the LOD and the local corrector problems have to be solved on patches which grow logarithmically with the wave number. The crucial observation is that the bilinear form is coercive on the kernel space of a suitable

*Institut für Analysis und Numerik, Westfälische Wilhelms-Universität Münster, Einsteinstr. 62, D-48149 Münster (barbara.verfuerth@uni-muenster.de).

interpolation operator. For Maxwell's equations this is not possible due to the large kernel of the curl-operator. However, a wave number independent inf-sup-stability of the bilinear form over the kernel of the interpolation operator is proved using a regular decomposition. This inf-sup-stability forces us to localize the corrector problem in a non-conforming manner, which leads to additional terms in the analysis and may be of independent interest. Still, we are able to define a well-posed localized numerical homogenization scheme which allow for order optimal (w.r.t. the mesh size) a priori estimates.

The paper is organized as follows. Section 2 introduces the model problem and the necessary notation for meshes and the interpolation operator. We introduce an ideal numerical homogenization scheme in Section 3. We localize the corrector operator, present the resulting main scheme and its a priori analysis in Section 4.

The notation $a \lesssim b$ denotes $a \leq Cb$ with a constant C independent of the mesh size H, the oversampling parameter m and the frequency ω. Bold face letters will indicate vector-valued quantities and all functions are complex-valued, unless explicitly mentioned. We study the high-frequency case, i.e. $\omega \gtrsim 1$ is assumed.

2. Problem setting.

2.1. Model problem. Let $\Omega \subset \mathbb{R}^3$ be an open, bounded, contractible domain with polyhedral Lipschitz boundary with outer unit normal $\mathbf{n}$. For any bounded subdomain $G \subset \Omega$, the spaces $\mathbf{H}(\mathrm{curl}, G)$, $\mathbf{H}_0(\mathrm{curl}, G)$ and $\mathbf{H}(\mathrm{div}, G)$ denote the usual curl- and div-conforming spaces; see [14] for details. We will omit the domain G if it is equal to the full domain Ω. In addition to the standard inner product, we equip $\mathbf{H}(\mathrm{curl}, G)$ with the following ω-dependent inner product

$$(\mathbf{v}, \mathbf{w})_{\mathrm{curl},\omega,G} := (\mathrm{curl}\, \mathbf{v}, \mathrm{curl}\, \mathbf{w})_{L^2(G)} + \omega^2 (\mathbf{v}, \mathbf{w})_{L^2(G)}.$$

Let $\mathbf{f} \in \mathbf{H}(\mathrm{div}, \Omega)$ and let $\mu^{-1} \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$ and $\varepsilon \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$ be uniformly elliptic. For any open subset $G \subset \Omega$, we define the bilinear form $\mathcal{B}_G : \mathbf{H}(\mathrm{curl}, G) \times \mathbf{H}(\mathrm{curl}, G) \to \mathbb{C}$ as

$$\mathcal{B}_G(\mathbf{v}, \boldsymbol{\psi}) := (\mu^{-1} \mathrm{curl}\, \mathbf{v}, \mathrm{curl}\, \boldsymbol{\psi})_{L^2(G)} - \omega^2 (\varepsilon \mathbf{v}, \boldsymbol{\psi})_{L^2(G)}, \tag{2.1}$$

and set $\mathcal{B} := \mathcal{B}_\Omega$. The form $\mathcal{B}_G$ is obviously continuous and the continuity constant is independent of ω if we use the norm $\|\cdot\|_{\mathrm{curl},\omega}$.

We now look for $\mathbf{u} \in \mathbf{H}_0(\mathrm{curl}, \Omega)$ such that

$$\mathcal{B}(\mathbf{u}, \boldsymbol{\psi}) = (\mathbf{f}, \boldsymbol{\psi})_{L^2(\Omega)} \quad \text{for all } \boldsymbol{\psi} \in \mathbf{H}_0(\mathrm{curl}, \Omega). \tag{2.2}$$

We implicitly assume that the above problem is a multiscale problem, i.e. the coefficients μ^{-1} and ε are rapidly varying on a very fine sale. Fredholm theory guarantees the existence of a unique solution $\mathbf{u}$ to (2.2) provided that ω is not an eigenvalue of curl-curl-operator, which we will assume from now on. This in particular implies that there is $\gamma(\omega) > 0$ such that $\mathcal{B}$ is inf-sup stable with constant $\gamma(\omega)$, i.e.

$$\inf_{\mathbf{v} \in \mathbf{H}_0(\mathrm{curl})\setminus\{0\}} \sup_{\boldsymbol{\psi} \in \mathbf{H}_0(\mathrm{curl})\setminus\{0\}} \frac{|\mathcal{B}(\mathbf{v}, \boldsymbol{\psi})|}{\|\mathbf{v}\|_{\mathrm{curl},\omega} \|\boldsymbol{\psi}\|_{\mathrm{curl},\omega}} \geq \gamma(\omega). \tag{2.3}$$

2.2. Mesh and interpolation operator. Let $\mathcal{T}_H$ be a regular partition of Ω into tetrahedra, such that $\cup \mathcal{T}_H = \overline{\Omega}$ and any two distinct $T, T' \in \mathcal{T}_H$ are either disjoint or share a common vertex, edge or face. We assume the partition $\mathcal{T}_H$ to

be shape-regular and quasi-uniform. The global mesh size H is defined as $H := \max\{\operatorname{diam}(T)|T \in \mathcal{T}_H\}$. $\mathcal{T}_H$ is a coarse mesh in the sense that it does not resolve the fine-scale oscillations of the parameters.

Given any subdomain $G \subset \overline{\Omega}$ define the patches via

$$\mathrm{N}^1(G) := \mathrm{N}(G) := \operatorname{int}(\cup\{T \in \mathcal{T}_H | T \cap \overline{G} \neq \emptyset\}) \qquad \text{and} \qquad \mathrm{N}^m(G) := \mathrm{N}(\mathrm{N}^{m-1}(G)).$$

We refer to [15], for instance, for a visualization of the patches. The shape regularity implies that there is a uniform bound $C_{\mathrm{ol},m}$ on the number of elements in the m-th order patch, i.e. $\max_{T\in\mathcal{T}_H} \operatorname{card}\{K \in \mathcal{T}_H | K \subset \overline{\mathrm{N}^m(T)}\} \leq C_{\mathrm{ol},m}$, and the quasi-uniformity implies that $C_{\mathrm{ol},m}$ depends polynomially on m. We abbreviate $C_{\mathrm{ol}} := C_{\mathrm{ol},1}$. We denote the lowest order Nédélec finite element, cf. [14, Section 5.5], by

$$\mathring{\mathcal{N}}(\mathcal{T}_H) := \{\mathbf{v} \in \mathbf{H}_0(\operatorname{curl}) | \forall T \in \mathcal{T}_H : \mathbf{v}|_T(\mathbf{x}) = \mathbf{a}_T \times \mathbf{x} + \mathbf{b}_T \text{ with } \mathbf{a}_T, \mathbf{b}_T \in \mathbb{C}^3\}.$$

We require an $\mathbf{H}(\operatorname{curl})$-stable interpolation operator (with some additional properties) for the numerical homogenization. The only suitable candidate is the Falk-Winter interpolation operator, see [4]. Some important properties are summarized below, see [5] for details and proofs.

PROPOSITION 2.1. *There exists a projection $\pi_H^E : \mathbf{H}_0(\operatorname{curl}) \to \mathring{\mathcal{N}}(\mathcal{T}_H)$ with the following local stability properties: For all $\mathbf{v} \in \mathbf{H}_0(\operatorname{curl})$ and all $T \in \mathcal{T}_H$ it holds that*

$$\|\pi_H^E(\mathbf{v})\|_{L^2(T)} \lesssim \big(\|\mathbf{v}\|_{L^2(\mathrm{N}(T))} + H\|\operatorname{curl}\mathbf{v}\|_{L^2(\mathrm{N}(T))}\big), \tag{2.4}$$

$$\|\operatorname{curl}\pi_H^E(\mathbf{v})\|_{L^2(T)} \lesssim \|\operatorname{curl}\mathbf{v}\|_{L^2(\mathrm{N}(T))}. \tag{2.5}$$

Moreover, for any $\mathbf{v} \in \mathbf{H}_0(\operatorname{curl}, \Omega)$, there are $\mathbf{z} \in \mathbf{H}_0^1(\Omega)$ and $\theta \in H_0^1(\Omega)$ such that $\mathbf{v} - \pi_H^E(\mathbf{v}) = \mathbf{z} + \nabla\theta$ with the local bounds for every $T \in \mathcal{T}_H$

$$\begin{aligned} &H^{-1}\|\mathbf{z}\|_{L^2(T)} + \|\nabla\mathbf{z}\|_{L^2(T)} \lesssim \|\operatorname{curl}\mathbf{v}\|_{L^2(\mathrm{N}^3(T))}, \\ &H^{-1}\|\theta\|_{L^2(T)} + \|\nabla\theta\|_{L^2(T)} \lesssim \big(\|\mathbf{v}\|_{L^2(\mathrm{N}^3(T))} + H\|\operatorname{curl}\mathbf{v}\|_{L^2(\mathrm{N}^3(T))}\big), \end{aligned} \tag{2.6}$$

where $\nabla\mathbf{z}$ stands for the Jacobi matrix of $\mathbf{z}$.

The stability estimates in particular imply that π_H^E is stable with respect to the $\|\cdot\|_{\operatorname{curl},\omega}$-norm if the condition $\omega H \lesssim 1$ is fulfilled:

$$\|\pi_H^E\mathbf{v}\|_{\operatorname{curl},\omega} \lesssim \|\mathbf{v}\|_{\operatorname{curl},\omega} \qquad \text{if} \quad \omega H \lesssim 1.$$

3. Ideal numerical homogenization. In this section we introduce an ideal numerical homogenization scheme which approximates the exact solution in $\mathbf{H}_0(\operatorname{curl})$ by a coarse part (which itself is a good approximation in $H^{-1}(\Omega)$) and a corrector contribution. The idea is based on the direct sum splitting $\mathbf{H}_0(\operatorname{curl}) = \mathring{\mathcal{N}}(\mathcal{T}_H) \oplus \mathbf{W}$ with $\mathbf{W} := \ker(\pi_H^E)$ the kernel of the Falk-Winther interpolation operator introduced in the previous section. The regular decomposition estimates (2.6) directly imply for any $\mathbf{w} \in \mathbf{W}$

$$\|\mathbf{w}\|_{\mathbf{H}(\operatorname{div})'} \lesssim H\|\mathbf{w}\|_{\mathbf{H}(\operatorname{curl})}. \tag{3.1}$$

¿From now on, we assume the resolution condition

$$\omega H \lesssim 1. \tag{3.2}$$

It reflects that a few degrees of freedom per wavelength are always required to represent a wave. The constant only depends on interpolation constants and the bounds

on the material coefficients. Under resolution condition (3.2), $\mathcal{B}$ is stable on $\mathbf{W}$, as details the next lemma.

LEMMA 3.1 (Properties of $\mathbf{W}$). *Let $\mathbf{w} \in \mathbf{W}$ be decomposed as $\mathbf{w} = \mathbf{z} + \nabla\theta$ and (3.2) be satisfied. Then*

- *we have a (ω-independent) norm equivalence between $\|\cdot\|_{\mathrm{curl},\omega}$ and $|||\mathbf{w}|||^2 := \|\operatorname{curl}\mathbf{z}\|^2 + \omega^2\|\nabla\theta\|^2$*
- *there is $\alpha > 0$ independent of ω such that*

$$\inf_{\mathbf{w}\in\mathbf{W}\setminus\{0\}} \sup_{\boldsymbol{\phi}\in\mathbf{W}\setminus\{0\}} \frac{|\mathcal{B}(\mathbf{w},\boldsymbol{\phi})|}{\|\mathbf{w}\|_{\mathrm{curl},\omega}\|\boldsymbol{\phi}\|_{\mathrm{curl},\omega}} \geq \alpha.$$

Proof. For the norm equivalence we obtain using (2.6) and $\operatorname{curl}\mathbf{w} = \operatorname{curl}\mathbf{z}$

$$|||\mathbf{w}|||^2 = \|\operatorname{curl}\mathbf{z}\|^2+\omega^2\|\nabla\theta\|^2 \lesssim \|\operatorname{curl}\mathbf{w}\|^2 + \omega^2\|\mathbf{w}\|^2 + \omega^2H^2\|\operatorname{curl}\mathbf{w}\|^2 \lesssim \|\mathbf{w}\|^2_{\mathrm{curl},\omega},$$
$$\|\mathbf{w}\|^2_{\mathrm{curl},\omega} \leq \|\mathbf{z}\|^2_{\mathrm{curl},\omega} + \|\nabla\theta\|^2_{\mathrm{curl},\omega} \lesssim \|\operatorname{curl}\mathbf{z}\|^2+\omega^2H^2\|\operatorname{curl}\mathbf{z}\|^2 + \omega^2\|\nabla\theta\|^2 \lesssim |||\mathbf{w}|||^2.$$

For the inf-sup-constant, define the sign-flip isomorphism $A(\mathbf{w}) := \mathbf{z} - \nabla\theta$. Observe that $\operatorname{curl}\pi_H^E\mathbf{z} = \operatorname{curl}\pi_H^E\mathbf{w} = 0$ because of the commuting property of π_H^E. Then

$$\begin{aligned}\Re\{\mathcal{B}(\mathbf{w},(\mathrm{id}-\pi_H^E)A(\mathbf{w}))\} \gtrsim{}& \|\operatorname{curl}\mathbf{z}\|^2 + \omega^2\|\nabla\theta\|^2 - \omega^2\|\mathbf{z}\|^2 - 2\omega^2|(\varepsilon\mathbf{z},\nabla\theta)| \\ &- 2\omega^2|(\varepsilon\mathbf{z},\pi_H^E\mathbf{z})| - 2\omega^2|(\varepsilon\nabla\theta,\pi_H^E\mathbf{z})|,\end{aligned}$$

where we used $\pi_H^E\nabla\theta = -\pi_H^E\mathbf{z}$ because of $\pi_H^E\mathbf{w} = 0$. Applying Young's inequality, the stability of π_H^E (2.4)–(2.5), estimate (2.6) and using the resolution condition (3.2), we arrive at

$$\Re\{\mathcal{B}(\mathbf{w},(\mathrm{id}-\pi_H^E)A(\mathbf{w}))\} \gtrsim \|\operatorname{curl}\mathbf{z}\|^2 + \omega^2\|\nabla\theta\|^2 \gtrsim \|\mathbf{w}\|^2_{\mathrm{curl},\omega}$$

because of the norm equivalence. The estimate $\|(\mathrm{id}-\pi_H^E)A(\mathbf{w})\|_{\mathrm{curl},\omega} \lesssim \|\mathbf{w}\|_{\mathrm{curl},\omega}$ finally gives the claim. ∎

In contrast to coercive problems, unique solvability is not guaranteed when $\mathcal{B}$ is restricted to subspaces. Therefore, the inf-sup-stability of $\mathcal{B}$ on $\mathbf{W}$ is the crucial ingredient to introduce a well-defined Corrector Green's Operator.

DEFINITION 3.2. *For $\mathbf{F} \in \mathbf{H}_0(\operatorname{curl})'$, we define the Corrector Green's Operator*

$$\mathcal{G} : \mathbf{H}_0(\operatorname{curl})' \to \mathbf{W} \qquad \text{by} \qquad \mathcal{B}(\mathcal{G}(\mathbf{F}),\mathbf{w}) = \mathbf{F}(\mathbf{w}) \quad \text{for all } \mathbf{w}\in\mathbf{W}. \tag{3.3}$$

Let $\mathcal{L} : \mathbf{H}_0(\operatorname{curl}) \to \mathbf{H}_0(\operatorname{curl})'$ denote the differential operator associated with $\mathcal{B}$ and set $\mathcal{K} := -\mathcal{G}\circ\mathcal{L}$. Inspired by the procedure in [5], an ideal numerical homogenization scheme consists in solving the variational problem over the "multiscale" space $(\mathrm{id}+\mathcal{K})\mathring{\mathcal{N}}(\mathcal{T}_H)$. The well-posedness of this scheme is proved in the next lemma.

LEMMA 3.3. *Under the resolution condition, we have with $\gamma(\omega)$ from (2.3) that*

$$\inf_{\mathbf{v}_H\in\mathring{\mathcal{N}}(\mathcal{T}_H)\setminus\{0\}} \sup_{\boldsymbol{\psi}_H\in\mathring{\mathcal{N}}(\mathcal{T}_H)\setminus\{0\}} \frac{|\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)|}{\|\mathbf{v}_H\|_{\mathrm{curl},\omega}\|\boldsymbol{\psi}_H\|_{\mathrm{curl},\omega}} \gtrsim \gamma(\omega). \tag{3.4}$$

Proof. Fix $\mathbf{v}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H)$. From (2.3), there exists $\boldsymbol{\psi} \in \mathbf{H}_0(\operatorname{curl})$ with $\|\boldsymbol{\psi}\|_{\mathrm{curl},\omega} = 1$ such that

$$|\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H,\boldsymbol{\psi})| \geq \gamma(\omega)\|(\mathrm{id}+\mathcal{K})\mathbf{v}_H\|_{\mathrm{curl},\omega}.$$

By the definition of $\mathcal{K}$, it holds that $(\mathrm{id}+\mathcal{K})\pi_H^E\boldsymbol{\psi} = (\mathrm{id}+\mathcal{K})\boldsymbol{\psi}$ and $\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H, \mathbf{w}) = 0$ for all $\mathbf{w} \in \mathbf{W}$. Thus, we obtain

$$\begin{aligned}|\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H, (\mathrm{id}+\mathcal{K})\pi_H^E\boldsymbol{\psi})| &= |\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H, (\mathrm{id}+\mathcal{K})\boldsymbol{\psi})| = |\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H, \boldsymbol{\psi})| \\ &\geq \gamma(\omega)\|(\mathrm{id}+\mathcal{K})\mathbf{v}_H\|_{\mathrm{curl},\omega}.\end{aligned}$$

The claim follows now by the norm equivalence

$$\|\mathbf{v}_H\|_{\mathrm{curl},\omega} = \|\pi_H^E(\mathrm{id}+\mathcal{K})\mathbf{v}_H\|_{\mathrm{curl},\omega} \lesssim \|(\mathrm{id}+\mathcal{K})\mathbf{v}_H\|_{\mathrm{curl},\omega},$$

which is a result of the stability of π_H^E. ∎

Before we introduce the ideal numerical homogenization scheme, we summarize the approximation and stability properties of the Corrector Green's Operator, cf. [5].

LEMMA 3.4 (Ideal corrector estimates). *Any* $\mathbf{F} \in \mathbf{H}_0(\mathrm{curl})'$ *and any* $\mathbf{f} \in \mathbf{H}(\mathrm{div})$ *satisfy*

$$H\|\mathcal{G}(\mathbf{F})\|_{\mathrm{curl},\omega} + \|\mathcal{G}(\mathbf{F})\|_{\mathbf{H}(\mathrm{div})'} \lesssim H\alpha^{-1}\|\mathbf{F}\|_{\mathbf{H}_0(\mathrm{curl})'} \tag{3.5}$$

$$H\|\mathcal{G}(\mathbf{f})\|_{\mathrm{curl},\omega} + \|\mathcal{G}(\mathbf{f})\|_{\mathbf{H}(\mathrm{div})'} \lesssim H^2\alpha^{-1}\|\mathbf{f}\|_{\mathbf{H}(\mathrm{div})}. \tag{3.6}$$

Collecting the results of the previous lemmas, we have the following result on our ideal numerical homogenization scheme.

THEOREM 3.5. *Let* $\mathbf{u}$ *denote the exact solution to* (2.2) *and* $\mathbf{u}_H = \pi_H^E\mathbf{u}$. *Then*

- *it holds that* $\mathbf{u} = \mathbf{u}_H + \mathcal{K}(\mathbf{u}_H) + \mathcal{G}(\mathbf{f})$
- *assuming* (3.2), $\mathbf{u}_H$ *is characterized as the unique solution to*

$$\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{u}_H, (\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H) = (\mathbf{f}, (\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H) \qquad \text{for all } \boldsymbol{\psi}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H) \tag{3.7}$$

- *assuming* (3.2), *it holds that*

$$\|\mathbf{u} - (\mathrm{id}+\mathcal{K})\mathbf{u}_H\|_{\mathrm{curl},\omega} + \|\mathbf{u} - \mathbf{u}_H\|_{\mathbf{H}(\mathrm{div})'} \lesssim H\|\mathbf{f}\|_{\mathbf{H}(\mathrm{div})}. \tag{3.8}$$

Proof. The proof of the first two items carries over verbatim from the elliptic case [5]. The a priori error estimate (3.8) follows from the first item and Lemma 3.4. ∎

The theorem shows that $(\mathrm{id}+\mathcal{K})\mathbf{u}_H$ approximates the analytical solution with linear rate without assumptions on the regularity of the problem. What is more, only the reasonable resolution condition $\omega H \lesssim 1$ is required, overcoming the pollution effect. However, the determination of $\mathcal{K}$ requires the solution of global problems, which limits the practical usability of the scheme.

4. Quasi-local numerical homogenization.

4.1. Exponential decay and localized corrector. The property that $\mathcal{K}$ can be approximated by local correctors is directly linked to the decay properties of $\mathcal{G}$ defined in (3.3). The following result states – loosely speaking – in which distance (measured in unit of the coarse mesh size H) from the support of the source term $\mathbf{F}$ the weighted $\mathbf{H}(\mathrm{curl})$-norm of $\mathcal{G}(\mathbf{F})$ becomes negligibly small. For that, recall the definition of element patches $\mathrm{N}^m(T)$ from Section 2.2.

PROPOSITION 4.1. *Let* $T \in \mathcal{T}_H$, $m \in \mathbb{N}$ *and* $\mathbf{F}_T \in \mathbf{H}_0(\mathrm{curl})'$ *be a local source functional, i.e.* $\mathbf{F}_T(\mathbf{v}) = 0$ *for all* $\mathbf{v} \in \mathbf{H}_0(\mathrm{curl})$ *with* $\mathrm{supp}(\mathbf{v}) \subset \Omega \setminus T$. *If* (3.2) *holds, there exists* $0 < \tilde{\beta} < 1$ *such that*

$$\|\mathcal{G}(\mathbf{F}_T)\|_{\mathrm{curl},\omega,\Omega\setminus\mathrm{N}^m(T)} \lesssim \beta^m\|\mathbf{F}_T\|_{\mathbf{H}_0(\mathrm{curl})'}. \tag{4.1}$$

Proof. The proof can be easily adapted from the elliptic case in [5] using the inf-sup-stability of $\mathcal{B}$ over $\mathbf{W}$ from Lemma 3.1. ∎

The result can be used to approximate $\mathcal{K}$, which has a non-local argument, via

$$\mathcal{K}(\mathbf{v}_H) = -\sum_{T\in\mathcal{T}_H} \mathcal{G}(\mathcal{L}_T(\mathbf{v}_H)),$$

where the localized differential operator $\mathcal{L}_T : \mathbf{H}(\mathrm{curl}, T) \to \mathbf{H}(\mathrm{curl}, \Omega)'$ is associated with $\mathcal{B}_T$, the restriction of $\mathcal{B}$ to the element T. Proposition 4.1 now suggests to truncate the computation of $\mathcal{G}(\mathbf{F}_T)$ to the patches $\mathrm{N}^m(T)$ and then collect the results from all elements T. Typically, m is referred to as *oversampling parameter*.

DEFINITION 4.2 (Localized Corrector Approximation). *For any element $T \in \mathcal{T}_H$ we define its patch $\Omega_T := \mathrm{N}^m(T)$. Let $\mathbf{F} \in \mathbf{H}_0(\mathrm{curl})'$ be the sum of local functionals, i.e. $\mathbf{F} = \sum_{T\in\mathcal{T}_H} \mathbf{F}_T$ with $\mathbf{F}_T$ as in Proposition 4.1. Denote by $\pi^E_{H,\Omega_T} : \mathbf{H}_0(\mathrm{curl}, \Omega) \to \mathring{\mathcal{N}}(\mathcal{T}_H(\Omega_T))$ the Falk-Winther interpolation operator which enforces essential boundary conditions (i.e. zero tangential traces) on $\partial\Omega_T$. We then define*

$$\mathbf{W}(\Omega_T) := \{\mathbf{w} \in \mathbf{H}_0(\mathrm{curl}) | \mathbf{w} = 0 \text{ outside } \Omega_T, \pi^E_{H,\Omega_T}\mathbf{w} = 0\} \not\subseteq \mathbf{W}. \tag{4.2}$$

We call $\mathcal{G}_{T,m}(\mathbf{F}_T) \in \mathbf{W}(\Omega_T)$ the localized corrector *if it solves*

$$\mathcal{B}(\mathcal{G}_{T,m}(\mathbf{F}_T), \mathbf{w}) = \mathbf{F}_T(\mathbf{w}) \qquad \text{for all } \mathbf{w} \in \mathbf{W}(\Omega_T). \tag{4.3}$$

The global corrector approximation is then given by

$$\mathcal{G}_m(\mathbf{F}) = \sum_{T\in\mathcal{T}_H} \mathcal{G}_{T,m}(\mathbf{F}_T).$$

Observe that problem (4.3) is only formulated on the patch Ω_T. Its well-posedness can be proved as in Lemma 3.1: For $\mathbf{w} \in \mathbf{W}(\Omega_T)$, use $(\mathrm{id} - \pi^E_{H,\Omega_T})A(\mathbf{w}) \in \mathbf{W}(\Omega_T)$ as test function (with the sign-flip isomorphism A). We emphasize that the definition of $\mathbf{W}(\Omega_T)$ via π^E_{H,Ω_T} is needed to make this test function a member of $\mathbf{W}(\Omega_T)$, otherwise the support would be enlarged. This is a non-conforming definition of the localized corrector (i.e. $\pi^E_H\mathcal{G}_m(\cdot) \neq 0$), so that additional terms appear in the error analysis. However, the non-conformity error only plays a role near the boundary of $\partial\Omega_T$ and can therefore be controlled very well.

THEOREM 4.3. *Let $\mathcal{G}(\mathbf{F})$ be the ideal Green's corrector and $\mathcal{G}_m(\mathbf{F})$ the localized corrector from Definition 4.2. Under* (3.2)*, there exists $0 < \beta < 1$ such that*

$$\|\mathcal{G}(\mathbf{F}) - \mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega} \lesssim \sqrt{C_{\mathrm{ol},m}}\,\beta^m \Big(\sum_{T\in\mathcal{T}_H} \|\mathbf{F}_T\|^2_{\mathbf{H}_0(\mathrm{curl})'}\Big)^{1/2}, \tag{4.4}$$

$$\|\pi^E_H\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega} \lesssim \sqrt{C_{\mathrm{ol},m}}\,\beta^m \Big(\sum_{T\in\mathcal{T}_H} \|\mathbf{F}_T\|^2_{\mathbf{H}_0(\mathrm{curl})'}\Big)^{1/2}. \tag{4.5}$$

The proof is postponed to Subsection 4.3.

4.2. The quasi-local numerical homogenization scheme. Following the above motivation, we define a quasi-local numerical homogenization scheme by replacing $\mathcal{K}$ in the ideal scheme (3.7) with $\mathcal{K}_m$.

DEFINITION 4.4. *Let $\mathcal{K}_m$ be defined as described in the previous subsection. The* quasi-local numerical homogenization scheme *seeks $\mathbf{u}_{H,m} \in \mathring{\mathcal{N}}(\mathcal{T}_H)$ such that*

$$\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{u}_{H,m}, (\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H) = (\mathbf{f}, (\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H) \qquad \text{for all } \mathbf{v}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H). \tag{4.6}$$

We observe that $\mathcal{K}_m$ can be computed by solving local decoupled problems, see [5] for details. Note that the spaces $\mathbf{W}(\Omega_T)$ are still infinite dimensional so that in practice, we require an additional fine-scale discretization of the corrector problems. We omit this step here and refer the reader to [5] for the elliptic case and [16] for the Helmholtz equation.

We now prove the well-posedness and the a priori error estimate for the quasi-local numerical homogenization scheme.

THEOREM 4.5 (Well-posedness of (4.6)). *If the resolution condition* (3.2) *and the oversampling condition*

$$m \gtrsim |\log\big(\gamma(\omega)/\sqrt{C_{\mathrm{ol},m}}\big)|/|\log(\beta)| \tag{4.7}$$

are fulfilled, $\mathcal{B}$ is inf-sup-stable over $(\mathrm{id}+\mathcal{K}_m)\mathring{\mathcal{N}}(\mathcal{T}_H)$, i.e.

$$\inf_{\mathbf{v}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H)\setminus\{0\}} \sup_{\boldsymbol{\psi}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H)\setminus\{0\}} \frac{|\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H, (\mathrm{id}+\mathcal{K}_m)\boldsymbol{\psi}_H)|}{\|\mathbf{v}_H\|_{\mathrm{curl},\omega}\|\boldsymbol{\psi}\|_{\mathrm{curl},\omega}} \geq \gamma_{\mathrm{LOD}}(\omega) \approx \gamma(\omega).$$

THEOREM 4.6 (A priori estimate). *Let $\mathbf{u}$ denote the analytical solution to* (2.2) *and $\mathbf{u}_{H,m}$ the solution to* (4.6). *If the resolution condition* (3.2) *and the oversampling condition*

$$m \gtrsim |\log\big(\gamma_{\mathrm{LOD}}(\omega)/\sqrt{C_{\mathrm{ol},m}}\big)|/|\log(\beta)| \tag{4.8}$$

are fulfilled, then

$$\|\mathbf{u} - (\mathrm{id}+\mathcal{K}_m)\mathbf{u}_{H,m}\|_{\mathrm{curl},\omega} \lesssim (H + \beta^m\gamma^{-1}(\omega))\|\mathbf{f}\|_{\mathbf{H}(\mathrm{div})}. \tag{4.9}$$

Note that the oversampling condition (4.8) is – up to constants independent of H and ω – the same as condition (4.7). Since $C_{\mathrm{ol},m}$ grows polynomially in m for quasi-uniform meshes, it is satisfiable and depends on the behavior of $\gamma(\omega)$. If $\gamma(\omega) \lesssim \omega^q$, we derive $m \approx \log(\omega)$, which is a better resolution condition than for a standard discretization. Note that in (4.9), we can replace $\gamma^{-1}(\omega)$ with $C_{\mathrm{stab}}(\omega)$, the stability constant of the original problem (2.2). This is exactly the same a priori estimate as for Helmholtz problems in [16]. To sum up, an oversampling parameter $m \approx |\log(\omega)|$ is sufficient for the stability of the LOD. Requiring additionally $m \approx |\log(H)|$, we obtain a linear convergence rate for the error.

4.3. Main proofs. Theorem 4.3 results from the exponential decay of $\mathcal{G}$ in Proposition 4.1.

Proof. [Proof of Theorem 4.3] We start by proving the following local estimate

$$\|\mathcal{G}(\mathbf{F}_T) - \mathcal{G}_{T,m}(\mathbf{F}_T)\|_{\mathrm{curl},\omega} \lesssim \tilde{\beta}^m\|\mathbf{F}_T\|_{\mathbf{H}_0(\mathrm{curl})'}. \tag{4.10}$$

By Strang's second Lemma we obtain

$$\begin{aligned}\|\mathcal{G}(\mathbf{F}_T) - \mathcal{G}_{T,m}(\mathbf{F}_T)\|_{\mathrm{curl},\omega} \lesssim{}& \inf_{\mathbf{w}_{T,m}\in\mathbf{W}(\Omega_T)} \|\mathcal{G}(\mathbf{F}_T) - \mathbf{w}_{T,m}\|_{\mathrm{curl},\omega} \\ &+ \sup_{\substack{\boldsymbol{\phi}_{T,m}\in\mathbf{W}(\Omega_T)\\ \|\boldsymbol{\phi}_{T,m}\|_{\mathrm{curl},\omega}=1}} |\mathcal{B}(\mathcal{G}(\mathbf{F}_T), \boldsymbol{\phi}_{T,m}) - \mathbf{F}_T(\boldsymbol{\phi}_{T,m})|.\end{aligned}$$

The first term can be estimated as in [5]. For the second term, we have due to (3.3) that it is equal to

$$\sup_{\boldsymbol{\phi}_{T,m}\in\mathbf{W}(\Omega_T),\|\boldsymbol{\phi}_{T,m}\|_{\mathrm{curl},\omega}=1} |\mathcal{B}(\mathcal{G}(\mathbf{F}_T),\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi})-\mathbf{F}_T(\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi})|$$

for any $\boldsymbol{\phi}\in\mathbf{W}$. Fixing $\boldsymbol{\phi}_{T,m}=\mathbf{z}_{T,m}+\nabla\theta_{T,m}$, we choose $\boldsymbol{\phi}=(\mathrm{id}-\pi_H^E)(\eta\mathbf{z}_{T,m}+\nabla(\eta\theta_{T,m}))$ with a cut-off function such that $\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi}=0$ in $\mathrm{N}^{m-2}(T)$. Then $\mathbf{F}_T(\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi})=0$ and we get with the stability of π_H^E and (2.6)

$$\begin{aligned}|\mathcal{B}(\mathcal{G}(\mathbf{F}_T),\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi})| &\lesssim \|\mathcal{G}(\mathbf{F}_T)\|_{\mathrm{curl},\omega,\Omega\setminus\mathrm{N}^{m-2}(T)}\|\boldsymbol{\phi}_{T,m}-\boldsymbol{\phi}\|_{\mathrm{curl},\omega}\\ &\lesssim \|\mathcal{G}(\mathbf{F}_T)\|_{\mathrm{curl},\omega,\Omega\setminus\mathrm{N}^{m-2}(T)}.\end{aligned}$$

Combination with Proposition 4.1 gives (4.10).

For (4.4), we split the error as

$$\|\mathcal{G}(\mathbf{F})-\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega}\le\|(\mathrm{id}-\pi_H^E)(\mathcal{G}(\mathbf{F})-\mathcal{G}_m(\mathbf{F}))\|_{\mathrm{curl},\omega}+\|\pi_H^E\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega}.$$

The first term can be estimated with the procedure from [5]. The second term is the left-hand side of (4.5) and thus, it suffices to prove (4.5). We observe that $\pi_H^E\mathcal{G}_{T,m}(\mathbf{F}_T)\neq 0$ only on a small ring $R\subset\mathrm{N}^{m+1}(T)$ because π_H^E and π_{H,Ω_T}^E only differ near the boundary of Ω_T. Hence, we get

$$\begin{aligned}\|\pi_H^E\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega}^2 &\le \sum_{T\in\mathcal{T}_H}|(\pi_H^E\mathcal{G}_m(\mathbf{F}),\pi_H^E\mathcal{G}_{T,m}(\mathbf{F}_T))_{\mathrm{curl},\omega}|\\ &\lesssim \sum_T \|\pi_H^E\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega,\mathrm{N}^{m+1}(T)}\|\pi_H^E(\mathcal{G}(\mathbf{F}_T)-\mathcal{G}_{T,m}(\mathbf{F}_T))\|_{\mathrm{curl},\omega}\\ &\lesssim \sqrt{C_{\mathrm{ol},m}}\|\pi_H^E\mathcal{G}_m(\mathbf{F})\|_{\mathrm{curl},\omega}\Big(\sum_T\|\mathcal{G}(\mathbf{F}_T)-\mathcal{G}_{T,m}(\mathbf{F}_T)\|_{\mathrm{curl},\omega}\Big)^{1/2}.\end{aligned}$$

Application of (4.10) gives the claim. □

The well-posedness of the quasi-local numerical scheme comes from the well-posedness of the ideal scheme (Theorem 3.5) and the fact that the localized corrector is exponentially close to the ideal corrector.

Proof. [Proof of Theorem 4.5] Fix $\mathbf{v}_H\in\mathring{\mathcal{N}}(\mathcal{T}_H)$ and set $\tilde{\mathbf{v}}_H=\pi_H^E(\mathrm{id}+\mathcal{K}_m)(\mathbf{v}_H)$. According to Theorem 3.5, there exists $\boldsymbol{\psi}_H\in\mathring{\mathcal{N}}(\mathcal{T}_H)$ with $\|\boldsymbol{\psi}_H\|_{\mathrm{curl},\omega}=1$ such that

$$|\mathcal{B}((\mathrm{id}+\mathcal{K})\tilde{\mathbf{v}}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)|\ge\gamma(\omega)\|\tilde{\mathbf{v}}_H\|_{\mathrm{curl},\omega}.$$

As $\mathcal{B}(\mathbf{w},(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)=0$ for all $\mathbf{w}\in\mathbf{W}$, we derive

$$\begin{aligned}\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H) &= \mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H-(\mathrm{id}-\pi_H^E)((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H),(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)\\ &= \mathcal{B}(\tilde{\mathbf{v}}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)=\mathcal{B}((\mathrm{id}+\mathcal{K})\tilde{\mathbf{v}}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H).\end{aligned}$$

This yields together with Theorem 4.3

$$\begin{aligned}&|\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H,(\mathrm{id}+\mathcal{K}_m)\boldsymbol{\psi}_H)|\\ &\quad=|\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H,(\mathcal{K}_m-\mathcal{K})\boldsymbol{\psi}_H)+\mathcal{B}((\mathrm{id}+\mathcal{K})\mathbf{v}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)|\\ &\quad=|\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H,(\mathcal{K}_m-\mathcal{K})\boldsymbol{\psi}_H)+\mathcal{B}((\mathrm{id}+\mathcal{K})\tilde{\mathbf{v}}_H,(\mathrm{id}+\mathcal{K})\boldsymbol{\psi}_H)|\\ &\quad\ge\gamma(\omega)\|\tilde{\mathbf{v}}_H\|_{\mathrm{curl},\omega}-C\sqrt{C_{\mathrm{ol},m}}\,\beta^m\|(\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H\|_{\mathrm{curl},\omega}.\end{aligned}$$

Moreover, we have

$$\|(\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H\|_{\mathrm{curl},\omega} \lesssim (1+\beta^m)\|\mathbf{v}_H\|_{\mathrm{curl},\omega} \lesssim \|\mathbf{v}_H\|_{\mathrm{curl},\omega},$$

since $\beta < 1$, and

$$\begin{aligned}\|\mathbf{v}_H\|_{\mathrm{curl},\omega} &= \|\pi_H^E(\mathrm{id}+\mathcal{K})\mathbf{v}_H\|_{\mathrm{curl},\omega} = \|\pi_H^E(\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H + \pi_H^E(\mathcal{K}-\mathcal{K}_m)\mathbf{v}_H\|_{\mathrm{curl},\omega}\\ &\lesssim \|\tilde{\mathbf{v}}_H\|_{\mathrm{curl},\omega} + C\sqrt{C_{\mathrm{ol},m}}\,\beta^m\|\mathbf{v}_H\|_{\mathrm{curl},\omega}.\end{aligned}$$

If m is large enough (indirectly implied by the oversampling condition), the second term can be hidden on the left-hand side. Thus, we finally obtain

$$|\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H, (\mathrm{id}+\mathcal{K}_m)\boldsymbol{\psi}_H)| \gtrsim (\gamma(\omega) - C\sqrt{C_{\mathrm{ol},m}}\,\beta^m)\|\mathbf{v}_H\|_{\mathrm{curl},\omega}.$$

Application of the oversampling condition (4.7) gives the assertion. ∎

The proof of the a priori error estimate is inspired by the procedure for the Helmholtz equation [16] and uses duality arguments.

Proof. [Proof of Theorem 4.6] Denote by $\mathbf{e}$ the error $\mathbf{u} - (\mathrm{id}+\mathcal{K}_m)\mathbf{u}_{H,m}$ and set $\mathbf{e}_{H,m} := (\mathrm{id}+\mathcal{K}_m)\pi_H^E(\mathbf{e})$. Let $\mathbf{z}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H)$ be the solution to the dual problem

$$\mathcal{B}((\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H, (\mathrm{id}+\mathcal{K}_m)\mathbf{z}_H) = (\mathbf{e}_{H,m}, (\mathrm{id}+\mathcal{K}_m)\mathbf{v}_H)_{\mathrm{curl},\omega} \qquad \text{for all } \mathbf{v}_H \in \mathring{\mathcal{N}}(\mathcal{T}_H).$$

Using the fact that $\mathcal{B}(\mathbf{w}, (\mathrm{id}+\mathcal{K})\mathbf{z}_H) = 0$ for all $\mathbf{w} \in \mathbf{W}$ and employing the Galerkin orthogonality $\mathcal{B}(\mathbf{e}, (\mathrm{id}+\mathcal{K}_m)\mathbf{z}_H) = 0$, we obtain that

$$\begin{aligned}\|\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega}^2 &= \mathcal{B}(\mathbf{e}_{H,m}, (\mathrm{id}+\mathcal{K}_m)\mathbf{z}_H)\\ &= \mathcal{B}(\mathbf{e}_{H,m}, (\mathcal{K}_m-\mathcal{K})\mathbf{z}_H) + \mathcal{B}(\mathbf{e}_{H,m}, (\mathrm{id}+\mathcal{K})\mathbf{z}_H)\\ &= \mathcal{B}(\mathbf{e}-\mathbf{e}_{H,m}, (\mathcal{K}-\mathcal{K}_m)\mathbf{z}_H) - \mathcal{B}(\pi_H^E(\mathbf{e}-\mathbf{e}_{H,m}), (\mathrm{id}+\mathcal{K})\mathbf{z}_H).\end{aligned}$$

Observe that $\pi_H^E(\mathbf{e}-\mathbf{e}_{H,m}) = \pi_H^E\mathcal{K}_m\pi_H^E(\mathbf{e})$. Theorem 4.3 and 4.5 yield

$$\begin{aligned}\|\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega}^2 &\lesssim \sqrt{C_{\mathrm{ol},m}}\,\beta^m\|\mathbf{e}-\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega}\|\mathbf{z}_H\|_{\mathrm{curl},\omega} + \sqrt{C_{\mathrm{ol},m}}\,\beta^m\|\mathbf{e}\|_{\mathrm{curl},\omega}\|\mathbf{z}_H\|_{\mathrm{curl},\omega}\\ &\lesssim \sqrt{C_{\mathrm{ol},m}}\,\beta^m\,\gamma_{\mathrm{LOD}}^{-1}(\omega)\,(\|\mathbf{e}-\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega} + \|\mathbf{e}\|_{\mathrm{curl},\omega})\|\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega}.\end{aligned}$$

The triangle inequality gives

$$\|\mathbf{e}\|_{\mathrm{curl},\omega} \le \|(\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m})\|_{\mathrm{curl},\omega} + \|\pi_H^E(\mathbf{e}-\mathbf{e}_{H,m})\|_{\mathrm{curl},\omega} + \|\mathbf{e}_{H,m}\|_{\mathrm{curl},\omega}.$$

The above computations and (4.5) imply with the resolution condition (4.8)

$$\|\mathbf{e}\|_{\mathrm{curl},\omega} \lesssim \|(\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m})\|_{\mathrm{curl},\omega}.$$

Observe that $\mathbf{e}-\mathbf{e}_{H,m} = \mathbf{u} - (\mathrm{id}+\mathcal{K}_m)\pi_H^E(\mathbf{u}) - (\mathrm{id}+\mathcal{K}_m)\pi_H^E\mathcal{K}_m\mathbf{u}_{H,m}$. Since $(\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m}) \in \mathbf{W}$, Lemma 3.1 gives $\mathbf{w} \in \mathbf{W}$ with $\|\mathbf{w}\|_{\mathrm{curl},\omega} = 1$ such that

$$\begin{aligned}&\|(\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m})\|_{\mathrm{curl},\omega}\\ &\lesssim |\mathcal{B}((\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m}), \mathbf{w})|\\ &= |\mathcal{B}(\mathbf{u},\mathbf{w}) - \mathcal{B}((\mathrm{id}+\mathcal{K}_m)\pi_H^E\mathbf{u},\mathbf{w}) - \mathcal{B}((\mathrm{id}+\mathcal{K}_m)\pi_H^E\mathcal{K}_m\mathbf{u}_{H,m},\mathbf{w}) - \mathcal{B}(\pi_H^E\mathcal{K}_m\pi_H^E\mathbf{e},\mathbf{w})|\\ &= |(\mathbf{f},\mathbf{w}) - \mathcal{B}((\mathcal{K}_m-\mathcal{K})\pi_H^E\mathbf{u},\mathbf{w}) - \mathcal{B}((\mathcal{K}_m-\mathcal{K})\pi_H^E\mathcal{K}_m\mathbf{u}_{H,m},\mathbf{w}) - \mathcal{B}(\pi_H^E\mathcal{K}_m\pi_H^E\mathbf{e},\mathbf{w})|.\end{aligned}$$

Theorems 4.3 and 4.5 now give together with the stability of π_H^E and (3.1)

$$\begin{aligned}&\|(\mathrm{id}-\pi_H^E)(\mathbf{e}-\mathbf{e}_{H,m})\|_{\mathrm{curl},\omega}\\ &\lesssim \big(H + \sqrt{C_{\mathrm{ol},m}}\,\beta^m\gamma^{-1}(\omega) + C_{\mathrm{ol},m}\,\beta^{2m}\gamma_{\mathrm{LOD}}^{-1}(\omega)\big)\|\mathbf{f}\|_{\mathbf{H}(\mathrm{div})} + \sqrt{C_{\mathrm{ol},m}}\,\beta^m\|\mathbf{e}\|_{\mathrm{curl},\omega}.\end{aligned}$$

The last term can be hidden on the left-hand side and the third term can be absorbed in the second term. ∎

Conclusion. In this paper, we presented and analyzed a numerical homogenization scheme for indefinite $\mathbf{H}(\mathrm{curl})$-problems, inspired by [5]. We showed that the indefinite bilinear form is inf-sup-stable for $\omega H \lesssim 1$ over the kernel of the Falk-Winther interpolation operator, which is crucial for the analysis. Under this reasonable resolution condition and the additional oversampling condition $m \approx |\log(\gamma(\omega))|$, the numerical homogenization method is stable and yields linear convergence (w.r.t. the mesh size) of the error in the $\mathbf{H}(\mathrm{curl})$-norm. These conditions are similar for the Helmholtz equation, suggesting that they are optimal. Incorporating impedance boundary conditions as well as numerical experiments are subject of future research.

Acknowledgments. Financial support by the DFG through project OH 98/6-1 is gratefully acknowledged. The author would like to thank Dietmar Gallistl (KIT) for fruitful discussion on the subject, in particular for suggesting the non-conforming definition of the localized spaces; and the anonymous referee for helpful remarks. Main ideas of this contribution evolved while the author enjoyed the kind hospitality of the Hausdorff Research Institute for Mathematics (HIM) in Bonn during the trimester program on multiscale problems.

REFERENCES

[1] A. Abdulle and P. Henning, *Localized orthogonal decomposition method for the wave equation with a continuum of scales*, Math. Comp., 86 (2017), pp. 549–587.
[2] I. M. Babuška and S. A. Sauter, *Is the pollution effect avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Rev., 42 (2000), pp. 451–484.
[3] P. Ciarlet Jr., S. Fliss, and C. Stohrer, *On the approximation of electromagnetic fields by edge finite elements. Part 2: A heterogeneous multiscale method for Maxwell's equations*, Comput. Math. Appl., 73 (2017), pp. 1900–1919.
[4] R. S. Falk and R. Winther, *Local bounded cochain projections*, Math. Comp., 83 (2014), pp. 2631–2656.
[5] D. Gallistl, P. Henning, and B. Verfürth, *Numerical homogenization of H(curl)-problems*, arXiv:1706.02966 (2017), preprint.
[6] D. Gallistl and D. Peterseim, *Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering*, Comp. Appl. Mech. Eng., 295 (2015), pp. 1–17.
[7] F. Hellman, P.Henning, and A. Målqvist, *Multiscale mixed finite elements*, Discr. Contin. Dyn. Syst. Ser. S, 9 (2016), pp. 1269–1298.
[8] P. Henning and A. Målqvist, Localized orthogonal decomposition techniques for boundary value problems, SIAM J. Sci. Comput., 36 (2014), pp. A1609–A1634.
[9] P. Henning, M. Ohlberger, and B. Verfürth, *A new Heterogeneous Multiscale Method for time-harmonic Maxwell's equations*, SIAM J. Numer. Anal., 54 (2016), pp. 3493–3522.
[10] P. Henning, M. Ohlberger, and B. Verfürth, *Analysis of multiscale methods for time-harmonic Maxwell's equations*, Pro. Appl. Math. Mech., 16 (2016), pp. 559–560.
[11] R. Hiptmair, *Maxwell equations: continuous and discrete*, in Computational Electromagnetism, A. Bermúdez de Castro and A. Valli, eds., Lecture Notes in Mathematics, Springer, Cham, 2015, pp. 1–58.
[12] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., 83 (2014), pp. 2583–2603.
[13] A. Moiola, *Trefftz-Discontinuous Galerkin methods for time-harmonic wave problems*, PhD thesis, ETH Zürich, 2011.
[14] P. Monk, *Finite element methods for Maxwell's equation*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2003.
[15] M. Ohlberger and B. Verfürth, *Localized Orthogonal Decomposition for two-scale Helmholtz-type problems*, AIMS Mathematics, 2 (2017), pp. 458–478.
[16] D. Peterseim, *Eliminating the pollution effect by local subscale correction*, Math. Comp., 86 (2017), pp. 1005–1036.
[17] N. Wellander and G. Kristensson, *Homogenization of the Maxwell equations at fixed frequency*, AIAM J. Appl. Math., 64 (2003), pp. 170–195.

Proceedings of EQUADIFF 2017
pp. 147–156

SINGULARLY PERTURBED SET OF PERIODIC FUNCTIONAL-DIFFERENTIAL EQUATIONS ARISING IN OPTIMAL CONTROL THEORY

VALERY Y. GLIZER*

Abstract. We consider the singularly perturbed set of periodic functional-differential matrix Riccati equations, associated with a periodic linear-quadratic optimal control problem for a singularly perturbed delay system. The delay is small of order of a small positive multiplier for a part of the derivatives in the system. A zero-order asymptotic solution to this set of Riccati equations is constructed and justified.

Key words. periodic linear-quadratic optimal control problem, singularly perturbed delay system, small delay, periodic functional-differential matrix Riccati equations, asymptotic solution

AMS subject classifications. 34H05, 34K13, 34K26, 35F50

1. Introduction. One of the fundamental results in control theory is the solution of the finite horizon linear-quadratic optimal control problem with fixed initial and free terminal states. Due to this result, the solution of the control problem is reduced to a terminal-value problem either for a matrix differential Riccati-type equation (finite dimensional case, [16]) or for an operator differential Riccati-type equation (infinite dimensional case, [2, 4, 6, 7, 8, 20, 23]). This result was extended to the finite horizon periodic linear-quadratic optimal control problem. Solution of this problem is reduced to a differential periodic matrix/operator Riccati-type equation (see e.g. [1, 3]).

If the controlled equation is a differential equation with a delay in the state, the operator Riccati-type equation is reduced to a set of matrix functional-differential equations with ordinary and partial derivatives (see e.g. [2, 6, 8, 18, 19, 23]).

If the controlled equation is singularly perturbed, the corresponding differential Riccati equation also is singularly perturbed. Singularly perturbed non-periodic matrix/operator Riccati equations were well studied in many works (see e.g. [11, 12, 14, 15, 17, 21, 24]). Singularly perturbed periodic matrix Riccati equations also were studied in the literature (see [9, 22]). However, to the best of our knowledge, singularly perturbed periodic operator Riccati equations have not yet been considered in the literature.

In this paper, we consider a finite horizon periodic linear-quadratic optimal control problem for a singularly perturbed system with small delays in the state. We construct an asymptotic solution to the set of periodic functional-differential matrix equations of Riccati type, associated with this problem by the control optimality conditions.

2. Problem statement.

2.1. Original optimal control problem. Consider the following linear system with delays in state variables

$$dx(t)/dt = A_1(t)x(t) + A_2(t)y(t) + H_1(t)x(t-\varepsilon h) + H_2(t)y(t-\varepsilon h)$$

*Department of Applied Mathematics, ORT Braude College of Engineering, P.O.B. 78, Karmiel 2161002, Israel (valery48@braude.ac.il, valgl120@gmail.com).

$$+\int_{-h}^{0}\Big[G_1(t,\eta)x(t+\varepsilon\eta)+G_2(t,\eta)y(t+\varepsilon\eta)\Big]d\eta+B_1(t)u(t)+f_1(t), \tag{2.1}$$

$$\begin{aligned}\varepsilon dy(t)/dt &= A_3(t)x(t)+A_4(t)y(t)+H_3(t)x(t-\varepsilon h)+H_4(t)y(t-\varepsilon h)\\ &+\int_{-h}^{0}\Big[G_3(t,\eta)x(t+\varepsilon\eta)+G_4(t,\eta)y(t+\varepsilon\eta)\Big]d\eta+B_2(t)u(t)+f_2(t),\end{aligned} \tag{2.2}$$

where $x(t)\in E^n$, $y(t)\in E^m$, $u(t)\in E^r$ (u is a control); $\varepsilon>0$ is a small parameter ($\varepsilon << 1$), $h>0$ is some constant independent of ε; the matrix-valued functions $A_i(t)$, $H_i(t)$, $B_j(t)$, $(i=1,...,4;\ j=1,2)$ and the vector-valued functions $f_j(t)$, $(j=1,2)$ are continuously differentiable in the interval $[0,T]$; the matrix-valued functions $G_i(t,\eta)$, $(i=1,...,4)$, are piece-wise continuous in $\eta\in[-h,0]$ for any $t\in[0,T]$, and these functions are continuously differentiable in $t\in[0,T]$ uniformly with respect to $\eta\in[-h,0]$; E^k is k-dimensional real Euclidean space.

In what follows, we assume that:

$$\begin{aligned}&A_i(0)=A_i(T),\ H_i(0)=H_i(T),\ G_i(0,\eta)=G_i(T,\eta),\ \eta\in[-h,0],\ i=1,...,4,\\ &B_j(0)=B_j(T),\quad f_j(0)=f_j(T),\quad j=1,2.\end{aligned} \tag{2.3}$$

The conditions (2.3) are called the T-periodicity conditions or, simply, the periodicity conditions of the corresponding functions.

The cost functional, evaluating the controlled process (2.1)-(2.2), is

$$J=\int_0^T\Big[x^{'}(t)D_1(t)x(t)+2x^{'}(t)D_2(t)y(t)+y^{'}(t)D_3(t)y(t)+u^{'}(t)M(t)u(t)\Big]dt \tag{2.4}$$

where the prime denotes the transposition; the matrix-valued functions $D_k(t)$, $(k=1,2,3)$ and $M(t)$ are continuously differentiable for $t\in[0,T]$ and satisfy the conditions:

$$\begin{aligned}&D_1^{'}(t)=D_1(t),\ D_3^{'}(t)=D_3(t),\ D(t)\stackrel{\triangle}{=}\begin{pmatrix} D_1(t) & D_2(t)\\ D_2^{'}(t) & D_3(t)\end{pmatrix}>0,\quad t\in[0,T],\\ &D_k(0)=D_k(T),\quad k=1,2,3,\end{aligned} \tag{2.5}$$

$$M^{'}(t)=M(t),\quad M(t)>0,\quad t\in[0,T],\quad M(0)=M(T). \tag{2.6}$$

The optimal control problem is to choose a control $u(t)\in L^2[0,T;E^r]$, satisfying the periodicity condition $u(0)=u(T)$ and minimizing the cost functional (2.4) along trajectories of the system (2.1)-(2.2) subject to the periodicity condition $x(\tau)=x(T+\tau)$, $y(\tau)=y(T+\tau)$, $\tau\in[-\varepsilon h,0]$. We call this problem the Original Optimal Control Problem (OOCP).

2.2. Control optimality conditions in the OOCP. Consider the following block-form matrices and vector

$$A(t,\varepsilon)=\begin{pmatrix} A_1(t) & A_2(t)\\ \varepsilon^{-1}A_3(t) & \varepsilon^{-1}A_4(t)\end{pmatrix},\quad H(t,\varepsilon)=\begin{pmatrix} H_1(t) & H_2(t)\\ \varepsilon^{-1}H_3(t) & \varepsilon^{-1}H_4(t)\end{pmatrix}, \tag{2.7}$$

$$G(t,\eta,\varepsilon)=\begin{pmatrix} G_1(\eta,t) & G_2(t,\eta)\\ \varepsilon^{-1}G_3(t,\eta) & \varepsilon^{-1}G_4(t,\eta)\end{pmatrix},\quad B(t,\varepsilon)=\begin{pmatrix} B_1(t)\\ \varepsilon^{-1}B_2(t)\end{pmatrix}, \tag{2.8}$$

$$S(t,\varepsilon)=B(t,\varepsilon)M^{-1}(t)B^{'}(t,\varepsilon)=\begin{pmatrix} S_1(t) & \varepsilon^{-1}S_2(t)\\ \varepsilon^{-1}S_2^{'}(t) & \varepsilon^{-2}S_3(t)\end{pmatrix},\ f(t,\varepsilon)=\begin{pmatrix} f_1(t)\\ \varepsilon^{-1}f_2(t)\end{pmatrix},$$

(2.9)

$S_1(t)=B_1(t)M^{-1}(t)B_1^{'}(t)$, $S_2(t)=B_1(t)M^{-1}(t)B_2^{'}(t)$, $S_3(t)=B_2(t)M^{-1}(t)B_2^{'}(t)$.

Also, let us consider the following set of functional-differential equations (ordinary and partial) with respect to the matrix-valued functions $P(t)$, $Q(t,\tau)$, $R(t,\tau,\rho)$ in the domain $\Omega_\varepsilon=\{(t,\tau,\rho): t\in[0,T],\tau\in[-\varepsilon h,0],\rho\in[-\varepsilon h,0]\}$:

$$\begin{aligned} dP(t)/dt=&-P(t)A(t,\varepsilon)-A^{'}(t,\varepsilon)P(t)+P(t)S(t,\varepsilon)P(t)\\ &-Q(t,0)-Q^{'}(t,0)-D(t),\end{aligned}\tag{2.10}$$

$$\begin{aligned} (\partial/\partial t-\partial/\partial\tau)Q(t,\tau)=&-\big[A(t,\varepsilon)-S(t,\varepsilon)P(t)\big]^{'}Q(t,\tau)\\ &-\varepsilon^{-1}P(t)G(t,\tau/\varepsilon,\varepsilon)-R(t,0,\tau),\end{aligned}\tag{2.11}$$

$$\begin{aligned} (\partial/\partial t-\partial/\partial\tau-\partial/\partial\rho)R(t,\tau,\rho)=&-\varepsilon^{-1}G^{'}(t,\tau/\varepsilon,\varepsilon)Q(t,\rho)\\ &-\varepsilon^{-1}Q^{'}(t,\tau)G(t,\rho/\varepsilon,\varepsilon)+Q^{'}(t,\tau)S(t,\varepsilon)Q(t,\rho).\end{aligned}\tag{2.12}$$

The set (2.10)-(2.12) is subject to the boundary conditions

$$Q(t,-\varepsilon h)=P(t)H(t,\varepsilon),\tag{2.13}$$

$$R(t,-\varepsilon h,\tau)=H^{'}(t,\varepsilon)Q(t,\tau),\quad R(t,\tau,-\varepsilon h)=Q^{'}(t,\tau)H(t,\varepsilon).\tag{2.14}$$

Based on the results of the works [3, 5, 8, 23], we have the lemma.

LEMMA 2.1. *Let for a given $\varepsilon>0$, any $t\in[0,T]$ and any complex λ with $\mathrm{Re}(\lambda)\geq 0$, the following equality is valid:*

$$\begin{aligned}\operatorname{rank}\left[A(t,\varepsilon)+H(t,\varepsilon)\exp(-\lambda\varepsilon h)+\int_{-h}^{0}G(t,\eta,\varepsilon)\exp(\lambda\varepsilon\eta)d\eta-\lambda I_{n+m},B(t,\varepsilon)\right]\\ =n+m.\end{aligned}\tag{2.15}$$

Then, the optimal state-feedback control in the OOCP has the form

$$u^*[t,z_{\varepsilon h}(t)]=-M^{-1}(t)B^{'}(t,\varepsilon)\left[P(t,\varepsilon)z(t)+\int_{-\varepsilon h}^{0}Q(t,\tau,\varepsilon)z(t+\tau)d\tau+\varphi(t,\varepsilon)\right],$$
$$z=\mathrm{col}(x,y),\quad z_{\varepsilon h}(t)=\{z(t+\tau),\ \tau\in[-\varepsilon h,0]\},$$

where $P(t,\varepsilon)$ and $Q(t,\tau,\varepsilon)$ are the components of the unique solution $\{P(t,\varepsilon)$, $Q(t,\tau,\varepsilon),R(t,\tau,\rho,\varepsilon)\}$ of the problem (2.10)-(2.14) satisfying the periodicity condition

$$P(0,\varepsilon)=P(T,\varepsilon),\quad Q(0,\tau,\varepsilon)=Q(T,\tau,\varepsilon),\quad R(0,\tau,\rho,\varepsilon)=R(T,\tau,\rho,\varepsilon),\tag{2.16}$$

and such that for any $t\in[0,T]$ the matrix $\begin{pmatrix} P(t,\varepsilon) & Q(t,\rho,\varepsilon)\\ Q^{'}(t,\tau,\varepsilon) & R(t,\tau,\rho,\varepsilon)\end{pmatrix}$ defines a linear bounded self-adjoint positive operator mapping the space $E^{n+m}\times L^2[-\varepsilon h,0;E^{n+m}]$ into itself. Moreover, the $(n+m)$-vector-valued function $\varphi(t,\varepsilon)$ is the unique periodic

solution ($\varphi(0,\varepsilon)=\varphi(T,\varepsilon)$) *of the equation*

$$d\varphi(t,\varepsilon)/dt=-\big[A(t,\varepsilon)-S(t,\varepsilon)P(t,\varepsilon)\big]'\varphi(t,\varepsilon)$$
$$-\left\{\begin{array}{ll} H'(t+\varepsilon h,\varepsilon)\varphi(t+\varepsilon h,\varepsilon), & t+\varepsilon h\le T\\ 0, & \text{otherwise}\end{array}\right\}$$
$$-\int_{-h}^{0}\left\{\begin{array}{ll} \widetilde{G}(t,\eta,\varepsilon)\varphi(t-\varepsilon\eta,\varepsilon), & t-\varepsilon\eta\le T\\ 0, & \text{otherwise}\end{array}\right\}d\eta$$
$$-P(t,\varepsilon)f(t,\varepsilon)-\left\{\begin{array}{ll} \int_{t-T}^{0}Q(t-\tau,\tau,\varepsilon)f(t-\tau,\varepsilon)d\tau, & t\in(T-\varepsilon h,T]\\ \int_{-\varepsilon h}^{0}Q(t-\tau,\tau,\varepsilon)f(t-\tau,\varepsilon)d\tau, & t\in[0,T-\varepsilon h]\end{array}\right\},$$

where $\widetilde{G}(t,\eta,\varepsilon)=\big[G(t-\varepsilon\eta,\eta,\varepsilon)-\varepsilon S(t-\varepsilon\eta,\varepsilon)Q(t-\varepsilon\eta,\varepsilon\eta,\varepsilon)\big]'$.

The objective of the present paper is to solve the set (2.10)-(2.12) subject to the conditions (2.13)-(2.14) and (2.16). The solution of this problem, mentioned in Lemma 2.1, satisfies the symmetry conditions $P'(t,\varepsilon)=P(t,\varepsilon)$, $R'(t,\tau,\rho,\varepsilon)=R(t,\rho,\tau,\varepsilon)$, $(t,\tau,\rho)\in\Omega_\varepsilon$. The system (2.10)-(2.12) consists of the three functional-differential Riccati-type matrix equations singularly depending on ε. One of these equations is ordinary, while the others are partial. The equations are with deviating arguments. All these features make the solving this set to be an extremely difficult task. An asymptotic approach turns out to be very helpful in solution of this set. This approach allows us to partition the original set of Riccati-type equations into several much simpler and ε-free subsets. Due to the latter circumstance, an approximate (asymptotic) solution to the original set of equations is derived once, while being valid for all sufficiently small values of ε.

3. Asymptotic solution of the problem (2.10)-(2.14),(2.16).

3.1. Equivalent transformation of (2.10)-(2.14),(2.16). To remove the singularities at $\varepsilon=0$ from the right-hand sides of (2.10)-(2.12), we represent the solution $\{P(t,\varepsilon),Q(t,\tau,\varepsilon),R(t,\tau,\rho,\varepsilon)\}$ to (2.10)-(2.14),(2.16) in the block form

$$P(t,\varepsilon)=\begin{pmatrix} P_1(t,\varepsilon) & \varepsilon P_2(t,\varepsilon)\\ \varepsilon P_2'(t,\varepsilon) & \varepsilon P_3(t,\varepsilon)\end{pmatrix},\quad Q(t,\tau,\varepsilon)=\begin{pmatrix} Q_1(t,\tau,\varepsilon) & Q_2(t,\tau,\varepsilon)\\ Q_3(t,\tau,\varepsilon) & Q_4(t,\tau,\varepsilon)\end{pmatrix},$$
$$R(t,\tau,\rho,\varepsilon)=(1/\varepsilon)\begin{pmatrix} R_1(t,\tau,\rho,\varepsilon) & R_2(t,\tau,\rho,\varepsilon)\\ R_2'(t,\rho,\tau,\varepsilon) & R_3(t,\tau,\rho,\varepsilon)\end{pmatrix},\tag{3.1}$$

where $P_k(t,\varepsilon)$ and $R_k(t,\tau,\rho,\varepsilon)$, $(k=1,2,3)$, are matrices of dimensions $n\times n, n\times m, m\times m$, respectively; $Q_i(t,\tau,\varepsilon),(i=1,...,4)$, are matrices of dimensions $n\times n, n\times m, m\times n, m\times m$, respectively. Substitution of the block representations for the matrices $D(t)$, $A(t,\varepsilon)$, $H(t,\varepsilon)$, $G(t,\eta,\varepsilon)$, $S(t,\varepsilon)$, $P(t,\varepsilon)$, $Q(t,\tau,\varepsilon)$, $R(t,\tau,\rho,\varepsilon)$ (see (2.5),(2.7),(2.8),(2.9),(3.1)) into the problem (2.10)-(2.14),(2.16) yields after some rearrangement the following equivalent problem (in this problem, for simplicity, we omit the designation of the dependence of the unknown matrices on ε):

$$dP_1(t)/dt=-P_1(t)A_1(t)-A_1'(t)P_1(t)-P_2(t)A_3(t)-A_3'(t)P_2'(t)$$
$$+P_1(t)S_1(t)P_1(t)+P_1(t)S_2(t)P_2'(t)+P_2(t)S_2'(t)P_1(t)$$
$$+P_2(t)S_3(t)P_2'(t)-Q_1(t,0)-Q_1'(t,0)-D_1(t),\tag{3.2}$$

$$\varepsilon dP_2(t)/dt=-P_1(t)A_2(t)-P_2(t)A_4(t)-\varepsilon A_1'(t)P_2(t)-A_3'(t)P_3(t)$$
$$+\varepsilon P_1(t)S_1(t)P_2(t)+P_1(t)S_2(t)P_3(t)+\varepsilon P_2(t)S_2'(t)P_2(t)$$
$$+P_2(t)S_3(t)P_3(t)-Q_2(t,0)-Q_3'(t,0)-D_2(t),\tag{3.3}$$

$$\begin{aligned}\varepsilon dP_3(t)/dt = &-\varepsilon P_2^{'}(t)A_2(t) - \varepsilon A_2^{'}(t)P_2(t) - P_3(t)A_4(t) - A_4^{'}(t)P_3(t)\\ &+\varepsilon^2 P_2^{'}(t)S_1(t)P_2(t) + \varepsilon P_2^{'}(t)S_2(t)P_3(t) + \varepsilon P_3(t)S_2^{'}(t)P_2(t)\\ &+P_3(t)S_3(t)P_3(t) - Q_4(t,0) - Q_4^{'}(t,0) - D_3(t),\end{aligned} \tag{3.4}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau)Q_1(t,\tau) = &-\varepsilon\Big[A_1^{'}(t) - P_1(t)S_1(t) - P_2(t)S_2^{'}(t)\Big]Q_1(t,\tau)\\ &-\Big[A_3^{'}(t) - P_1(t)S_2(t) - P_2(t)S_3(t)\Big]Q_3(t,\tau) - P_1(t)G_1(t,\tau/\varepsilon)\\ &-P_2(t)G_3(t,\tau/\varepsilon) - R_1(t,0,\tau),\end{aligned} \tag{3.5}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau)Q_2(t,\tau) = &-\varepsilon\Big[A_1^{'}(t) - P_1(t)S_1(t) - P_2(t)S_2^{'}(t)\Big]Q_2(t,\tau)\\ &-\Big[A_3^{'}(t) - P_1(t)S_2(t) - P_2(t)S_3(t)\Big]Q_4(t,\tau) - P_1(t)G_2(t,\tau/\varepsilon)\\ &-P_2(t)G_4(t,\tau/\varepsilon) - R_2(t,0,\tau),\end{aligned} \tag{3.6}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau)Q_3(t,\tau) = &-\varepsilon\Big[A_2^{'}(t) - \varepsilon P_2^{'}(t)S_1(t) - P_3(t)S_2^{'}(t)\Big]Q_1(t,\tau)\\ &-\Big[A_4^{'}(t) - \varepsilon P_2^{'}(t)S_2(t) - P_3(t)S_3(t)\Big]Q_3(t,\tau) - \varepsilon P_2^{'}(t)G_1(t,\tau/\varepsilon)\\ &-P_3(t)G_3(t,\tau/\varepsilon) - R_2^{'}(t,\tau,0),\end{aligned} \tag{3.7}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau)Q_4(t,\tau) = &-\varepsilon\Big[A_2^{'}(t) - \varepsilon P_2^{'}(t)S_1(t) - P_3(t)S_2^{'}(t)\Big]Q_2(t,\tau)\\ &-\Big[A_4^{'}(t) - \varepsilon P_2^{'}(t)S_2(t) - P_3(t)S_3(t)\Big]Q_4(t,\tau) - \varepsilon P_2^{'}(t)G_2(t,\tau/\varepsilon)\\ &-P_3(t)G_4(t,\tau/\varepsilon) - R_3(t,0,\tau),\end{aligned} \tag{3.8}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau - \partial/\partial\rho)R_1(t,\tau,\rho) = &-\varepsilon G_1^{'}(t,\tau/\varepsilon)Q_1(t,\rho) - \varepsilon Q_1^{'}(t,\tau)G_1(t,\rho/\varepsilon)\\ &-G_3^{'}(t,\tau/\varepsilon)Q_3(t,\rho) - Q_3^{'}(t,\tau)G_3(t,\rho/\varepsilon) + \varepsilon^2 Q_1^{'}(t,\tau)S_1(t)Q_1(t,\rho)\\ &+\varepsilon Q_3^{'}(t,\tau)S_2^{'}(t)Q_1(t,\rho) + \varepsilon Q_1^{'}(t,\tau)S_2(t)Q_3(t,\rho) + Q_3^{'}(t,\tau)S_3(t)Q_3(t,\rho),\end{aligned} \tag{3.9}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau - \partial/\partial\rho)R_2(t,\tau,\rho) = &-\varepsilon G_1^{'}(t,\tau/\varepsilon)Q_2(t,\rho) - \varepsilon Q_1^{'}(t,\tau)G_2(t,\rho/\varepsilon)\\ &-G_3^{'}(t,\tau/\varepsilon)Q_4(t,\rho) - Q_3^{'}(t,\tau)G_4(t,\rho/\varepsilon) + \varepsilon^2 Q_1^{'}(t,\tau)S_1(t)Q_2(t,\rho)\\ &+\varepsilon Q_3^{'}(t,\tau)S_2^{'}(t)Q_2(t,\rho) + \varepsilon Q_1^{'}(t,\tau)S_2(t)Q_4(t,\rho) + Q_3^{'}(t,\tau)S_3(t)Q_4(t,\rho),\end{aligned} \tag{3.10}$$

$$\begin{aligned}\varepsilon(\partial/\partial t - \partial/\partial\tau - \partial/\partial\rho)R_3(t,\tau,\rho) = &-\varepsilon G_2^{'}(t,\tau/\varepsilon)Q_2(t,\rho) - \varepsilon Q_2^{'}(t,\tau)G_2(t,\rho/\varepsilon)\\ &-G_4^{'}(t,\tau/\varepsilon)Q_4(t,\rho) - Q_4^{'}(t,\tau)G_4(t,\rho/\varepsilon) + \varepsilon^2 Q_2^{'}(t,\tau)S_1(t)Q_2(t,\rho)\\ &+\varepsilon Q_4^{'}(t,\tau)S_2^{'}(t)Q_2(t,\rho) + \varepsilon Q_2^{'}(t,\tau)S_2(t)Q_4(t,\rho) + Q_4^{'}(t,\tau)S_3(t)Q_4(t,\rho),\end{aligned} \tag{3.11}$$

$$\begin{aligned}Q_j(t,-\varepsilon h) &= P_1(t)H_j(t) + P_2(t)H_{j+2}(t), \quad j = 1,2,\\ Q_l(t,-\varepsilon h) &= \varepsilon P_2^{'}(t)H_{l-2}(t) + P_3(t)H_l(t), \quad l = 3,4,\end{aligned} \tag{3.12}$$

$$
\begin{aligned}
R_1(t,-\varepsilon h,\tau) &= \varepsilon H_1^{'} Q_1(t,\tau) + H_3^{'} Q_3(t,\tau),\\
R_1(t,\tau,-\varepsilon h) &= \varepsilon Q_1^{'}(t,\tau) H_1 + Q_3^{'}(t,\tau) H_3,\\
R_2(t,-\varepsilon h,\tau) &= \varepsilon H_1^{'} Q_2(t,\tau) + H_3^{'} Q_4(t,\tau)\\
R_2(t,\tau,-\varepsilon h) &= \varepsilon Q_1^{'}(t,\tau) H_2 + Q_3^{'}(t,\tau) H_4,\\
R_3(t,-\varepsilon h,\tau) &= \varepsilon H_2^{'} Q_2(t,\tau) + H_4^{'} Q_4(t,\tau),\\
R_3(t,\tau,-\varepsilon h) &= \varepsilon Q_2^{'}(t,\tau) H_2 + Q_4^{'}(t,\tau) H_4. \qquad (3.13)
\end{aligned}
$$

$$
P_k(0) = P_k(T), \quad Q_i(0,\tau) = Q_i(T,\tau), \quad R_k(0,\tau,\rho) = R_k(T,\tau,\rho), \qquad (3.14)
$$

where $k = 1,2,3;\ i = 1,...,4$. In the set (3.2)-(3.11), the equations (3.3)-(3.11) are with the small multiplier ε for the derivatives. Hence, (3.2)-(3.11) is singularly perturbed.

3.2. Formal construction of the zero-order asymptotic solution to the problem (3.2)-(3.14). In the sequel we assume:

(A1) $\operatorname{rank}\Big[A_4(t) + H_4(t)\exp(-h\lambda) + \int_{-h}^{0} G_4(t,\eta)\exp(\eta\lambda)d\eta - \lambda I_m,\ B_2(t)\Big] = m$ for any $t \in [0,T]$ and any complex number λ with $\mathrm{Re}\lambda \geq 0$.

We seek the zero-order asymptotic solution $\{P_{k0}(t,\varepsilon), Q_{i0}(t,\tau,\varepsilon), R_{k0}(t,\tau,\rho,\varepsilon), (k = 1,2,3;\ i = 1,...,4)\}$ of (3.2)-(3.14) in the form

$$
\begin{aligned}
&P_{k0}(t,\varepsilon) = \bar{P}_{k0}(t), \quad Q_{i0}(t,\tau,\varepsilon) = Q_{i0}^{\tau}(t,\eta), \quad R_{k0}(t,\tau,\rho,\varepsilon) = R_{k0}^{\tau,\rho}(t,\eta,\chi),\\
&\qquad\qquad \eta = \tau/\varepsilon, \quad \chi = \rho/\varepsilon \quad k = 1,2,3, \quad i = 1,...,4. \qquad (3.15)
\end{aligned}
$$

Equations and conditions for (3.15) are obtained by its substitution into (3.2)-(3.14) instead of $P_k(t)$, $Q_i(t,\tau)$, $R_k(t,\tau,\rho)$, $(k = 1,2,3;\ i = 1,...,4)$, and equating coefficients for ε^0 on both sides of the resulting equations. Thus, for the terms of the asymptotic solution, we obtain the set of 10 equations (8 differential and 2 algebraic ones) in the domain $\bar{\Omega} = \big\{(t,\eta,\chi) : t \in [0,T], \eta \in [-h,0], \chi \in [-h,0]\big\}$, and 11 boundary conditions. It is remarkable that this set of the equations and the conditions can be partitioned into four simpler problems solved successively. Since the problem (3.2)-(3.14) is t-periodic, its asymptotic solution consists only of the outer solution.

3.2.1. The first problem. This problem has the form

$$
\begin{aligned}
&\bar{P}_{30}(t)A_4(t) + A_4^{'}(t)\bar{P}_{30}(t) - \bar{P}_{30}(t)S_3(t)\bar{P}_{30}(t) + Q_{40}^{\tau}(t,0) + [Q_{40}^{\tau}(t,0)]^{'} + D_3(t) = 0,\\
&\partial Q_{40}^{\tau}(t,\eta)/\partial\eta = \big[A_4^{'}(t) - \bar{P}_{30}(t)S_3(t)\big]Q_{40}^{\tau}(t,\eta) + \bar{P}_{30}(t)G_4(t,\eta) + R_{30}^{\tau,\rho}(t,0,\eta),\\
&(\partial/\partial\eta + \partial/\partial\chi)R_{30}^{\tau,\rho}(t,\eta,\chi) = G_4^{'}(t,\eta)Q_{40}^{\tau}(t,\chi) + [Q_{40}^{\tau}(t,\eta)]^{'}G_4(t,\chi)\\
&\qquad\qquad -[Q_{40}^{\tau}(t,\eta)]^{'}S_3(t)Q_{40}^{\tau}(t,\chi),\\
&\qquad\qquad Q_{40}^{\tau}(t,-h) = \bar{P}_{30}(t)H_4(t),\\
&R_{30}^{\tau,\rho}(t,-h,\eta) = H_4^{'}(t)Q_{40}^{\tau}(t,\eta), \qquad R_{30}^{\tau,\rho}(t,\eta,-h) = [Q_{40}^{\tau}(t,\eta)]^{'}H_4(t). \qquad (3.16)
\end{aligned}
$$

Remark 1. *In the problem (3.16), $\eta \in [h,0]$, $\chi \in [h,0]$ are independent variables, while $t \in [0,T]$ is a parameter. Since the coefficients of this problem are T-periodic, then its solution (if it exists and is unique) also is T-periodic with respect to t.*

Based on the results of [7, 23] and using Remark 1, we have the lemma.

Lemma 3.1. *Let the assumption A1 be satisfied. Then for any $t \in [0,T]$:*

(i) *the First Problem has a solution $\{\bar{P}_{30}(t), Q_{40}^{\tau}(t,\eta), R_{30}^{\tau,\rho}(t,\eta,\chi),\ (\eta,\chi) \in [-h,0] \times$*

$[-h,0]\}$ *such that* $\bar{P}_{30}(t) \geq 0$ *and the matrix* $\begin{pmatrix} \bar{P}_{30}(t) & Q^{\tau}_{40}(t,\chi) \\ \left(Q^{\tau}_{40}(t,\eta)\right)' & R^{\tau,\rho}_{30}(t,\eta,\chi) \end{pmatrix}$ *defines a linear bounded self-adjoint positive operator mapping the space* $E^m \times L^2[-h,0;E^m]$ *into itself;*

(ii) *such a solution of the First Problem is unique;*

(iii) *all roots* λ *of the equation* $\det\Big[A_4(t) - S_3(t)\bar{P}_{30}(t) + H_4(t)\exp(-\lambda h) + \int_{-h}^{0}\Big(G_4(t,\eta) - S_3(t)Q^{\tau}_{40}(t,\eta)\Big)\exp(\lambda\eta)d\eta - \lambda I_m\Big] = 0$ *lie inside the left-hand half-plane;*

(vi) $\bar{P}_{30}(0) = \bar{P}_{30}(T)$, $Q^{\tau}_{40}(0,\eta) = Q^{\tau}_{40}(T,\eta)$, $R^{\tau,\rho}_{30}(0,\eta,\chi) = R^{\tau,\rho}_{30}(T,\eta,\chi)$, $(\eta,\chi) \in [-h,0]\times[-h,0]$.

By virtue of the results of [13], we have the corollary.

COROLLARY 3.2. *Let the assumption (A1) be satisfied. Then, the derivatives* $d\bar{P}_{30}(t)/dt$, $\partial Q^{\tau}_{40}(t,\eta)/\partial t$, $\partial R^{\tau,\rho}_{30}(t,\eta,\chi)/\partial t$ *exist and are continuous functions of* $t \in [0,T]$ *uniformly in* $(\eta,\chi) \in [h,0]\times[h,0]$.

3.2.2. The second problem. This problem has the form

$$\begin{aligned} \partial Q^{\tau}_{30}(t,\eta)/\partial\eta &= \big[A_4'(t) - \bar{P}_{30}(t)S_3(t)\big]Q^{\tau}_{30}(t,\eta) + \bar{P}_{30}(t)G_3(t,\eta) + [R^{\tau,\rho}_{20}(t,\eta,0)]', \\ (\partial/\partial\eta + \partial/\partial\chi)R^{\tau,\rho}_{20}(t,\eta,\chi) &= G_3'(t,\eta)Q^{\tau}_{40}(t,\chi) + [Q^{\tau}_{30}(t,\eta)]'G_4(t,\chi) \\ &\quad -[Q^{\tau}_{30}(t,\eta)]'S_3(t)Q^{\tau}_{40}(t,\chi), \\ Q^{\tau}_{30}(t,-h) &= \bar{P}_{30}(t)H_3(t), \\ R^{\tau,\rho}_{20}(t,-h,\eta) = H_3'(t)Q^{\tau}_{40}(t,\eta), &\quad R^{\tau,\rho}_{20}(t,\eta,-h) = [Q^{\tau}_{30}(t,\eta)]'H_4(t). \end{aligned} \tag{3.17}$$

REMARK 2. *Like in the First Problem (3.16), in the Second problem (3.17)* $t \in [0,T]$ *is a parameter. Moreover, similarly to the First Problem, the solution of the Second Problem (if it exists and is unique) is* T*-periodic with respect to* t.

Based on Lemma 3.1, Corollary 3.2 and the results of [11], we obtain the lemma.

LEMMA 3.3. *Under the assumption A1, for any* $t \in [0,T]$, *the Second Problem has the unique solution* $\{Q^{\tau}_{30}(t,\eta), R^{\tau,\rho}_{20}(t,\eta,\chi),\ (\eta,\chi) \in [-h,0]\times[-h,0]\}$, *where* $Q^{\tau}_{30}(t,\eta)$ *is the unique solution of the initial-value problem for the integral-differential equation*

$$\begin{aligned} \partial Q^{\tau}_{30}(t,\eta)/\partial\eta &= \big[A_4'(t) - \bar{P}_{30}(t)S_3(t)\big]Q^{\tau}_{30}(t,\eta) \\ &+ \int_{-h}^{\eta}\Big[G_4(t,s-\eta) - S_3(t)Q^{\tau}_{40}(t,s-\eta)\Big]'Q^{\tau}_{30}(t,s)ds + [Q^{\tau}_{40}(t,-\eta-h)]'H_3(t) \\ &+ \int_{-h}^{\eta}[Q^{\tau}_{40}(t,s-\eta)]'G_3(t,s)ds, \qquad Q^{\tau}_{30}(t,-h) = \bar{P}_{30}(t)H_3(t). \end{aligned} \tag{3.18}$$

The matrix-valued function $R^{\tau,\rho}_{20}(t,\eta,\chi)$ *has the explicit form*

$$\begin{aligned} R^{\tau,\rho}_{20}(t,\eta,\chi) &= \Phi_{20}(t,\eta,\chi) + \int_{\max(\eta-\chi-h,-h)}^{\eta}\Big[G_3'(t,s)Q^{\tau}_{40}(t,s-\eta+\chi) \\ &+[Q^{\tau}_{30}(t,s)]'G_4(t,s-\eta+\chi) - [Q^{\tau}_{30}(t,s)]'S_3(t)Q^{\tau}_{40}(t,s-\eta+\chi)\Big]ds \\ \Phi_{20}(t,\eta,\chi) &= \begin{cases} H_3'(t)Q^{\tau}_{40}(t,\chi-\eta-h), & -h \leq \eta-\chi \leq 0 \\ \big(Q^{\tau}_{30}(t,\eta-\chi-h)\big)'H_4(t), & 0 < \eta-\chi \leq h. \end{cases} \end{aligned} \tag{3.19}$$

Moreover, $Q^{\tau}_{3}(0,\eta) = Q^{\tau}_{3}(T,\eta)$, $R^{\tau,\rho}_{20}(0,\eta,\chi) = R^{\tau,\rho}_{20}(T,\eta,\chi)$, $(\eta,\chi) \in [-h,0]\times[-h,0]$, *and the derivatives* $\partial Q^{\tau}_{3}(t,\eta)/\partial t$, $\partial R^{\tau,\rho}_{20}(t,\eta,\chi)/\partial t$ *exist and are continuous functions of* $t \in [0,T]$ *uniformly in* $(\eta,\chi) \in [-h,0]\times[-h,0]$.

3.2.3. The third problem. This problem has the form

$$(\partial/\partial\eta+\partial/\partial\chi)R_{10}^{\tau;\rho}(t,\eta,\chi)=G_3^{'}(t,\eta)Q_{30}^{\tau}(t,\chi)+[Q_{30}^{\tau}(t,\eta)]^{'}G_3(t,\chi)$$
$$-[Q_{30}^{\tau}(t,\eta)]^{'}S_3(t)Q_{30}^{\tau}(t,\chi),$$
$$R_{10}^{\tau;\rho}(-h,\eta)=H_3^{'}Q_{30}^{\tau}(\eta),\quad R_{10}^{\tau;\rho}(\eta,-h)=[Q_{30}(\eta)]^{'}H_3. \tag{3.20}$$

REMARK 3. *Similarly to the First and Second Problems, the solution of the Third Problem (3.20) (if it exists and is unique) is T-periodic in the parameter t.*

Using Lemma 3.3 and the results of [11], we obtain the lemma.

LEMMA 3.4. *Under the assumption A1, for any $t\in[0,T]$, the Third Problem has the unique solution $R_{10}^{\tau;\rho}(t,\eta,\chi)$, $(\eta,\chi)\in[-h,0]\times[-h,0]$:*

$$R_{10}^{\tau;\rho}(t,\eta,\chi)=\Phi_{10}(t,\eta,\chi)+\int_{\max(\eta-\chi-h,-h)}^{\eta}\Big[G_3^{'}(t,s)Q_{30}^{\tau}(t,s-\eta+\chi)$$
$$+[Q_{30}^{\tau}(t,s)]^{'}G_3(t,s-\eta+\chi)-[Q_{30}^{\tau}(t,s)]^{'}S_3(t)Q_{30}^{\tau}(t,s-\eta+\chi)\Big]ds$$
$$\Phi_{10}(t,\eta,\chi)=\begin{cases}H_3^{'}(t)Q_{30}^{\tau}(t,\chi-\eta-h), & -h\le\eta-\chi\le 0\\ \big(Q_{30}^{\tau}(t,\eta-\chi-h)\big)^{'}H_3(t), & 0<\eta-\chi\le h.\end{cases} \tag{3.21}$$

Moreover, $R_{10}^{\tau;\rho}(0,\eta,\chi)=R_{10}^{\tau;\rho}(T,\eta,\chi)$, $(\eta,\chi)\in[-h,0]\times[-h,0]$, and the derivative $\partial R_{10}^{\tau;\rho}(t,\eta,\chi)/\partial t$ exists and is a continuous function of $t\in[0,T]$ uniformly in $(\eta,\chi)\in[-h,0]\times[-h,0]$.

3.2.4. The fourth problem. This problem has the form

$$d\bar P_{10}(t)/dt=-\bar P_{10}(t)A_1(t)-A_1^{'}(t)\bar P_{10}(t)-\bar P_{20}(t)A_3(t)-A_3^{'}(t)\bar P_{20}^{'}(t)$$
$$+\bar P_{10}(t)S_1(t)\bar P_{10}(t)+\bar P_{10}(t)S_2(t)\bar P_{20}^{'}(t)+\bar P_{20}(t)S_2^{'}(t)\bar P_{10}(t)$$
$$+\bar P_{20}(t)S_3(t)\bar P_{20}^{'}(t)-Q_{10}^{\tau}(t,0)-[Q_{10}^{\tau}(t,0)]^{'}-D_1(t),$$
$$\bar P_{10}(t)A_2(t)+\bar P_{20}(t)A_4(t)+A_3^{'}(t)\bar P_{30}(t)-\bar P_{10}(t)S_2(t)\bar P_{30}(t)$$
$$-\bar P_{20}(t)S_3(t)\bar P_{30}(t)+Q_{20}^{\tau}(t,0)+[Q_{30}^{\tau}(t,0)]^{'}+D_2(t)=0,$$
$$\partial Q_{10}^{\tau}(t,\eta)/\partial\eta=\Big[A_3^{'}(t)-\bar P_{10}(t)S_2(t)-\bar P_{20}(t)S_3(t)\Big]Q_{30}^{\tau}(t,\eta)$$
$$+\bar P_{10}(t)G_1(t,\eta)+\bar P_{20}(t)G_3(t,\eta)+R_{10}^{\tau;\rho}(t,0,\eta),$$
$$\partial Q_{20}^{\tau}(t,\eta)/\partial\eta=\Big[A_3^{'}(t)-\bar P_{10}(t)S_2(t)-\bar P_{20}(t)S_3(t)\Big]Q_{40}^{\tau}(t,\eta)$$
$$+\bar P_{10}(t)G_2(t,\eta)+\bar P_{20}(t)G_4(t,\eta)+R_{20}^{\tau;\rho}(t,0,\eta),$$
$$\bar P_{10}(0)=\bar P_{10}(T),\quad Q_{j0}^{\tau}(t,-h)=\bar P_{10}(t)H_j(t)+\bar P_{20}(t)H_{j+2}(t),\quad j=1,2. \tag{3.22}$$

REMARK 4. *In the differential equation with respect to $\bar P_{10}(t)$, $t\in[0,T]$ is an independent variable, while in the rest of the equations of the Fourth Problem (3.22) t is a parameter.*

Using the results of [11], we obtain the lemma.

LEMMA 3.5. *Under the assumption A1, the Fourth Problem is equivalent to the following set of equations:*

$$d\bar P_{10}(t)/dt=-\bar P_{10}(t)\bar A(t)-\bar A^{'}(t)\bar P_{10}(t)+\bar P_{10}(t)\bar S(t)\bar P_{10}(t)-\bar D(t),\ \bar P_{10}(0)=\bar P_{10}(T),$$
$$\bar P_{20}(t)=-\left(\bar P_{10}(t)L_1(t)+L_2(t)+\int_{-h}^{0}[Q_{30}^{\tau}(t,\eta)]^{'}d\eta\right),$$

$$Q^{\tau}_{j0}(t,\eta)=\bar{P}_{10}(t)H_j(t)+\bar{P}_{20}(t)H_{j+2}(t)$$
$$+[A_3^{'}(t)-\bar{P}_{10}(t)S_2(t)-\bar{P}_{20}(t)S_3(t)]\int_{-h}^{\eta}Q^{\tau}_{j+2,0}(t,\sigma)d\sigma$$
$$+\bar{P}_{10}(t)\int_{-h}^{\eta}G_j(t,\sigma)d\sigma+\bar{P}_{20}(t)\int_{-h}^{\eta}G_{j+2}(t,\sigma)d\sigma+\int_{-h}^{\eta}R^{\tau,\rho}_{j0}(t,0,\sigma)d\sigma, \tag{3.23}$$

where $j=1,2$, $\bar{A}(t)=\hat{A}_1(t)-L_1(t)\hat{A}_3(t)+S_2(t)L_2^{'}(t)-L_1(t)S_3(t)L_2^{'}(t)$, $\hat{A}_i(t)=A_i(t)+H_i(t)+\int_{-h}^{0}G_i(t,\eta)d\eta$, $(i=1,...,4)$, $\bar{S}(t)=\bar{B}(t)M^{-1}(t)\bar{B}^{'}(t)$, $\bar{B}(t)=B_1(t)-L_1(t)B_2(t)$, $\bar{D}(t)=D_1(t)-L_2(t)\hat{A}_3(t)-\hat{A}_3^{'}(t)L_2^{'}(t)-L_2(t)S_3(t)L_2^{'}(t)$, $L_1(t)=(\hat{A}_2(t)-S_2(t)N(t))K^{-1}(t)$, $L_2(t)=(\hat{A}_3^{'}(t)N(t)+D_2(t))K^{-1}(t)$, $K(t)=\hat{A}_4(t)-S_3(t)N(t)$, $N(t)=\bar{P}_{30}(t)+\int_{-h}^{0}Q^{\tau}_{40}(t,\eta)d\eta$.

In what follows, we assume:

(A2) $\operatorname{rank}\big[\bar{A}(t)-\lambda I_n,\bar{B}(t)\big]=n$ for any $t\in[0,T]$ and any complex λ with $\operatorname{Re}\lambda\geq 0$;

(A3) $\bar{D}(t)>0$ for any $t\in[0,T]$.

Corollary 3.6. *Under the assumptions A1-A3, the Fourth Problem has the unique solution* $\big\{\bar{P}_{10}(t),\bar{P}_{20}(t),Q^{\tau}_{10}(t,\eta),Q^{\tau}_{20}(t,\eta),t\in[0,T],\eta\in[-h,0]\big\}$ *such that* $\bar{P}_{10}(t)>0$, $t\in[0,T]$. *Moreover,* $\bar{P}_{20}(0)=\bar{P}_{20}(T)$, $Q^{\tau}_{10}(0,\eta)=Q^{\tau}_{10}(T,\eta)$, $Q^{\tau}_{20}(0,\eta)=Q^{\tau}_{20}(T,\eta)$, $\eta\in[-h,0]$, *and the derivatives* $d\bar{P}_{10}(t)/dt$, $d\bar{P}_{20}(t)/dt$, $\partial Q^{\tau}_{10}(t,\eta)/\partial t$, $\partial Q^{\tau}_{20}(t,\eta)/\partial t$ *exist and are continuous functions of* $t\in[0,T]$ *uniformly in* $\eta\in[-h,0]$.

Thus, the formal construction of the zero-order asymptotic solution to the problem (3.2)-(3.14) is completed.

3.3. Justification of the zero-order asymptotic solution to the problem (3.2)-(3.14). Consider the matrix

$$\begin{pmatrix} \bar{P}_{30}(t) & Q^{\tau}_{30}(\chi) & Q^{\tau}_{40}(\chi) \\ \big(Q^{\tau}_{30}(\eta)\big)^{'} & R^{\tau,\rho}_{10}(\eta,\chi) & R^{\tau,\rho}_{20}(\eta,\chi) \\ \big(Q^{\tau}_{40}(\eta)\big)^{'} & \big(R^{\tau,\rho}_{20}(\chi,\eta)\big)^{'} & R^{\tau,\rho}_{30}(\eta,\chi) \end{pmatrix}.$$

For any $t\in[0,T]$, this matrix defines a linear bounded self-adjoint operator $\mathcal{F}_t$ mapping the space $E^m\times L^2[-h,0;E^{n+m}]$ into itself. In what follows, we assume:

(A4) For any $t\in[0,T]$, the operator $\mathcal{F}_t$ is uniformly positive.

Using Lemmas 3.1, 3.3, 3.4, Corollaries 3.2, 3.6 and the results of [10, 11], we obtain the theorem.

Theorem 3.7. *Let the assumptions A1-A4 be valid. Then, there exists a number* $\varepsilon^*>0$ *such that for all* $\varepsilon\in(0,\varepsilon^*]$*:*

(I) the problem (3.2)-(3.14) has the unique solution $\big\{P_k(t,\varepsilon),Q_i(t,\tau,\varepsilon),R_k(t,\tau,\rho,\varepsilon),(k=1,2,3;i=1,...,4)\big\}$ *in the domain* Ω_ε *such that for any* $t\in[0,T]$ *the matrix* $\begin{pmatrix} P(t,\varepsilon) & Q(t,\rho,\varepsilon) \\ Q^{'}(t,\tau,\varepsilon) & R(t,\tau,\rho,\varepsilon)\end{pmatrix}$, *where* $P(t,\varepsilon)$, $Q(t,\tau,\varepsilon)$, $R(t,\tau,\rho,\varepsilon)$ *are given by (3.1), defines a linear bounded self-adjoint positive operator mapping the space* $E^{n+m}\times L^2[-\varepsilon h,0;E^{n+m}]$ *into itself;*

(II) this solution satisfies the inequalities $\big\|P_k(t,\varepsilon)-\bar{P}_{k0}(t)\big\|\leq a\varepsilon$, $\big\|Q_{i0}(t,\tau,\varepsilon)-Q^{\tau}_{i0}(t,\tau/\varepsilon)\big\|\leq a\varepsilon$, $\big\|R_k(t,\tau,\rho,\varepsilon)-R^{\tau,\rho}_{k0}(t,\tau/\varepsilon,\rho/\varepsilon)\big\|\leq a\varepsilon$, $(k=1,2,3;\ i=1,...,4)$, $(t,\tau,\rho)\in\Omega_\varepsilon$, *where* $a>0$ *is some constant independent of* ε.

Remark 5. *Note, that the* ε*-free assumptions A1-A2 yield the fulfilment of the equality (2.15) providing the existence and uniqueness of the corresponding solution to the problem (2.10)-(2.14),(2.16) for all* $\varepsilon\in(0,\varepsilon^*]$. *Moreover, these conditions, along with A3-A4, guarantee the validity of the inequalities presented in Theorem 3.7.*

REFERENCES

[1] S. Bittanti, A. Locatelli, and C. Maffezzoni, *Second-variation methods in periodic optimization*, J. Optim. Theory Appl., 14 (1974), pp. 31–48.

[2] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences, Vol. 8, Springer-Verlag, New York, NY, 1978.

[3] G. Da Prato and A. Ichikawa, *Quadratic control of linear periodic systems*, Appl. Math. Optim., 18 (1988), pp. 39–66.

[4] R. Datko, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[5] M. C. Delfour, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1976), pp. 101–162.

[6] M. C. Delfour, *The linear-quadratic optimal control problem with delays in state and control variables: a state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.

[7] M. C. Delfour, C. McCalla, and S. K. Mitter, *Stability and the infinite-time quadratic cost problem for linear hereditary differential systems*, SIAM J. Control, 13 (1975), pp. 48–88.

[8] M. C. Delfour and S. K. Mitter, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, SIAM J. Control, 10 (1972), pp. 298–328.

[9] M. G. Dmitriev, *On singular perturbations in a linear periodic optimal control problem with a quadratic functional*, in Proceedings of the 8th International Conference on Nonlinear Oscillations, Prague, 1978, pp. 861-866, (in Russian).

[10] V. Y. Glizer, *Infinite horizon quadratic control of linear singularly perturbed systems with small state delays: an asymptotic solution of Riccati-type equations*. IMA J. Math. Control Inform., 24 (2007), pp. 435–459.

[11] V. Y. Glizer, *Linear-quadratic optimal control problem for singularly perturbed systems with small delays*, in Nonlinear Analysis and Optimization II, A. Leizarowitz, B. S. Mordukhovich, I. Shafrir and A. J. Zaslavski, eds., Contemporary Mathematics Series, Vol. 514, American Mathematical Society, Providence, RI, 2010, pp. 155–188.

[12] V. Y. Glizer, *Stochastic singular optimal control problem with state delays: regularization, singular perturbation, and minimizing sequence*, SIAM J. Control Optim., 50 (2012), pp. 2862–2888.

[13] V. Y. Glizer, *Dependence on parameter of the solution to an infinite horizon linear-quadratic optimal control problem for systems with state delays*. Pure Appl. Funct. Anal., 2 (2017), pp. 259–283.

[14] V. Y. Glizer and M. G. Dmitriev, *Singular perturbations in a linear control problem with a quadratic functional*, Differ. Equ., 11 (1975), pp. 1427–1432.

[15] V. Y. Glizer and M. G. Dmitriev, *Asymptotic properties of the solution of a singularly perturbed Cauchy problem encountered in optimal control theory*, Differ. Equ., 14 (1978), pp. 423–432.

[16] R. E. Kalman, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mex., 5 (1960), pp. 102–119.

[17] P. V. Kokotovic and R. A. Yackel, *Singular perturbation of linear regulators: basic theorems*, IEEE Trans. Automat. Control, 17 (1972), pp. 29–37.

[18] V. B. Kolmanovskii and T. L. Maizenberg, *Optimal control of stochastic systems with aftereffect*, Autom. Remote Control, 34 (1973), pp. 39–52.

[19] H. J. Kushner and D. I. Barnea, *On the control of a linear functional-differential equation with quadratic cost*, SIAM J. Control, 8 (1970), pp. 257–272.

[20] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, NY, 1971.

[21] R. E. O'Malley and C. F. Kung, *On the matrix Riccati approach to a singularly perturbed regulator problem*, J. Differential Equations, 16 (1974), pp. 413–427.

[22] M. Osintcev and V. Sobolev, *Regularization of the matrix Riccati equation in optimal estimation problem with low measurement noise*, J. Phys.: Conf. Ser., 811 (2017), pp. 1-6.

[23] R. B. Vinter and R. H. Kwong, *The infinite time quadratic control problem for linear systems with state and control delays: an evolution equation approach*, SIAM J. Control Optim., 19 (1981), pp. 139–153.

[24] R. A. Yackel and P. V. Kokotovic, *A boundary layer method for the matrix Riccati equation*, IEEE Trans. Automat. Control, 18 (1973), pp. 17–24.

Proceedings of EQUADIFF 2017
pp. 157–162

NONEXISTENCE OF SOLUTIONS OF SOME INEQUALITIES WITH GRADIENT NONLINEARITIES AND FRACTIONAL LAPLACIAN*

EVGENY GALAKHOV† AND OLGA SALIEVA‡

Abstract. We obtain sufficient conditions for nonexistence of nontrivial solutions for some classes of nonlinear partial differential inequalities containing the fractional powers of the Laplace operator.

Key words. Nonexistence, nonlinear inequalities, fractional Laplacian.

AMS subject classifications. 35J61, 35J48, 35S05

1. Introduction. The necessary conditions of solvability of nonlinear partial differential equations and inequalities has been recently studied by many authors.

In particular, in [4, 1, 2] (see also references therein) such conditions were obtained for some classes of nonlinear elliptic and parabolic inequalities, in particular containing integer powers of the Laplacian, using the test function method developed by S. Pohozaev [5]. However, for similar inequalities with fractional powers of the Laplacian the problem remained open. For such inequalities with nonlinear terms of the form u^q it was considered in [6].

In the present paper we obtain sufficient conditions for nonexistence of solutions for a class of elliptic inequalities with fractional powers of the Laplacian and nonlinear terms of the form $|Du|^q$, as well as for elliptic systems of the same type.

The rest of the paper consists of three sections. In §2 we obtain some auxiliary estimates for the fractional Laplacian used further. In §3, we prove a nonexistence theorem for single elliptic inequalities with fractional powers of the Laplacian, and in §4, for systems of such inequalities.

2. Auxiliary estimates. We define the operator $(-\Delta)^s$ by the formula

$$(-\Delta)^s u(x) \stackrel{\text{def}}{=} c_{n,s} \cdot \text{p.v.} \int\limits_{\mathbb{R}^n} \frac{(-\Delta)^{[s]}u(y) - (-\Delta)^{[s]}u(x)}{|x-y|^{n+2\{s\}}}\, dy, \tag{2.1}$$

where

$$c_{n,s} \stackrel{\text{def}}{=} \frac{2^{\{s\}}\Gamma\left(\frac{n+\{s\}}{2}\right)}{\pi^{n/2}\left|\Gamma\left(-\frac{\{s\}}{2}\right)\right|}$$

(see, e.g., [3]).

We will use definition (2.1) for the proof of the following Lemmas.

*The publication was supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 05.Y09.21.0013 of May 19, 2017).

†Peoples Friendship University of Russia, ul. Miklukho-Maklaya 6, 117198, Moscow, Russia (egalakhov@gmail.com).

‡Moscow State Technological University Stankin, Vadkovsky lane 3a, 127055, Moscow, Russia (olga.a.salieva@gmail.com).

LEMMA 2.1. *Let $s \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$ and $q, q' > 1, \frac{1}{q} + \frac{1}{q'} = 1$. Consider a function $\varphi_1 : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\varphi_1(x) \stackrel{\text{def}}{=} \begin{cases} 1 & (|x| \le 1), \\ (2-|x|)^\lambda & (1 < |x| < 2), \\ 0 & (|x| \ge 2) \end{cases} \tag{2.2}$$

with $\lambda > [s] + 2q'$. Then one has

$$\int\limits_{\mathbb{R}^n} |(-\Delta)^s \varphi_1(x)|^{q'} (1+|x|)^{-\frac{\alpha q'}{q}} \varphi_1^{1-q'}(x)\, dx < \infty. \tag{2.3}$$

Remark. In the Mitidieri–Pohozaev approach such estimates were established by direct calculation of the iterated Laplacian of the test functions. This does not work for the fractional Laplacian, so we need to establish some additional estimates.

Proof. Let $\frac{3}{2} < |x| < 1$. Use (2.1) with notation $f(x,y) = \dfrac{\Delta^{[s]}\varphi_1(x) - \Delta^{[s]}\varphi_1(y)}{|x-y|^{n+2\{s\}}}$:

$$|(-\Delta)^s \varphi_1)(x)| = c_{n,s} \left| \int\limits_{\mathbb{R}^n} f(x,y)\, dy \right| = c_{n,s} \left| \sum_{i=1}^{2} \int\limits_{D_i} f(x,y)\, dy \right|, \tag{2.4}$$

where

$$D_1 \stackrel{\text{def}}{=} \{y \in \mathbb{R}^n : |x-y| \ge (2-|x|)/2\},$$
$$D_2 \stackrel{\text{def}}{=} \{y \in \mathbb{R}^n : |x-y| < (2-|x|)/2\}$$

(here and below the singular integrals are understood in the sense of the Cauchy principal value).

For any $\varepsilon \in (0, 2\{s\})$, since we have $|x-y| \ge (2-|x|)/2$ in D_1, we get

$$\begin{aligned} &\int\limits_{D_1} f(x,y)\, dy = \int\limits_{D_1} \frac{(-\Delta)^{[s]}\varphi_1(x) - (-\Delta)^{[s]}\varphi_1(y)}{|x-y|^{n+2\{s\}}}\, dy \le \\ &\le (-\Delta)^{[s]}\varphi_1(x) \int\limits_{D_1} \frac{dy}{|x-y|^{n+2\{s\}}} \le \\ &\le (-\Delta)^{[s]}\varphi_1(x) \cdot \left(\frac{2-|x|}{2}\right)^{\varepsilon-2s} \int\limits_{D_1} \frac{dy}{|x-y|^{n+\varepsilon}} \le c_1 (2-|x|)^{\lambda+\varepsilon-2s} \end{aligned} \tag{2.5}$$

with some constant $c_1 > 0$.

Finally, the Lagrange Mean Value Theorem implies that

$$\begin{aligned} &\int\limits_{D_2} f(x,y)\, dy = \\ &= \frac{1}{2} \int\limits_{\tilde{D}_2} \frac{2(-\Delta)^{[s]}\varphi_1(x) - (-\Delta)^{[s]}\varphi_1(x+z) + (-\Delta)^{[s]}\varphi_1(x-z)}{|z|^{n+2s}}\, dz \le \\ &\le c_2 \cdot \max_{z \in \tilde{D}_2} |((2-|x+z|)^{\lambda-[s]})''| \int\limits_{\tilde{D}_2} \frac{|z|^2}{|z|^{n+2\{s\}}\, dy} = \\ &= c_3 \cdot \max_{z \in \tilde{D}_2} (2-|x+z|)^{\lambda-[s]-2} \cdot \int\limits_{\tilde{D}_2} \frac{dz}{|z|^{n+2\{s\}-2}}, \end{aligned}$$

where $\tilde{D}_2 = \{z \in \mathbb{R}^n : |z| < (2 - |x|)/2\}$, with constants c_2, $c_3 > 0$.

For $z \in \tilde{D}_2$ we have

$$2 - |x + z| = 2 - |x| + |x| - |x + z| \le (2 - |x|) + |z| \le \frac{3}{2}(2 - |x|).$$

Hence

$$\int_{D_2} f(x, y)\, dy \le c_4(2 - |x|)^{\lambda - [s] - 2} \tag{2.6}$$

with some constant $c_4 > 0$.

Combining (2.4)–(2.6), we obtain

$$|(-\Delta)^s \varphi_1(x)| \le c_5(2 - |x|)^{\lambda - [s] - 2} \tag{2.7}$$

and consequently

$$|(-\Delta)^s \varphi_1(x)|^{q'} (1 + |x|)^{-\frac{\alpha q'}{q}} \varphi_1^{1-q'}(x) \le$$

$$\le c_6(2 - |x|)^{(\lambda - [s] - 2)q' - \lambda(1 - q')} = c_6(2 - |x|)^{\lambda - ([s]+2)q'}$$

with some constants $c_5, c_6 > 0$ independent of x, which implies (2.3). □

LEMMA 2.2. *Let $s \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$ and $q, q' > 1, \dfrac{1}{q} + \dfrac{1}{q'} = 1$. For a family of functions $\varphi_R(x) = \varphi_1\left(\dfrac{x}{R}\right)$, where $R > 0$, one has*

$$\int_{\mathbb{R}^n} |(-\Delta)^s \varphi_R(x)|^{q'} (1 + |x|)^{-\frac{\alpha q'}{q}} \varphi_R^{1-q'}(x)\, dx \le cR^{n - 2q's - \frac{\alpha q'}{q}} \tag{2.8}$$

for any $R > 0$ and some $c > 0$ independent of R.

Proof. By (2.1) and a change of variables $\tilde{y} = \dfrac{y}{R}$, we have

$$(-\Delta)^s \varphi_R(x) = R^{-2s}(-\Delta)^s \varphi_1(x). \tag{2.9}$$

Substituting (2.9) into the left-hand side of (2.8) and applying Lemma 2.1, we obtain the claim. □

3. Single elliptic inequalities. Now consider the nonlinear elliptic inequality

$$(-\Delta)^s u \ge c|Du|^q (1 + |x|)^\alpha \quad (x \in \mathbb{R}^n), \tag{3.1}$$

where $s > 1$, $c > 0$, $q > 1$ and α are real numbers.

DEFINITION 3.1. *A weak solution of inequality* (3.1) *is a function $u \in W^{1,q}_{\mathrm{loc}}(\mathbb{R}^n)$ such that for any nonnegative function $\varphi \in C_0^\infty(\mathbb{R}^n)$ there holds the inequality*

$$-\int_{\mathbb{R}^n} (Du, D((-\Delta)^{s-1}\varphi))\, dx \ge c \int_{\mathbb{R}^n} |Du|^q (1 + |x|)^\alpha \varphi\, dx. \tag{3.2}$$

THEOREM 3.2. *Inequality* (3.1) *has no nontrivial (i.e., distinct from a constant a.e.) weak solutions for* $\alpha > 1 - 2s$ *and*

$$1 < q \le \frac{n+\alpha}{n-2s+1}. \tag{3.3}$$

Proof. Introduce a test function $\varphi_R(x) = \varphi_1\left(\frac{x}{R}\right)$, where $\varphi_1 \in C_0^\infty(\mathbb{R}^n)$ is nonnegative and

$$\varphi_1(x) = \begin{cases} 1 & (|x| \le 1), \\ 0 & (|x| \ge 2). \end{cases} \tag{3.4}$$

Substituting $\varphi(x) = \varphi_R(x)$ into (3.1) and applying the Hölder inequality, we get

$$\begin{aligned} & c\int\limits_{\mathbb{R}^n} |Du|^q (1+|x|)^\alpha \varphi_R\,dx \le -\int\limits_{\mathbb{R}^n} (Du, D((-\Delta)^{s-1}\varphi))\varphi_R\,dx \le \\ & \le \int\limits_{\mathbb{R}^n} |Du|\cdot|D((-\Delta)^{s-1}\varphi_R)|\,dx \le \left(\int\limits_{\mathbb{R}^n} |Du|^q(1+|x|)^\alpha\varphi_R\,dx\right)^{\frac{1}{q}} \times \\ & \times \left(\int\limits_{\operatorname{supp}|D\varphi_R|} |(-\Delta)^s\varphi_R|^{q'}(1+|x|)^{\frac{\alpha q'}{q}}\varphi_R^{1-q'}\,dx\right)^{\frac{1}{q}}, \end{aligned} \tag{3.5}$$

where $\dfrac{1}{q} + \dfrac{1}{q'} = 1$. Hence,

$$\int\limits_{\mathbb{R}^n} |Du|^q(1+|x|)^\alpha\varphi_R\,dx \le c\int\limits_{\mathbb{R}^n} |D((-\Delta)^{s-1}\varphi_R)|^{q'}(1+|x|)^{\frac{\alpha q'}{q}}\varphi_R^{1-q'}\,dx. \tag{3.6}$$

From Lemma 2.2 we have

$$\begin{aligned} & \int\limits_{\mathbb{R}^n} |(-\Delta)^s\varphi_R|^{q'}(1+|x|)^{\frac{\alpha q'}{q}}\varphi_R^{1-q'}\,dx \le \\ & cR^{n-q'(2s-1)-\frac{\alpha q'}{q}} \int\limits_{\mathbb{R}^n} |(-\Delta)^s\varphi_1(y)|^{q'}(1+|y|)^{\frac{\alpha q'}{q}}\varphi_1^{1-q'}(y)\,dy, \end{aligned} \tag{3.7}$$

where $y = \dfrac{x}{R}$. Combining (3.6) and (2.3), since the integral on the right-hand side of (3.7) converges for an appropriate choice of $\varphi_1(y)$, we obtain

$$\int\limits_{\mathbb{R}^n} |Du|^q(1+|x|)^\alpha\varphi_R\,dx \le cR^{n-q'(2s-1)-\frac{\alpha q'}{q}}.$$

Taking $R \to \infty$, in case of strict inequality in (3.3) we come to a contradiction, which proves the claim. In case of equality, we have

$$\int\limits_{\mathbb{R}^n} |Du|^q(1+|x|)^\alpha\,dx < \infty,$$

whence

$$\int\limits_{\mathrm{supp}|D\varphi_R|} |Du|^q(1+|x|)^\alpha \varphi_R\,dx \to 0 \text{ for } R\to\infty$$

and by (3.5)

$$\int\limits_{\mathbb{R}^n} |Du|^q(1+|x|)^\alpha\,dx = 0,$$

which completes the proof in this case as well. □

4. Systems of elliptic inequalities. Here we consider a system of nonlinear elliptic inequalities

$$\begin{cases} (-\Delta)^{s_1}u \ge c_1|Dv|^{q_1}(1+|x|)^{\alpha_1} & (x\in\mathbb{R}^n),\\ (-\Delta)^{s_2}u \ge c_2|Du|^{q_2}(1+|x|)^{\alpha_2} & (x\in\mathbb{R}^n), \end{cases} \tag{4.1}$$

where $s_1>1$, $s_2>1$, $q_1>1$, $q_2>1$, α_1 and α_2 are real numbers.

DEFINITION 4.1. *A weak solution of system of inequalities (3.7) is a pair of functions* $(u,v)\in W^{1,q_2}_{\mathrm{loc}}(\mathbb{R}^n)\times W^{1,q_1}_{\mathrm{loc}}(\mathbb{R}^n)$ *such that for any nonnegative function* $\varphi\in C_0^\infty(\mathbb{R}^n)$ *there hold the inequalities*

$$\begin{aligned} \int\limits_{\mathbb{R}^n}(Du, D((-\Delta)^{s_1}\varphi))\,dx &\ge c_1\int\limits_{\mathbb{R}^n}|Dv|^{q_1}(1+|x|)^{\alpha_1}\varphi\,dx,\\ \int\limits_{\mathbb{R}^n}(Dv, D((-\Delta)^{s_2}\varphi))\,dx &\ge c_2\int\limits_{\mathbb{R}^n}|Du|^{q_2}(1+|x|)^{\alpha_2}\varphi\,dx. \end{aligned} \tag{4.2}$$

Denote

$$\begin{aligned} \beta_1 &= q_1((2s_2-1)q_2-(2s_1-1)-\alpha_2)-\alpha_1,\\ \beta_2 &= q_2((2s_1-1)q_1-(2s_2-1)-\alpha_2)-\alpha_2. \end{aligned}$$

We will prove the following

THEOREM 4.2. *System* (4.1) *has no nontrivial (i.e., distinct from constants a.e.) weak solutions for*

$$n(q_1q_2-1)\le\max\{\beta_1,\beta_2\}. \tag{4.3}$$

Proof. Introduce a test function $\varphi_R(x)$ as in the proof of the previous theorem. Similarly to (3.5), we get

$$\begin{aligned} c_1\int\limits_{\mathbb{R}^n} v^{q_1}(1+|x|)^{\alpha_1}\varphi_R\,dx \le &\left(\int\limits_{\mathbb{R}^n}|Du|^{q_2}(1+|x|)^{\alpha_2}\varphi_R\,dx\right)^{\frac{1}{q_2}}\times\\ \times&\left(\int\limits_{\mathrm{supp}|D\varphi_R|}|D((-\Delta)^{s_2}\varphi_R)|^{q_2'}(1+|x|)^{\frac{\alpha_2q_2'}{q_2}}\varphi_R^{1-q_2'}\,dx\right)^{\frac{1}{q_2'}}, \end{aligned}$$

$$c_2 \int\limits_{\mathbb{R}^n} u^{q_2}(1+|x|)^{\alpha_2}\varphi_R\,dx \le \left(\int\limits_{\mathbb{R}^n} |Dv|^{q_1}(1+|x|)^{\alpha_1}\varphi_R\,dx\right)^{\frac{1}{q_1}} \times$$
$$\times \left(\int\limits_{\operatorname{supp}|D\varphi_R|} |D((-\Delta)^{s_1}\varphi_R)|^{q_1'}(1+|x|)^{\frac{\alpha_1 q_1'}{q_1}}\varphi_R^{1-q_1'}\,dx\right)^{\frac{1}{q_1'}},$$

where $\dfrac{1}{q_1}+\dfrac{1}{q_1'}=\dfrac{1}{q_2}+\dfrac{1}{q_2'}=1$. Estimating the second factors on the right-hand sides of the obtained inequalities similarly to (2.3), we get

$$(4.4)\int\limits_{\mathbb{R}^n} |Dv|^{q_1}(1+|x|)^{\alpha_1}\varphi_R\,dx \le cR^{\frac{n}{q_2'}-(2s_2-1)-\frac{\alpha_2}{q_2}}\left(\int\limits_{\mathbb{R}^n} u^{q_2}(1+|x|)^{\alpha_2}\varphi_R\,dx\right)^{\frac{1}{q_2}},$$

$$(4.5)\int\limits_{\mathbb{R}^n} |Du|^{q_2}(1+|x|)^{\alpha_2}\varphi_R\,dx \le cR^{\frac{n}{q_1'}-(2s_1-1)-\frac{\alpha_1}{q_1}}\left(\int\limits_{\mathbb{R}^n} v^{q_1}(1+|x|)^{\alpha_1}\varphi_R\,dx\right)^{\frac{1}{q_1}}$$

and, substituting (4.5) into (4.4) and vice versa,

$$\int\limits_{\mathbb{R}^n} |Dv|^{q_1}(1+|x|)^{\alpha_1}\varphi_R\,dx \le cR^{n-\frac{q_1((2s_2-1)q_2-(2s_1-1)-\alpha_2)-\alpha_1}{q_1q_2-1}},$$

$$\int\limits_{\mathbb{R}^n} |Du|^{q_2}(1+|x|)^{\alpha_2}\varphi_R\,dx \le cR^{n-\frac{q_2((2s_1-1)q_1-(2s_2-1)-\alpha_1)-\alpha_2}{q_1q_2-1}}.$$

Passing to the limit as $R\to\infty$, we complete the proof of the theorem similarly to the previous one. □

REFERENCES

[1] E. Galakhov and O. Salieva, *Blow-up for some nonlinear inequalities with singularities on unbounded sets*, Math. Notes, 98 (2015), pp. 222-229.

[2] E. Galakhov and O. Salieva, *On blow-up of solutions to differential inequalities with singularities on unbounded sets*, J. Math. Anal. Appl., 408 (2013), pp. 102–113.

[3] M. Kwaśnicki, *Ten equivalent definitions of the fractional Laplace operator*, Frac. Calc. Appl. Anal., 20 (2017), pp. 7–51.

[4] E. Mitidieri and S. Pohozaev, *A priori estimates and nonexistence fo solutions of nonlinear partial differential equations and inequalities*, Proc. Steklov Math. Inst., 234 (2001), pp. 3-383.

[5] S. Pohozaev, *Essentially nonlinear capacities induced by differential operators*, Dokl. RAN, 357 (1997), pp. 592-594.

[6] O. Salieva, *Nonexistence of solutions of some nonlinear inequalities with fractional powers of the Laplace operator*, Math. Notes, 101 (2017), pp. 699–703.

Proceedings of EQUADIFF 2017
pp. 163–172

SEMI-ANALYTICAL APPROACH TO INITIAL PROBLEMS FOR SYSTEMS OF NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS WITH CONSTANT DELAY.

HELENA ŠAMAJOVÁ *

Abstract. This paper deals with the differential transform method for solving of an initial value problem for a system of two nonlinear functional partial differential equations of parabolic type. We consider non-delayed as well as delayed types of coupling and the different variety of initial functions are thought over. The convergence of solutions and the error estimation to the presented procedure is studied. Two numerical examples for non-delayed and delayed systems are included.

Key words. nonlinear partial differential equation, parabolic type equation, delayed equation, system of partial differential equation, initial problem

AMS subject classifications. 35K55, 35K51 35K61

1. Introduction. We consider a system of two nonlinear functional partial differential equations of parabolic type with constant delays

$$\begin{aligned}\frac{\partial y_1(x,t)}{\partial t} =& \frac{\partial^2 y_1(x,t)}{\partial x^2} + K_1(y_2(x,t-\tau_1) - y_1(x,t)) + \eta_1 y_1^3(x,t)\\ \frac{\partial y_2(x,t)}{\partial t} =& \frac{\partial^2 y_2(x,t)}{\partial x^2} + K_2(y_1(x,t-\tau_2) - y_2(x,t)) + \eta_2 y_2^3(x,t)\end{aligned} \tag{1.1}$$

with the given initial function $\tilde{\psi}_i(x,t)$, constant delays τ_i , constants η_i , and K_i where $i = 1, 2$.

We may rewrite the system (1.1) into the vector form

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + \kappa_1 u(x,t) + \kappa_2 \hat{u}(x,t) + \eta \tilde{u}(x,t) \tag{1.2}$$

where we consider square matrix

$$\kappa_1 = \begin{pmatrix} -K_1 & 0 \\ 0 & -K_2 \end{pmatrix}; \qquad \kappa_2 = \begin{pmatrix} 0 & K_1 \\ K_2 & 0 \end{pmatrix}; \qquad \eta = \begin{pmatrix} \eta_1 & 0 \\ 0 & \eta_2 \end{pmatrix} \tag{1.3}$$

and the vector form of functions

$$u(x,t) = \begin{pmatrix} u_1(x,t) \\ u_2(x,t) \end{pmatrix};\ \hat{u}(x,t) = \begin{pmatrix} u_1(x,t-\tau_1) \\ u_2(x,t-\tau_2) \end{pmatrix};\ \tilde{u}(x,t) = \begin{pmatrix} u_1^3(x,t) \\ u_2^3(x,t) \end{pmatrix}. \tag{1.4}$$

We consider the system where the time of response may be 0 or different from 0. A real time of response causes that solutions do not affect each other in the same time.

*Dept. of Applied Mathematics, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 1, Slovakia (helena.samajova@fstroj.uniza.sk).

Some types of nonlinear parabolic equation with a constant delay are exactly solved in [5] by functional constraints method. This method brings exact solutions that are supposed to be in the generalized separable form

$$u(x,t) = \sum_{n=1}^{N} \varphi_n(x)\psi_n(t)$$

where $N \in \mathbb{N}$. Functions $\varphi_n(x)$ and $\psi_n(x)$ are established by additional functional constrains given by difference or functional equation. The results in the cited paper are extended to a class of nonlinear partial differential-difference equations with linear differential operators which are defined as separated differential operators with respect to the independent variables x, t and to some partial functional differential equations with time delay. The presented way of solution in [5] requires an assumption that initial functions to an initial problem of a delayed equation are obliged to satisfy the considered equation.

An approach established in this paper enables us to use different types of initial functions that need not indispensable to fulfill the system (1.1).

2. Main Properties of 2D Differential Transform Method (DTM). In the next it is proposed a procedure which allows us to combine DTM and method of steps to obtain semi-analytical solutions for given system of two equations (1.1). This method is used for example in [2, 6, 7] and the references given therein.

The two dimensional Differential transformation method (DTM) for a function $g(x,t)$ is defined by

$$G(m,n) = \frac{1}{m!n!}\left[\frac{\partial^{m+n} g(x,t)}{\partial x^m \partial t^n}\right]_{x=x_0,t=t_0}.$$

An inverse transform of $G(m,n)$ leads to

$$g(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} G(m,n)(x-x_0)^m(t-t_0)^n$$

and if $x = 0, t = 0$ then

$$g(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} G(m,n)x^m t^n.$$

The main properties of the DTM are given in the overview:

Let functions $G, G_i(n), \quad i = 1,2,3$ are differential transforms of the functions $g, g_i(n), \quad i = 1,2,3$, constants $r, s \in \mathbb{N}$, and $\alpha, \beta \in \mathbb{R}$

1. $g(x,t) = \alpha g_1(x,t) + \beta g_2(x,t) \qquad G(m,n) = \alpha G_1(m,n) + \beta G_2(m,n)$
2. $g(x,t) = x^r t^s \qquad G(m,n) = \delta(m-r, n-s) = \delta(m-r)\delta(n-s)$
3. $g(x,t) = \mathrm{e}^{\alpha x+\beta t} \qquad G(m,n) = \frac{\alpha^m \beta^n}{m!n!}$
4. $g(x,t) = \sin(\alpha x)t^s \qquad G(m,n) = \dfrac{\alpha^m}{m!}\sin(\dfrac{m\pi}{2})\delta(n-s)$

5. $g(x,t)=\cos(\alpha x)t^s \qquad G(m,n)=\dfrac{\alpha^m}{m!}\cos(\dfrac{m\pi}{2})\delta(n-s)$
6. $g(x,t)=g_1(x,t)g_2(x,t)g_3(x,t)$
   $G(m,n)=\sum_{i=0}^{m}\sum_{j=0}^{m-i}\sum_{k=0}^{n}\sum_{l=0}^{n-k}G_1(i,n-k-l)G_2(j,k)G_3(m-i-j,l)$
7. $g(x,t)=\dfrac{\partial g_1(x,t)}{\partial x}\dfrac{\partial g_2(x,t)}{\partial t}$
   $G(m,n)=\sum_{i=0}^{m}\sum_{j=0}^{n}(m-i+1)(n-j+1)G_1(m-i+1,j)G_2(i,n-j+1).$

   For delayed functions in the next we suppose $N\to\infty$
8. $g(x,t)=g_1(x,t+\tau) \qquad G(m,n)=\sum_{h=n}^{N}\binom{h}{n}\tau^{h-n}G_1(m,h)$

where $\delta(n)$ is the Kronecker delta symbol and $N\in\mathbb{N}$.
The main steps of the DTM, as a tool for solving different classes of nonlinear problems, are the following. First, we apply the differential transform to the presented problem, and then the functions $G(m,n)$ are given by the recurrence relations. In the second, the iterative solution of this relations and using the inverse differential transform, lead to the solution of the problem as polynomials of two independent variables.
Applying this rules for system (1.1) one obtains following recurrence relations for $\tau=0$

$$\begin{aligned}
Y_1(m,n+1)&=\frac{1}{n+1}\left[(m+2)Y_1(m+2,n)+K_1\left[Y_2(m,n)-Y_1(m,n)\right]\right.\\
&\left.+\eta_1\sum_{r_1=0}^{m}\sum_{r_2=0}^{m-r_1}\sum_{s_1=0}^{n}\sum_{s_2=0}^{n-s_1}Y_1(r_1,n-s_1-s_2)Y_1(r_2,s_1)Y_1(m-r_1-r_2,s_2)\right]\\
Y_2(m,n+1)&=\frac{1}{n+1}\left[(m+2)Y_2(m+2,n)+K_2\left[Y_1(m,n)-Y_2(m,n)\right]\right.\\
&\left.+\eta_2\sum_{r_1=0}^{m}\sum_{r_2=0}^{m-r_1}\sum_{s_1=0}^{n}\sum_{s_2=0}^{n-s_1}Y_2(r_1,n-s_1-s_2)Y_2(r_2,s_1)Y_2(m-r_1-r_2,s_2)\right].
\end{aligned}\tag{2.1}$$

2.1. Initial problem for systems of delayed functions. If we suppose delayed system with $\tau_i>0,\ i=1,2$ the system is considered with the known initial functions $\psi_i,\ i=1,2$

$$\psi_i(x,t)=\begin{cases}0, & t<-\tau;\\ \tilde{\psi}_i(x,t), & t\in\langle-\tau,0\rangle;\\ 0, & t>0.\end{cases}\tag{2.2}$$

We consider different types of functions on the intervals $(-\tau_i,0)$, as an initial functions for unknown solutions $y_i(x,t),\ i=1,2$ to the system (1.1).

A different types of initial functions produce appertaining initial conditions for the system (1.1) and some of them are presented in the table below.

- Calculations are valid on minimum length of the intervals $(0,\tau_i),\ i=1,2$
- $\tilde{\psi}_1(x,t),\ \tilde{\psi}_2(x,t)$ are considered as constant, polynomial, exponential, sin, cos functions
- Recurrent relations are used for evaluations of coefficients $Y_1(m,n),\ Y_2(m,n)$

TABLE 2.1
Types of initial functions.

Initial functions $\Psi_i(x,t)$	**Initial condition** $\Psi_i(x,0)$
$\Psi_i(x,t) = x^r t^s$	$\Psi_i(x,0) = 0$
$\Psi_i(x,t) = x^r \epsilon^{st} \quad r \neq 0$	$\Psi_i(x,0) = x^r$
$\Psi_i(x,t) = t^s \epsilon^{rx} \quad s \neq 0$	$\Psi_i(x,0) = 0$
$\Psi_i(x,t) = x^r \cos st$	$\Psi_i(x,0) = x^r$
$\Psi_i(x,t) = x^r \sin st$	$\Psi_i(x,0) = 0$

- An individual evaluation for the initial functions and initial conditions is required
- Functions $y_1(x, t-\tau_2)$, $y_2(x, t-\tau_1)$ are replaced by the initial functions $\tilde{\psi}_1(x,t)$, $\tilde{\psi}_2(x,t)$ on the intervals $(-\tau_2, 0)$, $(-\tau_1, 0)$ respectively
- The multi-step differential transform method (MsDTM) given in [1, 3] may be used to extend the domain for the obtained solutions.

In the Table 2.1 we give some examples of types of the initial functions and the initial conditions connected to the initial functions.

The DT method applied to the system (1.1) with $\tau_i > 0$ gives

$$
\begin{aligned}
Y_1(m,n+1) = \frac{1}{n+1}\Bigg[&(m+2)Y_1(m+2,n) + K_1\left(\sum_{h=n}^{N}\binom{h}{n}\tau_1^{h-n}\Psi_2(m,h) - Y_1(m,n)\right) \\
&+\eta_1 \sum_{r_1=0}^{m}\sum_{r_2=0}^{m-r_1}\sum_{s_1=0}^{n}\sum_{s_2=0}^{n-s_1} Y_1(r_1, n-s_1-s_2)Y_1(r_2,s_1)Y_1(m-r_1-r_2,s_2)\Bigg] \\
Y_2(m,n+1) = \frac{1}{n+1}\Bigg[&(m+2)Y_2(m+2,n) + K_2\left(\sum_{h=n}^{N}\binom{h}{n}\tau_2^{h-n}\Psi_1(m,h) - Y_2(m,n)\right) \\
&+\eta_2 \sum_{r_1=0}^{m}\sum_{r_2=0}^{m-r_1}\sum_{s_1=0}^{n}\sum_{s_2=0}^{n-s_1} Y_2(r_1, n-s_1-s_2)Y_2(r_2,s_1)Y_2(m-r_1-r_2,s_2)\Bigg].
\end{aligned}
\tag{2.3}
$$

3. Convergence of the 2D Differential Transform Method. In this section, the convergence of the 2-dimensional DTM when applied to a system of partial differential equations is studied. Moreover there is given the sufficient condition for a convergence of the vector function.

This condition of the convergence leads to an estimation of the maximum absolute error of the approximate solutions.

Let consider functions $f_1(x,t) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, $\quad f_2(x,t) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$

$$f_1(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_1(m,n)(x-x_0)^m(t-t_0)^n;$$

$$f_2(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_2(m,n)(x-x_0)^m(t-t_0)^n;$$

and

$$\vec{f}(x,t) = \begin{pmatrix} f_1(x,t) \\ f_2(x,t) \end{pmatrix}.$$

For the vector function we define the vector norm L^∞

$$\|\vec{f}\|_\infty = \max_i |f_i|, \qquad i \in \{1,2\}.$$

The theorem stated below is a special case of the Banach fixed point theorem [4]. In the next this theorem is adapted for 2D DTM.

THEOREM 3.1.

Let there exist two series for functions

$$f_1(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_1(m,n)(x-x_0)^m(t-t_0)^n$$

$$f_2(x,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_2(m,n)(x-x_0)^m(t-t_0)^n.$$

Then the vector series $\vec{f}(x,t)$ converges if there exists $0<\alpha<1$ such that

$$\|\vec{f}_{k+1}(x,t)\| \le \alpha \|\vec{f}_k(x,t)\|$$

for any $k \ge k_0$, for some $k_0 \in \mathbb{N}$.

The estimation of the error of the vector series is a part of the proof of Theorem 3.1.

Proof. We denote $(C(A), \|.\|)$ the Banach space of all continuous vector functions on a domain A with the norm $\|f(x,t)\| = \max_{(x,t)\in A}\|f(x,t)\|$ where $A = [x_0-\varepsilon, x_0+\varepsilon]\times[t_0-\tau, t_0+\tau]$.

Denote individual terms $\varphi^1_{(m,n)}(x,t)$, $\varphi^2_{(m,n)}(x,t)$, $\Phi_{(m,n)}(x,t)$ as

$$\varphi^i_{(m,n)}(x,t) = F_i(m,n)(x-x_0)^m(t-t_0)^n \qquad i=1,2\ ,$$

$$\Phi_{(m,n)}(x,t) = \begin{pmatrix} F_1(m,n)(x-x_0)^m(t-t_0)^n \\ F_2(m,n)(x-x_0)^m(t-t_0)^n \end{pmatrix} = \begin{pmatrix} \varphi^1_{(m,n)}(x,t) \\ \varphi^2_{(m,n)}(x,t) \end{pmatrix}.$$

We define the sequence of vector partial sums $\{S_n\}_{n=0}^\infty$ as follows

$$S_n = \Phi_{(0,0)}(x,t)+\Phi_{(1,0)}(x,t)+\Phi_{(0,1)}(x,t)+\Phi_{(2,0)}(x,t)+\Phi_{(1,1)}(x,t)+\Phi_{(0,2)}(x,t)+\ldots+$$

$$\Phi_{(n,0)}(x,t)+\Phi_{(n-1,1)}(x,t)+\ldots+\Phi_{(1,n-1)}(x,t)+\Phi_{(0,n)}(x,t) =$$

$$\sum_{j=0}^{n}\sum_{i=0}^{j}\Phi_{(i,j-i)}(x,t).$$

In the next we will show that $\{S_n\}_{n=0}^\infty$ is a Cauchy sequence in the Banach space. For this purpose

$$\|S_{n+1}-S_n\| = \left\|\sum_{i=0}^{n+1}\Phi_{(i,n+1-i)}(x,t)\right\| \le \alpha\left\|\sum_{i=0}^{n}\Phi_{(i,n-i)}(x,t)\right\| \le \ldots \le$$

$$\leq \alpha^{n-k_0+1} \left\| \sum_{i=0}^{k_0} \Phi_{i,k_0-i}(x,t) \right\| =$$

$$= \alpha^{n-k_0+1} \max_{(x,t)\in A} \left\{ \sum_{i=0}^{k_0} \left| \varphi^1_{(i,k_0-i)}(x,t) \right|, \sum_{i=0}^{k_0} \left| \varphi^2_{(i,k_0-i)}(x,t) \right| \right\}.$$

For any $i, j \in \mathbb{N},\ i > j > k_0$ we have

$$\|S_i - S_j\| = \left\| \sum_{l=j}^{i-1} (S_{l+1} - S_l) \right\| \leq \sum_{l=j}^{i-1} \|(S_{l+1} - S_l)\|$$

$$\leq \sum_{l=j}^{i-1} \alpha^{l-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \|\Phi_{s,k_0-s}(x,t)\|$$

$$= \frac{1-\alpha^{i-j}}{1-\alpha} \alpha^{j-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \|\Phi_{s,k_0-s}(x,t)\|$$

and whereas $0 < \alpha < 1$, we obtain

$$\lim_{i,j\to\infty} \|(S_i - S_j\| = 0.$$

Therefore, $\{S_n\}_{n=0}^{\infty}$ is a Cauchy sequence in the Banach space $(C(A), \|.\|)$ and the vector series

$$\begin{pmatrix} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \varphi^1_{(m,n)}(x,t) \\ \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \varphi^2_{(m,n)}(x,t) \end{pmatrix}$$

converges. The proof is complete. □

Under the condition that there exists $\alpha \in (0,1)$ such that

$$\sum_{s=0}^{k+1} \left\| \Phi_{(s,k+1-s)}(x,t) \right\| \leq \alpha \sum_{s=0}^{k} \left\| \Phi_{(s,k-s)}(x,t) \right\|$$

for any $k \geq k_0$ where $k_0 \in \mathbb{N}$, power series solution converges to the exact solution.

We define constants α_k for any $k \geq k_0$

$$\alpha_{k+1} = \begin{cases} \dfrac{\sum_{s=0}^{k+1} \left\| \Phi_{(s,k+1-s)}(x,t) \right\|}{\sum_{s=0}^{k} \left\| \Phi_{(s,k-s)}(x,t) \right\|} & \text{for} \quad \sum_{s=0}^{k} \left\| \Phi_{(s,k-s)}(x,t) \right\| \neq 0; \\ 0 & \text{for} \quad \sum_{s=0}^{k} \left\| \Phi_{(s,k-s)}(x,t) \right\| = 0. \end{cases}$$

If $\forall k > k_0 :\ 0 \leq \alpha_k < 1$, then an approximate solution in the form of finite series converges to the exact solution $\vec{u}(x,t)$.

THEOREM 3.2. *Let the approximate solution be in the form*

$$\vec{f}(x,t) = \begin{pmatrix} \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_1(m,n)(x-x_0)^m(t-t_0)^n \\ \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} F_2(m,n)(x-x_0)^m(t-t_0)^n \end{pmatrix}$$

and converges to the solution

$$\vec{u}(x,t) = \begin{pmatrix} u_1(x,y) \\ u_2(x,y) \end{pmatrix}.$$

If the finite series

$$\begin{pmatrix} \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} F_1(m,n)(x-x_0)^m(t-t_0)^n \\ \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} F_2(m,n)(x-x_0)^m(t-t_0)^n \end{pmatrix}$$

is considered as an approximation to the solution, then the estimation of the absolute error is given as

$$\left\| \vec{u}(x,t) - \begin{pmatrix} \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} \varphi^1_{(m,n)}(x,t) \\ \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} \varphi^2_{(m,n)}(x,t) \end{pmatrix} \right\| \leq$$

$$\leq \frac{1}{1-\alpha}\alpha^{j-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \|\Phi_{(s,k_0-s)}(x,t)\| \tag{3.1}$$

where $j = \min\{\mu,\nu\}$, $\mu,\nu \in \mathbb{N}$.

Proof. From the Theorem 3.1 we obtained

$$\|S_i - S_j\| \leq \frac{1-\alpha^{i-j}}{1-\alpha}\alpha^{j-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \left\|\Phi_{(s,k_0-s)}(x,t)\right\|$$

Since the term $(1-\alpha^{i-j}) < 1$ under the condition that there exists an $\alpha \in (0,1)$ and $k_0 \leq j \leq i$, the inequality above can be simplify to

$$\|S_i - S_j\| \leq \frac{1}{1-\alpha}\alpha^{j-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \left\|\Phi_{(s,k_0-s)}(x,t)\right\|.$$

If we consider that $i \to \infty$ then $S_i \to \vec{u}(x,t)$ - two dimensional power series vector solution converges to the vector solution and the estimation of an absolute error is determined by (3.1). □

In accordance with Theorem 3.2 the estimation of the absolute error is given by the inequality below

$$\left\| \vec{u}(x,t) - \begin{pmatrix} \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} F_1(m,n)(x-x_0)^m(t-t_0)^n \\ \sum_{m=0}^{\mu}\sum_{n=0}^{\nu} F_2(m,n)(x-x_0)^m(t-t_0)^n \end{pmatrix} \right\| \leq$$

$$\frac{1}{1-\beta}\beta^{j-k_0+1} \max_{(x,t)\in A} \sum_{s=0}^{k_0} \left\|\Phi_{(s,k_0-s)}(x,t)\right\|,$$

where $\beta = \max\{\alpha_k,\ k = k_0+1, k_0+2, \ldots, j+1\}$.

As an example of non-delayed and delayed coupling there are given pairs of figures of solutions $y_1(x,t)$ and $y_2(x,t)$. For different types of initial functions the Figures (3.1) and (3.3) represent non-delayed coupling, the Figures (3.2) and (3.4) delayed coupling. For calculation the system Mathematica was used.

For parameters $K_1 = 0.5$, $K_2 = 1.1$, $\eta_1 = 0.5$, $\eta_2 = 0.3$, $N = 6$ and initial functions $\tilde{\psi}_1 = 0$ and $\tilde{\psi}_2 = \cos x$ the solutions to the system (1.1) for a non-delayed case are on Fig. 3.1

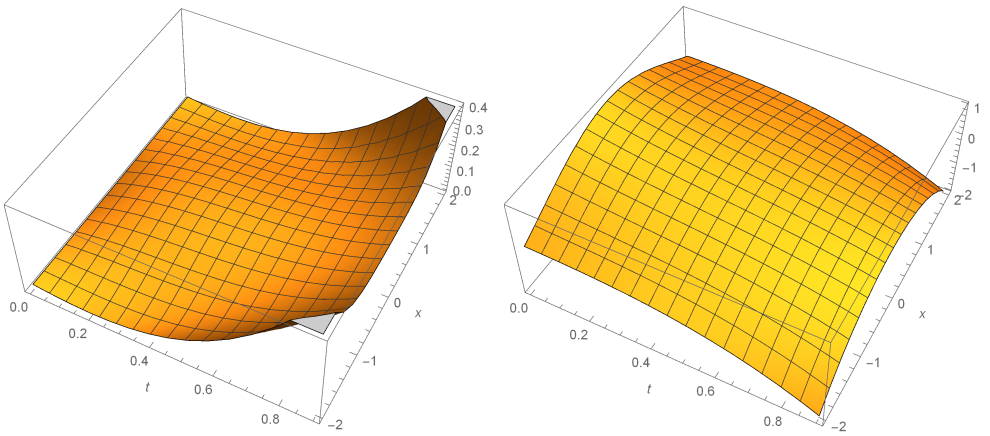


FIG. 3.1. *Solutions from left:* $y_1(x,t)$, $y_2(x,t)$, $\tau = 0$.

where

$$\begin{aligned} y_1(x,t) =& 0.5t - 0.9t^2 + 0.6967t^3 - 0.25tx^2 + 0.2833t^2x^2 + 0.0208tx^4 \\ y_2(x,t) =& 1 - 2.1t + 2.1467t^2 - 1.5438t^3 - 0.5x^2 + 0.7167tx^2 - 0.64t^2x^2 + 0.0417x^4 \\ & -0.0542tx^4 - 0.0014x^6. \end{aligned}$$

For a delayed case with $\tau_i = 0.8$ the solutions are in Fig. 3.2

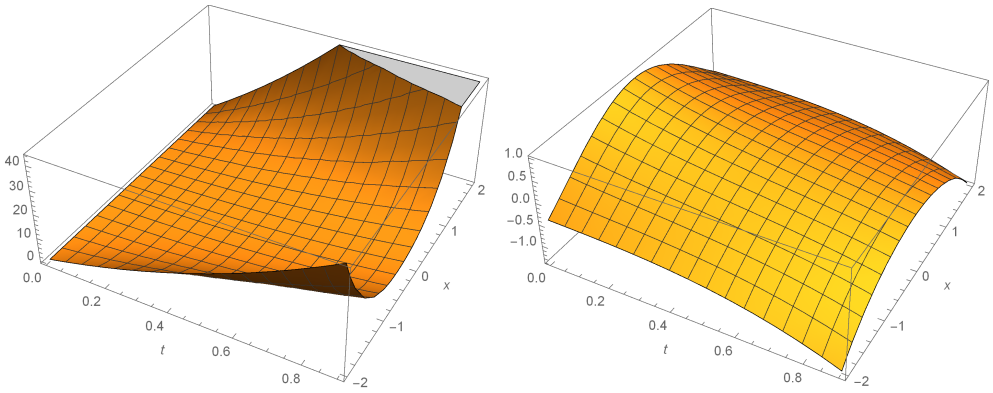


FIG. 3.2. *Solutions from left:* $y_1(x,t)$, $y_2(x,t)$, $\tau_1 = \tau_2 = 0.8$.

where

$$\begin{aligned} y_1(x,t) =& 2.5t + 3.75t^2 + 4.7917t^3 + 2.5tx + 5t^2x + 2.5tx^2 + 6.25t^2x^2 \\ & +2.5tx^3 + 2.5tx^4 \\ y_2(x,t) =& 1 - 2.1t + 1.8717t^2 - 1.0213t^3 - 0.5x^2 + 0.7167tx^2 - 0.5025t^2x^2 \\ & +0.0417x^4 - 0.0542tx^4 - 0.0014x^6. \end{aligned}$$

For parameters $K = 0.2,\ \eta_1 = 1.5,\ \eta_2 = 0.3$ and initial functions $\tilde{\psi}_1 = \cos x$ and $\tilde{\psi}_2 = \sin x$, a non-delayed case is on Fig. 3.3

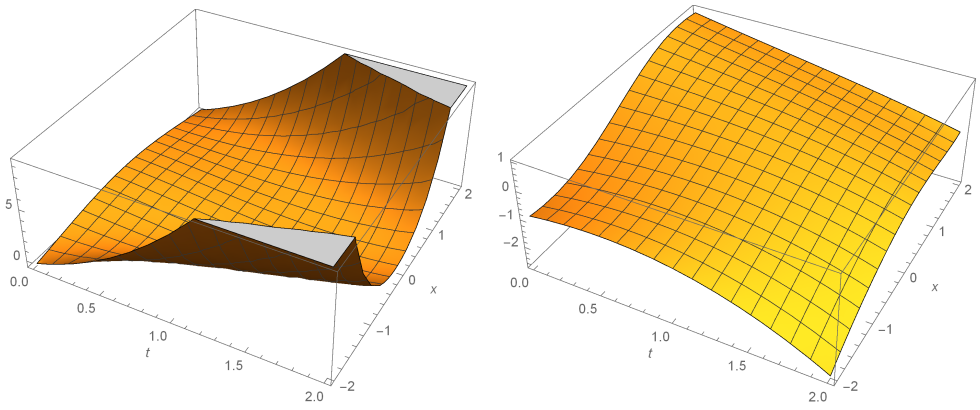


FIG. 3.3. *Solutions from left:* $y_1(x,t),\ y_2(x,t),\quad \tau = 0.$

where

$$\begin{aligned} y_1(x,t) =& 1 - 0.7t - 0.9183t^2 + 0.5294t^3 + 0.2tx + 0.01t^2x - 0.5x^2 \\ & -0.4833tx^2 + 1.0425t^2x^2 - 0.0333tx^3 + 0.0417x^4 + 0.4208tx^4 \\ & -0.0014x^6 \\ y_2(x,t) =& 0.2t - 0.19t^2 - 0.063t^3 + x - 0.7tx + 0.2025t^2x - 0.1tx^2 \\ & -0.0217t^2x^2 - 0.0083x^3 + 0.075tx^3 + 0.0083tx^4 + 0.0083x^5. \end{aligned}$$

Solutions for a delayed case with $\tau_i = 0.8$ are on Fig.3.4

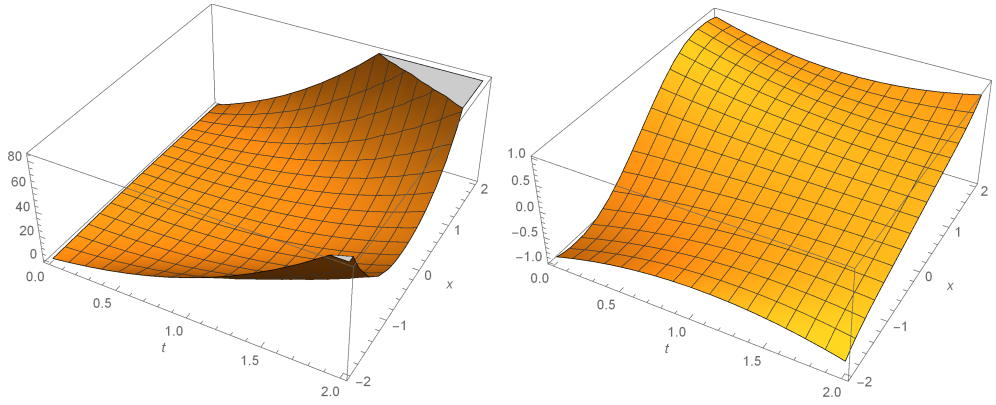


FIG. 3.4. *Solutions from left:* $y_1(x,t),\ y_2(x,t),\quad \tau_1 = \tau_2 = 0.8.$

where

$$\begin{aligned} y_1(x,t) =& 1 + 0.3t + 1.2117t^2 + 3.2051t^3 + tx + 2.65t^2x - 0.5x^2 \\ & +0.5167tx^2 + 3.4525t^2x^2 + tx^3 + 0.0417x^4 + 1.4208tx^4 \\ & -0.0014x^6 \\ y_2(x,t) =& x - 0.7tx + 0.1825t^2x - 0.1667x^3 + 0.075tx^3 + 0.0083x^5. \end{aligned}$$

REFERENCES

[1] Benhammouda, B., Vazquez-Leal, H. *A new multi-step technique with differential transform method for analytical solution of some nonlinear variable delay differential equations* SpringerPlus, (2016), 5, 1723. DOI 10.1186/s40064-016-3386-8

[2] Khan Y., Svoboda Z., Šmarda Z., *Solving certain classes of Lane-Emden type equations using the differential transformation method*, Advances in difference equations, 174, (2012).

[3] Odibat, Z. M., Bertelle C., Aziz-Alaouic M.A., Duchampd H. E. G. *A multi-step differential transform method and application to non-chaotic or chaotic systems* Computers and Mathematics with Applications, 59, (2010), pp 1462-1472.

[4] Odibat, Z. M., Kumar S., Shawagfeh N., Alsaedi A., Hayat T. *A study on the convergence conditions of generalized differential transform method* Mathematical Methods in the Applied Sciences, 40, (2017), pp 40-48.

[5] Polyanin A. D., Zhurov A. I. *Functional constraints method for constructing exact solutions to delay reactiondiffusion equations and more complex nonlinear equations* Commun. Nonlinear Sci. Numer. Simulat., 19, (2014), pp 417-430.

[6] Rebenda J., Šmarda Z., *A differential transformation approach for solving functional differential equations with multiple delays*, Commun. Nonlinear Sci. Numer. Simulat., 48, (2017), pp 246-257.

[7] Rebenda J., Šmarda Z., Khan Y., *A New Semi-analytical Approach for Numerical Solving of Cauchy Problem for Differential Equations with Delay*, FILOMAT, 31, (2017), pp 4725-4733.

Proceedings of EQUADIFF 2017
pp. 173–180

VISCO-ELASTO-PLASTIC MODELING *

JANA KOPFOVÁ, MÁRIA MINÁROVÁ AND JOZEF SUMEC

Abstract. In this paper we deal with the mathematical modelling of rheological models with applications in various engineering disciplines and industry. We study the mechanical response of visco-elasto-plastic materials. We describe the basic rheological elements and focus our attention to the specific model of concrete, for which we derive governing equations and discuss its solution. We provide an application of rheological model involving rigid-plastic element as well - mechanical and mathematical model of failure of one dimensional construction member, straight beam. Herein, the physical model is considered with a homogeneous isotropic material of the beam, quasi static regime is supposed.

Key words. rheological elements, constitutive equation, large deformations, hysteresis, dissipated energy

AMS subject classifications. 35J86, 00A79

Introduction. In mechanics, the constitutive relation between the stress σ and the strain ε, is essential. Rheology deals with problems concerning deformation processes of materials exhibiting different kinds of material response, e.g. elastic, viscous and plastic behavior. Time dependent mechanical behavior is governed by constitutive equations describing the relations between stress and strain variables and their time derivatives. There exist materials that behave in a different way during loading and unloading, some are and some are not able to recover. This phenomenon is called *hysteresis*. Herein, and this it is very well known fact [9, 7, 2], the potential energy plays important role. There are elementary matters, called also members or elements involved in each model. Very elegant presentation of basic rheological models can an interested reader find in the monograph [2]. There, the main focus is on elasto-plastic materials, hysteresis phenomena being the main area of interest. For the description of visco-plastic materials we refer to the book [1]. Corresponding models in electricity are studied in detail in [3].

There are specific tests executed on material models by prescribed stress or strain load action. The creep or relaxation of stress is recorded. Creep is a deformation change in time under constant stress load being maintained, relaxation is a stress change in time when a constant deformation is maintained. Boltzmann theory using hereditary integrals is exerted, as well, [5].

In the paper the basic phenomena of rheology models are introduced together with constitutive relation derivation techniques. Involving a viscous member in the model yields the presence of derivatives in physical equations, plastic element brings a variational inequality. Finally, the three of them - elastic, plastic and viscous members are involved in a very simple model of concrete. The constitutive relation is derived.

1. Fundamental elements, compositions, relations. In agreement with Definition 1.1 in [2] we call a rheological element a system consisting of a constitutive relation between stress σ and strain ε and a potential energy $U \geq 0$. Along this paper

*This work was supported by Grant No.: VEGA 1/0456/17, by GAČR Grant 15-12227S, and by the institutional support for the development of research organizations IČ 47813059.

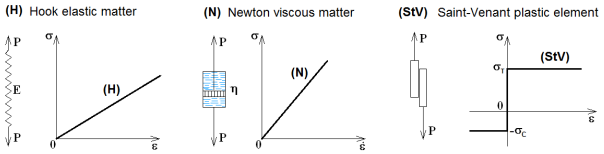


FIG. 1.1. *Stress - strain dependence in an a) elastic, b) viscous, c) rigid-plastic element.*

we will deal with uniaxial thermodynamically consistent rheological models, which means that the quantity called dissipation rate

$$\dot{q} = \langle \dot{\varepsilon}, \sigma \rangle - \dot{U} \tag{1.1}$$

will be supposed to be non-negative in sense of distributions for all ε, σ, [2].

1.1. Fundamental elements of a visco-elasto-plastic model, physical properties. [4, 2]

There are Newton's viscous (N), Hook's elastic (H) and Saint-Venant (StV) rigid-plastic elements involved in a visco-elasto-plastic model.

Elastic element (H) is represented by an ideally elastic spring, where the stress - strain relation is linear:

$$\sigma = \mathbf{A}\varepsilon, \tag{1.2}$$

with $\mathbf{A}$ an elastic modulus matrix, in the case of homogeneous isotropic material it is replaced by a real number E - Young elastic modulus. In more dimensions it includes both volumetric and deviatoric change. (H) is completely reversible, i.e. all inner potential energy U gathered in the loading process is conserved and no energy is dissipated. After loading stops all energy is used to reverse the previous position. Potential energy is given by $U = \frac{1}{2}E\varepsilon^2$ and it can be easily checked that the thermodynamical consistency of the model is fulfilled.

Viscous element (N) is symbolized graphically by a piston. Here we have is a linear relation between stress and strain rate, which can be expressed both in deviatoric and volumetric components $\sigma_{dev} = \eta\dot{\varepsilon}_{dev}, \quad \sigma_{vol} = \zeta\dot{\varepsilon}_{vol}$, with η and ζ being deviatoric and volumetric proportional coefficient respectively. For incompressible liquids only deviatoric component comes into play and the stress-strain relation can be expressed simply as

$$\sigma = \eta\,\dot{\varepsilon} \tag{1.3}$$

No potential energy is stored, i.e. $U = 0$, the deformation process is irreversible. Viscous elements act as dashpots.

Rigid plastic element (StV). Its graphical symbol is depicted as two touching plates with certain friction between. When a (StV) is exposed to a load, it remains rigid as long as the instantaneous stress does not reach the threshold. If so, material becomes plastic immediately.

Let Z be the space of all admissible stress values with all thresholds situated in its boundary ∂Z. The plasticity is governed by the following physical principles:

- $\sigma \in int(Z)$ ensures the rigidity persisting of the body
- $\sigma \in \partial Z$ (the plastic behavior is triggered)
- $\langle \dot{\varepsilon}, \sigma - \tilde{\sigma} \rangle \geq 0, \quad \forall \tilde{\sigma} \in Z$

The last principle, the variational inequality, is called the maximal dissipation rate principle with regard to admissible stress values. It states that while the threshold is not reached, the deformation does not change, i.e. $\forall \sigma \in int(Z) \implies \dot{\varepsilon} = 0$, [9, 2].

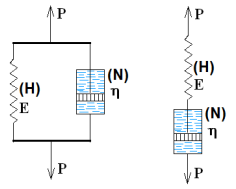


FIG. 1.2. *Parallel and serial combination of fundamental elements.*

In the uniaxial case $Z = \langle -\sigma_C, \sigma_T \rangle$ and $\partial Z = \{-\sigma_C, \sigma_T\}$, where we assume that σ_C, σ_T are two positive constants, so $0 \in Z$, which corresponds to the natural hypotheses that no deformation occurs for $\sigma = 0$. This condition is essential for the thermodynamic consistency of the model.

In Fig 1.1c) an uniaxial representation of rigid-plastic body is performed. The polygonal line graph is called three branch diagram. Herein, as Z is an interval, its boundary are the endpoints called compressive threshold $-\sigma_C$ and tension threshold σ_T. In general $\sigma_C \neq \sigma_T$. When a threshold is reached, plasticity proceeds and takes place until the magnitude decreases again and the rigidity comes back, permanent (plastic) deformation persists. No potential energy is stored, i.e. $U = 0$ and no recovery occurs. It has been observed that during plastic deformation the volume change is negligible. [6]

In Fig. 1.1 the graphical symbols representing particular elementary matters and graphical interpretation of the stress - strain relations are shown. Here P denotes the tension force.

1.2. Configuration, geometry and corresponding relations. There are two possible ways of connecting any couple of fundamental elements - either serially or in parallel, by using two auxiliary rigid slabs for this sake, as depicted in Fig. 1.2 a), the two slabs are represented by the upper and lower thick lines connecting (H) and (N).

- *Serial connection* of elementary members
  Under a load P, the resulting deformation of the system of serially connected elementary matters is the sum of the deformations of particular members, stress is distributed among the members equally:

$$\varepsilon = \varepsilon_H + \varepsilon_N, \quad \sigma = \sigma_H = \sigma_N. \tag{1.4}$$

  Sub-indexes H, N, then SvV indicate an incidence with Hook elastic, Newton viscous and Saint-Venant rigid-plastic matters.
- *Parallel connection* of elementary members
  As two linking slabs are shifting vertically up or down without any rotation, the deformation is the same, while the stress of the entire model is the sum of stresses of particular members:

$$\varepsilon = \varepsilon_H = \varepsilon_N, \quad \sigma = \sigma_H + \sigma_N. \tag{1.5}$$

Having the two or more elementary matters at hand and utilizing both parallel and serial connection, and considering the fundamental elements as simplest rheological models, we can proceed in composing visco-elasto-plastic models recursively. By connecting two simpler models serially or in parallel we compose the new, more complex one. When we couple the geometry equations yielded by configuration with the

fundamental element constitutive relations into account, we can derive the resulting constitutive equation of the entire model.
For the sake of clear notation it is worth utilizing an abbreviations of such models. Having, beside (N), (H) and (Stv) marks standing for the particular fundamental elements, the vertical line standing for parallel and the horizontal line standing for serial, we can assign a structural formula to each model. Accordingly, the structural formulas of the two-element models in Fig. 1.2 are: $(H)|(N)$ for the left one and $(H)-(N)$ for the right one respectively.

2. Creep and relaxation tests. Creep and relaxation tests are typical for testing materials with the aim of their mechanical response prediction and materials' mechanical behavior comparison. The special load is imposed to the material and the response is recorded and monitored. Roughly speaking, creep-deformation change in time under the constant stress load is maintained or relaxation-stress change in time when a constant deformation is maintained.

Creep test is executed by inflicting an instantaneous stress keeping it constant in a given time period. The immediate change is obviously followed by subsidiary one - the creep. Resulting deformation response is recorded.

Relaxation test is executed by carrying out an instantaneous strain, keeping it constant during a given time period. The immediate change of stress is obviously followed by subsidiary one - the relaxation. Changes of stress are recorded.

3. Elasto-plasticity resulting in Hysteresis. Let us examine what happens when we combine elastic and plastic element. First for the combination in paralel we get

$$\varepsilon = \varepsilon_H = \varepsilon_{StV}, \tag{3.1}$$
$$\sigma = \sigma_H + \sigma_{StV}, \tag{3.2}$$
$$\sigma_H = E\,\varepsilon_H. \tag{3.3}$$

Let us employ the Saint Venant variational inequality for the rigid-plastic matter

$$\dot{\varepsilon}_{StV}\big(\sigma_{StV} - \tilde{\sigma}\big) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle \tag{3.4}$$

and we have

$$\dot{\varepsilon}\,(\sigma - E\varepsilon - \tilde{\sigma}) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle. \tag{3.5}$$

For the potential energy we get $U = \frac{1}{2} E \varepsilon_H^2$ (the only contribution comes from the elastic element) and the thermodynamical consistency of the model follows actually from the variational inequality (3.5) (with $\tilde{\sigma} = 0$).

For the combination in series we get similarly

$$\varepsilon = \varepsilon_H + \varepsilon_{StV}, \tag{3.6}$$
$$\sigma = \sigma_H = \sigma_{StV}, \tag{3.7}$$
$$\sigma_H = E\,\varepsilon_H. \tag{3.8}$$

Let us now employ again the Saint Venant variational inequality for the rigid-plastic matter

$$\dot{\varepsilon}_{StV}\big(\sigma_{StV} - \tilde{\sigma}\big) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle \tag{3.9}$$

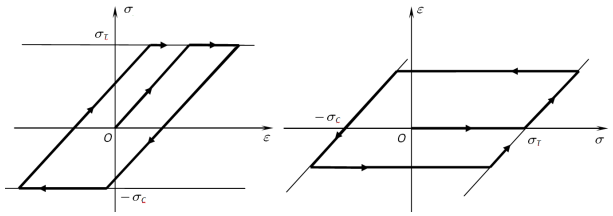


FIG. 3.1. *a) Stop and b) Play operators*

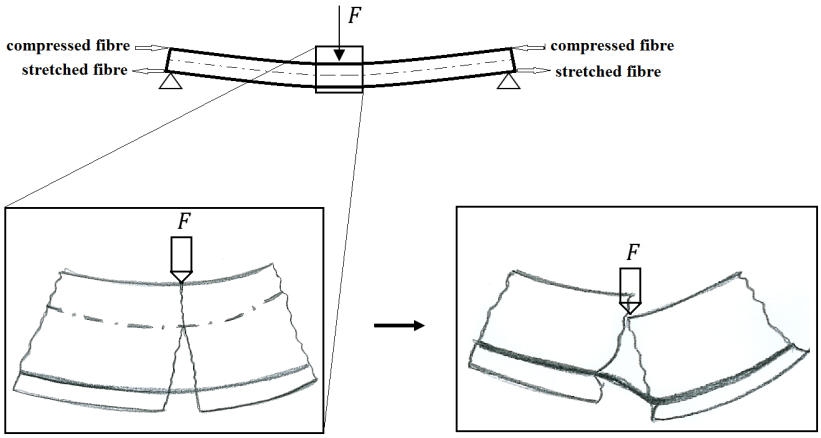


FIG. 4.1. *Concrete beam under heavy transversal load, compressed and stretched fibres, crack, rupture*

and we have as a consequence

$$\left(\dot{\varepsilon} - \frac{1}{E}\dot{\sigma}\right)(\sigma - \tilde{\sigma}) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle. \tag{3.10}$$

The potential energy is again given by $U = \frac{1}{2}E\varepsilon_H^2$ and the thermodynamical consistency follows from (3.10) (again taking $\tilde{\sigma} = 0$).

The variational inequalities obtained in both cases (series or paralel) are both of the same type and it was shown e.g. in [2], Theorem 1.9, that there exists a unique solution of these variational inequalities, which is given by a hysteresis operator - the so-called play or stop operator respectively. Hysteresis operators exhibit memory effects (the current state depends on the previous history of the system) and they are rate independent (this property allows us to draw diagrams as on Fig.3.1. For more details in this direction we refer to [2] and the references therein.

4. Rheological model of concrete. There exists an exceptional group within the rheology of composite materials on a silicate basis, group of concretes and reinforced concretes. Due to mechanical, chemo-mechanical or thermo-mechanical load acting in concrete or steel-concrete constructions, some immediate, short-term and long-term deformations evolve, the change lasting up to several years. When the concrete mixture is poured into a form, it initiates the solidification together with a chemical processes resulting in a volume contraction regardless the load imposed. And, on the other hand, an imposed load activates a creep, hysteresis response in-

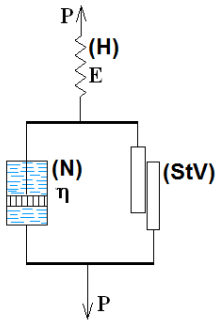


FIG. 4.2. *Parallel and serial connection of fundamental elements.*

volved. Creep is the essential phenomenon that has to be investigated carefully as the mechanical behavior of the designed structures and constructions can be predicted.

4.1. Concrete behavior under frequent heavy load. When the imposed load is of magnitude within the range obvious in concrete constructions, the resulting deformation as a consequence of creep will be several times greater than the initial (immediate) one. In this context, the notion "aging of concrete" is often used. [8] Nevertheless, the mechanical response of the concrete construction is proportional to the subjected load, accordingly the habitual operating load response of a construction is derived by using the superposition principle. However, once unloaded, a permanent deformation remains, [9].

Another essential fact concerning concrete has to be mentioned: Compressive strength of concrete is much higher then tensile strength. Hence, concrete mechanical response to tensile and to the compressive load of the same magnitude differs significantly. That is why the reinforcement with material strong in tension is placed where a tensile stress is supposed. In Fig. 4.1 the reinforcement is placed at the bottom of the beam. Namely, it is supposed to be doubly supported at the ends and loaded transversally by a pressure.

4.2. Simplified model of concrete - graphical representation structural form, geometrical relations. As proposed by [7], the simplest model of concrete can be set up by connecting viscous and rigid-plastic element in parallel and connect an elastic member with this couple in series. The structural form is $(H)-[(N)|(StV)]$. The physical relations are considered as given in Section 1.1, E being Young elastic modulus of (H), η the viscous coefficient of (N), σ_T and σ_C the stress tensile and compressive thresholds of the stress in (StV).

In the following considerations, the subindexes of stress and strain variables will be used to indicate the incidence with particular elementary members; e.g. ε_H will stand for partial deformation of Hook elastic matter, etc. Geometric equations of the $(H)-[(N)|(StV)]$ model are:

$$\varepsilon = \varepsilon_H + \varepsilon_N \tag{4.1}$$

$$\varepsilon_N = \varepsilon_{StV} \tag{4.2}$$

$$\sigma = \sigma_H = \sigma_N + \sigma_{StV}, \tag{4.3}$$

where $\sigma_{StV} \in \langle -\sigma_C, \sigma_T \rangle$.

Energy audit yields that the energy of elastic member is the only nonzero part of the potential energy of the whole system

$$U = \frac{1}{2} E \varepsilon_H^2$$

and the thermodynamical consistency of the model

$$\dot{\varepsilon}\sigma - \frac{1}{2} E \dot{\varepsilon}_H^2 \geq 0$$

follows from the variational inequality (4.9) bellow, taking $\tilde{\sigma} = 0$.

In the following we will deduce the constitutive relation of the model. We are looking for the $\sigma \sim \varepsilon$ equation, describing the dependence between global stress and global strain employing merely physical parameters of particular elementary members. It means we want to exclude the sub-indexed stress and strain variables from the dependence forms. Reminding elementary physical relations embedded in Section 1.1, we can proceed in the following way:

$$\sigma_H = E\, \varepsilon_H, \quad \sigma_N = \eta\, \dot{\varepsilon_N} \implies \varepsilon_N = \varepsilon - \varepsilon_H = \varepsilon - \frac{1}{E}\sigma_H, \tag{4.4}$$

$$\sigma = E\varepsilon_H = \eta\, \dot{\varepsilon}_N + \sigma_{StV}, \tag{4.5}$$

$$\sigma_{StV} = \sigma - \eta\, \dot{\varepsilon_N} = \sigma - \ \eta\dot{\varepsilon} + \frac{\eta}{E}\, \dot{\sigma}. \tag{4.6}$$

Let us employ the variational inequality of the Saint-Venant rigid-plastic matter

$$\dot{\varepsilon}_{StV}\, (\sigma_{StV} - \tilde{\sigma}) \geq 0, \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle. \tag{4.7}$$

Let us recall that

$$\varepsilon_{StV} = \varepsilon_N = \varepsilon - \varepsilon_H = \varepsilon - \frac{\sigma}{E}. \tag{4.8}$$

As a result we get the following variational inequality

$$\left(\dot{\varepsilon} - \frac{\dot{\sigma}}{E}\right)\left(\sigma - \eta\dot{\varepsilon} + \frac{\eta}{E}\dot{\sigma} - \tilde{\sigma}\right) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle \tag{4.9}$$

When we denote $\dot{v} = \dot{\varepsilon} - \frac{\dot{\sigma}}{E}$, we can rewrite (4.9) in the form

$$\dot{v}\big(\sigma - \eta\dot{v} - \tilde{\sigma}\big) \geq 0 \quad \forall \tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle \tag{4.10}$$

Let us have a closer look at the variational inequality (4.10). First of all, if $\sigma - \eta\dot{v}$ is in the open interval $(-\sigma_C, \sigma_T)$, the second bracket can take positive or negative values as $\tilde{\sigma}$ changes. Therefore we must have $\dot{v} = 0$.

Let us now aim our attention to the compressive marginal value $\sigma - \eta\dot{v} = -\sigma_C$. In such a case, it is apparent for all $\tilde{\sigma} \in \langle -\sigma_C, \sigma_T \rangle$ that the expression in brackets on the left hand side of (4.10) is non-positive. This implies that in order to hold the inequality (4.10) it must be hold $\dot{v} \leq 0$, so:

$$\dot{\varepsilon} - \frac{\dot{\sigma}}{E} \leq 0 \tag{4.11}$$

And accordingly, if $\sigma - \eta\dot{v} = -\sigma_T$ then it must be $\dot{v} \geq 0$ or

$$\dot{\varepsilon} - \frac{\dot{\sigma}}{E} \geq 0. \tag{4.12}$$

This can be described by a single relation in terms of v

$$\dot{v} = \frac{1}{\eta}(\sigma - \mathcal{P}(\sigma)), \tag{4.13}$$

where $\mathcal{P}$ denotes the projection on the interval $\langle -\sigma_C, \sigma_T \rangle$ (in the sense of convex analyses).

Alternatively in terms of σ and ε we have

$$\dot{\varepsilon} = \frac{\dot{\sigma}}{E} + \frac{1}{\eta}(\sigma - \mathcal{P}(\sigma)). \tag{4.14}$$

This is the constitutive relation we were looking for. It involves both stress on strain and strain on stress dependence. It means that with such a constitutive relation of the model at hand we can operate further and investigate and predict material behavior depending on the kind and magnitude of the load. Either we impose stress load, solving the corresponding linear non-homogenous differential equation in sense of deformation, or vice versa, i.e. we impose a strain and compute stress response. The initial condition have to be posed as well. The first equation is of course much simpler to solve, we can get the solution by simple integration. Alternatively the obtained differential equations can be solved easily e.g. numerically.

Creep and relaxation test are examples of material behavior investigation.

5. Conclusion. Nowadays, a lot of new material is developed and used in industry. Undoubtedly, the investigation prior to their usage is inevitable. Avoiding or predicting the failure due to heavy or repeating load is essential. For this sake the models with time dependent material behavior are utilized, each material matched with its appropriate models. Then by using mathematical tools various theoretical tests can be executed and response vs. load can be traced. Constitutive equations are essential, visco-elasto-plastic models being of great importance and interest within them.

REFERENCES

[1] I. R. Ionescu, M. Sofonea, *Functional and Numerical Methods in Viscoplasticity*, Oxford University Press, (1993).

[2] P. Krejčí, *Hysteresis, Convexity and Dissipation in Hyperbolic Equations*, Gakuto Intern. Ser. Math. Sci. Appl., Vol. **8**, Gakkotōsho, Tokyo (1996).

[3] K. Kuhnen , *Inverse Steuerung piezoelektrischer Aktoren mit Hysterese-, Kriech- und Superpositionsoperatoren (Berichte aus der Steuerungs- und Regelungstechnik)* , Gebundene Ausgabe, (2001).

[4] Minárová, M. Sumec.J., *Constitutive equations for selected rheological models in linear viscoelasticity*, Advances and Trends in Engineering Sciences and Technologies II :Proc. of ESAT conference, CRC Press, (2016), pp.207-212.

[5] Minárová, M. , *Mathematical Modeling of Phenomenological Material Properties - Differential Operator Forms of Constitutive equations*, in Slovak Journal of Civil Engineering Vol.22/4 (2014).

[6] Rabotnov, R.J., *Creep of Structural Elements* Moscow: Nauka, (1966).

[7] Sobotka, Z., *Reologie hmot a konstrukc*, Praha, Academia publishing house (1981).

[8] Sumec, J., *Mechanics - Mathematical Modeling of Materials Whose Physical Properties are Time Dependent*, Internal Research Report, III-3-4/9.4 USTARCH-SAV Bratislava (1983).

[9] Sumec, J. et al., *Elasticity and Plasticity in Civil Engineering*, Slovak Technical University Bratislava (2007).

Proceedings of EQUADIFF 2017
pp. 181–190

CROSS-DIFFUSION SYSTEMS WITH ENTROPY STRUCTURE*

ANSGAR JÜNGEL†

Abstract. Some results on cross-diffusion systems with entropy structure are reviewed. The focus is on local-in-time existence results for general systems with normally elliptic diffusion operators, due to Amann, and global-in-time existence theorems by Lepoutre, Moussa, and co-workers for cross-diffusion systems with an additional Laplace structure. The boundedness-by-entropy method allows for global bounded weak solutions to certain diffusion systems. Furthermore, a partial result on the uniqueness of weak solutions is recalled, and some open problems are presented.

Key words. Strongly coupled parabolic systems, local existence of solutions, global existence of solutions, gradient flow, duality method, boundedness-by-entropy method, nonlinear Aubin-Lions lemma, Kullback-Leibler entropy.

AMS subject classifications. 35K51, 35K57, 35B65.

1. Introduction. Multi-species systems from physics, biology, chemistry, etc. can be modeled by reaction-diffusion equations. When the gradient of the density of one species induces a flux of another species, cross diffusion occurs. Mathematically, this means that the diffusion matrix involves nonvanishing off-diagonal elements. In many applications, it turns out that the diffusion matrix is neither symmetric nor positive definite, which considerably complicates the mathematical analysis (see the examples in Section 2 and [25, Section 4.1]). In recent years, some progress has been made in this analysis by identifying a structural condition, namely a formal gradient-flow or entropy structure, allowing for a mathematical treatment. In this review, we report on selected results obtained from several researchers.

The cross-diffusion equations have the form

$$\partial_t u_i - \sum_{j=1}^n \operatorname{div}(A_{ij}(u)\nabla u_j) = f_i(u) \quad \text{in } \Omega,\ t>0,\ i=1,\ldots,n, \tag{1.1}$$

where $u_i(x,t)$ is the density or concentration or volume fraction of the ith species of a multicomponent mixture, $u=(u_1,\ldots,u_n)$, $A_{ij}(u)$ are the diffusion coefficients, $f_i(u)$ is the reaction term of the ith species, and $\Omega\subset\mathbb{R}^d$ ($d\ge 1$) is a bounded domain with smooth boundary. We impose no-flux and initial conditions

$$\sum_{j=1}^n A_{ij}\nabla u_j\cdot\nu = 0 \quad \text{on } \partial\Omega,\ t>0, \quad u_i(0)=u_i^0 \quad \text{in } \Omega,\ i=1,\ldots,n, \tag{1.2}$$

with the exterior normal unit vector ν on $\partial\Omega$, but Dirichlet or mixed Dirichlet-Neumann boundary conditions could be considered as well [20]. Setting $A(u)=(A_{ij}(u))$ and $f(u)=(f_1(u),\ldots,f_n(u))$, we may write (1.1) more compactly as

$$\partial_t u - \operatorname{div}(A(u)\nabla u) = f(u) \quad \text{in } \Omega,\ t>0.$$

*The author acknowledges partial support from the Austrian Science Fund (FWF), grants P27352, P30000, F65, and W1245.

†Institute for Analysis and Scientific Computing, Vienna University of Technology, Wiedner Hauptstraße 8–10, 1040 Wien, Austria (juengel@tuwien.ac.at).

In contrast to scalar parabolic equations, generally there do not exist maximum principles or a regularity theory for diffusion systems. For instance, there exist Hölder continuous solutions to certain parabolic systems that develop singularities in finite time [37]. Here, the situation is even worse: The diffusion matrix $A(u)$ is generally neither symmetric nor positive definite such that coercivity theory cannot be applied. Our approach is to assume a structure inspired from thermodynamics: We suppose that there exists a convex function $h:\mathbb{R}^n\to\mathbb{R}$, called an entropy density, such that the (possibly nonsymmetric) matrix product $h''(u)A(u)$ is positive semidefinite (in the sense $z^\top h''(u)A(u)z\geq 0$ for all $z\in\mathbb{R}^n$). Here, $h''(u)$ denotes the Hessian of h at the point u. We say that A has a *strict* entropy structure if $h''(u)A(u)$ is positive definite for all u. Then the entropy $\mathcal{H}[u]=\int_\Omega h(u)dx$ is a Lyapunov functional along solutions to (1.1)-(1.2) if $f(u)\cdot h'(u)\leq 0$ for all u:

$$\frac{d\mathcal{H}}{dt}=\int_\Omega \partial_t u\cdot h'(u)dx=-\int_\Omega \nabla u : h''(u)A(u)\nabla u dx+\int_\Omega f(u)\cdot h'(u)dx\leq 0, \quad (1.3)$$

where ":" denotes the Frobenius matrix product. If $h''(u)A(u)$ is positive definite, this yields gradient estimates needed for the global existence analysis.

Introducing the entropy variables $w_i=\partial h/\partial u_i$ or $w=h'(u)$, we may write (1.1) equivalently as

$$\partial_t u(w)-\operatorname{div}(B(w)\nabla w)=f(u(w)),\quad B(w):=A(u(w))h''(u(w))^{-1}, \quad (1.4)$$

where $u(w)=(h')^{-1}(w)$ is interpreted as a function of $w=(w_1,\ldots,w_n)$ and $h''(u)^{-1}$ is the inverse of the Hessian of h. By assumption, $B(w)$ is positive semidefinite, which indicates a (nonstandard) parabolic structure.

The entropy structure will be made more explicit for two examples in Section 2. In Sections 3 and 4, the local and global in time existence of solutions, respectively, will be reviewed. Furthermore, we comment in Section 5 on uniqueness results, and we close in Section 6 with some open problems.

2. Examples. We present two prototypic examples.

Example 1 (Maxwell-Stefan equations). The dynamics of a fluid mixture of $n=3$ components with volume fractions u_1, u_2, $u_3=1-u_1-u_2$ can be described by the Maxwell-Stefan equations [38], defined by (1.1) with

$$A(u)=\frac{1}{a(u)}\begin{pmatrix} d_2+(d_0-d_2)u_1 & (d_0-d_1)u_1 \\ (d_0-d_2)u_2 & d_1+(d_0-d_1)u_2\end{pmatrix},$$

where $d_i>0$ and $a(u)=d_1d_2(1-u_1-u_2)+d_0(d_1u_1+d_2u_2)>0$. The model can be generalized to $n\geq 3$ components; see [4, 26]. For simplicity, we set $f\equiv 0$. Define the entropy density

$$h(u)=\sum_{i=1}^2 u_i(\log u_i-1)+(1-u_1-u_2)\big(\log(1-u_1-u_2)-1\big),$$

where $u=(u_1,u_2)$, and the entropy $\mathcal{H}[u]=\int_\Omega h(u)dx$. A formal computation shows that

$$\frac{d\mathcal{H}}{dt}+\int_\Omega\frac{1}{a(u)}\bigg(d_2\frac{|\nabla u_1|^2}{u_1}+d_1\frac{|\nabla u_2|^2}{u_2}+d_0\frac{|\nabla(u_1+u_2)|^2}{1-u_1-u_2}\bigg)dx=0,$$

and in particular, $h''(u)A(u)$ is positive definite for $u_i > 0$. The entropy variables become $w_i = \partial h/\partial u_i = \log(u_i/(1-u_1-u_2))$ with inverse $u_i(w) = e^{w_i}/(1+e^{w_1}+e^{w_2})$, which lies in the triangle $G = \{u \in \mathbb{R}^2 : u_1, u_2 > 0,\ 1-u_1-u_2 > 0\}$. This property makes sense since u_i are volume fractions and they are expected to be bounded. This property can be exploited in the existence analysis to obtain *bounded* solutions without using a maximum principle (which generally cannot be applied). □

EXAMPLE 2 (Population model). The evolution of two interacting species may be modeled by equations (1.1) with the diffusion matrix

$$A(u) = \begin{pmatrix} a_{10} + a_{11}u_1 + a_{12}u_2 & a_{12}u_1 \\ a_{21}u_2 & a_{20} + a_{21}u_1 + a_{22}u_2 \end{pmatrix},$$

where $a_{ij} \geq 0$ [36]. We neglect the environmental potential and source terms, so $f \equiv 0$. The entropy is given by $\mathcal{H}[u] = \int_\Omega h(u)dx$, where $h(u) = a_{21}u_1(\log u_1 - 1) + a_{12}u_2(\log u_2 - 1)$. A formal computation shows that

$$\frac{d\mathcal{H}}{dt} + \int_\Omega \left\{ \left(\frac{a_{10}}{u_1} + a_{21}a_{11} \right) |\nabla u_1|^2 + \left(\frac{a_{20}}{u_2} + a_{12}a_{22} \right) |\nabla u_2|^2 + 4|\nabla \sqrt{u_1 u_2}|^2 \right\} dx = 0. \tag{2.1}$$

The entropy variables are $w_1 = a_{21}\log u_1$, $w_0 = a_{12}\log u_2$. Then the population densities are $u_1 = e^{w_1/a_{21}}$, $u_2 = e^{w_2/a_{12}} > 0$. An upper bound cannot be expected.

The model can be generalized to $n \geq 2$ species with diffusion coefficients

$$A_{ij}(u) = \delta_{ij}\left(a_{i0} + \sum_{k=1}^n a_{ik}u_k \right) + a_{ij}u_i, \quad i,j = 1,\ldots,n. \tag{2.2}$$

The entropy structure is more delicate than in the two-species case. Indeed, assume that there exist numbers $\pi_i > 0$ such that the equations

$$\pi_i a_{ij} = \pi_j a_{ji}, \quad i,j = 1,\ldots,n, \tag{2.3}$$

are satisfied. Then $h(u) = \sum_{i=1}^n \pi_i u_i(\log u_i - 1)$ is an entropy density, i.e. $d\mathcal{H}/dt \leq 0$ [8]. Equations (2.3) are recognized as the detailed-balance condition for the Markov chain with transition rates a_{ij}, and $\pi = (\pi_1,\ldots,\pi_n)$ is the corresponding invariant measure [25, Section 5.1]. □

3. Local existence of classical solutions. A very general result on the local-in-time existence of classical solutions to diffusion systems was proved by Amann (see [2, Section 1] or [3, Theorem 14.1]). A special version reads as follows.

THEOREM 3.1 (Amann [2]). *Let $G \subset \mathbb{R}^n$ be open, A_{ij}, $f_i \in C^\infty(G)$, all eigenvalues of $A(u)$ have positive real parts for all $u \in G$, and $u^0 \in V := \{v \in W^{1,p}(\Omega;\mathbb{R}^n) : v(\overline{\Omega}) \subset G\}$, where $p > d$. Then there exists a unique maximal solution u to* (1.1)-(1.2) *satisfying $u \in C^0([0,T^*);V) \cap C^\infty(\overline{\Omega}\times(0,T^*);\mathbb{R}^n)$, where $0 < T^* \leq \infty$.*

An elliptic operator $u \mapsto \operatorname{div}(A(u)\nabla u)$ with the property that all eigenvalues of $A(u)$ have positive real parts is called *normally elliptic*. We claim that any cross-diffusion system with strict entropy structure is normally elliptic.

LEMMA 3.2 (Eigenvalues of A). *Let $A \in \mathbb{R}^{n\times n}$. We assume that there exists a symmetric, positive definite matrix $H \in \mathbb{R}^{n\times n}$ such that HA is positive definite. Then every eigenvalue of A has a positive real part.*

In the context of cross-diffusion systems, H stands for the Hessian $h''(u)$.

Proof. Let $\lambda = \xi + i\eta$ with ξ, $\eta \in \mathbb{R}$ be an eigenvalue of A with eigenvector $u = v + iw$, where v, $w \in \mathbb{R}^n$ with $v \neq 0$ or $w \neq 0$. It follows from $Au = \lambda u$

that $Av = \xi v - \eta w$, $Aw = \eta v + \xi w$. We multiply both equations by $v^\top H$, $w^\top H$, respectively:

$$0 < v^\top HAv = \xi v^\top Hv - \eta v^\top Hw, \quad 0 < w^\top HAw = \eta w^\top Hv + \xi w^\top Hw.$$

Since H is symmetric, we have $v^\top Hw = w^\top Hv$. Therefore, adding both identities,

$$0 < v^\top HAv + w^\top HAw = \xi(v^\top Hv + w^\top Hw).$$

We infer from the positive definiteness of H that $\xi > 0$, proving the claim. □

4. Global existence of weak solutions. The classical solution of Amann can be continued for all time under some assumptions [3, Theorem 15.3].

THEOREM 4.1 (Amann [3]). *Let u be the classical maximal solution to* (1.1)-(1.2) *on $[0,T^*)$. Assume that $u|_{[0,T]}$ is bounded away from ∂G for each $T > 0$ and that there exists $\alpha > 0$ such that $\|u(t)\|_{C^{0,\alpha}} \le C(T)$ for all $0 \le t \le T < \infty$, $t < T^*$. Then $T^* = \infty$.*

Unfortunately, it is not easy to derive a uniform bound in the Hölder norm. A possibility is to show that the gradient $\nabla u_i(t)$ satisfies some higher integrability, namely $L^p(\Omega)$ for $p > d$, since $W^{1,p}(\Omega)$ embeds continuously into $C^{0,\alpha}(\overline{\Omega})$ for $\alpha = 1 - p/d > 0$. Estimates in the $W^{1,p}$ norm with $p > d$ for a particular system were derived in, e.g., [23, 29].

Another approach is to find weak solutions using the entropy method as outlined in the introduction. The key elements of the existence proof are the definition of an approximate problem and a compactness argument. We are aware of two approaches in the literature. In both approaches, the time derivative is replaced by the implicit Euler discretization. This avoids issues with the (low) time regularity. To define the change of unknowns $u(w)$, we need bounded approximate solutions w. The first approach regularizes the equations by adding a weak form of $\varepsilon((-\Delta)^s w + w)$. Since $H^s(\Omega) \hookrightarrow L^\infty(\Omega)$ for $s > d/2$, this yields bounded weak solutions. The second approach formulates the implicit Euler scheme as a fixed-point equation involving the solution operator $(M - \Delta)^{-1}$ for sufficiently large $M > 0$. This allows one to exploit the regularization property of the solution operator $(M - \Delta)^{-1} : L^p(\Omega) \to W^{2,p}(\Omega)$, and the continuous embedding $W^{2,p}(\Omega) \hookrightarrow L^\infty(\Omega)$ for $p > d$ yields bounded solutions. We detail both approaches in the following subsections.

4.1. Boundedness-by-entropy method. This method does not only give the global existence of solutions but it also yields L^∞ bounds. It was first used in [5] and made systematic in [24]. The first key assumption is that the derivative $h' : G \to \mathbb{R}^n$ is invertible, where $G \subset \mathbb{R}^n$ is a bounded set. Then $u(w(x,t)) = (h')^{-1}(w(x,t)) \in G$ yields lower and upper bounds for the densities u_i; see Example 1. The second key assumption is the positive definiteness of $h''(u)A(u)$. Applications indicate that this property does not hold uniformly in u. Therefore, we impose a weaker condition.

(H1) $h \in C^2(G;[0,\infty))$ is convex with invertible derivative $h' : G \to \mathbb{R}^n$.

(H2) $G \subset (0,1)^n$ and for $z = (z_1,\ldots,z_n)^\top \in \mathbb{R}^n$ and $u = (u_1,\ldots,u_n) \in G$,

$$z^\top h''(u)A(u)z \ge \kappa \sum_{i=1}^n u_i^{2m-2} z_i^2, \quad \text{where } m \ge \frac{1}{2},\ \kappa > 0. \tag{4.1}$$

(H3) $A = (A_{ij}) \in C^0(G;\mathbb{R}^{n\times n})$ and $|A_{ij}(u)| \le C_A|u_j|^a$ for all $u \in G$, $i,j = 1,\ldots,n$, where C_A, $a > 0$.

(H4) $f \in C^0(G;\mathbb{R}^n)$ and $\exists\ C_f > 0$: $\forall\ u \in G$: $f(u) \cdot h'(u) \le C_f(1 + h(u))$.

Hypothesis (4.1) is satisfied with $m = \frac{1}{2}$ in Examples 1 and 2 if $a_{10} > 0$, $a_{20} > 0$ and $m = 1$ in Example 2 if $a_{11} > 0$, $a_{22} > 0$. The following theorem is proved in [24, Theorem 2]; also see [25, Section 4.4].

THEOREM 4.2 (Global existence [24]). *Let (H1)-(H4) hold and let $u^0 \in L^1(\Omega;\mathbb{R}^n)$ be such that $u^0(\Omega) \subset \overline{G}$. Then there exists a bounded weak solution u to (1.1)-(1.2) satisfying $u(\Omega,t) \subset \overline{G}$ for all $t > 0$ and $u \in L^2_{\rm loc}(0,\infty;H^1(\Omega;\mathbb{R}^n))$, $\partial_t u \in L^2_{\rm loc}(0,\infty;H^1(\Omega;\mathbb{R}^n)')$, for all $T > 0$ and $\phi \in L^2(0,T;H^1(\Omega;\mathbb{R}^n))$,*

$$\int_0^T \langle \partial_t u, \phi\rangle dt + \int_0^T \int_\Omega \nabla\phi : A(u)\nabla u dx dt = \int_0^T \int_\Omega f(u)\cdot\phi dx dt,$$

where $\langle\cdot,\cdot\rangle$ denotes the dual pairing of $H^1(\Omega)'$, and $u(0) = u^0$ holds in $H^1(\Omega;\mathbb{R}^n)'$.

The idea of the proof is to solve first for given u^{k-1} the regularized problem

$$\begin{aligned}\frac{1}{\tau}\int_\Omega \big(u(w^k) - u(w^{k-1})\big)\cdot\phi dx + \int_\Omega \nabla\phi : B(w^k)\nabla w^k dx \\ + \int_\Omega \bigg(\sum_{|\alpha|=s} D^\alpha w^k \cdot D^\alpha\phi + w^k\cdot\phi\bigg)dx = \int_\Omega f(u(w^k))\cdot\phi dx\end{aligned} \tag{4.2}$$

for $\phi \in H^s(\Omega;\mathbb{R}^n)$, where $s > d/2$, $\alpha = (\alpha_1,\ldots,\alpha_n) \in \mathbb{N}_0^n$ with $|\alpha| = \alpha_1 + \cdots + \alpha_n = s$ is a multiindex, $D^\alpha = \partial^s/(\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n})$ is a partial derivative of order m, $u(w) := (h')^{-1}(w)$, and w^k is an approximation of $w(\cdot,k\tau)$ with the time step $\tau > 0$. This problem is solved by the Leray-Schauder theorem. Uniform estimates are derived from a discrete version of the entropy-production identity (1.3) and Hypothesis (H2).

Let $u^{(\tau)}(x,t) = u(w^k(x))$ for $x \in \Omega$ and $t \in ((k-1)\tau, k\tau]$, $k = 1,\ldots,N$, be piecewise constant functions in time. If $t = 0$, we set $u^{(\tau)}(\cdot,0) = u^0$. We also need the time shift operator $(\sigma_\tau u^{(\tau)})(\cdot,t) = u(w^{k-1})$ for $t \in ((k-1)\tau, k\tau]$. It follows from the boundedness and the discrete entropy-production inequality that [25, Section 4.4]

$$\|u^{(\tau)}\|_{L^\infty(0,T;L^1(\Omega))} \le C, \tag{4.3}$$

$$\tau^{-1}\|u^{(\tau)} - \sigma_\tau u^{(\tau)}\|_{L^2(0,T;H^s(\Omega)')} + \|(u^{(\tau)})^m\|_{L^2(0,T;H^1(\Omega))} \le C, \tag{4.4}$$

where $C > 0$ is independent of ε and τ. (In fact, we have even a bound for $(u^{(\tau)})$ in $L^\infty(0,T;L^\infty(\Omega))$.) If $m = 1$, we deduce relative compactness for $(u^{(\tau)})$ in $L^2(Q_T)$ (where $Q_T = \Omega\times(0,T)$) from the discrete Aubin-Lions lemma in the version of [15]. When $m \ne 1$, we need the nonlinear version of [8, 11, 39].

LEMMA 4.3 (Nonlinear Aubin-Lions). *Let $T > 0$, $m > 0$, and let $(u^{(\tau)})$ be a family of nonnegative functions that are piecewise constant in time with uniform time step $\tau > 0$. Assume that there exists $C > 0$ such that (4.4) holds for all $\tau > 0$.*

- *Let $m > 1$ and let $(u^{(\tau)})$ be bounded in $L^\infty(Q_T)$. Then $(u^{(\tau)})$ is relatively compact in $L^p(Q_T)$ for any $p < \infty$ [39, Lemma 9].*
- *Let $1/2 \le m \le 1$. Then $(u^{(\tau)})$ is relatively compact in $L^{2m}(0,T;L^{pm}(\Omega))$, where $p \ge 1/m$ and $H^1(\Omega) \hookrightarrow L^p(\Omega)$ is continuous [11, Theorem 3].*
- *Let $\max\{0, 1/2 - 1/d\} < m < 1/2$ and let (4.3) hold. Then $(u^{(\tau)})$ is relatively compact in $L^1(0,T;L^{d/(d-1)}(\Omega))$ [8, Theorem 22].*

Another version of the nonlinear Aubin-Lions lemma is shown in [31].

Theorem 4.2 can be directly applied to the Maxwell-Stefan equations from Example 1 yielding the global existence of bounded weak solutions.

4.2. Cross-diffusion system with Laplace structure. Theorem 4.2 can be only applied to situations in which the densities are bounded (volume fractions). However, the method of proof can be adapted to cases, in which the domain G is not bounded. The main difference is that we cannot work in $L^\infty(\Omega)$ anymore but only in $L^p(\Omega)$ for suitable $p<\infty$. The precise value of p depends on m in Hypothesis (H2), and a global existence result can be proved under certain growth conditions on $A_{ij}(u)$ and $f_i(u)$. As an example, consider the population model from Example 2 for $n\geq 2$ species. The following theorem was proved in [8].

THEOREM 4.4 (Population model, linear A_{ij} [8]). *Let $u_i^0\geq 0$ be such that $\int_\Omega h(u^0)dx<\infty$ and let the detailed-balance condition (2.3) and $a_{ii}>0$ hold. Then there exists a weak solution $u=(u_1,\ldots,u_n)$ to (1.1)-(1.2) with diffusion matrix (2.2) satisfying $u_i\geq 0$ in Ω, $t>0$, and $u_i\in L^2_{\rm loc}(0,\infty;H^1(\Omega))$, $\partial_t u_i\in L^{q'}_{\rm loc}(0,T;W^{1,q}(\Omega)')$, where $q=2d+2$ and $q'=(2d+2)/(2d+1)$.*

We have assumed that there is self-diffusion $a_{ii}>0$, yielding an L^2 estimate for ∇u_i, which is stronger than the L^2 estimate for ∇u_i^m with $m<1$. An existence result with vanishing self-diffusion $a_{ii}=0$ was shown in [7] for the two-species model. Here, we only have an L^2 bound for $\nabla\sqrt{u_i}$. The lack of regularity for ∇u_i can be compensated by exploiting the gradient estimate for $\nabla\sqrt{u_1u_2}$ in (2.1) and an $L^2\log L^2$ estimate coming from the Lotka-Volterra reaction terms.

The detailed-balance condition can be replaced by a "weak cross-diffusion" assumption which is automatically satisfied if (A_{ij}) is symmetric; see [8, Formula (12)].

Another generalization concerns *nonlinear* diffusion coefficients

$$A_{ij}(u)=\delta_{ij}\bigg(a_{i0}+\sum_{k=1}^n a_{ik}u_k^{s_k}\bigg)+s_j a_{ij}u_i u_j^{s_j-1},\quad i,j=1,\ldots,n, \tag{4.5}$$

for $s_i\geq 0$. The corresponding cross-diffusion system can be analyzed by the method of the previous subsection. However, improved results can be obtained by exploiting the Laplace structure, meaning that (1.1) with (4.5) writes as

$$\partial_t u_i-\Delta(u_i p_i(u))=f_i(u),\quad \text{where } p_i(u)=\alpha_{i0}+\sum_{j=1}^n\alpha_{ij}u_j^{s_j}, \tag{4.6}$$

and $\alpha_{ij}=a_{ij}$ for $i\neq j$ and $\alpha_{ii}=(s_i+1)a_{ii}$. Let $a_{ii}>0$ and $s_i\leq 2$. Then, by the entropy-production inequality, $\nabla u_i^{s_i/2}$ is bounded in $L^2(Q_T)$, and the Gagliardo-Nirenberg inequality with $q=2+4/(ds_i)$ and $\theta=ds_i/(2+ds_i)$ shows that

$$\begin{aligned}\|u_i^{s_i/2}\|_{L^q(Q_T)}^q&=\int_0^T\|u_i^{s_i/2}\|_{L^q(\Omega)}^q dt\leq\int_0^T\|u_i^{s_i/2}\|_{H^1(\Omega)}^{q\theta}\|u_i^{s_i/2}\|_{L^{2/s_i}(\Omega)}^{q(1-\theta)}dt\\&\leq\|u_i\|_{L^\infty(0,T;L^1(\Omega))}^{qs_i(1-\theta)/2}\int_0^T\|u_i^{s_i/2}\|_{H^1(\Omega)}^2dt\leq C.\end{aligned}$$

We deduce that u_i is bounded in $L^{s_i+2/d}(Q_T)$. Using the duality method of Pierre [35], an improved regularity result can be derived. Indeed, set $\bar u=\sum_{i=1}^n u_i$ and $\mu=\sum_{i=1}^n u_i p_i(u)/\bar u$. If $f_i(u)$ grows at most linearly in u_i, we find that $\bar u$ solves $\partial_t\bar u-\Delta(\mu\bar u)\leq C\bar u$ for some constant $C>0$ depending on f_i. Then (see, e.g., [30, Lemma 1.2] or the review [34])

$$\int_0^T\int_\Omega\mu\bar u^2dxdt\leq C(T,u^0). \tag{4.7}$$

We infer that $u_i^2 p_i(u)$ is uniformly bounded in $L^1(Q_T)$, giving a bound for u_i in $L^{s_i+2}(Q_T)$. For $d>1$, this bound is better than the bound in $L^{s_i+2/d}(Q_T)$ derived above. The improved regularity is a key element in proving the global existence of solutions [30, Theorem 1.10] (also see the precursor versions in [12, 13]). We define the entropy density $h(u)=\sum_{i=1}^n h_i(u_i)$, where

$$h_i(u_i)=\begin{cases}(u_i^{s_i}-s_iu_i+s_i-1)/(s_i-1) & \text{if } s_i\neq 1,\\ u_i(\log u_i-1)+1 & \text{if } s_i=1.\end{cases}$$

THEOREM 4.5 (Population model, nonlinear A_{ij} [30]). *Assume that $s_i>0$, $s_is_j\le 1$ for $i\neq j$, let the detailed-balance condition (2.3) hold, and $f_i(u)=b_{i0}-\sum_{j=1}^n b_{ij}u_j^{\alpha_{ij}}$ for $b_{ij}\ge 0$ and $\alpha_{ij}<1$. Finally, let $u_i^0\in L^1(\Omega)\cap H^1(\Omega)'$, $\int_\Omega h_i(u_i^0)<\infty$. Then there exists an integrable solution $u_i\ge 0$ to (4.6) and (1.2) such that for all smooth test functions ϕ satisfying $\nabla\phi_i\cdot\nu=0$ on $\partial\Omega$,*

$$-\int_0^\infty\int_\Omega u\cdot\partial_t\phi dxdt-\int_0^\infty\int_\Omega\sum_{i=1}^n u_ip_i(u)\Delta\phi_i dxdt$$
$$=\int_0^\infty\int_\Omega f(u)\cdot\phi dxdt+\int_\Omega u^0(x)\cdot\phi(x,0)dx.$$

It is an open problem to show the same result for arbitrary $s_i>0$.

The key idea of the proof is to formulate the implicit Euler scheme

$$\tau^{-1}(u_i^k-u_i^{k-1})=\Delta F_i(u^k)+f_i(u^k),\quad \text{where } F_i(u^k)=u_i^kp_i(u^k),$$

as the fixed-point equation

$$u^k=F^{-1}\Big((M-\Delta)^{-1}\big(u^{k-1}-u^k+MF(u^k)\big)\Big),$$

where $F=(F_1,\ldots,F_n)$ and $M>0$ is a sufficiently large number. In fact, if M is large and $u_i^{k-1}>0$, we can show that $v:=u_i^{k-1}-u_i^k+MF_i(u_i^k)>0$, and by the maximum principle, $(M-\Delta)^{-1}v>0$. Then, if F is a homeomorphism on $[0,\infty)^n$, $u_i^k>0$, which yields positivity. Moreover, elliptic regularity theory implies that for $v\in L^p(\Omega)$ with $p>d/2$, we have $(M-\Delta)^{-1}v\in W^{2,p}(\Omega)\hookrightarrow L^\infty(\Omega)$. This shows that u_i^k is bounded in L^∞ and it defines a fixed-point operator on $L^\infty(\Omega;\mathbb{R}^n)$.

The main assumption is that F is a homeomorphism. Under this assumption, Theorem 4.5 can be considerably generalized; see [30, Theorem 1.7] for details.

5. Uniqueness of weak solutions. The uniqueness of weak solutions to diffusion systems is a delicate topic. One of the first uniqueness results was shown in [1], assuming that the elliptic operator is linear and the time derivative of u_i is integrable. The latter hypothesis was relaxed in [32] allowing for finite-energy solutions but to scalar equations only. The uniqueness of solutions was shown in [33] for a cross-diffusion system with a strictly positive definite diffusion matrix. For cross-diffusion systems with entropy structure (and not necessarily positive definite $A(u)$), there are much less papers. The first result was for a special two-species population model [27], later extended to a volume-filling system [39], and generalized in [9] for a class of cross-diffusion systems. In this section, we report on the result of [9].

We allow for cross-diffusion systems involving drift terms,

$$\partial_tu_i=\operatorname{div}\sum_{j=1}^n\big(A_{ij}(u)\nabla u_j+B_{ij}(u)\nabla\phi\big),\quad i=1,\ldots,n, \tag{5.1}$$

where ϕ is a potential solving the Poisson equation

$$-\Delta\phi = u_0 - f(x) \quad \text{in } \Omega, \quad u_0 := \sum_{i=1}^n a_i u_i, \tag{5.2}$$

$a_i \geq 0$ are some constants, and $f(x)$ is a given background density. The equations are complemented by (1.2) and $\nabla\phi\cdot\nu = 0$ on $\partial\Omega$, $t > 0$. For consistency, we need to impose the condition $\int_\Omega \sum_{i=1}^n a_i u_i^0 dx = \int_\Omega f(x)dx$.

The uniqueness proof only works for a special class of coefficients, namely

$$A_{ij}(u) = p(u_0)\delta_{ij} + a_j u_i q(u_0), \quad B_{ij}(u) = r(u_0)u_i\delta_{ij}, \quad i,j = 1,\ldots,n, \tag{5.3}$$

for some functions p, q, and r. The main result is as follows.

THEOREM 5.1 (Uniqueness of bounded weak solutions [9]). *Let $u^0 \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. Let (u,ϕ) be a weak solution to (5.1)-(5.3), (1.2) such that $u_0(\Omega,t) \subset [0,L]$ for some $L > 0$. Assume that there exists $M > 0$ such that for all $s \in [0,L]$,*

$$p(s) \geq 0, \quad p(s) + q(s)s \geq 0, \tag{5.4}$$

$$r(s)s \in C^1([0,L]), \quad \frac{(r(s)+r'(s)s)^2}{p(s)+q(s)s} \leq M. \tag{5.5}$$

Then (u,ϕ) is unique in the class of solutions satisfying $\int_\Omega \phi dx = 0$, $\nabla\phi \in L^\infty(0,T;L^\infty(\Omega))$, and $u_i \in L^2(0,T;H^1(\Omega))$, $\partial_t u_i \in L^2(0,T;H^1(\Omega)')$ for $i = 1,\ldots,n$. In the case $r \equiv 0$, the boundedness of u_0 is not needed, provided that $\sqrt{p(u_0)}\nabla u_i$, $\sqrt{|q(u_0)|}\nabla u_i \in L^2(\Omega\times(0,T))$.

The proof is based on the H^{-1} method and the entropy method of Gajewski [19]. First, we show the uniqueness of $u_0 = \sum_{i=1}^n a_i u_i$, solving

$$\partial_t u_0 = \operatorname{div}\big(\nabla Q(u_0) + R(u_0)\nabla\phi\big),$$

where $Q(s) = \int_0^s (p(z)+q(z)z)dz$ and $R(s) = r(s)s$. Sine Q is nondecreasing, the use of the H^{-1} technique seems to be natural. Given two solutions (u,ϕ) and (v,ψ), the idea is to use the test function χ that solves the dual problem $-\Delta\chi = u_0 - v_0$ in Ω, $\nabla\chi\cdot\nu = 0$ on $\partial\Omega$ and to show that $\frac{d}{dt}\|\nabla\chi\|_{L^2(\Omega)}^2 \leq C\|\nabla\chi\|_{L^2(\Omega)}^2$, using the monotonicity of Q. This implies that $u_0 = v_0$ and $\phi = \psi$. Second, we differentiate (a regularized version of) the semimetric

$$d(u,v) = \sum_{i=1}^n \int_\Omega \left(h(u_i) + h(v_i) - 2h\left(\frac{u_i+v_i}{2}\right)\right)dx,$$

where $h(s) = s(\log s - 1) + 1$. Computing the time derivative of $d(u(t),v(t))$, it turns out that the drift terms cancel and we end up with $\frac{d}{dt}d(u,v) \leq 0$ implying that $u = v$.

Gajewski's semimetric is related to the relative entropy or Kullback-Leibler entropy $\mathcal{H}[u|v] = \mathcal{H}[u] - \mathcal{H}[v] - \mathcal{H}'[v]\cdot\mathcal{H}(u-v)$ used in statistics [28]. In fact, the proof of Theorem 5.1 can be performed as well with the symmetrized relative entropy $d_0(u,v) = \mathcal{H}[u|v] + \mathcal{H}[v|u]$. Both distances $d(u,v)$ and $d_0(u,v)$ behave like $|u-v|^2$ for "small" $|u-v|$, but they lead to different expressions when computed explicitly. The Kullback-Leibler entropy was also employed to derive explicit exponential convergence rates to equilibrium [6] and to prove weak-strong uniqueness results for (diagonal) reaction-diffusion systems [18].

6. Open problems. We mention some open questions.

- *Reaction terms:* Hypothesis (H4) excludes reaction terms which grow superlinearly. The global existence of solutions to cross-diffusion systems with, for instance, quadratic reactions is an open problem. One approach could be to consider renormalized instead of weak solutions, as done in [17] for (diagonal) reaction-diffusion systems. This is currently under development [9]. Another idea is to exploit the entropy techniques devised for reaction-diffusion systems [16].
- *n-species population model:* It is an open problem to find global solutions to the population model with diffusion matrix (2.2) and $n \geq 3$ without detailed balance or "weak cross-diffusion". Numerical experiments indicate that standard choices like the Boltzmann entropy, relative entropy, etc. are not Lyapunov functionals. So, the problem to find a priori estimates is open.
- *Uniqueness of solutions:* The uniqueness result presented in Theorem 5.1 is rather particular. One may ask whether weak-strong uniqueness of solutions can be shown like in [18] for diagonal diffusion systems. In fact, uniqueness of weak solutions is known to be delicate even for drift-diffusion equations; see, e.g., [14].
- *Regularity theory:* The duality method yields global L^p regularity results for cross-diffusion systems with Laplace structure (see (4.7)). Another approach is to apply maximal L^p regularity theory as done in [21] for Maxwell-Stefan systems, at least for local solutions. The (open) question is to what extent this theory can be applied to general systems with entropy structure?
- *Entropies:* Given a cross-diffusion system, a major open question is how an entropy structure can be detected. In thermodynamics, often the entropy (more precisely: free energy) and entropy production are given and the system of partial differential equations follows from these quantities. Furthermore, it is an open question how large is the class of cross-diffusion systems with entropy structure. Are there diffusion systems with normally elliptic operator, which have no entropy structure?

REFERENCES

[1] H.-W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.* 183 (1983), 311-341.

[2] H. Amann. Dynamic theory of quasilinear parabolic equations. II. Reaction-diffusion systems. *Diff. Int. Eqs.* 3 (1990), 13-75.

[3] H. Amann. Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems. In: H.J. Schmeisser and H. Triebel (editors), *Function Spaces, Differential Operators and Nonlinear Analysis*, pp. 9-126. Teubner, Stuttgart, 1993.

[4] D. Bothe. On the Maxwell-Stefan equations to multicomponent diffusion. In: *Progress in Nonlinear Differential Equations and their Applications*, pp. 81-93. Springer, Basel, 2011.

[5] M. Burger, M. Di Francesco, J.-F. Pietschmann, and B. Schlake. Nonlinear cross-diffusion with size exclusion. *SIAM J. Math. Anal.* 42 (2010), 2842-2871.

[6] J. A. Carrillo, A. Jüngel, P. Markowich, G. Toscani, and A. Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatsh. Math.* 133 (2001), 1-82.

[7] L. Chen and A. Jüngel. Analysis of a parabolic cross-diffusion population model without self-diffusion. *J. Diff. Eqs.* 224 (2006), 39-59.

[8] X. Chen, E. Daus, and A. Jüngel. Global existence analysis of cross-diffusion population systems for multiple species. To appear in *Arch. Ration. Mech. Anal.*, 2017. arXiv:1608.03696.

[9] X. Chen and A. Jüngel. A note on the uniqueness of weak solutions to a class of cross-diffusion systems. Submitted for publication, 2017. arXiv:1706.08812.

[10] X. Chen and A. Jüngel. Global renormalized solutions to reaction-cross-diffusion systems. Work in progress, 2017.

[11] X. Chen, A. Jüngel, and J.-G. Liu. A note on Aubin-Lions-Dubinskiĭ lemmas. *Acta Appl. Math.* 133 (2014), 33-43.

[12] L. Desvillettes, T. Lepoutre, and A. Moussa. Entropy, duality, and cross diffusion. *SIAM J. Math. Anal.* 46 (2014), 820-853.
[13] L. Desvillettes, T. Lepoutre, A. Moussa, and A. Trescases. On the entropic structure of reaction-cross diffusion systems. *Commun. Partial Diff. Eqs.* 40 (2015), 1705-1747.
[14] J. I. Díaz, G. Galiano, and A. Jüngel. On a quasilinear degenerate system arising in semiconductor theory. Part I: existence and uniqueness of solutions. *Nonlin. Anal. RWA* 2 (2001), 305-336.
[15] M. Dreher and A. Jüngel. Compact families of piecewise constant functions in $L^p(0,T;B)$. *Nonlin. Anal.* 75 (2012), 3072-3077.
[16] K. Fellner, W. Prager, B. Q. Tang. The entropy method for reaction-diffusion systems without detailed balance: first order chemical reaction networks. *Kinetic Related Models* 10 (2017), 1055-1087.
[17] J. Fischer. Global existence of renormalized solutions to entropy-dissipating reaction-diffusion systems. *Arch. Ration. Mech. Anal.* 218 (2015), 553-587.
[18] J. Fischer. Weak-strong uniqueness of solutions to entropy-dissipating reaction-diffusion equations. Submitted for publication, 2017. arXiv:1703.00730.
[19] H. Gajewski. On a variant of monotonicity and its application to differential equations. *Nonlin. Anal. TMA* 22 (1994), 73-80.
[20] A. Gerstenmayer and A. Jüngel. Analysis of a degenerate parabolic cross-diffusion system for ion transport. Submitted for publication, 2017. arXiv:1706.07261.
[21] M. Herberg, M. Meyries, J. Prüss, and M. Wilke. Reaction-diffusion systems of Maxwell-Stefan type with reversible mass-action kinetics. *Nonlin. Anal.* 159 (2017), 264-284.
[22] S. Hittmeir and A. Jüngel. Cross diffusion preventing blow up in the two-dimensional Keller–Segel model. *SIAM J. Math. Anal.* 43 (2011), 997-1022.
[23] L. Hoang, T. Nguyen, T. V. Phan. Gradient estimates and global existence of smooth solutions to a cross-diffusion system. *SIAM J. Math. Anal.* 47 (2015), 2122-2177.
[24] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity* 28 (2015), 1963-2001.
[25] A. Jüngel. *Entropy Methods for Diffusive Partial Differential Equations*. BCAM SpringerBriefs, 2016.
[26] A. Jüngel and I. Stelzer. Entropy structure of a cross-diffusion tumor-growth model. *Math. Models Meth. Appl. Sci.* 22 (2012), 1250009, 26 pages.
[27] A. Jüngel and N. Zamponi. Qualitative behavior of solutions to cross-diffusion systems from population dynamics. *J. Math. Anal. Appl.* 440 (2016), 794-809.
[28] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.* 22 (1951), 79-86.
[29] D. Le. Global existence for a class of strongly coupled parabolic systems. *Ann. Matem.* 185 (2006), 133-154.
[30] T. Lepoutre and A. Moussa. Entropic structure and duality for multiple species cross-diffusion systems. *Nonlin. Anal.* 159 (2017), 298-315.
[31] A. Moussa. Some variants of the classical Aubin-Lions lemma. *J. Evol. Eqs.* 16 (2016), 65-93.
[32] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Diff. Eqs.* 131 (1996), 20-38.
[33] D. Pham and R. Temam. A result of uniqueness of solutions of the Shigesada-Kawasaki-Teramoto equations. *Adv. Nonlin. Anal.*, online first, 2017. arXiv:1703.10544.
[34] M. Pierre. Global existence in reaction-diffusion systems with control of mass: a survey. *Milan J. Math.* 78 (2010), 417-455.
[35] M. Pierre and D. Schmitt. Blow up in reaction-diffusion systems with dissipation of mass. *SIAM J. Math. Anal.* 28 (1997), 259-269.
[36] N. Shigesada, K. Kawasaki, and E. Teramoto. Spatial segregation of interacting species. *J. Theor. Biol.* 79 (1979), 83-99.
[37] J. Stará and O. John. Some (new) counterexamples of parabolic systems. *Comment. Math. Univ. Carolin.* 36 (1995), 503-510.
[38] J. Wesselingh and R. Krishna. *Mass Transfer in Multicomponent Mixtures*. Delft University Press, Delft, 2000.
[39] N. Zamponi and A. Jüngel. Analysis of degenerate cross-diffusion population models with volume filling. *Ann. Inst. H. Poincaré Anal. Non Lin.* 34 (2017), 1-29.

Proceedings of EQUADIFF 2017
pp. 191–200

NUMERICAL MODELING OF HEAT EXCHANGE AND UNSATURATED-SATURATED FLOW IN POROUS MEDIA*

JOZEF KAČUR†, PATRIK MIHALA‡, AND MICHAL TÓTH§

Abstract. We discuss the numerical modeling of heat exchange between the infiltrated water and porous media matrix. An unsaturated-saturated flow is considered with boundary conditions reflecting the external driven forces. The developed numerical method is efficient and can be used for solving the inverse problems concerning determination of transmission coefficients for heat energy exchange inside and also on the boundary of porous media. Numerical experiments support our method.

Key words. porous media infiltration, water and heat transport, heat energy exchange, numerical modeling of nonlinear system

AMS subject classifications. 65M08, 65M32, 76S05

1. Introduction. In this contribution we discuss the heat transported by infiltrated water into porous media taking into account the heat exchange between infiltrated water and the porous media matrix assuming the flow is unsaturated. This is motivated by an analysis of hygrothermal insulation properties of building facades. The influence of external weather conditions is included in the considered model. We focus especially on the determination of model parameters in a complex mathematical model. Solution of corresponding inverse problems relies on measurements in laboratory conditions using real 3D samples.

The mathematical model consists of the coupled system of strongly nonlinear PDE of elliptic-parabolic type. The flow of water in unsaturated-saturated porous media is governed by Richard's equation. The heat energy transported by infiltrated water is subject to the convection, molecular diffusion, and dispersion, which are driven by external forces due to water and heat fluxes caused by weather conditions. Mathematical models are well known and presented in many monographs, e.g., [1], with very complex list of quotations. Fundamentals of heat and mass transfer with many applications are discussed in [9]. In our setting the heat energy transmission from water in pores to the porous media matrix is treated analogously to the reversible adsorption of contaminant in unsaturated porous media, see e.g. [8],[2]. Additionally, we take into account the heat conduction of the porous media matrix itself. Thus, soluted contaminant in water is replaced by heat energy. Solving the heat conduction of porous media (without water in pores) is difficult task and is modeled by homogenization method. In our setting we assume very simple heat conduction in matrix, where heat permeability is obtained separately by solving a corresponding inverse problem and using practical measurements. We also determine both transmission coefficient

*This work was supported by the Slovak Research and Development Agency APVV-15-0681 and VEGA 1/0565/16.

†SvF STU and FMFI UK Bratislava, Radlinského 11, 810 05, Slovakia (Jozef.Kacur@fmph.uniba.sk).

‡FMFI UK Bratislava, Department of Mathematics, Mlynská dolina, 842 48, Slovakia (pmihala@gmail.com).

§FMFI UK Bratislava, Department of Mathematics, Mlynská dolina, 842 48, Slovakia (m.toth82@gmail.com).

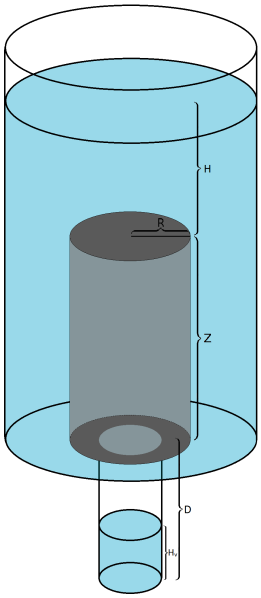


FIG. 1.1. *Sample*

and heat permeability in matrix via the solution of the inverse problem.

Recently, we have discussed in [6] determination of soil parameters in porous media flow model, based on empirical van Genuchten/Mualem capillary/pressure model. There, we have used radially symmetric 3D sample using inflow/outflow measurements. The main reason was that 1D samples (in form of thin tubes) used before suffer from preferential stream lines arising in experiments, especially using centrifugation. We have significantly eliminated this effect by suitable infiltration scenario with cylindrical sample, where infiltration flux from sample mantle is orthogonal to gravitational force. Moreover, the infiltration area is substantially larger then the area of the top of the tube. Thus, we obtained more reliable results in determination of soil parameters. In Fig. 1.1 we sketch the cylinder sample used in experiments.

In this manuscript we present the experiment scenarios to determine transmission, heat conduction and heat boundary transmission coefficients. The heat energy exchange is modeled by temperature gap, water saturation and transmission coefficient. It is almost impossible to measure the temperature gap between the water in pores and in matrix inside the porous media, but we can measure the consequences of heat energy exchange. To determine the heat transmission coefficient and heat conduction coefficient of the matrix we suggest the following experiment scenario. The cylindrical sample is initially uniformly low saturated (almost dry). The temperature of water and matrix is the same, e.g. 20C. Then we let to infiltrate water (through the cylinder mantle) with lower temperature, e.g., 5C. The top and bottom boundary of the cylinder are isolated. We measure the time evolution of temperature in the middle of the top of the cylinder. We note that temperatures of water and matrix in this point (even on whole axis of the cylinder) are the same for long time interval in experiment. The reason is that the infiltrating water has a very sharp front and slowly progresses towards the cylinder axis. Simultaneously the heat is conducted by the matrix and due to the heat exchange the temperature of water and the matrix

are almost the same in the neighborhood of the axis and decreases. This observation is supported by our numerical experiments. Thus, the time evolution of temperature in the top point of axis is the main information in determination of transmission coefficient and, moreover, also heat conduction coefficient of the matrix.

To determine the boundary heat transmission coefficient we consider saturated sample with constant temperature field in porous media. The sample boundary is flow isolated (except of the top and bottom). The external temperature is constant and different from the initial temperature of the sample. We measure the time evolution temperature of the cumulated outflow water. The initial sample temperature and the temperature of infiltrated water from the top is the same. Water infiltrates from the cylindrical chambre with constant water level. In this scenario we have simple flow model and outflow boundary condition. This experimental scenario is used also in analysis of temperature isolation propertis of material applied on mantel surface in thin film form.

In our model setting we do not assume the temperature influence on the water flow, but it could be included. However, the heat transport and its mutual transmission with the porous matrix strongly depend on the water saturation in pores.

In the heat and mass transfer problem in facades we consider 2D problem which represents a cross-section of the facade, or cylindrical sample. The parallel vertical boundaries of the rectangle represent the building and outdoor environment contacts. In the case of cylindrical sample the left vertical boundary corresponds to its axis.

In the numerical method we use operator splitting method where we successively along small time interval separately solve water flow, then heat transport in water and then in matrix including heat exchange. In the solution of water flow we follow the approximation strategy introduced in [5] and also used in well known software Hydrus (see [3]). To control the correctness of our numerical results we have developed also an approximation scheme (see [8] used only for 1D) based on the reduction of the governing parabolic equations to a stiff system of ordinary differential equations. This approximation solves simultaneously whole system, but computational time is significantly larger. The main reason is that the system is stiff and too large when using necessary space discretization. Comparisons justify our method which is significantly quicker and therefore applicable in the solution of inverse problems in mathematical model scaling. Moreover, present method could be efficiently used also for solving 3D problems.

2. Mathematical model.

2.1. Water flow model. Water saturation $\theta \in (\theta_r, \theta_s)$ (θ_r is irreducible saturation and θ_s is porosity) we rescale to effective saturation

$$\theta_{ef}(h) = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r}.$$

Here, h ([cm]) is head and the fundamental empirical relation between saturation θ and h in terms of van Genuchten/Mualem empirical model (capillary/pressure law) is

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{(1 + (\alpha h)^n)^m}, \ \theta_{ef}(h) = 1 \quad \text{for } h \geq 0, \tag{2.1}$$

where $\alpha < 0, n > 1$ and $m (m = 1 - 1/n)$ are soil parameters. Hydraulic permeability K is modeled by

$$K(h) = K_s k(\theta_{ef}), \ \ k(\theta_{ef}) = \theta_{ef}^{\frac{1}{2}}(1 - (1 - \theta_{ef}^{\frac{1}{m}})^m)^2, \tag{2.2}$$

where K_s is hydraulic permeability for fully saturated porous media. Water flux $\vec{q}$

$$\vec{q} = (q^x, q^y), \quad \vec{q}(h) = -K(h)(\nabla h - e_y),$$

where e_y is a unit vector in direction y representing gravitational driving force. Richards equation is of the form

$$\partial_t \theta(h) - div(K(h)(\nabla h - e_y)) = 0 \tag{2.3}$$

with the corresponding boundary conditions which will be specified in numerical experiments.

2.2. Heat energy in the water. Conservation of water heat energy is expressed in PDE

$$c_v \partial_t(\theta T_w) - div(-c_v \vec{q} T_w + (D_o \theta + \bar{D})\nabla T_w) = \sigma\theta(T_w - T_m) \tag{2.4}$$

where T_w is temperature of water, c_v is heat capacity of unite water volume, σ is transmission coefficient at the heat exchange with the matrix. Convective part is $c_v \vec{q} T_w$ and the diffusion/dispersion are characterized by molecular diffusion coefficient D_o and dispersion matrix $\bar{D}$, where

$$\bar{D} = \begin{pmatrix} D_{1,1} & D_{1,2} \\ D_{2,1} & D_{2,2} \end{pmatrix} = \begin{pmatrix} \alpha_L((q^x)^2 + \alpha_T((q^y)^2 & (\alpha_L - \alpha_T)(q^x q^y) \\ (\alpha_L - \alpha_T)(q^x q^y) & \alpha_L((q^y)^2 + \alpha_T((q^x)^2 \end{pmatrix} \frac{1}{|\vec{q}|}$$

Here, α_L, α_T are longitudinal and transversal dispersion coefficients. The corresponding initial and boundary conditions will be specified in the numnerical experiments.

2.3. Heat conduction in the matrix. We assume the simple heat conduction model in the matrix

$$c_m \partial_t T_m - \lambda \Delta T_m = \sigma\theta(T_w - T_m). \tag{2.5}$$

where T_m - matrix temperature, λ - heat conduction coefficient and c_m- heat capacity of the matrix.

For simplicity, we assume that on the boundary there are presribed fluxes or values of the unknown h, T_w, T_m and a combination of them.

2.4. Boundary conditions. Our solution domain Ω is a rectangle $(x, y) \in (0, X) \times (0, Y)$ and $t \in (0, \Upsilon)$. We consider the following boundary and initial conditions in our numerical solution drawn in Fig. 2.1

$$\begin{aligned} &\partial_y T_m = 0, \quad QT^y = 0, \quad h = h_0 && \text{on } (0, X) \times \{0\} \times (0, \Upsilon) \\ &\partial_y T_m = 0, \quad QT^y = 0, \quad q^y = 0 && \text{on } (0, X) \times \{Y\} \times (0, \Upsilon) \\ &T_m = 20, \quad T_w = 20, \quad h = -200 && \text{on } (0, X) \times (0, Y) \times \{0\}. \end{aligned}$$

The boundary conditions on $\{X\} \times (0, Y) \times (0, \Upsilon)$ are in the form

$$\partial_x T_m = \sigma_{m,r}(TM_r - T_m), \; QT^x = \sigma_{w,r}(TW_r - T_m), \; q^x = \sigma_{ww,r}(HW_r - h)$$

and on $\{0\} \times (0, Y) \times (0, \Upsilon)$

$$-\partial_x T_m = \sigma_{m,l}(TM_l - T_m), \; -QT^x = \sigma_{w,l}(TW_l - T_w), \; -q^x = \sigma_{ww,l}(HW_l - h),$$

where TM, TW, TH are external temperature and pressure sources, and $\sigma_{.,r}$, $\sigma_{.,l}$ are corresponding boundary transmission coefficients. The boundary flux conditions could be changed to Dirichlet boundary conditions.

2.5. Mathematical model in cylindrical coordinates. Consider the cylinder with radius R and height Y. Using cylindrical coordinates (r, y) in (2.3), (2.4) , (2.5) we obtain

$$\partial_t \theta(h) = \frac{1}{r}\partial_r(rK(h)\partial_r h) + \partial_y(K(h)(\partial_y h - 1)) \tag{2.6}$$

for water flow

$$c_v \partial_t(\theta T_w) - \left(\frac{1}{r}\partial_r(rQT^r) + \partial_y(QT^y)\right) = \sigma\theta(T_w - T_m) \tag{2.7}$$

for heat transport in water and

$$c_m \partial_t T_m - \lambda\left(\frac{1}{r}\partial_r(rQT_m^r) + \partial_y(\partial_y T_m)\right) = \sigma\theta(T_w - T_m).$$

for heat conduction in the matrix, where

$$\mathbf{q} = -(q^r, q^y)^T,\ q^r = K(h)\partial_r h,\ q^y = K(h)(\partial_z h - 1), \tag{2.8}$$

$$QT^r = -q^r T_w + \theta(D_{1,1}\partial_r T_w + D_{1,2}\partial_y T_w + D_o\theta, \tag{2.9}$$

$$QT^y = -q^y T_w + \theta(D_{2,1}\partial_r T_w + D_{2,2}\partial_y T_w + D_o\theta,\ QT_m^r = \partial_r T_m. \tag{2.10}$$

2.6. Model data and corresponding numerical solution for cylindrical sample. Our solution domain $R = X = 10, Y = 10$. In our numerical experiments we assume the following model data ([CGS] units) as "standard data" : $\theta_0 = 0.38$, $\theta_r = 0$, $K_s = 2.4\ 10^{-4}$, $\alpha = 0.0189$, $n = 2.81$, $D_o = 0.03$, $\lambda = 0.3$, $\alpha_L = 1$, $\alpha_T = \frac{1}{10}$, $c_v = c_m = 1$ and $\sigma = 1$. These data correspond to a limestone.
We consider the boundaries $(0, 10) \times 0$ and $(0, 10) \times 10$ with zero heat and flow fluxes (isolation). On the boundary $10 \times (0, 10)$ the hydrostatic pressure $h = (Y - y), y \in (0, Y)$ is prescribed and $T_w = 0$. On the axis $10 \times (0, 10)$ we have $q^r = 0,\ QT^r = 0$. We consider heat isolation for matrix boundaries.
The initial conditions are $h = -200$ and $T_w = T_m = 20$. In the Fig. 2.1 we draw the corresponding flow and temperature fields for the cylinder cross-section at the time moment $t = 60s$.

3. Numerical method. In our approximation scheme we apply a flexible time stepping and a finite volume method in space variables. We consider uniform partition of the domain with $(N_x, N_y) = (31, 31)$ grid points $(x_i, y_j) = (i\Delta x, j\Delta y),\ i, j = 0, 1, ..., 30,\ \Delta x = \frac{X}{N_x - 1}, \Delta y = \frac{Y}{N_y - 1}$. The time derivative we approximate by backwards difference and then we integrate our system over the angular control volume $V_{i,j}$ with the corners $x_{i\pm\frac{1}{2}}, y_{j\pm\frac{1}{2}}$ and with the length $(\Delta x, \Delta y)$ of the edges. Then, our approximation linked with the inner grid point (x_i, y_j) at the time $t = t_k$ is

$$\Delta x\Delta y\frac{\theta(h) - \theta(h^{k-1})}{\tau} - \Delta y\left[\frac{K(h_{i+1}) + K(h)}{2}\left(\frac{h_{i+1} - h}{\Delta x}\right)\right]$$
$$+\Delta y\left[\frac{K(h) + K(h_{i-1})}{2}\left(\frac{h - h_{i-1}}{\Delta x}\right)\right] - \Delta x\left[\frac{K(h_{j+1}) + K(h)}{2}\left(\frac{h_{j+1} - h}{\Delta y} - 1\right)\right]$$
$$+\Delta x\left[\frac{K(h) + K(h_{j-1})}{2}\left(\frac{h - h_{j-1}}{\Delta y} - 1\right)\right] = 0.$$

Omitted indices are of values $\{i, j, k\}$.

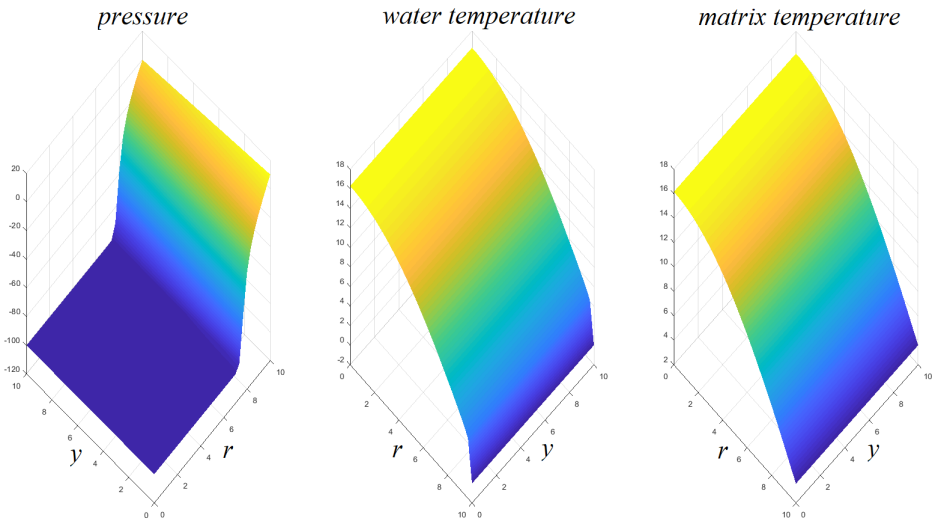


FIG. 2.1. *Water pressure h and temperatures T_w, T_m in cylinder at $t = 60s$*

3.1. Quasi-Newton linearization. In each (x_i, y_j) we linearize θ in terms of h iteratively (with iteration parameter l) following [Cellia at all][5] in the following way

$$\frac{\theta(h^{k,l+1}) - \theta(h^{k-1})}{\tau} = C^{k,l}\frac{h^{k,l+1} - h^{k,l}}{\tau} + \frac{\theta^{k,l} - \theta^{k-1}}{\tau} \tag{3.1}$$

where

$$C^{k,l} = \frac{\partial\theta^{k,l}}{\partial h^{k,l}} = (\theta_s - \theta_r)(1-n)\alpha(\alpha h^{k,l})^{n-1}(1 + (\alpha h^{k,l})^n)^{-(m+1)}$$

for $h^{k,l} < 0$, else $C^{k,l} = 0$. We stop iterations for $l = l^*$, when

$$|h^{k,l^*} - h^{k,l^*-1}| \leq Tollerance, \text{ then we put } h^k := h^{k,l^*}.$$

Finally. we replace the nonlinear term $K(h^k)$ by $K(h^{k,l})$. Our approximation scheme then becomes linear in terms of $h^{k,l+1}$. Generally, we speed up the iteration by a special construction of starting point $h^{k,0} \approx h^{k-1}$ and eventually using suitable damping parameter in solving corresponding linearized system.
The complex system is solved by **operator splitting method**. To obtain an approximate solution for temperatures in water and matrix at the time section $t = t_k$, starting from $t = t_{k-1}$ we use flow characteristics obtained at $t = t_k$ for $\theta^k, h^k, \vec{q}^k$, and D^k.

3.2. Approximation scheme for water temperature. For $T_w(\equiv T), T_m$ at (x_i, y_j) for $t = t_k$ we obtain by finite volume

$$
\begin{aligned}
&c_v\theta\frac{T-T^{k-1}}{\tau}\Delta x\Delta y\\
&-\Delta y\left[-c_v q^x_{i+\frac12}\frac{T_{i+1}+T_i}{2}+D_{1,1,i+\frac12}\frac{T_{i+1}-T_i}{\Delta x}+D_{1,2,i+\frac12}\frac{T_{i+1,j+1}+T_{i,j+1}-T_{i+1,j-1}-T_{i,j-1}}{4\Delta y}\right]\\
&+\Delta y\left[-c_v q^x_{i-\frac12}\frac{T_i+T_{i-1}}{2}+D_{1,1,i-\frac12}\frac{T_i-T_{i-1}}{\Delta x}+D_{1,2,i-\frac12}\frac{T_{i,j+1}+T_{i-1,j+1}-T_{i,j-1}-T_{i-1,j-1}}{4\Delta y}\right]\\
&-\Delta x\left[-c_v q^y_{j+\frac12}\frac{T_{j+1}+T_j}{2}+D_{2,2,j+\frac12}\frac{T_{j+1}-T_j}{\Delta y}+D_{2,1,j+\frac12}\frac{T_{i+1,j+1}+T_{i+1,j}-T_{i-1,j+1}-T_{i-1,j}}{4\Delta x}\right]\\
&+\Delta x\left[-c_v q^y_{j-\frac12}\frac{T_j+T_{j-1}}{2}+D_{2,2,j-\frac12}\frac{T_j-T_{j-1}}{\Delta y}+D_{2,1,j-\frac12}\frac{T_{i+1,j}+T_{i+1,j-1}-T_{i-1,j}-T_{i-1,j-1}}{4\Delta x}\right]\\
&=\Delta x\Delta y\sigma\theta^k(T_m-T_w).
\end{aligned}
\tag{3.2}
$$

3.3. Approximation of $\vec{q}$ and D in middle points. We approximate $\vec{q}$ and D in middle points by

$$
q^x_{i\pm\frac12}=-\frac{K(h_{i\pm1})+K(h_i)}{2}\left(\frac{\pm h_{i\pm1}\mp h_i}{\Delta x}\right),\ q^y_{j\pm\frac12}=-\frac{K(h_{j\pm1})+K(h_j)}{2}\left(\frac{\pm h_{j\pm1}\mp h_j}{\Delta y}-1\right)
$$

$$
\begin{aligned}
q^y_{i\pm\frac12}&=-\frac{K(h_{i\pm1})+K(h_i)}{2}\left(\frac{h_{i\pm1,j+1}+h_{i,j+1}-h_{i\pm1,j-1}-h_{i,j-1}}{4\Delta y}-1\right)\\
q^x_{j\pm\frac12}&=-\frac{K(h_{j\pm1})+K(h_i)}{2}\left(\frac{h_{i+1,j\pm1}+h_{i+1,j}-h_{i-1,j\pm1}-h_{i-1,j}}{4\Delta x}\right),\\
D_{1,1,i\pm\frac12}&=\left(\alpha_L(q^x_{i\pm\frac12})^2+\alpha_T(q^y_{i\pm\frac12})^2\right)\frac{1}{\sqrt{(q^x_{i\pm\frac12})^2+(q^y_{i\pm\frac12})^2}}+\lambda_t\theta_{i\pm\frac12}\\
D_{1,2,i\pm\frac12}&=(\alpha_L-\alpha_T)q^x_{i\pm\frac12}q^y_{i\pm\frac12}\frac{1}{\sqrt{(q^x_{i\pm\frac12})^2+(q^y_{i\pm\frac12})^2}}
\end{aligned}
$$

and analogously

$$
D_{2,2,j\pm\frac12}=D_{1,1,i\pm\frac12}\ (i\leftrightarrow j;\alpha_L\leftrightarrow\alpha_T);\ D_{2,1,j\pm\frac12}=D_{1,2,i\pm\frac12}\ (i\leftrightarrow j).
$$

3.4. Approximation scheme for matrix temperature. The governing PDE with BC and IC

$$
c_m\partial_t T_m-\lambda\Delta T_m=\sigma\theta(T_w-T_m)\tag{3.3}
$$

$T_m=T^k_m$ is approximated by FVM ((x_i, y_j), $t = t_k$) to

$$
\begin{aligned}
&c_m\frac{T_m-T^{k-1}_m}{\tau}\Delta x\Delta y-\Delta y\lambda\left[\frac{T_{m\ i+1}-T_{m\ i}}{\Delta x}-\frac{T_{m\ i}-T_{m\ i-1}}{\Delta x}\right]\\
&-\Delta x\lambda\left[\frac{T_{m\ j+1}-T_{m\ j}}{\Delta y}-\frac{T_{m\ j}-T_{m\ j-1}}{\Delta y}\right]=\Delta x\Delta y\sigma\theta^k(T_m-T_w).
\end{aligned}
\tag{3.4}
$$

3.5. Solution of the inverse problem to determine σ, λ. Measuring the temperature of water and matrix in the sample is a difficult task. So we propose the infiltration scenario which enables us to measure σ. There is uniformly distributed small amount of water $h = -100$ in the sample and the initial temperature of water and matrix are the same $T_w = T_m = 20C$. The top and bottom of the sample are isolated (zero water and temperature fluxes). The top boundary is free and from the bottom we let water to infiltrate. Water infiltrates by hydraulic pressure from the mantel. The temperature of infiltrating water is $0C$ We are measuring the water temperature on the top boundary. The model data are the same as data in Fig. 2.1. The time evolution of the computed temperature on the top is presented in the Fig. 3.1 (blue line).

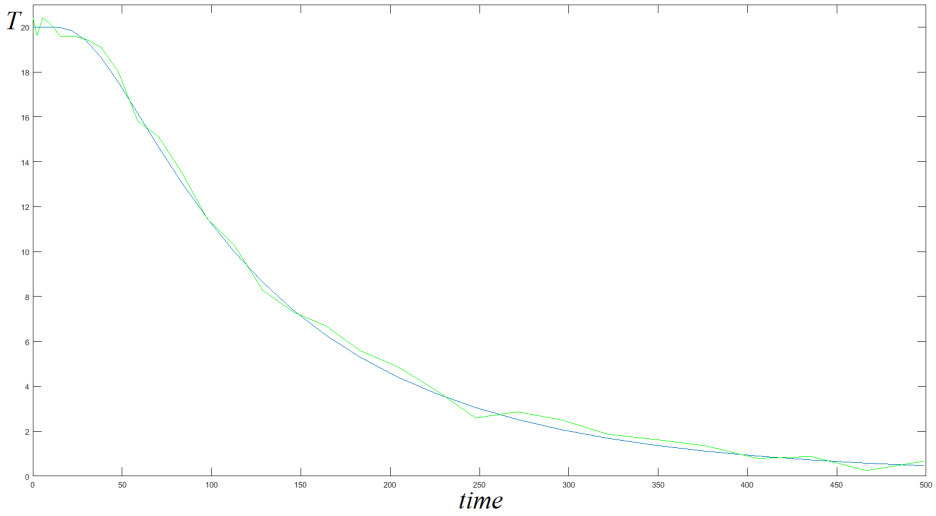


FIG. 3.1. *Time evolution of temperature in top of the axis*

The computed data have been perturbed by the $0.5C$ of noise using the random function (green line). These were considered to be measured data. Then, we forgot our transmission coefficient σ $(= 1)$, and matrix heat conduction coefficient λ $(= 0.3)$ and we used an iteration procedure to minimize the discrepancy between the measured data and the computed data.
The optimal point σ_{opt}, λ_{opt} (with respect to the given tolerance) is taken for the required transmission and conduction coefficients. We verify its stability with respect to the choice of starting points in iteration procedure. All measured data correspond to different random noises. During the measured time interval $t \in (0, 500)$ we have used only 31 time moments.

The obtained results with different starting points are collected in Table 3.1. Used starting points are combinations of $\sigma \in \{0.5, 1.5\}$ and $\lambda \in \{0.1, 0.5\}$. The final value does not depend on the starting point. However, it can be noticed that values of parameters σ, λ slightly mutually interfer. There is always one that is lower than the exact value while the other is changed in a opposite way.

3.6. Solution of the inverse problem to determine $\sigma_{m,r}$, $\sigma_{w,r}$. In the fact, we shall assume that $\sigma_{m,r} = \sigma_{w,r} := \sigma_B$ and in this case we obtain the stable results. When they differ, the determination procedure is unstable and one can note the substitution effects between them. In the determination procedure we proceed as in the

TABLE 3.1
Optimal values of λ, σ

start	σ, λ	σ, λ
$[0.5, 0.1]$	$[1.0475, 0.2974]$	$[0.9842, 0.3082]$
$[1.5, 0.1]$	$[0.9547, 0.3277]$	$[1.0212, 0.2945]$
$[0.5, 0.5]$	$[1.0324, 0.2821]$	$[0.9771, 0.3114]$
$[1.5, 0.5]$	$[0.9685, 0.2854]$	$[1.0389, 0.2901]$

previous section determining (σ, λ). In our experiment we use the "standard" model data with the following changes. The sample is fully saturated and the temperature of the sample is $20C$. The mantel of the cylinder is water flow isolated. The temperature outside the cylinder matle is $15C$. The temperature field in the sample is drawn in the Fig. 3.2 at the time moment $t = 1500s$. Water is infiltrated into sample throught the smaller circle on top ($R_{in} = 5$) with pressure 5 and temperature 20. Water flows out from the sample bottom. Temperature time evolution of the cumulated outflow water is drawn in Fig. 3.3.

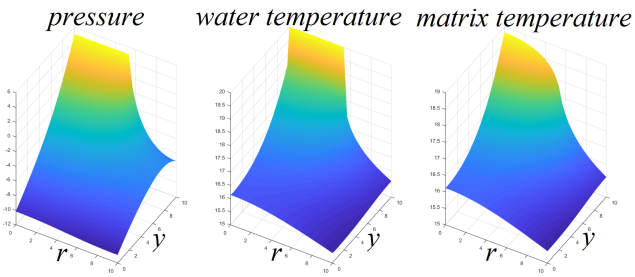


FIG. 3.2. *Pressure and temperatures at* $t = 1500$

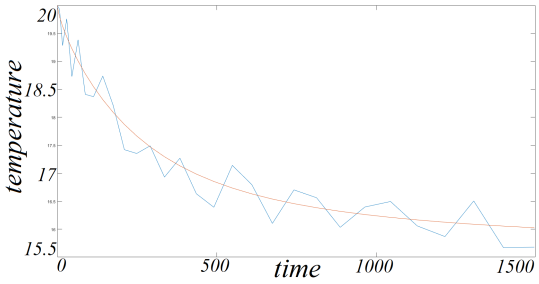


FIG. 3.3. *Temperature of cumulated outflow water*

To determine σ_B $(= 1)$ we are using measurements from the blue line in Fig. 3.3 which correspond to $0.5C$ of noise using the random function. As a starting point for σ_B we use $\{0.5, 11.5\}$ with two different noise applications. The obtained optimal values are collected in Table 3.2. We present there also experiments corresponding to 0.1 noise.

Originally, we have solved our system using its approximation by a corresponding stiff system of ODE, where space discretizarion is realized by finite volume method. In this case we do not use operator splitting method. The solution coincides with that one obtained by present method up to the first two digits when ruther dense grid points are used. But the present method is signifficantly faster and more suitable in

TABLE 3.2
Optimal values of α_L

start	σ_B (0.5)	σ_B (0.5) —	σ_B (0.1)	σ_B (0.1)
0.5	0.90322	1.13708	1.03359	0.98857
1.5	1.18073	1.11269	0.96027	1.01879

solving inverse problems. In 1D we have compared solutions with that ones in [8], [3] and [4].

4. Summary.

- Numerical modeling of heat exchange arising in water infiltration in unsaturated porous media is discussed.
- Efficient numerical method is developed on the base of time stepping, operator splitting and FVM.
- An infiltration scenario is proposed to determine the heat transmission coefficient inside the porous media by solution of inverse problem.
- Also a scenario is proposed to determine the boundary heat transmission coefficient.
- The efficiency of the numerical method is demonstrated by numerical experiments.

REFERENCES

[1] J. BEAR, A. H.-D. CHENG, *Modeling Groundwater Flow and Contaminant Transport*, Springer 2010, V.23 .

[2] D. CONSTALES, J. KAČUR, *Determination of soil parameters via the solution of inverse problems in infiltration*, Computational Geosciences 5 (2004), pp. 25–46.

[3] J. ŠIMUNEK, M. ŠEJNA, H. SAITO, M. SAKAI, M. TH. VAN GENUCHTEN, *The Hydrus-1D Software Package for Simulating the Movement of Water*, Heat, and Multiple Solutes in Variably Saturated Media, (2013).

[4] J. ŠIMUNEK, J. R. NIMO, *Estimating soil hydraulic parameters from transient flow experiments in a centrifuge using parameter optimization technique*, Water Resour. Res. 41, W04015 (2005).

[5] M. A. CELIA, Z. BOULOUTAS, *A general mass-conservative numerical solution for the unsaturated flow equation*, Water Resour. Res. 26 (1990), pp. 1483–1496.

[6] J. KAČUR, P. MIHALA, M. TÓTH, *Determination of soil parameters under gravitation and centrifugal forces in 3D infiltration*, Vseas transactions on heat and mass tranfer, Vol. 11, (2016), pp. 115–120.

[7] D. CONSTALES, J. KAČUR, B. MALENGIER, *A precise numerical scheme for contaminant transport in dual-well flow*, Water Resources Research, vol. 39(10), (2003), pp. 1292–1303,

[8] J. KAČUR , J. MINÁR, *A benchmark solution for infiltration and adsorption of polluted water into unsaturated-saturated porous media*, Transport in porous media, vol. 97, (2013), pp. 223–239.

[9] T. L. BERGMAN, A. S. LAVINE, F. P. INCROPERA, D. P. DEWITT, *Fundamentals of heat and mass transfer*, John Wiley and Sons, 7th edition, (2011), ISBN 13 978-0470-50197-9.

Proceedings of EQUADIFF 2017
pp. 201–210

A WELL-POSEDNESS RESULT FOR A STOCHASTIC MASS CONSERVED ALLEN-CAHN EQUATION WITH NONLINEAR DIFFUSION.*

PERLA EL KETTANI†, DANIELLE HILHORST‡, AND KAI LEE§

Abstract. In this paper, we prove the existence and uniqueness of the solution of the initial boundary value problem for a stochastic mass conserved Allen-Cahn equation with nonlinear diffusion together with a homogeneous Neumann boundary condition in an open bounded domain of $\mathbb{R}^n$ with a smooth boundary. We suppose that the additive noise is induced by a Q-Brownian motion.

Key words. stochastic nonlocal reaction-diffusion equation, monotonicity method, conservation of mass.

AMS subject classifications. 35K55, 35K57, 60H15, 60H30

1. Introduction.

In this paper, we study the problem

$$(P)\begin{cases} \dfrac{\partial \varphi}{\partial t} = \operatorname{div}(A(\nabla\varphi)) + f(\varphi) - \dfrac{1}{|D|}\displaystyle\int_D f(\varphi)dx + \dfrac{\partial W}{\partial t}, & x \in D, \quad t \geq 0 \\ A(\nabla\varphi).\nu = 0, & x \in \partial D, t \geq 0 \\ \varphi(0,x) = \varphi_0(x), & x \in D \end{cases}$$

where:

- D is an open bounded set of $\mathbb{R}^n$ with a smooth boundary ∂D;
- ν is the outer normal vector to ∂D;
- The function f is given by $f(s) = s - s^3$;
- We assume that $A = \nabla_v \Psi(v) : \mathbb{R}^n \to \mathbb{R}^n$ for some strictly convex function $\Psi \in C^{1,1}(\mathbb{R}^n)$ (i.e. $\Psi \in C^1(\mathbb{R}^n)$ and $\nabla\Psi$ is Lipschitz continuous) which satisfies

$$\begin{cases} A(0) = \nabla\Psi(0) = 0, \Psi(0) = 0 \\ \|D^2\Psi\|_{L^\infty(\mathbb{R}^n;\mathbb{R}^{n\times n})} \leq c_1, \end{cases} \tag{1.1}$$

for some constant $c_1 > 0$. We remark that (1.1) implies that

$$|A(a) - A(b)| \leq C|a - b| \tag{1.2}$$

for all $a, b \in \mathbb{R}^n$,where C is a positive constant, and that the strict convexity of Ψ implies that A is strictly monotone, namely there exists a positive constant C_0 such that

$$(A(a) - A(b))(a - b) \geq C_0|a - b|^2, \tag{1.3}$$

*This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH.

†University of Paris-Sud.
‡CNRS and University of Paris-Sud.
§University of Tokyo, Japan.

for all $a, b \in \mathbb{R}^n$.

We remark that if A is the identity matrix, the nonlinear diffusion operator $-div(A(\nabla u))$ reduces to the linear operator $-\Delta u$.

- The function $W = W(x,t)$ is a Q-Brownian motion. More precisely, let Q be a nonnegative definite symmetric operator on $L^2(D)$, $\{e_l\}_{l\geq 1}$ be an orthonormal basis in $L^2(D)$ diagonalizing Q, and $\{\lambda_l\}_{l\geq 1}$ be the corresponding eigenvalues, so that

$$Qe_l = \lambda_l e_l, \quad \text{for all } l \geq 1.$$

We suppose that Q satisfies

$$\operatorname{Tr} Q = \sum_{l=1}^{\infty} \langle Qe_l, e_l \rangle_{L^2(D)} = \sum_{l=1}^{\infty} \lambda_l \leq \Lambda_0. \tag{1.4}$$

for some positive constant Λ_0. We suppose furthermore that $e_l \in H^1(D) \cap L^\infty(D)$ for $l = 1, 2...$ and that there exist positive constants Λ_1 and Λ_2 such that

$$\sum_{l=1}^{\infty} \lambda_l \|e_l\|^2_{L^\infty(D)} \leq \Lambda_1, \tag{1.5}$$

$$\sum_{l=1}^{\infty} \lambda_l \|\nabla e_l\|^2_{L^2(D)} \leq \Lambda_2. \tag{1.6}$$

- Next we define the following spaces:

$$H = \left\{ v \in L^2(D), \int_D v = 0 \right\}, \quad V = H^1(D) \cap H \quad \text{and} \quad Z = V \cap L^4(D)$$

where $\|\cdot\|$ corresponding to the space H.
We also define $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{Z^*, Z}$ as the duality product between Z and its dual space $Z^* = V^* + L^{\frac{4}{3}}(D)$ ([3], p.175).

The corresponding deterministic equation with linear diffusion has been introduced by Rubinstein and Sternberg [10] as a model for phase separation in a binary mixture. The well-posedness and the stabilization of the solution for large times for the corresponding Neumann problem were proved by Boussaïd, Hilhorst and Nguyen [4]. They assumed that the initial function was bounded in $L^\infty(D)$ and proved the existence of the solution in an invariant set using the Galerkin method together with a compactness method.

A singular limit of a rescaled version of Problem (P) with linear diffusion has been studied by Antonopoulou, Bates, Blömker and Karali [1] to model the motion of a droplet. However, they left open the problem of proving the existence and uniqueness of the solution, which we address here. The proof of the existence of the solution of Problem (P) is based on a Galerkin method together with a monotonicity argument similar to that used in [9] for a deterministic reaction-diffusion equation, and that in [8] for a stochastic problem. We refer to the forthcoming article [6] for more details

and for the proofs.

The organization of this paper is as follows. In section 2, we present regularity properties of the solution W_A of the nonlinear stochastic heat equation with a homogeneous Neumann boundary condition and initial condition zero. In section 3, we prove the existence of a solution of Problem (P). To that purpose we take the function $\varphi - W_A$ as the new unknown function. Finally, we prove the uniqueness of the solution in section 4.

2. An auxiliary problem. We consider the Neumann boundary value problem for the stochastic nonlinear heat equation

$$(P_1)\quad \begin{cases} \dfrac{\partial W_A}{\partial t} = \operatorname{div}(A(\nabla W_A)) + \dfrac{\partial W}{\partial t}, & x \in D,\ t \geq 0 \\ A(\nabla W_A).\nu = 0, & x \in \partial D,\ t \geq 0 \\ W_A(0,x) = 0, & x \in D. \end{cases}$$

First we define a solution of Problem (P_1).

DEFINITION 2.1. *We say that W_A is a solution of Problem (P_1) if :*

1. $W_A \in L^\infty(0,T;L^2(\Omega \times D)) \cap L^2(\Omega \times (0,T);H^1(D))$;
2. $div(A(\nabla W_A)) \in L^2(\Omega \times (0,T);(H^1(D))')$;
3. *W_A satisfies almost surely the problem*

$$\begin{cases} W_A(t) = \displaystyle\int_0^t \operatorname{div}(A(\nabla W_A(s)))ds + W(t), \text{ in the sense of distributions,} \\ A(\nabla W_A).\nu = 0, \quad \text{in the sense of distributions on } \partial D \times \mathbb{R}^+. \end{cases} \tag{2.1}$$

Using ideas from Krylov & Rosovskii [8] we prove in [6] that this problem possesses a unique solution W_A. We are interested in further regularity properties of the solution W_A. A first step is to apply a result of Gess [7] who proves the existence and uniqueness of a solution in the sense of $L^2(D)$, namely almost everywhere in D. More precisely, he defines a strong solution as follows (cf. [7], Definition 1.3).

DEFINITION 2.2. **(Strong solution)** *We say that W_A is a strong solution of Problem (P_1) if :*

1. *W_A is a solution in the sense of Krylov and Rosovskii;*
2. $W_A \in L^2(\Omega;C([0,T];L^2(D)))$;
3. $\operatorname{div}(A(\nabla W_A)) \in L^2(\Omega \times (0,T);L^2(D))$;
4. *W_A satisfies a.s. for all $t \in (0,T)$ the problem*

$$\begin{cases} W_A(t) = \displaystyle\int_0^t \operatorname{div}(A(\nabla W_A(s)))ds + W(t), \text{ in } L^2(D), \\ A(\nabla W_A(t)).\nu = 0, \quad \text{in a suitable sense of trace on } \partial D. \end{cases} \tag{2.2}$$

We will show in [6] the boundedness of W_A in $L^\infty(0,T;L^q(\Omega \times D))$ for all $q \geq 2$. The proof of this result is based on an article by Bauzet, Vallet, Wittbold [2] where a similar result is proved for a convection-diffusion equation with a multiplicative noise involving a standard adapted one-dimensional Brownian motion.

THEOREM 2.3. *Let W_A be a solution of Problem (P_1); then $W_A \in L^\infty(0,T;L^q(\Omega\times D))$, for all $q\geq 2$.*

3. Existence and uniqueness of the solution of Problem (P). To begin with, we perform the change of functions $u(t) := \varphi(t) - W_A(t)$; then φ is a solution of (P) if and only if u satisfies:

$$(P_2)\quad \begin{cases} \dfrac{\partial u}{\partial t} = \operatorname{div}[A(\nabla(u+W_A)) - A(\nabla W_A)] + f(u+W_A) \\ \qquad\qquad -\dfrac{1}{|D|}\displaystyle\int_D f(u+W_A)dx, & x\in D,\quad t\geq 0,\\ A(\nabla(u+W_A)).\nu = 0, & x\in\partial D, t\geq 0,\\ u(0,x)=\varphi_0(x), & x\in D. \end{cases}$$

We remark that Problem (P_2) has the form of a deterministic problem; however it is stochastic since the random function W_A appears in the parabolic equation for u.

DEFINITION 3.1. *We say that u is a solution of Problem (P_2) if :*

1. *$u\in L^\infty(0,T;L^2(\Omega\times D))\cap L^2(\Omega\times(0,T);H^1(D))\cap L^4(\Omega\times(0,T)\times D)$; $div[A(\nabla(u+W_A))]\in L^2(\Omega\times(0,T);(H^1(D))')$;*
2. *u satisfies almost surely the problem: for all $t\in[0,T]$*

$$\begin{cases} u(t) = \varphi_0 + \displaystyle\int_0^t \operatorname{div}[A(\nabla(u+W_A)) - A(\nabla W_A)]\ ds + \int_0^t f(u+W_A)ds \\ -\displaystyle\int_0^t \frac{1}{|D|}\int_D f(u+W_A)dxds, \textit{ in the sense of distributions},\\ A(\nabla(u+W_A)).\nu = 0, \quad \textit{in the sense of distributions on } \partial D\times\mathbb{R}^+. \end{cases} \tag{3.1}$$

The conservation of mass property holds, namely

$$\int_D u(x,t)dx = \int_D \varphi_0(x)dx, \quad \text{a.s. for a.e. } t\in\mathbb{R}^+.$$

The main result of this section is the following.

THEOREM 3.2. *There exits a unique solution of Problem (P_2).*

Proof. In this section we apply the Galerkin method to prove the existence of solution of Problem (P_2). Denote by $0<\gamma_1<\gamma_2\leq ...\leq\gamma_{\tilde{k}}\leq ...$ the eigenvalues of the operator $-\Delta$ with homogeneous Neumann boundary conditions, and by $w_{\tilde{k}}, \tilde{k}=0,...$ the corresponding unit eigenfunctions in $L^2(D)$. Note that they are smooth functions. We remark that the functions $\{w_j\}$ are an orthonormal basis of $L^2(D)$ and satisfy

$$\int_D w_j w_0 = 0 \quad \text{for all } j\neq 0 \ \text{ and } \ w_0 = \frac{1}{\sqrt{|D|}}.$$

We look for an approximate solution of the form

$$u_m(x,t) - M = \sum_{i=1}^m u_{im}(t)w_i = \sum_{i=1}^m \langle u_m(t), w_i\rangle w_i,$$

where $M = \frac{1}{|D|}\int_D \varphi_0(x)dx$; the function u_m satisfies the equations

$$\int_D \frac{\partial}{\partial t}(u_m(x,t)-M)w_j dx = -\int_D [A(\nabla(u_m - M + W_A)) - A(\nabla W_A)]\nabla w_j dx + \int_D f(u_m + W_A)w_j dx - \frac{1}{|D|}\int_D \big(\int_D f(u_m + W_A)dx\big)w_j dx, \tag{3.2}$$

for all w_j, $j = 1, ..., m$. We remark that $u_m(x,0) = M + \sum_{i=1}^{m}(\varphi_0, w_i)w_i$ converges strongly to φ_0 in $L^2(D)$ as $m \to \infty$.

Problem (3.2) is an initial value problem for a system of m ordinary differential equations, so that it has a unique solution u_m on some interval $(0, T_m)$, $T_m > 0$; in fact the following a priori estimates show that this solution is global in time.

First we remark that the contribution of the nonlocal term vanishes. Indeed for all $j = 1, ..., m$

$$-\frac{1}{|D|}\int_D \big(\int_D f(u_m + W_A(t))dx\big)w_j dx = -\frac{1}{|D|}\big(\int_D f(u_m + W_A(t))dx\big) \times \int_D w_j dx = 0. \tag{3.3}$$

We substitute (3.3) into (3.2), we multiply (3.2) by $u_{jm} = u_{jm}(t)$, sum on $j = 1, ..., m$ and use property (1.3) to deduce that

$$\frac{1}{2}\frac{d}{dt}\int_D (u_m - M)^2 dx + C_0\int_D |\nabla(u_m - M)|^2 dx + C_1\int_D (u_m - M)^4 dx \le C_2\int_D |W_A(t)|^4 dx + \tilde{C}_2(M)|D|. \tag{3.4}$$

3.1. A priori estimates and passing to the limit. In what follows, we derive a priori estimates for the function u_m.

LEMMA 3.1. *There exists a positive constant $\mathcal{C}$ such that*

$$\sup_{t\in[0,T]} \mathbb{E}\int_D (u_m - M)^2 dx \le \mathcal{C}, \quad \mathbb{E}\int_0^T\int_D |\nabla(u_m - M)|^2 dxdt \le \mathcal{C}, \tag{3.5}$$

$$\mathbb{E}\int_0^T\int_D (u_m - M)^4 dxdt \le \mathcal{C}, \tag{3.6}$$

$$\mathbb{E}\int_0^T\int_D (f(u_m + W_A))^{\frac{4}{3}} dxdt \le \mathcal{C}, \tag{3.7}$$

$$\mathbb{E}\int_0^T \|\operatorname{div}(A(\nabla(u_m + W_A)))\|^2_{(H^1(D))'} dt \le \mathcal{C}. \tag{3.8}$$

Proof. Integrating (3.4) from 0 to t and taking the expectation we deduce for all $t \in [0,T]$

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\int_D (u_m - M)^2(t)dx + C_0\mathbb{E}\int_0^t\int_D |\nabla(u_m - M)|^2 dxds + C_1\mathbb{E}\int_0^t\int_D (u_m - M)^4 dxds \\
&\leq \frac{1}{2}\int_D (u_m(0) - M)^2 dx + C_2\mathbb{E}\int_0^t\int_D |W_A(t)|^4 dxds + \tilde{C}_2(M)|D|T \\
&\leq K,
\end{aligned}$$

where we have used Theorem 2.3 of section 2. Therefore u_m is bounded independently of m in $L^\infty(0,T,L^2(\Omega\times D)) \cap L^2(\Omega\times(0,T);H^1(D)) \cap L^4(\Omega\times(0,T)\times D))$.

Moreover we have that

$$\begin{aligned}
\mathbb{E}\|f(u_m + W_A)\|^{\frac{4}{3}}_{L^{\frac{4}{3}}((0,T)\times D)} &\leq c_2\mathbb{E}\int_0^T\int_D |u_m - M|^4 dxdt + c_2\mathbb{E}\int_0^T\int_D |W_A|^4 dxdt \\
&\quad + C_5|D|T \\
&\leq K_1,
\end{aligned}$$

by (3.6) and Theorem 2.3 in section 2.

Finally one can show that the elliptic term is bounded in the sense of distributions. □

Hence there exists a subsequence which we denote again by $\{u_m - M\}$ and a function $u - M \in L^2(\Omega\times(0,T);V) \cap L^4(\Omega\times(0,T)\times D) \cap L^\infty(0,T;L^2(\Omega\times D))$ such that

$$u_m - M \rightharpoonup u - M \quad \text{weakly in} \quad L^2(\Omega\times(0,T);V) \text{ and } L^4(\Omega\times(0,T)\times D) \tag{3.9}$$

$$u_m - M \rightharpoonup u - M \quad \text{weakly star in } L^\infty(0,T;L^2(\Omega\times D)) \tag{3.10}$$

$$f(u_m + W_A) \rightharpoonup \chi \quad \text{weakly in} \quad L^{\frac{4}{3}}(\Omega\times(0,T)\times D) \tag{3.11}$$

$$\operatorname{div}(A(\nabla(u_m + W_A))) \rightharpoonup \Phi \quad \text{weakly in} \quad L^2(\Omega\times(0,T);(H^1(D))') \tag{3.12}$$

as $m \to \infty$.

Next, we pass to the limit as $m \to \infty$. To that purpose, we integrate in time the equation (3.2), we recall that the nonlocal term vanishes in (3.2) and multiply the equation by the product $y\psi$, where $y(\omega)$ is any a.s. bounded random variable and $\psi(t)$ is a bounded function on (0,T); we finally integrate between 0 and T and take

the expectation, which yields for all $j = 1, .., m$

$$\begin{aligned}
&\mathbb{E}\int_0^T\int_D y\psi(t)(u_m(t)-M)w_j dxdt\\
&=\mathbb{E}\int_0^T\int_D y\psi(t)(u_m(0)-M)w_j dxdt\\
&+\mathbb{E}\int_0^T y\psi(t)\{\int_0^t \langle \mathrm{div}[A(\nabla(u_m-M+W_A))], w_j\rangle ds\}dt\\
&-\mathbb{E}\int_0^T y\psi(t)\{\int_0^t \langle \mathrm{div}[A(\nabla W_A)], w_j\rangle ds\}dt\\
&+\mathbb{E}\int_0^T y\psi(t)\{\int_0^t\int_D f(u_m+W_A)w_j dxds\}dt. \qquad (3.13)
\end{aligned}$$

Passing to the limit in (3.13) by using Lebesgue-dominated convergence theorem, we deduce that for a.e. $(t,\omega)\in(0,T)\times\Omega$ and for all $\tilde{w}\in V\cap L^4(D)$.

$$\langle u(t)-M,\tilde{w}\rangle = \langle \varphi_0 - M,\tilde{w}\rangle + \int_0^t \langle \Phi+\chi-\mathrm{div}(A(\nabla W_A)),\tilde{w}\rangle ds. \qquad (3.14)$$

LEMMA 3.2. *The function u is such that $u\in C([0,T];L^2(D))$ a.s.*

Proof. Apply Lemma 1.2 p. 260 in [11]. □

It remains to prove that $\langle \Phi+\chi,\tilde{w}\rangle = \langle \mathrm{div}(A(\nabla(u-M+W_A)))+f(u+W_A(t)),\tilde{w}\rangle$ for all $\tilde{w}\in V\cap L^4(D)$.

3.2. Monotonicity argument. Let w be such that $w-M\in L^2(\Omega\times(0,T);V)\cap L^4(\Omega\times D\times(0,T))$ and let c be a constant such that $c\geq 2$. We define

$$\begin{aligned}
\mathcal{O}_m = \mathbb{E}\big[\int_0^T e^{-cs}\{&2\langle \mathrm{div}\left(A(\nabla(u_m-M+W_A))-A(\nabla W_A)\right)\\
&-\mathrm{div}\left(A(\nabla(w-M+W_A))-A(\nabla W_A)\right), u_m-M-(w-M)\rangle_{Z^*,Z}\\
&+2\langle f(u_m+W_A)-f(w+W_A), u_m-M-(w-M)\rangle_{Z^*,Z}\\
&-c\|u_m-M-(w-M)\|^2\}\big]ds.
\end{aligned}$$

We have the following result

LEMMA 3.3. $O_m\leq 0$.

We write $\mathcal{O}_m$ in the form $\mathcal{O}_m=\mathcal{O}_m^1+\mathcal{O}_m^2$ where

$$\begin{aligned}
\mathcal{O}_m^1 = \mathbb{E}\big[\int_0^T e^{-cs}\{&2\langle \mathrm{div}\left(A(\nabla(u_m-M+W_A))-A(\nabla W_A)\right), u_m-M\rangle_{Z^*,Z}\\
&+2\langle f(u_m+W_A), u_m-M\rangle_{Z^*,Z}-c\|u_m-M\|^2\}\big]ds. \qquad (3.15)
\end{aligned}$$

We integrate the equation (3.2) between 0 and T and recall that the nonlocal term in (3.2) vanishes, we apply a chain rule formula and take the expectation to obtain

$$
\begin{aligned}
&\mathbb{E}[e^{-cT}\|u_m(T)-M\|^2]\\
&=\mathbb{E}[\|u_m(0)-M\|^2]-c\mathbb{E}[\int_0^T e^{-cs}\|u_m(s)-M\|^2ds]\\
&+2\mathbb{E}[\int_0^T e^{-cs}\langle \mathrm{div}[A(\nabla(u_m-M+W_A))-A(\nabla W_A)],u_m-M\rangle_{Z^*,Z}\\
&+2\mathbb{E}[\int_0^T e^{-cs}\langle f(u_m+W_A),u_m-M\rangle_{Z^*,Z}]. \qquad (3.16)
\end{aligned}
$$

It follows from (3.15) and (3.16) that

$$
\lim_{m\to\infty}\sup \mathcal{O}_m^1=\mathbb{E}[e^{-cT}\|u(T)-M\|^2]-\mathbb{E}[\|u(0)-M\|^2]+\delta e^{-cT}, \qquad (3.17)
$$

where $\delta=\lim_{m\to\infty}\sup\mathbb{E}[\|u_m(T)-M\|^2]-\mathbb{E}[\|u(T)-M\|^2]\geq 0$.

On the other hand, the equation (3.14) implies that a.s. in $Z^*=V^*+L^{\frac{4}{3}}(D)$:

$$
u(t)-M=\varphi_0-M+\int_0^t\Phi-\mathrm{div}(A(\nabla W_A))+\int_0^t\chi,\quad \forall t\in[0,T]. \qquad (3.18)
$$

Applying a chain rule formula we deduce that

$$
\begin{aligned}
\mathbb{E}[e^{-cT}\|u(T)-M\|^2]=\mathbb{E}[\|u(0)-M\|^2]&-c\mathbb{E}[\int_0^T e^{-cs}\|u(s)-M\|^2ds]\\
&+2\mathbb{E}\int_0^T e^{-cs}\langle\Phi-\mathrm{div}(A(\nabla W_A)),u-M\rangle_{Z^*,Z}\\
&+2\mathbb{E}[\int_0^T e^{-cs}\langle\chi,u-M\rangle_{Z^*,Z}].
\end{aligned}
$$

which we combine with (3.17) to deduce that

$$
\begin{aligned}
\lim_{m\to\infty}\sup O_m^1=2\mathbb{E}[&\int_0^T e^{-cs}\langle\Phi-\mathrm{div}(A(\nabla W_A)),u-M\rangle_{Z^*,Z}]\\
&+2\mathbb{E}\int_0^T e^{-cs}\langle\chi,u-M\rangle_{Z^*,Z}-c\mathbb{E}[\int_0^T e^{-cs}\|u(s)-M\|^2ds]+\delta e^{-cT}. \qquad (3.19)
\end{aligned}
$$

It remains to compute the limit of $\mathcal{O}_m^2$; in view of (3.9), (3.11) and (3.12), we deduce that

$$
\begin{aligned}
&\lim_{m\to\infty}\mathcal{O}_m^2\\
&=\mathbb{E}\int_0^T e^{-cs}\{-2\langle\mathrm{div}[A(\nabla(w-M+W_A))-A(\nabla W_A)],u-M\rangle_{Z^*,Z}\\
&-2\langle\Phi-\mathrm{div}(A(\nabla W_A))-\mathrm{div}[A(\nabla(w-M+W_A))-A(\nabla W_A)],w-M\rangle_{Z^*,Z}\\
&-2\langle f(w+W_A),u-M\rangle_{Z^*,Z}-2\langle\chi-f(w+W_A),w-M\rangle_{Z^*,Z}\\
&-c\|w-M\|^2+2c\langle u-M,w-M\rangle\}ds. \qquad (3.20)
\end{aligned}
$$

Combining (3.19) and (3.20), and remembering that $O_m \leq 0$, yields

$$\mathbb{E}\int_0^T e^{-cs}\{2\langle \Phi - \operatorname{div}\left(A\nabla(w-M+W_A)\right), u-M-(w-M)\rangle_{Z^*,Z}$$
$$+2\langle \chi - f(w+W_A), u-M-(w-M)\rangle_{Z^*,Z} - c\|u-M-(w-M)\|^2\}ds + \delta e^{-cT} \leq 0.$$

Let $v \in L^2(\Omega\times(0,T);V)\cap L^4(\Omega\times(0,T)\times D)$ be arbitrary and set

$$w - M = u - M - \lambda v, \text{ with } \lambda \in \mathbb{R}_+.$$

Dividing by λ and letting $\lambda \to 0$, we find that for all $v \in L^2(\Omega\times(0,T);V)\cap L^4(\Omega\times(0,T)\times D)$

$$\mathbb{E}\int_0^T \langle \Phi+\chi, v\rangle_{Z^*,Z} = \mathbb{E}\int_0^T \langle \operatorname{div}[A(\nabla(u-M+W_A))] + f(u+W_A), v\rangle_{Z^*,Z}.$$

One finally concludes that u satisfies Definition 3.1.

4. Uniqueness of the solution of Problem (P_2). Let ω be given such that two pathwise solutions of Problem (P_2), $u_1 = u_1(\omega,x,t)$ and $u_2 = u_2(\omega,x,t)$ satisfy

$$u_i(\cdot,\cdot,\omega) \in L^\infty(0,T;L^2(D))\cap L^2(0,T;H^1(D))\cap L^4((0,T)\times D),$$
$$f(u_i+W_A) \in L^{\frac{4}{3}}((0,T)\times D),$$
$$\operatorname{div}(A(\nabla(u_i+W_A)) \in L^2((0,T);(H^1(D))')$$

for $i=1,2$, and $u_1(\cdot,0) = u_2(\cdot,0) = \varphi_0$. The difference $u_1 - u_2$ satisfies the equation

$$\begin{aligned} u_1(t) - u_2(t) = &\int_0^t \operatorname{div}(A(\nabla(u_1+W_A))) - div(A(\nabla(u_2+W_A))) \\ &+ \int_0^t [f(u_1+W_A) - f(u_2+W_A)] \\ &- \frac{1}{|D|}\int_0^t [\int_D f(u_1+W_A) - \int_D f(u_2+W_A)dx], \end{aligned}$$

in $L^2((0,T);V^*) + L^{\frac{4}{3}}((0,T)\times D)$.
We take the duality product of the equation for the difference u_1-u_2 with $u_1-u_2 \in L^2((0,T);V^*)\cap L^{\frac{4}{3}}((0,T)\times D)$, we use (1.3) to obtain

$$\begin{aligned} \|u_1-u_2\|^2_{L^2(D)} \leq &-C_0\int_0^t\int_D |\nabla(u_1-u_2)|^2 \\ &+ \int_0^t \langle f(u_1+W_A) - f(u_2+W_A)), u_1-u_2\rangle_{Z^*,Z} \\ &- \int_0^t \langle \frac{1}{|D|}\int_D (f(u_1+W_A) - f(u_2+W_A))dx, u_1-u_2\rangle_{Z^*,Z}. \end{aligned}$$

Since $\int_D u_1(x,t)dx = \int_D u_2(x,t)dx = \int_D \varphi_0(x)dx$, it follows that the nonlocal term vanishes. Using the fact that $f' \leq 1$ we obtain

$$\int_D (u_1-u_2)^2(x,t)dx \leq \int_0^t\int_D (u_1-u_2)^2(x,t)dxds, \quad \text{for all} \quad t\in(0,T).$$

which in turn implies by Gronwall's Lemma that $u_1 = u_2$ a.e. in $D\times(0,T)$.
□

Acknowledgment. The authors would like to thank Professor T. Funaki and Professor M. Hofmanova for invaluable discussions and the GDRI ReaDiNet for financial support.

REFERENCES

[1] D. C. Antonopoulou, and P. W. Bates, D. Blömker and G. D. Karali, *Motion of a droplet for the stochastic mass-conserving Allen-Cahn equation*, in SIAM J. Math. Anal. 48 (2016), pp. 670–708.
[2] C. Bauzet, G. Vallet and P. Wittbold, *The Cauchy problem for conservation laws with a multiplicative stochastic perturbation*, J. Hyperbolic Differ. Equ. 9,4 (2012), pp. 661-709.
[3] C. Bennett and R. Sharpley, *Interpolation of Operators*, academic press,Vol. 129 (1988).
[4] S. Boussaïd, D. Hilhorst and T. Nguyen, *Convergence to steady state for the solutions of a nonlocal reaction-diffusion equation*, Evol. Equ. Control Theory 4,1 (2015), pp. 39-59.
[5] G. Da Prato, J.Zabczyk, *Stochastic equations in infinite dimensions*, Second edition. Encyclopedia of Mathematics and its Applications, 152 (2014), Cambridge University Press, Cambridge.
[6] P. El Kettani, D. Hilhorst and K.Lee, *A stochastic mass conserved reaction-diffusion equation with nonlinear diffusion*, preprint.
[7] B. Gess, *Strong solutions for stochastic partial differential equations of gradient type*, Journal of Functional Analysis, vol. 263, no 8 (2012), pp. 2355-2383.
[8] N. V. Krylov and B. L. Rozovskii, *Stochastic evolution equations*, Journal of Soviet Mathematics, vol. 14 (1981), pp. 1233-1277.
[9] M. Marion, *Attractors for reaction-diffusion equations: existence and estimate of their dimension*, Applicable Analysis: An International Journal, 25:1-2 (1987), pp. 101-147.
[10] J. Rubinstein and P. Sternberg, *Nonlocal reaction-diffusion equations and nucleation*, IMA Journal of Applied Mathematics, 48 (1992), pp. 249-264.
[11] R.Temam, *Navier-stokes equations*, Amsterdam: North-Holland, Vol. 2, revised edition (1979).

Proceedings of EQUADIFF 2017
pp. 211–220

VECTORIAL QUASILINEAR DIFFUSION EQUATION WITH DYNAMIC BOUNDARY CONDITION

RYOTA NAKAYASHIKI*

Abstract. In this paper, we consider a class of initial-boundary value problems for quasilinear PDEs, subject to the dynamic boundary conditions. Each initial-boundary problem is denoted by $(\mathrm{S})_\varepsilon$ with a nonnegative constant ε, and for any $\varepsilon \geq 0$, $(\mathrm{S})_\varepsilon$ can be regarded as a vectorial transmission system between the quasilinear equation in the spatial domain Ω, and the parabolic equation on the boundary $\Gamma := \partial\Omega$, having a sufficient smoothness. The objective of this study is to establish a mathematical method, which can enable us to handle the transmission systems of various vectorial mathematical models, such as the Bingham type flow equations, the Ginzburg–Landau type equations, and so on. On this basis, we set the goal of this paper to prove two Main Theorems, concerned with the well-posedness of $(\mathrm{S})_\varepsilon$ with the precise representation of solution, and ε-dependence of $(\mathrm{S})_\varepsilon$, for $\varepsilon \geq 0$.

Key words. vectorial parabolic equation, quasilinear diffusion, dynamic boundary condition

AMS subject classifications. 35K40, 35K59, 35R35.

1. Introduction. Let $0 < T < \infty$ and $\kappa > 0$ be fixed constants, and let $m \in \mathbb{N}$ and $1 < N \in \mathbb{N}$ be fixed constants of dimensions. Let Ω be a bounded spatial domain in $\mathbb{R}^N$ with a smooth boundary $\Gamma := \partial\Omega$, and let n_Γ be the unit outer normal to Γ. Besides, we denote by $Q := (0,T) \times \Omega$ the product space of a time interval $(0,T)$ and the spatial domain Ω, and we put $\Sigma := (0,T) \times \Gamma$.

In this paper, we take a constant $\varepsilon \geq 0$, and consider the following initial-boundary value problem:

$$(\mathrm{S})_\varepsilon: \quad \begin{cases} \partial_t u - \mathrm{div}\left(\dfrac{\nabla u}{\|\nabla u\|} + \kappa^2 \nabla u\right) \ni \theta \text{ in } Q, \\ \partial_t u_\Gamma - \Delta_\Gamma(\varepsilon^2 u_\Gamma) + \left(\frac{\nabla u}{\|\nabla u\|} + \kappa^2 \nabla u\right)_{|\Gamma} n_\Gamma \ni \theta_\Gamma \text{ and } u_{|\Gamma} = u_\Gamma \text{ on } \Sigma, \\ u(0,\cdot) = u_0 \text{ in } \Omega \text{ and } u_\Gamma(0,\cdot) = u_{\Gamma,0} \text{ on } \Gamma, \end{cases}$$

for the vectorial unknowns $u : Q \to \mathbb{R}^m$ and $u_\Gamma : \Sigma \to \mathbb{R}^m$. In the context, $\theta : Q \to \mathbb{R}^m$ and $\theta_\Gamma : \Sigma \to \mathbb{R}^m$ are given forcing terms, and $u_0 : \Omega \to \mathbb{R}^m$ and $u_{\Gamma,0} : \Gamma \to \mathbb{R}^m$ are given initial data for u and u_Γ, respectively. "$_{|\Gamma}$" denotes the trace of a Sobolev function on Ω, and Δ_Γ denotes the Laplace–Beltrami operator on Γ.

The boundary condition of $(\mathrm{S})_\varepsilon$ is given in the form of the so-called "*dynamic boundary condition*". In particular, since we can use the equation $u_{|\Gamma} = u_\Gamma$ on Σ to resemble a kind of the transmission condition, we can say that the problem $(\mathrm{S})_\varepsilon$ is a vectorial transmission system between the quasilinear equation in Ω, and the parabolic equation on Γ.

The objective of this study is to establish a mathematical method, which enables us to handle various nonlinear phenomena described by vectorial unknowns. In this regard, the study on $(\mathrm{S})_\varepsilon$, for any $\varepsilon \geq 0$, is aimed at the mathematical analysis for quasilinear transmission systems, associated with the Bingham type flow equations, the Ginzburg–Landau type equations, and so on.

*Department of Mathematics and Informatics, Graduate School of Science, Chiba University, 1–33, Yayoi-cho, Inage-ku, Chiba, 263–8522, Japan.

In view of such backgrounds, we set the goal to obtain some generalized results for the previous works [4, 10], which dealt with the scalar-valued cases of quasilinear transmission systems. On this basis, our principal results will be stated in forms of the following two Main Theorems, which will be to verify the qualitative properties of the systems, for every $\varepsilon \geq 0$.

Main Theorem 1 : the well-posedness for $(\mathrm{S})_\varepsilon$ with the precise expression of solutions, for any $\varepsilon \geq 0$.

Main Theorem 2 : the continuous dependence of solutions to $(\mathrm{S})_\varepsilon$ with respect to $\varepsilon \geq 0$.

The content of this paper is as follows. Main Theorems are stated in Section 3 and these are discussed on the basis of the preliminaries prepared in Section 2. The keypoints of the proofs are specified in Section 4, and the proofs of the Main Theorems are provided in the last Section 5.

2. Preliminaries. In this section, we outline some basic notations.

Abstract Notations. For an abstract Banach space X, we denote by $|\cdot|_X$ the norm of X, and denote by $\langle\,\cdot\,,\,\cdot\,\rangle_X$ the duality pairing between X and the dual space X^* of X. Let $\mathcal{I}_X : X \to X$ be the identity map from X onto X. In particular, when X is a Hilbert space, we denote by $(\,\cdot\,,\,\cdot\,)_X$ the inner product of X.

For any proper lower semi-continuous (l.s.c. from now on) and convex function Ψ defined on a Hilbert space X, we denote by $D(\Psi)$ its effective domain, and denote by $\partial\Psi$ its subdifferential. The subdifferential $\partial\Psi$ is a set-valued map corresponding to a weak differential of Ψ, and it turns out to be a maximal monotone graph in the product space $X^2 := X \times X$ (see [1–3, 7], for details). More precisely, for each $z_0 \in X$, the value $\partial\Psi(z_0)$ is defined as a set of all elements $z_0^* \in X$ which satisfy the following variational inequality:

$$(z_0^*, z - z_0)_X \leq \Psi(z) - \Psi(z_0), \text{ for any } z \in D(\Psi).$$

The set $D(\partial\Psi) := \{z \in X \,|\, \partial\Psi(z) \neq \emptyset\}$ is called the domain of $\partial\Psi$. We often use the notation "$[z_0, z_0^*] \in \partial\Psi$ in X^2", to mean that "$z_0^* \in \partial\Psi(z_0)$ in X with $z_0 \in D(\partial\Psi)$", by identifying the operator $\partial\Psi$ with its graph in X^2.

Additionally, in this study, we use the following notion of convergence, called "Mosco-convergence", for sequences of convex functions.

Definition 2.1 (Mosco-convergence: cf. [9]). *Let X be an abstract Hilbert space. Let $\Psi : X \to (-\infty, \infty]$ be a proper l.s.c. and convex function, and let $\{\Psi_n\}_{n=1}^\infty$ be a sequence of proper l.s.c. and convex functions $\Psi_n : X \to (-\infty, \infty]$, $n \in \mathbb{N}$. Then, it is said that $\Psi_n \to \Psi$ on X, in the sense of Mosco, as $n \to \infty$, iff. the following two conditions are fulfilled.*

(M1) Lower-bound condition: $\underline{\lim}_{n\to\infty} \Psi_n(\check{z}_n) \geq \Psi(\check{z})$, *if* $\check{z} \in X$, $\{\check{z}_n\}_{n=1}^\infty \subset X$, *and* $\check{z}_n \to \check{z}$ *weakly in* X *as* $n \to \infty$.

(M2) Optimality condition: *for any* $\hat{z} \in D(\Psi)$, *there exists a sequence* $\{\hat{z}_n\}_{n=1}^\infty \subset X$ *such that* $\hat{z}_n \to \hat{z}$ *in* X *and* $\Psi_n(\hat{z}_n) \to \Psi(\hat{z})$, *as* $n \to \infty$.

Notations in real analysis. Let $d \in \mathbb{N}$ be any fixed dimension. Then, we simply denote by $a \cdot b$ and $|a|$ the standard scalar product of $a, b \in \mathbb{R}^d$ and the Euclidean norm of $a \in \mathbb{R}^d$, respectively. Besides, for arbitrary d-dimensional vectors $a = [a_i]$, $b = [b_i] \in \mathbb{R}^d$ with components $a_i, b_i \in \mathbb{R}$ $(i = 1, \ldots, d)$, we define:

$$a \otimes b := a^{\mathrm{t}} b = \begin{bmatrix} a_1 b_1 & \cdots & a_1 b_d \\ \vdots & \ddots & \vdots \\ a_d b_1 & \cdots & a_d b_d \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

For any $d \in \mathbb{N}$, the d-dimensional Lebesgue measure is denoted by $\mathcal{L}^d$, and d-dimensional Hausdorff measure is denoted by $\mathcal{H}^d$. Unless otherwise specified, the measure theoretical phrases, such as "a.e.", "dt", "dx", and so on, are with respect to the Lebesgue measure in each corresponding dimension. Also, in the observation on a smooth surface S, the phrase "a.e." is with respect to the Hausdorff measure in each corresponding Hausdorff dimension, and the area element on S is denoted by dS.

Notations of surface-differentials. Throughout this paper, let $1 < N \in \mathbb{N}$ be a fixed dimension, let $\Omega \subset \mathbb{R}^N$ be a bounded domain with C^∞-boundary $\Gamma := \partial\Omega$, and let $n_\Gamma \in C^\infty(\Gamma; \mathbb{R}^N)$ be the unit outer normal on Γ. Besides, we suppose that the distance function $x \in \mathbb{R}^N \mapsto d_\Gamma(x) := \inf_{y \in \Gamma} |x - y| \in \mathbb{R}$ forms a C^∞-function on a neighborhood of Γ. Based on these, we define:

$$L^2_{\tan}(\Gamma) := \{\ \tilde{\omega} \in L^2(\Gamma; \mathbb{R}^N) \ \big|\ \tilde{\omega} \cdot n_\Gamma = 0 \text{ on } \Gamma\ \},$$

and we define the so-called Laplace–Beltrami operator Δ_Γ as the composition $\Delta_\Gamma := \mathrm{div}_\Gamma \circ \nabla_\Gamma : C^\infty(\Gamma) \to C^\infty(\Gamma)$ of the *surface-gradient*:

$$\varphi \in C^1(\Gamma) \mapsto \nabla_\Gamma \varphi := \nabla \varphi^{\mathrm{ex}} - (\nabla d_\Gamma \otimes \nabla d_\Gamma) \nabla \varphi^{\mathrm{ex}} \in L^2_{\tan}(\Gamma) \cap C(\Gamma; \mathbb{R}^N),$$

and the *surface-divergence*:

$$\omega \in C^1(\Gamma; \mathbb{R}^N) \mapsto \mathrm{div}_\Gamma\, \omega := \mathrm{div}\, \omega^{\mathrm{ex}} - \nabla(\omega^{\mathrm{ex}} \cdot \nabla d_\Gamma) \cdot \nabla d_\Gamma \in C(\Gamma).$$

As is well-known (cf. [11]), the surface-gradient ∇_Γ can be extended to a linear operator from the Sobolev space $H^1(\Gamma)$ into $L^2_{\tan}(\Gamma)$, and the extension (also denoted by ∇_Γ) define the inner product of the Hilbert space $H^1(\Gamma)$ as follows:

$$(\varphi, \psi)_{H^1(\Gamma)} := (\varphi, \psi)_{L^2(\Gamma)} + (\nabla_\Gamma \varphi, \nabla_\Gamma \psi)_{L^2(\Gamma; \mathbb{R}^N)}, \ \text{ for all } \varphi, \psi \in H^1(\Gamma).$$

Also, the surface-divergence div_Γ can be extended to an operator from $L^2(\Gamma; \mathbb{R}^N)$ into $H^{-1}(\Gamma)$, and as a consequence, the composition $-\mathrm{div}_\Gamma \circ \nabla_\Gamma = -\Delta_\Gamma : H^1(\Gamma) \to H^{-1}(\Gamma)$ provides a duality map, such that:

$$\langle -\Delta_\Gamma \varphi, \psi \rangle_{H^1(\Gamma)} = (\nabla_\Gamma \varphi, \nabla_\Gamma \psi)_{L^2(\Gamma; \mathbb{R}^N)}, \ \text{ for all } \varphi, \psi \in H^1(\Gamma).$$

Notations in tensor analysis. Let $m \in \mathbb{N}$ be another dimension (besides N). For arbitrary $(m \times N)$-matrices $A = [a_{ij}], B = [b_{ij}] \in \mathbb{R}^{m \times N}$ with components $a_{ij}, b_{ij} \in \mathbb{R}$ $(i = 1, \dots m,\ j = 1, \dots, N)$, we denote by $A : B$ and $\|A\|$ the scalar product of A and B and the Frobenius norm of A, respectively, i.e.:

$$A : B := \sum_{j=1}^{N} \sum_{i=1}^{m} a_{ij} b_{ij} \in \mathbb{R} \text{ and } \ \|A\| := \sqrt{A : A} \in \mathbb{R}, \ \text{ for all } A, B \in \mathbb{R}^{m \times N}.$$

For any vectorial function $z = [z_i] \in L^2(\Omega; \mathbb{R}^m)$, we denote by ∇z the (distributional) gradient of z, defined as:

$$\nabla z := {}^{\mathrm{t}}[\nabla z_1, \dots, \nabla z_m] = \begin{bmatrix} \partial_1 z_1 & \cdots & \partial_N z_1 \\ \vdots & \ddots & \vdots \\ \partial_1 z_m & \cdots & \partial_N z_m \end{bmatrix} \in \mathcal{D}'(\Omega)^{m \times N},$$

and, for any matrix-valued function $Z = [z_{ij}] \in L^2(\Omega; \mathbb{R}^{m\times N})$, we denote by $\mathrm{div} Z$ the (distributional) divergence of Z, defined as:

$$\mathrm{div} Z := \left[\sum_{k=1}^{N} \partial_k z_{ik} \right] \in \mathcal{D}'(\Omega)^m.$$

Similarly, for any vectorial function $z = [z_i] \in H^1(\Gamma; \mathbb{R}^m)$, we define the surface-gradient $\nabla_\Gamma z$ of z by $\nabla_\Gamma z := {}^{\mathrm{t}}[\nabla_\Gamma z_1, \dots, \nabla_\Gamma z_m] \in L^2_{\tan}(\Gamma)^m$, and we define $\Delta_\Gamma z := [\Delta_\Gamma z_i] \in H^{-1}(\Gamma; \mathbb{R}^m)$.

REMARK 1 (cf. [8, Proposition 1.6]). The mapping $M \in H^1(\Omega; \mathbb{R}^{m\times N}) \mapsto M_{|\Gamma} n_\Gamma \in H^{\frac{1}{2}}(\Gamma; \mathbb{R}^m)$ can be extended as a linear and continuous operator $[\,\cdot\,, n_\Gamma]_\Gamma$ from

$$L^2_{\mathrm{div}}(\Omega) := \left\{ \ \tilde{M} \in L^2(\Omega; \mathbb{R}^{m\times N}) \ \middle|\ \mathrm{div}\tilde{M} \in L^2(\Omega; \mathbb{R}^m) \ \right\}$$

into $H^{-\frac{1}{2}}(\Gamma; \mathbb{R}^m)$, such that:

$$\left\langle [M, n_\Gamma]_\Gamma, z_{|\Gamma} \right\rangle_{H^{\frac{1}{2}}(\Gamma;\mathbb{R}^m)} := \int_\Omega \mathrm{div} M \cdot z \, dx + \int_\Omega M : \nabla z \, dx, \quad \text{for all } M \in L^2_{\mathrm{div}}(\Omega) \text{ and } z \in H^1(\Omega; \mathbb{R}^m). \tag{2.1}$$

3. Main Theorems. Let us set

$$\mathscr{H} := L^2(\Omega; \mathbb{R}^m) \times L^2(\Gamma; \mathbb{R}^m),$$

and for any $\varepsilon \geq 0$, let us set:

$$\mathscr{V}_\varepsilon := \left\{ \ [v, v_\Gamma] \in \mathscr{H} \ \middle|\ \begin{array}{l} v \in H^1(\Omega; \mathbb{R}^m),\ v_\Gamma \in H^{\frac{1}{2}}(\Gamma; \mathbb{R}^m), \\ \varepsilon v_\Gamma \in H^1(\Gamma; \mathbb{R}^m), \text{ and } v_{|\Gamma} = v_\Gamma, \text{ a.e. on } \Gamma \end{array} \ \right\}.$$

Note that $\mathscr{H}$ is a Hilbert space endowed with the inner product:

$$([z_1, z_{\Gamma,1}], [z_2, z_{\Gamma,2}])_{\mathscr{H}} := (z_1, z_2)_{L^2(\Omega;\mathbb{R}^m)} + (z_{\Gamma,1}, z_{\Gamma,2})_{L^2(\Gamma;\mathbb{R}^m)}, \quad \text{for } [z_k, z_{\Gamma,k}] \in \mathscr{H},\ k = 1, 2.$$

Also, if $\varepsilon > 0$ (resp. $\varepsilon = 0$), then the corresponding class $\mathscr{V}_\varepsilon$ (resp. $\mathscr{V}_0$) is a closed linear space in $H^1(\Omega; \mathbb{R}^m) \times H^1(\Gamma; \mathbb{R}^m)$ (resp. in $H^1(\Omega; \mathbb{R}^m) \times H^{\frac{1}{2}}(\Gamma; \mathbb{R}^m)$), and hence, it is a Hilbert space endowed with the standard inner product of $H^1(\Omega; \mathbb{R}^m) \times H^1(\Gamma; \mathbb{R}^m)$ (resp. $H^1(\Omega; \mathbb{R}^m) \times H^{\frac{1}{2}}(\Gamma; \mathbb{R}^m)$). Furthermore, for any $\varepsilon \geq 0$, $\mathscr{V}_\varepsilon$ is dense in $\mathscr{H}$, i.e. $\overline{\mathscr{V}_\varepsilon} = \mathscr{H}$, and the embedding $\mathscr{V}_\varepsilon \subset \mathscr{H}$ is compact.

By using the above notations, we define the solution to $(\mathrm{S})_\varepsilon$, for $\varepsilon \geq 0$, as follows.

DEFINITION 3.1. *Let $\varepsilon \geq 0$ be a fixed constant. Then, a pair of functions $[u, u_\Gamma] \in L^2(0, T; \mathscr{H})$ is called a solution to $(S)_\varepsilon$, iff. the following conditions are fulfilled.*

(S1) $[u, u_\Gamma] \in C([0, T]; \mathscr{H}) \cap W^{1,2}_{\mathrm{loc}}((0, T]; \mathscr{H}) \cap L^2(0, T; \mathscr{V}_\varepsilon) \cap L^\infty_{\mathrm{loc}}((0, T]; \mathscr{V}_\varepsilon)$, $[u(0), u_\Gamma(0)] = [u_0, u_{\Gamma,0}]$ *in* $\mathscr{H}$.

(S2) *There exists a function $M_u : Q \to \mathbb{R}^{m\times N}$, such that:*

$$M_u(t) \in L^2_{\mathrm{div}}(\Omega), \ \text{a.e. } t \in (0, T) \text{ and } [\nabla u, M_u] \in \partial\|\cdot\| \text{ in } [\mathbb{R}^{m\times N}]^2, \text{ a.e. in } Q,$$

and

$$\begin{aligned}&\int_\Omega \partial_t u(t)\cdot z\,dx+\int_\Omega (M_u(t)+\kappa^2\nabla u(t)):\nabla z\,dx\\&\quad+\int_\Gamma \partial_t u_\Gamma(t)\cdot z_\Gamma\,d\Gamma+\int_\Gamma \nabla_\Gamma(\varepsilon u_\Gamma(t)):\nabla_\Gamma(\varepsilon z_\Gamma)\,d\Gamma\\&=\int_\Omega \theta(t)\cdot z\,dx+\int_\Gamma \theta_\Gamma(t)\cdot z_\Gamma\,d\Gamma,\quad \textit{for any } [z,z_\Gamma]\in\mathscr{V}_\varepsilon,\ \textit{and a.e. } t\in(0,T),\end{aligned}$$

where $\partial\|\cdot\|\subset[\mathbb{R}^{m\times N}]^2$ *denotes the subdifferential of the Frobenius norm* $\|\cdot\|$ *on* $\mathbb{R}^{m\times N}$.

Based on this, the Main Theorems of this paper are stated as follows.

MAIN THEOREM 1 (Well-posedness). *Let* $\varepsilon\ge 0$ *be a fixed constant. Then, the following two items hold.*

(I-1) *(Solvability) For every* $[\theta,\theta_\Gamma]\in L^2(0,T;\mathscr{H})$ *and* $[u_0,u_{\Gamma,0}]\in\mathscr{H}$*, the system* $(S)_\varepsilon$ *admits a unique solution* $[u,u_\Gamma]$.

(I-2) *(Continuous dependence) For* $k=1,2$*, let* $[u_k,u_{\Gamma,k}]$ *be two solutions to* $(S)_\varepsilon$*, corresponding to the forcing pairs* $[\theta_k,\theta_{\Gamma,k}]\in L^2(0,T;\mathscr{H})$ *and the initial pairs* $[u_{0,k},u_{\Gamma,0,k}]\in\mathscr{H}$*, respectively. Then, it follows that:*

$$\begin{aligned}&|[u_1-u_2,u_{\Gamma,1}-u_{\Gamma,2}]|^2_{C([0,T];\mathscr{H})}\\&\le \mathrm{e}^T\big(|[\theta_1-\theta_2,\theta_{\Gamma,1}-\theta_{\Gamma,2}]|^2_{L^2(0,T;\mathscr{H})}+|[u_{0,1}-u_{0,2},u_{\Gamma,0,1}-u_{\Gamma,0,2}]|^2_{\mathscr{H}}\big).\end{aligned}$$

MAIN THEOREM 2 (Continuous dependence with respect to $\varepsilon\ge 0$). *Let* $\varepsilon_0\ge 0$ *be a fixed constant. Let* $\{[\theta_\varepsilon,\theta_{\Gamma,\varepsilon}]\}_{\varepsilon\ge0}\subset L^2(0,T;\mathscr{H})$ *be a sequence of the forcing pair, let* $\{[u_{0,\varepsilon},u_{\Gamma,0,\varepsilon}]\}_{\varepsilon\ge0}\subset\mathscr{H}$ *be a sequence of the initial pair, and for any* $\varepsilon\ge0$*, let* $[u_\varepsilon,u_{\Gamma,\varepsilon}]$ *be a solution to* $(S)_\varepsilon$ *corresponding to the forcing pair* $[\theta_\varepsilon,\theta_{\Gamma,\varepsilon}]$ *and the initial pair* $[u_{0,\varepsilon},u_{\Gamma,0,\varepsilon}]$*. Here, if:*

$$\begin{cases}[\theta_\varepsilon,\theta_{\Gamma,\varepsilon}]\to[\theta_{\varepsilon_0},\theta_{\Gamma,\varepsilon_0}]\ \textit{weakly in } L^2(0,T;\mathscr{H}),\\ [u_{0,\varepsilon},u_{\Gamma,0,\varepsilon}]\to[u_{0,\varepsilon_0},u_{\Gamma,0,\varepsilon_0}]\ \textit{in } \mathscr{H},\end{cases}\qquad \textit{as } \varepsilon\to\varepsilon_0,$$

then:

$$[u_\varepsilon,u_{\Gamma,\varepsilon}]\to[u_{\varepsilon_0},u_{\Gamma,\varepsilon_0}]\ \textit{in } C([0,T];\mathscr{H}),\ \textit{and in } L^2(0,T;\mathscr{V}_0)\ \textit{as } \varepsilon\to\varepsilon_0,\tag{3.1}$$

and in particular, if $\varepsilon_0>0$*, then:*

$$u_{\Gamma,\varepsilon}\to u_{\Gamma,\varepsilon_0}\ \textit{in } L^2(0,T;H^1(\Gamma;\mathbb{R}^m)),\ \textit{as } \varepsilon\to\varepsilon_0.\tag{3.2}$$

4. Keypoints of the proofs. In this section, we specify the keypoints in the proofs of Main Theorems. Roughly summarized, we will prove the Main Theorems by reformulating our system $(\mathrm{S})_\varepsilon$ to the following Cauchy problem for an evolution equation, denoted by $(\mathrm{CP})_\varepsilon$:

$$(\mathrm{CP})_\varepsilon\quad\begin{cases}U'(t)+\partial\Phi_\varepsilon(U(t))\ni\Theta(t)\ \text{in } \mathscr{H},\ \text{a.e. } t\in(0,T),\\ U(0)=U_0\ \text{in } \mathscr{H},\end{cases}\quad\text{for } \varepsilon\ge0.$$

In the context, the unknown $U\in L^2(0,T;\mathscr{H})$ corresponds to the solution $[u,u_\Gamma]$ to the system $(\mathrm{S})_\varepsilon$, and $\Theta:=[\theta,\theta_\Gamma]$ in $L^2(0,T;\mathscr{H})$ and $U_0:=[u_0,u_{\Gamma,0}]$ in $\mathscr{H}$ correspond to

the pair of the forcing terms and the pair of the initial data, respectively. $\partial\Phi_\varepsilon$ denotes the subdifferential of a proper l.s.c. and convex function $\Phi_\varepsilon : \mathscr{H} \to [0,\infty]$, defined as:

$$\begin{aligned} U &= [u, u_\Gamma] \in \mathscr{H} \mapsto \Phi_\varepsilon(U) = \Phi_\varepsilon(u, u_\Gamma) \\ &:= \begin{cases} \displaystyle\int_\Omega \left(\|\nabla u\| + \frac{\kappa^2}{2}\|\nabla u\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2 \, d\Gamma, & \text{for } \varepsilon \geq 0. \\ \qquad \text{if } U = [u, u_\Gamma] \in \mathscr{V}_\varepsilon, \\ \infty, \quad \text{otherwise}, \end{cases} \end{aligned}$$

Now, the essential keypoint is to show the following Key-Lemma, which is to sustain a certain association between the system $(\mathrm{S})_\varepsilon$ and the Cauchy problem $(\mathrm{CP})_\varepsilon$, for any $\varepsilon \geq 0$.

KEY-LEMMA 1 (The representation of $\partial\Phi_\varepsilon$) *For any $\varepsilon \geq 0$, the following two items are equivalent.*

(I) *$[u, u_\Gamma] \in D(\partial\Phi_\varepsilon)$ and $[u^*, u_\Gamma^*] \in \partial\Phi_\varepsilon(u, u_\Gamma)$ in $\mathscr{H}$.*

(II) *$[u, u_\Gamma] \in D(\Phi_\varepsilon)$ and there exists $M_u^* \in L^\infty(\Omega; \mathbb{R}^{m\times N})$, such that:*

$$[\nabla u, M_u^*] \in \partial\|\cdot\| \text{ in } [\mathbb{R}^{m\times N}]^2, \text{ a.e. in } \Omega, \tag{4.1}$$

$$\begin{cases} M_u^* + \kappa^2 \nabla u \in L^2_{\mathrm{div}}(\Omega), \\ -\Delta_\Gamma(\varepsilon^2 u_\Gamma) + [(M_u^* + \kappa^2\nabla u), n_\Gamma]_\Gamma \in L^2(\Gamma; \mathbb{R}^m), \end{cases} \tag{4.2}$$

$$\begin{cases} u^* = -\mathrm{div}(M_u^* + \kappa^2 \nabla u) \text{ in } L^2(\Omega; \mathbb{R}^m), \\ u_\Gamma^* = -\Delta_\Gamma(\varepsilon^2 u_\Gamma) + [(M_u^* + \kappa^2\nabla u), n_\Gamma]_\Gamma \text{ in } L^2(\Gamma; \mathbb{R}^m). \end{cases} \tag{4.3}$$

For the proof of the Key-Lemma, we prepare a class of relaxed convex functions $\{\, \Phi_\varepsilon^\delta \,|\, \varepsilon \geq 0,\ 0 < \delta \leq 1\}$, defined as:

$$\begin{aligned} U &= [u, u_\Gamma] \in \mathscr{H} \mapsto \Phi_\varepsilon^\delta(U) = \Phi_\varepsilon^\delta(u, u_\Gamma) \\ &:= \begin{cases} \displaystyle\int_\Omega \left(\sqrt{\|\nabla u\|^2 + \delta^2} + \frac{\kappa^2}{2}\|\nabla u\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2 \, d\Gamma, \\ \qquad \text{if } U = [u, u_\Gamma] \in \mathscr{V}_\varepsilon, \\ \infty, \quad \text{otherwise}, \end{cases} \\ &\text{for all } \varepsilon \geq 0 \text{ and } 0 < \delta \leq 1. \end{aligned}$$

Similar relaxation methods have been adopted in some previous results (e.g. [4, Key-Lemma 1-2 and Lemma 4.1]), and referring to some of these, we can verify the following facts.

(Fact 1) Let us fix all $\varepsilon > 0$, $0 < \delta \leq 1$, and let us define:

$$\mathscr{D}_\varepsilon^\delta := \left\{ [z, z_\Gamma] \in \mathscr{H} \;\middle|\; \begin{array}{l} \frac{\nabla z}{\sqrt{\|\nabla z\|^2+\delta^2}} + \kappa^2 \nabla z \in L^2_{\mathrm{div}}(\Omega), \\ -\Delta_\Gamma(\varepsilon^2 z_\Gamma) + [(\frac{\nabla z}{\sqrt{\|\nabla z\|^2+\delta^2}} + \kappa^2\nabla z), n_\Gamma]_\Gamma \in L^2(\Gamma; \mathbb{R}^m) \end{array} \right\},$$

and let us define a single-valued operator $\mathcal{A}_\varepsilon^\delta : \mathscr{D}_\varepsilon^\delta \subset \mathscr{H} \to \mathscr{H}$, by letting:

$$\begin{aligned} [z, z_\Gamma] &\in \mathscr{D}_\varepsilon^\delta \mapsto \mathcal{A}_\varepsilon^\delta [z, z_\Gamma] \\ &:= {}^{\mathrm{t}}\begin{bmatrix} -\mathrm{div}(\frac{\nabla z}{\sqrt{\|\nabla z\|^2+\delta^2}} + \kappa^2 \nabla z) \\ -\Delta_\Gamma(\varepsilon^2 z_\Gamma) + [(\frac{\nabla z}{\sqrt{\|\nabla z\|^2+\delta^2}} + \kappa^2\nabla z), n_\Gamma]_\Gamma \end{bmatrix} \in \mathscr{H}. \end{aligned}$$

Then, $\partial\Phi_\varepsilon^\delta \subset \mathscr{H}^2$ coincides with the (graph of) operator $\mathcal{A}_\varepsilon^\delta$, i.e.:

$$\partial\Phi_\varepsilon^\delta = \mathcal{A}_\varepsilon^\delta \ \text{ in } \mathscr{H}^2, \text{ for all } \varepsilon > 0 \text{ and } 0 < \delta \leq 1.$$

(Fact 2) Let $\varepsilon_0 \geq 0$, and let $\{\varepsilon_n\}_{n=1}^\infty \subset [0,\infty)$ and $\{\delta_n\}_{n=1}^\infty \subset (0,1]$ be arbitrary sequences, which fulfill that $\varepsilon_n \to \varepsilon_0$ and $\delta_n \to 0$, as $n \to \infty$. Then, the sequence of convex functions $\{\Phi_{\varepsilon_n}^{\delta_n}\}_{n=1}^\infty$, converges to the convex function Φ_{ε_0} on $\mathscr{H}$, in the sense of Mosco, as $n \to \infty$.

(Fact 3) A sequence of convex functions:

$$W \in L^2(\Omega;\mathbb{R}^{m\times N}) \mapsto \int_\Omega \sqrt{\|W\|^2 + \delta^2}\, dx \in [0,\infty), \text{ for any } 0 < \delta \leq 1$$

converges to the convex function:

$$W \in L^2(\Omega;\mathbb{R}^{m\times N}) \mapsto \int_\Omega \|W\|\, dx \in [0,\infty)$$

on $L^2(\Omega;\mathbb{R}^{m\times N})$, in the sense of Mosco, as $\delta \to 0$.

Finally, in the rest of this section, we give the proof of the Key-Lemma.

Proof of Key-Lemma 1. Let us take a constant $\varepsilon \geq 0$, and let us set:

$$\mathscr{D}_\varepsilon := \left\{ \ [u,u_\Gamma] \in \mathscr{V}_\varepsilon \ \middle|\ \text{there exists } M_u^* \in L^\infty(\Omega;\mathbb{R}^{m\times N}), \text{ such that (4.1)–(4.2)} \ \right\},$$

and let us define a set-valued operator $\mathcal{A}_\varepsilon$, by putting:

$$\begin{aligned}[u,u_\Gamma] \quad &\in \quad \mathscr{D}_\varepsilon \mapsto \mathcal{A}_\varepsilon[u,u_\Gamma] \\ &:= \left\{ \ [u^*,u_\Gamma^*] \in \mathscr{H} \ \middle|\ \begin{array}{l}\text{(4.3) holds, for some } M_u^* \in L^\infty(\Omega;\mathbb{R}^{m\times N}),\\ \text{fulfilling (4.1)–(4.2)}\end{array} \ \right\}.\end{aligned}$$

Then, the assertion of Key-Lemma 1 can be rephrased as follows:

$$\partial\Phi_\varepsilon = \mathcal{A}_\varepsilon \ \text{ in } \mathscr{H}^2, \text{ for any } \varepsilon \geq 0. \tag{4.4}$$

We prove the above (4.4) via the following two Claims.

Claim #1: $\mathcal{A}_\varepsilon \subset \partial\Phi_\varepsilon$ in $\mathscr{H}^2$, for any $\varepsilon \geq 0$.

Let us assume that $[u,u_\Gamma] \in \mathscr{D}_\varepsilon$ and $[u^*,u_\Gamma^*] \in \mathcal{A}_\varepsilon[u,u_\Gamma]$ in $\mathscr{H}$. Then, by (2.1) and the definition of the subdifferential, we can verify that:

$$\begin{aligned}&([u^*,u_\Gamma^*],[z,z_\Gamma]-[u,u_\Gamma])_{\mathscr{H}} \\ &= (-\mathrm{div}(M_u^* + \kappa^2\nabla u), z-u)_{L^2(\Omega;\mathbb{R}^m)} \\ &\qquad + (-\Delta_\Gamma(\varepsilon^2 u_\Gamma) + [(M_u^* + \kappa^2\nabla u), n_\Gamma]_\Gamma, z_\Gamma - u_\Gamma)_{L^2(\Gamma;\mathbb{R}^m)} \\ &= \int_\Omega (M_u^* + \kappa^2\nabla u) : \nabla(z-u)\, dx + \int_\Gamma \nabla_\Gamma(\varepsilon u_\Gamma) : \nabla_\Gamma(\varepsilon(z_\Gamma - u_\Gamma))\, d\Gamma \\ &\leq \int_\Omega \left(\|\nabla z\| + \frac{\kappa^2}{2}\|\nabla z\|^2 \right) dx - \int_\Omega \left(\|\nabla u\| + \frac{\kappa^2}{2}\|\nabla u\|^2 \right) dx \\ &\qquad + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon z_\Gamma)\|^2\, d\Gamma - \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2\, d\Gamma \\ &= \Phi_\varepsilon(z,z_\Gamma) - \Phi_\varepsilon(u,u_\Gamma), \quad \text{for any } [z,z_\Gamma] \in \mathscr{V}_\varepsilon.\end{aligned}$$

Claim #2: $(\mathcal{A}_\varepsilon + \mathcal{I}_{\mathscr{H}})\mathscr{H} = \mathscr{H}$.

It is sufficient to show $(\mathcal{A}_\varepsilon + \mathcal{I}_{\mathscr{H}}) \supset \mathscr{H}$, because the other inclusion is trivial. Let $[w, w_\Gamma] \in \mathscr{H}$ be any pair of functions. Then, owing to (Fact 1) and Minty's Theorem, we can configure a class of functions $\{[u_\delta, u_{\Gamma,\delta}]\}_{0<\delta\le 1} \subset \mathscr{H}$, such that:

$$[u_\delta, u_{\Gamma,\delta}] := (\mathcal{A}_\varepsilon^\delta + \mathcal{I}_{\mathscr{H}})^{-1}[w, w_\Gamma] \text{ in } \mathscr{H}, \text{ for any } 0 < \delta \le 1,$$

and by taking any $[z, z_\Gamma] \in \mathscr{V}_\varepsilon$, we can see that:

$$\begin{aligned} &\int_\Omega \Big(\frac{\nabla u_\delta}{\sqrt{\|\nabla u_\delta\|^2+\delta^2}} + \kappa^2 \nabla u_\delta\Big) : \nabla z \, dx + \int_\Gamma \nabla_\Gamma(\varepsilon u_{\Gamma,\delta}) : \nabla_\Gamma(\varepsilon z_\Gamma)\, d\Gamma \\ &\quad = (w - u_\delta, z)_{L^2(\Omega;\mathbb{R}^m)} + (w_\Gamma - u_{\Gamma,\delta}, z_\Gamma)_{L^2(\Gamma;\mathbb{R}^m)}, \text{ for any } 0 < \delta \le 1. \end{aligned} \tag{4.5}$$

Here, let us put $[z, z_\Gamma] = [u_\delta, u_{\Gamma,\delta}] \in \mathscr{V}_\varepsilon$ in (4.5). Then, by using Young's inequality, we deduce that:

$$|[u_\delta, u_{\Gamma,\delta}]|^2_{\mathscr{H}} + 2\big(\kappa^2 |\nabla u_\delta|^2_{L^2(\Omega;\mathbb{R}^m)} + |\nabla_\Gamma(\varepsilon u_{\Gamma,\delta})|^2_{L^2(\Gamma;\mathbb{R}^m)}\big) \le |[w, w_\Gamma]|^2_{\mathscr{H}} + \delta \mathcal{L}^N(\Omega), \quad \text{for any } 0 < \delta \le 1.$$

The above estimation may suppose that $\{[u_\delta, u_{\Gamma,\delta}]\}_{0<\delta\le 1}$ is bounded in $\mathscr{V}_\varepsilon$, and is compact in $\mathscr{H}$. Therefore, we can find a sequence $\{\delta_n\}_{n=1}^\infty \subset \{\delta\}$ and a pair of functions $[u, u_\Gamma] \in \mathscr{V}_\varepsilon$, such that:

$$[u_n, u_{\Gamma,n}] := [u_{\delta_n}, u_{\Gamma,\delta_n}] \to [u, u_\Gamma] \ \text{ in } \mathscr{H} \text{ and weakly in } \mathscr{V}_\varepsilon, \text{ as } n \to \infty. \tag{4.6}$$

Additionally, since

$$\left| \frac{\nabla u_n}{\sqrt{\|\nabla u_n\|^2+\delta_n^2}} \right| \le 1, \ \text{a.e. in } \Omega, \text{ for any } n \in \mathbb{N},$$

there exists a function $M_u^* \in L^\infty(\Omega; \mathbb{R}^{m\times N})$, such that:

$$\frac{\nabla u_n}{\sqrt{\|\nabla u_n\|^2+\delta_n^2}} \to M_u^*, \ \text{weakly-}* \text{ in } L^\infty(\Omega;\mathbb{R}^{m\times N}), \text{ as } n \to \infty, \tag{4.7}$$

by taking more one subsequence if necessary.

Now, with (4.6)–(4.7) in mind, let us take any function $\varphi_0 \in H_0^1(\Omega;\mathbb{R}^m)$, and let us put $[z, z_\Gamma] = [\varphi_0, 0] \in \mathscr{V}_\varepsilon$ in (4.5). Then, putting $\delta = \delta_n$ with $n \in \mathbb{N}$, and letting $n \to \infty$ in (4.5) yield that:

$$\int_\Omega (M_u^* + \kappa^2 \nabla u) : \nabla \varphi_0 \, dx = (w - u, \varphi_0)_{L^2(\Omega;\mathbb{R}^m)}.$$

It implies that:

$$-\mathrm{div}(M_u^* + \kappa^2 \nabla u) = w - u \in L^2(\Omega;\mathbb{R}^m) \text{ in } \mathcal{D}'(\Omega)^m. \tag{4.8}$$

As well as, putting $\delta = \delta_n$, letting $n \to \infty$ in (4.5) and applying (2.1) and (4.8) lead to:

$$(w_\Gamma - u_\Gamma, z_\Gamma)_{L^2(\Gamma;\mathbb{R}^m)} = \big\langle -\Delta_\Gamma(\varepsilon^2 u_\Gamma) + [(M_u^* + \kappa^2 \nabla u), n_\Gamma]_\Gamma, z_\Gamma \big\rangle_{H^1(\Gamma;\mathbb{R}^m)}, \quad \text{for any } z_\Gamma \in H^1(\Gamma;\mathbb{R}^m).$$

Therefore, we can observe that:

$$-\Delta_\Gamma(\varepsilon^2 u_\Gamma) + [(M_u^* + \kappa^2\nabla u), n_\Gamma]_\Gamma = w_\Gamma - u_\Gamma \in L^2(\Gamma;\mathbb{R}^m) \text{ in } H^{-1}(\Gamma;\mathbb{R}^m). \quad (4.9)$$

Finally, from (Fact 2)–(Fact 3), it is immediately seen that:

$$\begin{cases} \displaystyle\varliminf_{n\to\infty} \int_\Omega \sqrt{\|\nabla u_n\|^2 + \delta_n^2}\, dx \geq \int_\Omega \|\nabla u\|\, dx, \\ \displaystyle\varliminf_{n\to\infty} \left(\frac{\kappa^2}{2} \int_\Omega \|\nabla u_n\|^2\, dx \right) \geq \frac{\kappa^2}{2} \int_\Omega \|\nabla u\|^2\, dx, \\ \displaystyle\varliminf_{n\to\infty} \left(\frac{1}{2} \int_\Gamma \|\nabla_\Gamma(\varepsilon u_{\Gamma,n})\|^2\, d\Gamma \right) \geq \frac{1}{2} \int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2\, d\Gamma. \end{cases} \quad (4.10)$$

Then, by putting $[z, z_\Gamma] = [u_n - u, u_{\Gamma,n} - u_\Gamma] \in \mathscr{V}_\varepsilon$ in (4.5), we can compute that:

$$\begin{aligned} &\int_\Omega \left(\sqrt{\|\nabla u_n\|^2 + \delta_n^2} + \frac{\kappa^2}{2}\|\nabla u_n\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_{\Gamma,n})\|^2\, d\Gamma \\ &\leq \int_\Omega \left(\sqrt{\|\nabla u\|^2 + \delta_n^2} + \frac{\kappa^2}{2}\|\nabla u\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2\, d\Gamma \\ &\qquad + (w - u_n, u_n - u)_{L^2(\Omega;\mathbb{R}^m)} + (w_\Gamma - u_{\Gamma,n}, u_{\Gamma,n} - u_\Gamma)_{L^2(\Gamma;\mathbb{R}^m)}. \end{aligned}$$

Based on these, we take the limit of the above inequality, and infer that:

$$\begin{aligned} &\overline{\varlimsup_{n\to\infty}} \left(\int_\Omega \left(\sqrt{\|\nabla u_n\|^2 + \delta_n^2} + \frac{\kappa^2}{2}\|\nabla u_n\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_{\Gamma,n})\|^2\, d\Gamma \right) \\ &\leq \int_\Omega \left(\|\nabla u\| + \frac{\kappa^2}{2}\|\nabla u\|^2 \right) dx + \frac{1}{2}\int_\Gamma \|\nabla_\Gamma(\varepsilon u_\Gamma)\|^2\, d\Gamma. \end{aligned} \quad (4.11)$$

By virtue of (4.6)–(4.7), (4.10)–(4.11) and the uniform convexity of the L^2-based topologies, it is further deduced that:

$$\nabla u_n \to \nabla u \ \text{ in } L^2(\Omega;\mathbb{R}^{m\times N}), \text{ as } n \to \infty. \quad (4.12)$$

On account of (4.12), (Fact 3), [1, Proposition 3.59 and Theorem 3.66], [3, Proposition 2.16] and [5, Appendix], we can obtain that:

$$M_u^* \in \{\ \tilde{M} \in L^2(\Omega;\mathbb{R}^{m\times N}) \ \big|\ [\nabla u, \tilde{M}] \in \partial\|\cdot\| \text{ in } [\mathbb{R}^{m\times N}]^2, \text{ a.e. in } \Omega\ \}. \quad (4.13)$$

As a consequence of (4.8)–(4.9) and (4.13), we verify Claim #2.

Now, with Claims #1–#2 and the maximality of the subdifferential $\partial\Phi_\varepsilon \subset \mathscr{H}^2$ in mind, we can deduce the coincidence (4.4), and we conclude Key-Lemma 1. □

5. Proofs of Main Theorems. In this section, we will prove the Main Theorems 1–2 on the basis of Key-Lemma 1 and (Fact 1)–(Fact 3) as in the previous sections.

Proof of Main Theorem 1. First, we show the item (I-1). In the Cauchy problem $(\mathrm{CP})_\varepsilon$, let us first confirm that:

$$\Theta := [\theta, \theta_\Gamma] \in L^2(0,T;\mathscr{H}) \text{ and } U_0 := [u_0, u_{\Gamma,0}] \in \overline{D(\Phi_\varepsilon)} = \overline{\mathscr{V}_\varepsilon} = \mathscr{H}.$$

Then, by applying the general theories of evolution equations, e.g. [2, Theorem 4.1], [3, Proposition 3.2], [6, Section 2], and [7, Theorem 1.1.2], we immediately have the existence and uniqueness of the solution $U = [u, u_\Gamma] \in L^2(0,T;\mathscr{H})$ to $(\mathrm{CP})_\varepsilon$, such that:

$$U \in C([0,T];\mathscr{H}) \cap L^2(0,T;\mathscr{H}) \cap W^{1,2}_{\mathrm{loc}}((0,T];\mathscr{H}) \text{ and } \Phi_\varepsilon(U) \in L^1(0,T) \cap L^\infty_{\mathrm{loc}}((0,T]).$$

Now, by Key-Lemma 1, we observe that the solution $U = [u, u_\Gamma]$ to $(\mathrm{CP})_\varepsilon$ coincides with that to the system $(\mathrm{S})_\varepsilon$, and hence, we verify the item (I-1).

In the meantime, the equivalence between $(\mathrm{S})_\varepsilon$ and $(\mathrm{CP})_\varepsilon$ enables us to conclude the other item (I-2) by applying the standard methods for evolution equations: more precisely, by taking the difference between the two evolution equations, multiplying its both sides by the difference of solutions, and using Gronwall's lemma. □

Proof of Main Theorem 2. For any $\varepsilon \geq 0$, let us simply put $\Theta_\varepsilon := [\theta_\varepsilon, \theta_{\Gamma,\varepsilon}] \in L^2(0,T;\mathscr{H})$ and $U_{0,\varepsilon} := [u_{0,\varepsilon}, u_{\Gamma,0,\varepsilon}] \in \mathscr{H}$, and let us denote by U_ε the solution $[u_\varepsilon, u_{\Gamma,\varepsilon}]$ to $(\mathrm{S})_\varepsilon$ corresponding to the forcing term $\Theta_\varepsilon = [\theta_\varepsilon, \theta_{\Gamma,\varepsilon}]$ and the initial data $U_{0,\varepsilon} = [u_{0,\varepsilon}, u_{\Gamma,0,\varepsilon}]$. Then, by the equivalence between $(\mathrm{S})_\varepsilon$ and $(\mathrm{CP})_\varepsilon$, we can apply some of analytic techniques for nonlinear evolution equations, e.g. [7, Theorem 2.7.1] and [4, Main Theorem 2], and we can derive the following convergences:

$$U_\varepsilon \to U_{\varepsilon_0} \text{ in } C([0,T];\mathscr{H}), \int_0^T \Phi_\varepsilon(U_\varepsilon(t))\,dt \to \int_0^T \Phi_{\varepsilon_0}(U_{\varepsilon_0}(t))\,dt, \text{ as } \varepsilon \to \varepsilon_0. \quad (5.1)$$

Now, the required convergences (3.1)–(3.2) will be obtained as straightforward convergences of (5.1) and the uniform convexity of L^2-based topologies. □

REFERENCES

[1] Attouch, H. *Variational Convergence for Functions and Operators.* Applicable Mathematics Series. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[2] Barbu, V. *Nonlinear Differential Equations of Monotone Types in Banach Spaces.* Springer Monographs in Mathematics. Springer, New York, 2010.

[3] Brézis, H. *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert.* North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).

[4] Colli, P.; Gilardi, G.; Nakayashiki, R.; Shirakawa, K. A class of quasi-linear Allen–Cahn type equations with dynamic boundary conditions. *Nonlinear Anal.*, **158**:32–59, 2017.

[5] Giga, Y.; Kashima, Y.; Yamazaki, N. Local solvability of a constrained gradient system of total variation. *Abstr. Appl. Anal.*, (8):651–682, 2004.

[6] Ito, A.; Yamazaki, N.; Kenmochi, N. Attractors of nonlinear evolution systems generated by time-dependent subdifferentials in Hilbert spaces. *Discrete Contin. Dynam. Systems*, (Added Volume I):327–349, 1998. Dynamical systems and differential equations, Vol. I (Springfield, MO, 1996).

[7] Kenmochi, N. Solvability of nonlinear evolution equations with time-dependent constraints and applications. *Bull. Fac. Education, Chiba Univ. http://ci.nii.ac.jp/naid/110004715232*, **30**:1–87, 1981.

[8] Kenmochi, N. Pseudomonotone operators and nonlinear elliptic boundary value problems. *J. Math. Soc. Japan*, **27**:121–149, 1975.

[9] Mosco, U. Convergence of convex sets and of solutions of variational inequalities. *Advances in Math.*, **3**:510–585, 1969.

[10] Nakayashiki, R.; Shirakawa, K. Weak formulation for singular diffusion equations with dynamic boundary condition. Springer INdAM Series. to appear, 2017.

[11] Savaré, G.; Visintin, A. Variational convergence of nonlinear diffusion equations: applications to concentrated capacity problems with change of phase. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl.*, **8**(1):49–89, 1997.

Proceedings of EQUADIFF 2017
pp. 221–228

REMARKS ON THE QUALITATIVE BEHAVIOR OF THE UNDAMPED KLEIN-GORDON EQUATION *

JORGE A. ESQUIVEL-AVILA †

Abstract. We present sufficient conditions on the initial data of an undamped Klein-Gordon equation in bounded domains with homogeneous Dirichlet boundary conditions to guarantee the blow up of weak solutions. Our methodology is extended to a class of evolution equations of second order in time. As an example, we consider a generalized Boussinesq equation. Our result is based on a careful analysis of a differential inequality. We compare our results with the ones in the literature.

Key words. Klein-Gordon equation, Blow up, High energies, Abstract wave equation, Generalized Boussinesq equation

AMS subject classifications. 35L70, 35B35, 35B40

1. Functional framework and previous results. For the Cauchy problem associated to any evolution equation on a Banach space, we have the usual questions in terms on the initial data:

- Existence and uniqueness of solutions.
- Non global existence: maximal time of existence $\equiv T_{MAX} < \infty$.
- Global existence: $T_{MAX} = \infty$.
- In the latter case, the behavior of the solution as time approaches infinity.

Here, we present a short overview paper presenting recent advances, published in [1, 2], on the non global existence of solutions corresponding to a non-linear Klein-Gordon equation and to abstract wave equations, in particular to a generalized Boussinesq equation.

We shall first consider the following problem for a Klein-Gordon equation

$$(\mathbf{P}) \begin{cases} u_{tt}(x,t) - \Delta u(x,t) + m^2 u(x,t) = f(u(x,t)), & (x,t) \in \Omega \times (0,T), \\ u(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T), \\ u(x,0) = u_0(x),\ u_t(x,0) = v_0(x), & x \in \Omega. \end{cases}$$

where $m \neq 0$ is a real constant, which is assumed to be equal to one without loss of generality, and $\Omega \subset R^n$ is a bounded and open set with sufficiently smooth boundary. We assume that the source term f, is locally Lipschitz continuous and satisfies

$$|f(s)| \leq \mu |s|^{r-1}, \ \ sf(s) - rF(s) \geq 0, \ \forall s \in R,$$

where $F(s) \equiv \int_0^s f(t)dt$, and $r > 2, \mu > 0$, are constants. For this problem, Ball [3, 4] proved the following theorem about existence, uniqueness and continuation of weak solutions.

THEOREM 1.1. *Assume that $r \leq 2(n-1)/(n-2)$ if $n \geq 3$. For every initial data $(u_0, v_0) \in H \equiv H_0^1(\Omega) \times L_2(\Omega)$, there exists a unique (local) weak solution*

*This work was supported by UAM Azacpotzalco

†Departamento de Ciencias Básicas, Análisis Matemático y sus Aplicaciones, UAM-Azcapotzalco, Av. San Pablo 180, Col. Reynosa Tamaulipas, 02200 México, D. F., México (jaea@correo.azc.uam.mx).

$(u_0, v_0) \mapsto (u(t), v(t))$, $v(t) \equiv \frac{d}{dt}u(t)$, *of problem* **(P)**, *that is*

$$\frac{d}{dt}(v(t), w)_2 + (\nabla u(t), \nabla w)_2 + (u(t), w)_2 = (f(u(t)), w)_2,$$

a. e. in (0,T) and for every $w \in H_0^1(\Omega)$, *such that* $(u, v) \in C([0, T); H)$. *Here,* $(\cdot, \cdot)_2$ *denotes the inner product in* $L_2(\Omega)$. *Furthermore, the following energy equation holds*

$$E(u(t_0), v(t_0)) = E(u(t), v(t)) \equiv \frac{1}{2}\|v(t)\|_2^2 + J(u(t)), \quad \forall t \geq t_0 \geq 0,$$
$$J(u(t)) \equiv \frac{1}{2}\left(\|u(t)\|_2^2 + \|\nabla u(t)\|_2^2\right) - \int_\Omega F(u(t))dx.$$

Here, $\|\cdot\|_q$ *is the norm in* $L_q(\Omega)$. *Finally, if the maximal time of existence* $T_{MAX} < \infty$, *then the solution blows up in finite time. That is,*

$$\lim_{t \nearrow T_{MAX}} \|(u(t), v(t))\|_H^2 \equiv \lim_{t \nearrow T_{MAX}} \|u(t)\|_2^2 + \|\nabla u(t)\|_2^2 + \|v(t)\|_2^2 = \infty.$$

Moreover, by the energy equation,

$$\lim_{t \nearrow T_{MAX}} \|u(t)\|_r = \infty.$$

REMARK 1. *Problem* **(P)** *is invariant if we reverse the time direction:* $t \mapsto -t$. *The solution backwards* $(u(t), v(t)), t < 0$, *with initial conditions* (u_0, v_0) *corresponds to the solution forwards* $(u(-t), -v(-t)), -t > 0$ *with initial conditions* $(u_0, -v_0)$. *Then, the local existence and uniqueness Theorem 1.1 holds backwards and the results presented in this work for positive times have the corresponding for backwards solutions.*

If the solution u is independent of time, it is called an equilibrium and satisfies

$$(\nabla u, \nabla w)_2 + (u, w)_2 = (f(u), w)_2,$$

for every $w \in H_0^1(\Omega)$. In particular, for $w = u$,

$$I(u) \equiv \|\nabla u\|_2^2 + \|u\|_2^2 - (f(u), u)_2 = 0.$$

We notice that $u = 0$ is an equilibrium. The set of equilibria $u \neq 0$, with minimal energy is called the ground state, and the corresponding value of the energy is positive and denoted by d. This number is the mountain pass level of the functional J, see [5]. For initial energies $E(u_0, v_0) < d$, a characterization of the qualitative properties of the solutions in terms of the sign of $I(u_0)$ has been proved in [6] by the potential well method. Indeed, if $I(u_0) \geq 0$, respectively $I(u(t_0)) < 0$, the corresponding solution is global and uniformly bounded in H, respectively the solution blows up in finite time. For high values of the initial energy the sign of $I(u_0)$ is not sufficient in order to prove qualitative properties of the solution. Certainly, for $E(u_0, v_0) > d$ and for a source term of the form $f(u) \equiv |u|^{r-2}u, r > 2$, in [7] is proved that the solution blows up if $I(u_0) < 0$, $(u_0, v_0)_2 \geq 0$, and $\|u_0\|_2 \geq \sup\{\|u\|_2 : I(u)) = 0, J(u) \leq E(u_0, v_0)\}$. For $E(u_0, v_0) = d$ and $f(u) \equiv |u|^{r-2}u$, in [8] the following is proved: (i) the solution blows up if $I(u_0) < 0$ and $(u_0, v_0)_2 \geq 0$ and (ii) the solution is global and uniformly bounded in H if $I(u_0) > 0$. Recently, several works have proved blow up of solutions with one o several source terms of the form $|u|^{r-2}u$, under sufficient conditions that involve upper bounds on the initial energy as follows.

THEOREM 1.2. *For every solution of problem* **(P)** *with*

$$(u_0, v_0)_2 > 0, \|u_0\|_2 > 0,$$

the solution blows up in finite time if one of the following holds:

$$(Wang\ [9]). \qquad E(u_0, v_0) \leq \frac{r-2}{2r}\|u_0\|_2^2, \ \ I(u_0) < 0, \tag{1.1}$$

$$(Korpusov\ [10]). \qquad E(u_0, v_0) < \frac{1}{2}P(u_0, v_0), \tag{1.2}$$

$$(Kutev, et\ al.\ [11]). \qquad E(u_0, v_0) < \frac{r-2}{2r}\|u_0\|_2^2 + \frac{1}{2}P(u_0, v_0), \tag{1.3}$$

$$(Dimova, et\ al.\ [12]). \qquad E(u_0, v_0) \leq \frac{r-2}{2r}\|u_0\|_2^2 + \frac{1}{2}P(u_0, v_0)$$

$$+ \frac{\|u_0\|_2^2}{r}\left[1 - \left\{1 + \frac{P(u_0, v_0)}{\|u_0\|_2^2}\right\}^{-(\frac{r-2}{2})}\right], \tag{1.4}$$

where $P(u_0, v_0) \equiv \frac{|(v_0, u_0)_2|^2}{\|u_0\|_2^2}$.

REMARK 2. *For the proof of anyone of the items in this theorem, some differential inequality is employed to prove that the solution only exists up to a finite time:* $T < \infty$. *The estimate of the maximal time of existence by this means is not always optimal, that is,* $T > T_{MAX}$. *See [13, 3, 4] for more discussion. The technique described above belongs to the so called functional method. That is, some functional in terms of a norm of the solution well defined in the sense of Theorem 1.1, satisfies a differential inequality that necessarily implies that such norm blows up in finite time. Consequently, the solution can not be global. This method has been used for many authors to show nonexistence of solutions, see for instance [14] for an early reference where a concavity argument is used.*

REMARK 3. *In [11] is proved that any one of the sufficient conditions (1.1) or (1.2) imply (1.3), and that the contrary is not true. We notice that (1.3) implies (1.4) but the opposite does not occur. In next section we easily show this implication and by this means we propose a new condition to get blow up of the solution in finite time.*

2. Main result. In this section we consider solutions with any positive value of the initial energy, in particular with $E(u_0, v_0) \geq d$. The understanding of the complete dynamics in this case is an open question and very much complicated. Here, we limit ourselves to study blow up and give sufficient conditions on $(u_0, v_0) \in H$ and $E(u_0, v_0) > 0$.

We first notice that the right hand-side of (1.3) and (1.4) have the following form

$$\eta_q(u, v) \equiv \frac{1}{2}\Phi(u, v) - \frac{1}{r}\Psi(u)\left(\frac{\Psi(u)}{\Phi(u, v)}\right)^q,$$

where $q \geq 0$ and

$$\Phi(u, v) \equiv \Psi(u) + P(u, v), \qquad \Psi(u) \equiv \|u\|_2^2, \qquad P(u, v) \equiv \frac{|(v, u)_2|^2}{\|u\|_2^2}.$$

The functional P comes from the orthogonal decomposition of the velocity, introduced in [11]. That is,

$$v = \frac{(v, u)_2}{\|u\|_2^2}u + h, \qquad \|v\|_2^2 = \|h\|_2^2 + P(u, v),$$

where $(u,h)_2 = 0$. Indeed, the one in (1.4) is equal to $\eta_{\frac{r-2}{2}}(u_0,v_0)$. We notice that the function $q \mapsto \eta_q(u_0,v_0)$, is strictly increasing for $q \geq 0$, whenever $P(u_0,v_0) > 0$, and that $\eta_0(u_0,v_0)$ is equal to the right-hand side of condition (1.3). Hence, (1.4) is implied by (1.3) but not the contrary. Now, we define a strictly decreasing function $\lambda \mapsto \mu_\lambda(u,v)$, for $0 < \lambda < 1$, by

$$\mu_\lambda(u,v) \equiv \frac{1}{2}\Phi(u,v) - \frac{1}{r}\Psi(u,v)\left(\frac{r-2}{r-2\lambda}\,\frac{\Psi(u)}{\Phi(u,v)}\right)^{\frac{r-2}{2}},$$

with the property that $\mu_\lambda(u,v) \to \eta_{\frac{r-2}{2}}(u,v)$ if $\lambda \to 1$. That is, $\eta_{\frac{r-2}{2}}(u,v) < \mu_\lambda(u,v)$.

Next we present our blow up result whose proof is based on a careful analysis of a differential inequality satisfied by $\Psi(u)$, and where $P(u,v)$ and $\mu_\lambda(u,v)$ are essential to improve the previous results given in Theorem 1.2.

THEOREM 2.1. *(Esquivel-Avila [1]). Consider any solution of problem* **(P)**. *Assume that*

$$\|u_0\|_2 > 0,\ (u_0,v_0)_2 > 0. \tag{2.1}$$

Hence, $P(u_0,v_0) > 0$, and there exists a nonempty interval

$$\mathcal{I}_{P(u_0,v_0)} \equiv \left(\alpha_{P(u_0,v_0)}, \beta_{P(u_0,v_0)}\right) \subset \left(0, \frac{1}{2}\Phi(u_0,v_0)\right),$$

such that if $E(u_0,v_0) \in \mathcal{I}_{P(u_0,v_0)}$, then the solution blows up in finite time. Moreover, for fixed $\Psi(u_0)$,

$$\lim_{P(u_0,v_0)\to\infty}\left|\beta_{P(u_0,v_0)} - \frac{1}{2}\Phi(u_0,v_0)\right| = 0 = \lim_{P(u_0,v_0)\to\infty}\alpha_{P(u_0,v_0)}.$$

REMARK 4. *We observe that $\beta_{P(u_0,v_0)} = \mu_{\lambda^*}(u_0,v_0)$, where $\lambda^* \in (0,1)$, is uniquely defined by*

$$\lambda^* \equiv \left(\frac{\Psi(u_0)}{\Phi(u_0,v_0)}\right)^{\frac{r}{2}}\left(\frac{r-2}{r-2\lambda^*}\right)^{\frac{r-2}{2}}.$$

Hence, Theorem 2.1 improves the condition on the upper bound of the initial energy given in Theorem 1.2, (1.1)-(1.4).

If $\mu_{\lambda^*}(u_0,v_0) \leq E(u_0,v_0) \leq \mu_\lambda(u_0,v_0)$, for $\lambda \leq \lambda^*$, the qualitative behavior of the solution is unknown. However, given any positive value of the initial energy, if (2.1) holds and $P(u_0,v_0)$ is large enough, then we can always have that $E(u_0,v_0) \in \mathcal{I}_{P(u_0,v_0)}$. Consequently, the corresponding solution blows up in finite time.

REMARK 5. *For small energies, the result in [6] characterizes blow up of any solution under the condition $I(u_0) < 0$. For high energies, blow up follows from $I(u_0) < 0$ and additional conditions on the initial data, see [7]-[9]. Under the hypotheses of Theorem 2.1, $I(u_0) < 0$ follows if $P(u_0,v_0) > 0$ is sufficiently large, see [1].*

3. Evolution equations of second order in time. We extend Theorem 2.1 to the following class of abstract wave equations:

$$(\mathbf{P_M}) \begin{cases} Mu_{tt}(t) + Au(t) = \mathcal{F}(u(t)), & t \in (0,T), \\ u(0) = u_0,\ u_t(0) = v_0, \end{cases}$$

where $M : H_{\mathcal{M}} \to H_{\mathcal{M}}^{'}$ and $A : V \to V^{'}$, are linear, positive and symmetric operators, and $V \subset H_{\mathcal{M}} \subset H$ are linear subspaces of the Hilbert space H with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, and $H = H^{'} \subset H_{\mathcal{M}}^{'} \subset V'$ are the dual spaces. Hence, we define the bilinear forms and corresponding inner products and norms

$$\mathcal{M} : H_{\mathcal{M}} \times H_{\mathcal{M}} \to R, \quad \mathcal{M}(u,w) \equiv (Mu,w)_{H_{\mathcal{M}} \times H'_{\mathcal{M}}},$$
$$(u,w)_{\mathcal{M}} \equiv \mathcal{M}(u,w), \quad \|u\|^2_{\mathcal{M}} \equiv (u,u)_{\mathcal{M}}, \quad \forall u, w \in H_{\mathcal{M}}$$

and

$$\mathcal{A} : V \times V \to R, \quad \mathcal{A}(u,w) \equiv (Au,w)_{V \times V'},$$
$$(u,w)_V \equiv \mathcal{A}(u,w), \quad \|u\|^2_V \equiv (u,w)_V, \quad \forall u, w \in V.$$

We assume that there exists some constant $c > 0$, such that

$$\|u\|^2_V \geq c\|u\|^2_{\mathcal{M}},\ \forall u \in V. \tag{3.1}$$

Also, we assume that the nonlinear term $\mathcal{F} : V \subset H \to H$, is a potential operator with potential $\mathcal{G} : V \to R$, and

$$\mathcal{F}(0) = 0,\ \ (\mathcal{F}(u),u) - r\mathcal{G}(u) \geq 0,\ \forall u \in V, \tag{3.2}$$

where $r > 2$ is a constant.

We consider solutions in the following functional framework.

For every initial data $(u_0, v_0) \in \mathcal{H} \equiv V \times H_{\mathcal{M}}$, there exists $T > 0$, and a unique local solution $(u_0, v_0) \mapsto (u, v) \in C([0,T); \mathcal{H}),\ v(t) \equiv \frac{d}{dt}u(t)$, of the problem $(\mathbf{P_M})$ in the following sense

$$\frac{d}{dt}\mathcal{M}(v(t),w) + \mathcal{A}(u(t),w) = (\mathcal{F}(u(t)),w),$$

a. e. in $(0,T)$ and for every $w \in V$. Furthermore, the following energy equation holds

$$E(u(t_0),v(t_0)) = E(u(t),v(t)) \equiv \frac{1}{2}\|v(t)\|^2_{\mathcal{M}} + J(u(t)),\ \ t \in [t_0,T),\ t_0 \geq 0,$$
$$J(u(t)) \equiv \frac{1}{2}\|u(t)\|^2_V - \mathcal{G}(u(t)).$$

We define

$$\Phi(u,v) \equiv c\Psi(u) + P_{\mathcal{M}}(u,v),\ \Psi(u) \equiv \|u\|^2_{\mathcal{M}},\ P_{\mathcal{M}}(u,v) \equiv \frac{|\mathcal{M}(v,u)|^2}{\|u\|^2_{\mathcal{M}}}.$$

Then, we have the following result.

THEOREM 3.1. *(Esquivel-Avila [2]). Consider any solution of problem* $(\mathbf{P_M})$. *Assume that*

$$\|u_0\|_{\mathcal{M}} > 0, \quad \mathcal{M}(u_0,v_0) > 0. \tag{3.3}$$

Then, there exists a nonempty open interval

$$\mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)} \equiv \left(\alpha_{P_{\mathcal{M}}(u_0,v_0)}, \beta_{P_{\mathcal{M}}(u_0,v_0)}\right) \subset \left(0, \frac{1}{2}\Phi(u_0,v_0)\right),$$

such that if $E(u_0,v_0) \in \mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)}$, *then the solution is not global. Moreover, for fixed* $\Psi(u_0)$,

$$\lim_{P_{\mathcal{M}}(u_0,v_0)\to\infty} \left|\beta_{P_{\mathcal{M}}(u_0,v_0)} - \frac{1}{2}\Phi(u_0,v_0)\right| = 0 = \lim_{P_{\mathcal{M}}(u_0,v_0)\to\infty} \alpha_{P_{\mathcal{M}}(u_0,v_0)}.$$

Here, $\beta_{P_{\mathcal{M}}(u_0,v_0)} = \mu_{\lambda^*}(u_0,v_0)$, *where* $\lambda^* \in (0,1)$ *is uniquely defined by*

$$\lambda^* \equiv \left(\frac{c\Psi(u_0)}{\Phi(u_0,v_0)}\right)^{\frac{r}{2}} \left(\frac{r-2}{r-2\lambda^*}\right)^{\frac{r-2}{2}},$$

and

$$\mu_\lambda(u_0,v_0) \equiv \frac{1}{2}\Phi(u_0,v_0) - \frac{c}{r}\Psi(u_0,v_0)\left(\frac{r-2}{r-2\lambda}\,\frac{c\Psi(u_0)}{\Phi(u_0,v_0)}\right)^{\frac{r-2}{2}}.$$

Furthermore, given any positive value of the initial energy we can always find initial data satisfying (3.3) with $P_{\mathcal{M}}(u_0,v_0)$ *sufficiently large so that* $E(u_0,v_0) \in \mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)}$ *and hence the corresponding solution exists only up to a finite time.*

We can apply Theorem 3.1 to several problems, in particular here we present the following Cauchy problem associated to a generalized Boussinesq equation.

$$(\mathbf{P_B}) \begin{cases} u_{tt}(x,t) - \beta_1\Delta u(x,t) - \beta_2\Delta u_{tt}(x,t) + \beta_3\Delta^2 u(x,t) & \\ \qquad\qquad +mu(x,t) + \Delta\mathcal{F}(u(x,t)) = 0, & (x,t) \in R^n \times (0,T), \\ u(x,0) = u_0(x),\ u_t(x,0) = v_0(x), & x \in R^n, \end{cases}$$

where $\beta_i > 0,\ i = 1,2,3,\ m > 0$ are constants and the source term, that satisfies (3.2), is

$$\mathcal{F}(u) \equiv \mu|u|^{r-2}u,\ \mu > 0, r > 2.$$

Applying $(-\Delta)^{-1}$ to the equation above, we obtain the form of the problem $(\mathbf{P_M})$, where we identify the operators

$$Mu = ((-\Delta)^{-1} + \beta_2 I_d)u, \quad Au = (-\beta_3\Delta + m(-\Delta)^{-1} + \beta_1 I_d)u,$$

and the spaces

$$H = L_2(R^n), \quad H_{\mathcal{M}} = \{u \in L_2(R^n) : (-\Delta)^{-\frac{1}{2}}u \in L_2(R^n)\},$$

and

$$V = \{u \in H^1(R^n) : (-\Delta)^{-\frac{1}{2}}u \in L_2(R^n)\}.$$

If

$$(u,w)_* \equiv ((-\Delta)^{-\frac{1}{2}}u, (-\Delta)^{-\frac{1}{2}}w)_2, \quad \|u\|_*^2 \equiv (u,u)_*,$$

then the bilinear forms, inner products and norms are

$$(u,w)_{\mathcal{M}} \equiv \mathcal{M}(u,w) \equiv (u,w)_* + \beta_2(u,w)_2, \quad \|u\|^2_{\mathcal{M}} \equiv (u,u)_{\mathcal{M}},$$

and

$$(u,w)_V \equiv \mathcal{A}(u,w) \equiv \beta_3(\nabla u, \nabla w)_2 + m(u,w)_* + \beta_1(u,w)_2, \quad \|u\|^2_V \equiv (u,u)_V.$$

Hence, (3.1) holds with $c \equiv \min\{m, \frac{\beta_1}{\beta_2}\}$. Fortunately, there exists an existence and uniqueness result in our functional framework and nonexistence of global solutions is due to blow up, see for instance [15, 16]. Then, by Theorem 3.1, if the initial data satisfy

$$\|u_0\|^2_* + \beta_2\|u_0\|^2_2 > 0, \ (u_0,v_0)_* + \beta_2(u_0,v_0)_2 > 0, \tag{3.4}$$

and the initial energy is such that $E(u_0,v_0) \in \mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)}$, where

$$E(u,v) \equiv \frac{1}{2}\left(\|v\|^2_* + \beta_2\|v\|^2_2 + \beta_3\|\nabla u\|^2_2 + m\|u\|^2_* + \beta_1\|u\|^2_2\right) - \frac{\mu}{r}\|u\|^r_r,$$

then the solution blows up in finite time in the norm of $\mathcal{H}$ and, by the energy equation, also in the $L_r(R^n)$ norm. This result improves the ones known in the literature in the following sense. In [17, 18] blow up is proved by means of the analysis of a differential inequality and by the construction of invariant sets, if (3.4) holds and the initial energy is such that

$$E(u_0,v_0) \le \eta_0(u_0,v_0) \equiv \frac{r-2}{2r}c\left(\|u_0\|^2_* + \beta_2\|u_0\|^2_2\right) + \frac{1}{2}\frac{|(u_0,v_0)_* + \beta_2(u_0,v_0)_2|^2}{\|u_0\|^2_* + \beta_2\|u_0\|^2_2}.$$

We notice that $\eta_0(u_0,v_0) = \frac{1}{2}\Phi(u_0,v_0) - \frac{c}{r}\Psi(u_0,v_0) \in \mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)}$. Then, Theorem 3.1 agrees with the result in [17, 18] and states that blow up occur even for larger initial energies, that is, if

$$\eta_0(u_0,v_0) < E(u_0,v_0) < \mu_{\lambda^*}(u_0,v_0).$$

Moreover, given any positive value of the initial energy there exist initial data satisfying (3.4) and with

$$P_{\mathcal{M}}(u_0,v_0) \equiv \frac{|(u_0,v_0)_* + \beta_2(u_0,v_0)_2|^2}{\|u_0\|^2_* + \beta_2\|u_0\|^2_2},$$

sufficiently large, so that $E(u_0,v_0) \in \mathcal{I}_{P_{\mathcal{M}}(u_0,v_0)}$ holds and consequently the corresponding solution blows up in finite time.

REMARK 6. *For each concrete example of* $(\mathbf{P_M})$*, if the potential well method is applicable as it is in* $(\mathbf{P})$*, then there are conditions to get blow up when* $E(u_0,v_0) < d$*. Theorem 3.1 gives sufficient conditions for* $\alpha_{P_{\mathcal{M}}(u_0,v_0)} < E(u_0,v_0) < \beta_{P_{\mathcal{M}}(u_0,v_0)}$*. In case that* $E(u_0,v_0) \le \alpha_{P_{\mathcal{M}}(u_0,v_0)}$ *the blow up problem is resolved as follows. (i) If* $E(u_0,v_0) < \min\{\alpha_{P_{\mathcal{M}}(u_0,v_0)}, d\}$*, by the potential well method. (ii) If* $d \le E(u_0,v_0) \le \alpha_{P_{\mathcal{M}}(u_0,v_0)}$*, by the techniques in [17, 18].*

Acknowledgements

I thank the referee and the editor for their valuable suggestions that contributed to the final form of this work.

REFERENCES

[1] Esquivel-Avila, J., *Remarks on the qualitative behavior of the undamped Klein-Gordon equation*, Math. Meth. Appl. Sci. (2017) 9 pp., DOI: 10.1002/mma.4598.

[2] Esquivel-Avila, J., *Nonexistence of global solutions of abstract wave equations with high energies*, Journal of Inequalities and applications, 2017:268 (2017) 14 pp., DOI 10.1186/s13660-017-1546-1.

[3] Ball J., *Finite blow up in nonlinear problems*, in Nonlinear Evolution Equations, M. G. Crandall Editor, Academic Press, 1978, pp. 189-205.

[4] Ball J., *Remarks on blow up and nonexistence theorems for nonlinear evolution equations*, Quart. J. Math. Oxford 28 (1977) 473-486.

[5] Willem, M., Minimax Theorems, Progress in Nonlinear Differential Equations and Applications, Vol. 24, Birkhäuser, 1996.

[6] Payne, L. E., D. H. Sattinger, *Saddle points and instability of nonlinear hyperbolic equations*, Israel J. Math. 22 (1975) 273-303.

[7] Gazzola, F., M. Squassina, *Global solutions and finite time blow up for damped semilinear wave equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire 23 (2006) 185-207.

[8] Esquivel-Avila, J., *Blow up and asymptotic behavior in a nondissipative nonlinear wave equation*, Appl. Anal. 93 (2014) 1963-1978.

[9] Wang, Y., *A sufficient condition for finite time blow up of the nonlinear Klein-Gordon equations with arbitrary positive initial energy*, Proc. Amer. Math. Soc. 136 (2008) 3477-3482.

[10] Korpusov, M. O., *Blowup of a positive-energy solution of model wave equations in nonlinear dynamics*, Theoret. and Math. Phys. 171 421-434 (2012).

[11] Kutev, N., N. Kolkovska, M. Dimova, *Sign-preserving functionals and blow up to Klein-Gordon equation with arbitrary high energy*, Appl. Anal. 95 (2016) 860-873.

[12] Dimova, M, N. Kolkovska, N. Kutev, *Revised concavity method and application to Klein-Gordon equation*, Filomat 30 (2016) 831-839.

[13] Alinhac, S., Blow up for nonlinear hyperbolic equations, Progress in Nonlinear Differential Equations and Applications 17, Birkhäuser, 1995.

[14] Levine, H.A., *Instability and nonexistence of global solutions to nonlinear wave equations of the form* $Pu_{tt} = -Au + \mathcal{F}(u)$. Trans. Am. Math. Soc. 192 (1974) 1-21.

[15] Wang, S., H. Xuek, *Global solution for a generalized Boussinesq equation*, Appl. Math. Comput. 204 (2008) 130-136.

[16] Xu, R., Y. Liu, *Global existence and nonexistence of solution for Cauchy problem of multidimensional double dispersion equations*, J. Math. Anal. Appl. 359 (2009) 739-751.

[17] Kutev, N., N. Kolkovska, M. Dimova, *Nonexistence of global solutions to new ordinary differential inequality and applications to nonlinear dispersive equations*, Math. Meth. Appl. Sci. 39 (2016) 2287-2297.

[18] Kutev, N., N. Kolkovska, M. Dimova, *Finite time blow up of the solutions to Boussinesq equation with linear restoring force and arbitrary positive energy*, Acta Math. Scientia 36B (2016) 881-890.

Proceedings of EQUADIFF 2017
pp. 229–236

TWO APPROACHES FOR THE APPROXIMATION OF THE NONLINEAR SMOOTHING TERM IN THE IMAGE SEGMENTATION *

MATÚŠ TIBENSKÝ † AND ANGELA HANDLOVIČOVÁ ‡

Abstract. Purpose of the paper is to study nonlinear smoothing term initiated in [3], [4], [6] and [7] for problems of image segmentation and missing boundaries completion. The generalization of approach presented in [1] is proposed and applied in the field of image segmentation. So called regularised Riemannian mean curvature flow equation is studied and the construction of the numerical scheme based on the finite volume method approach is explained. The principle of the level set, for the first time given in [2], is used. We mention two different approaches for the approximation of the nonlinear smoothing term in the equation and known theoretical results for both of them. We provide the numerical tests for both schemes. It the last section we discuss obtained results and propose possibilities for the future research.

Key words. image segmentation, level set, regularised Riemannian mean curvature flow equation, finite volume method, approximation of the nonlinear smoothing term

1. Introduction. The main goal of the image segmentation is to divide given image to the parts called regions, to identify the pixels segmented object contains of or to add the boundary to the object, where it is missing. The errors we have to face with are mainly missing boundaries and noise. The range of application areas is wide and contains medicine, traffic control systems, recognition tasks and overall object detection and computer vision.

There are lot of techniques used in segmentation based on the various principles as statistical analysis, graph theory or machine learning. We are considering the approach based on the partial differential equations and especially so called level set methods based on the level set function introduced in [2].

2. Studied equation and assumptions on the data. We consider following problem arising in image segmentation as a generalisation of the approach given in [1], find an approximate solution to the equation

$$u_t - f_1(|\nabla u|)\nabla \cdot \left(g(|\nabla G_S * I^0|)\frac{\nabla u}{f(|\nabla u|)} \right) = r, \quad a.e.\ (x,t) \in \Omega \times (0,T). \tag{2.1}$$

Here the $u(x,t)$ is an unknown (segmentation) function defined in $Q_T \equiv [0,T] \times \Omega$, where Ω is bounded rectangular domain, $[0,T]$ is a time interval and I^0 is a given image, typically on this image is an object we want to segment.

We consider zero Dirichlet boundary condition

$$u = 0, \quad a.e.\ (x,t) \in \partial\Omega \times [0,T] \tag{2.2}$$

and initial condition

$$u(x,0) = u_0(x), \quad a.e.\ x \in \Omega. \tag{2.3}$$

*This work was supported by grants APVV 15-0522 and VEGA 1/0728/15.

†Dpt. of Mathematics, Slovak University of Technology in Bratislava, Radlinského 11, 810 05 Bratislava, Slovakia, (matus.tibensky@stuba.sk).

‡Dpt. of Mathematics, Slovak University of Technology in Bratislava, Radlinského 11, 810 05 Bratislava, Slovakia, (angela.handlovicova@stuba.sk).

The assumptions on the data in (2.1)-(2.3) are similar as in [1] and [3]. We can summarize them into the following hypothesis:

Hypothesis H

- (H1) Ω is a finite connected open subset of $\mathbb{R}^d$, $d \in \mathbb{N}$, with boundary $\partial\Omega$,
- (H2) $u_0 \in L^\infty(\Omega)$,
- (H3) $r \in L^2(\Omega \times (0,T))$ for all $T > 0$,
- (H4) $f \in C^0(\mathbb{R}_+; [a,b])$ is a Lipschitz continuous (non-strictly) increasing function, such that the function $x \mapsto x/f(x)$ is strictly increasing on $\mathbb{R}_+$. For practical application we are using $f(s) = \min(\sqrt{s^2+a^2}, b)$, where a and b are given positive parameters,
- (H5) $f_1 \in C^0(\mathbb{R}_+; [a_1, b_1])$, in general $a_1 \neq a$, $b_1 \neq b$, but for now in our model we consider the case $a_1 = a$ and $b_1 = b$,
- (H6) $g \in C^0(\mathbb{R}_+; [0,1])$ is decreasing function, $g(0) = 1$, $g(s) \to 0$ for $s \to \infty$. For practical numerical computation we use $g(s) = \frac{1}{1+Ks^2}$, where K is constant of sensitivity of function g and we choose it,
- (H7) $G_S \in C^\infty(\mathbb{R}^d)$ is a smoothing kernel (Gauss function), with width of the convolution mask S and such that $\int_{\mathbb{R}^d} G_S(x)dx = 1$, $\int_{\mathbb{R}^d} |G_S|dx \leq C_S, C_S \in \mathbb{R}$, $G_S(x) \to \delta_x$ for $S \to 0$, where δ_x is Dirac measure at point x and

$$(\nabla G_S * I^0)(x) = \int_{\mathbb{R}^d} \nabla G_S(x-\xi)\tilde{I^0}(\xi)d\xi, \tag{2.4}$$

where $\tilde{I^0}$ is extension of image I^0 to $\mathbb{R}^d$ given by periodic reflection through boundary of Ω and for which

$$1 \geq g^S(x) = g(|\nabla G_S * I^0|)(x) \geq \nu_S > 0 \tag{2.5}$$

for $\forall x \in \Omega$ due to properties of the convolution. The ν_S is a constant depending only on width of the convolution mask S.

Definition of the numerical scheme and the space discretisation of the equation we are generalising in this paper could be found in [1]. We apply method presented in [1] in the field of image segmentation, but in addition we have function g and convolution of the initial image with smoothing kernel in our approach (see [3] or [4]). For now just remark that discretisation of Ω, denoted by $\mathcal{D}$, is defined as the triplet $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where $\mathcal{M}$ is a finite family of non-empty connected open disjoint subsets of Ω (the "control volumes") with measure marked by $|p|$, $\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh) with measure marked by $|\sigma|$ and $\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (x_p)_{p\in\mathcal{M}}$, such that for all $p \in \mathcal{M}$, $x_p \in p$ and p is assumed to be x_p-star-shaped so for all $x \in p$ the inclusion $[x_p, x] \subset p$ holds.

We say that $(\mathcal{D}, \tau)$ is a space-time discretisation of $\Omega \times (0,T)$ if $\mathcal{D}$ is a space discretisation of Ω in the sense we mentioned above and if there exists $N_T \in \mathbb{N}$ with $T = (N_T + 1)\tau$, where τ is a symbol for the time step.

Another important assumption on the discretisation we make is that

$$d_{p\sigma} n_{p,\sigma} = x_\sigma - x_p, \ \forall p \in \mathcal{M}, \ \forall \sigma \in \mathcal{E}_p, \tag{2.6}$$

where $\mathcal{E}_p$ denotes the set of the edges of the control volume p, $x_\sigma \in \sigma$, $d_{p\sigma}$ is a symbol for the Euclidean distance between x_p and hyperplane including σ (it is assumed that $d_{p\sigma} > 0$) and $n_{p,\sigma}$ denotes the unit vector normal to σ outward to p.

We define the set $H_{\mathcal{D}} \subset \mathbb{R}^{|\mathcal{M}|} \times \mathbb{R}^{|\mathcal{E}|}$ such that $u_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\text{ext}}$ (the set of boundary interfaces). We define the following functions on $H_{\mathcal{D}}$:

$$N_p(u)^2 = \frac{1}{|p|} \sum_{\sigma \in \mathcal{E}_p} \frac{|\sigma|}{d_{p\sigma}} (u_\sigma - u_p)^2, \ \forall p \in \mathcal{M}, \ \forall u \in H_{\mathcal{D}}, \tag{2.7}$$

where u_p is defined as $u_p = u(x_p)$ and u_σ is defined as $u_\sigma = u(x_\sigma)$.

Let us recall that

$$\|u\|_{1,\mathcal{D}}^2 = \sum_{p \in \mathcal{M}} |p| N_p(u)^2 \tag{2.8}$$

defines a norm on $H_{\mathcal{D}}$ (see [1] and references there).

Under the above mentioned assumptions and notations the semi-implicit scheme is defined by

$$u_p^0 = u_0(x_p), \ \forall p \in \mathcal{M}, \tag{2.9}$$

$$u_\sigma^0 = u_0(x_\sigma), \ \forall \sigma \in \mathcal{E}, \tag{2.10}$$

$$r_p^{n+1} = \int_{n\tau}^{(n+1)\tau} \int_p r(x,t) dx dt, \ \forall p \in \mathcal{M}, \ \forall n \in \mathbb{N}, \tag{2.11}$$

$$u_\sigma^{n+1} = 0, \ \forall \sigma \in \mathcal{E}_{\text{ext}}, \ \forall n \in \mathbb{N}, \tag{2.12}$$

and

$$\begin{aligned} &\frac{|p|}{\tau \ f_1(N_p(u^n))}(u_p^{n+1} - u_p^n) - \frac{1}{f(N_p(u^n))} \sum_{\sigma \in \mathcal{E}_p} g_{\mathcal{D}}^S \frac{|\sigma|}{d_{p\sigma}} (u_\sigma^{n+1} - u_p^{n+1}) = \\ &\quad = \frac{r_p^{n+1}}{\tau \ f_1(N_p(u^n))}, \forall p \in \mathcal{M}, \ \forall n \in \mathbb{N}, \end{aligned} \tag{2.13}$$

where the following relation is given for the interior edges

$$\frac{u_\sigma^{n+1} - u_p^{n+1}}{f(N_p(u^n)) \ d_{p\sigma}} + \frac{u_\sigma^{n+1} - u_q^{n+1}}{f(N_q(u^n)) \ d_{q\sigma}} = 0, \tag{2.14}$$

$\forall n \in \mathbb{N}$, $\forall \sigma \in \mathcal{E}_{\text{int}}$ (the set of interior interfaces) where σ is the edge between p and q.

For the explanation of the selection of u_p^0 and u_σ^0, which impacts the assumptions given on function u_0 in (H2) see [8].

There are two options how to choose $g_{\mathcal{D}}^S$ (approximation of g^S) in (2.13) considered in this paper. First one, we will label it **(APR1)**, is for $\forall \sigma \in \mathcal{E}$ defined by

$$g_\sigma^S := g^S(x_\sigma) = g(|\int_{\mathbb{R}^d} \nabla G_S(x_\sigma - \xi) \tilde{I^0}(\xi) d\xi|). \tag{2.15}$$

It means that the convolution of the initial image with Gaussian kernel is done in the points x_σ on the border of the control volume, which is exactly the point where it, from (2.13), should be done.

The second one, labeled as **(APR2)**, is $\forall p \in \mathcal{M}$ defined by

$$g_p^S := g^S(x_p) = g(|\int_{\mathbb{R}^d} \nabla G_S(x_p - \xi) \tilde{I^0}(\xi) d\xi|). \tag{2.16}$$

This means that the convolution is done in the points x_p inside the control volume, so we are making an error. The problem we are interested in is the impact of this approximation error on the final model and it segmentation ability.

3. Theoretical results. Theoretical properties for the scheme (2.13) - (2.14) with the approximation **(APR2)** as the stability estimates on the numerical solution, the uniqueness of the numerical solution, the convergence of the numerical scheme to the weak solution and the convergence of the approximation of the numerical gradient were proven in [5].

For the approximation **(APR1)** the case is more complex and the stability of the scheme is conditional, the time step and the space step have to be the same order to guarantee the stability estimates on the numerical solution and all of the other theoretical features mentioned above.

If we summarize, from the perspective of the theory the approximation **(APR2)** is better as we are able to prove unconditional stability for the scheme (2.13) - (2.14). On the other hand with this choice of approximation we are making bigger approximation error than for **(APR1)**. How big impact does this error have on the computations we test in the next section.

4. Numerical experiments. For testing of the difference between **(APR1)** and **(APR2)** we chose following benchmark example (see Figure 4.1) - noised square with missing boundaries as an example of the object with both typical errors occuring in the image segmentation - noise and missing boundaries. On the other hand with square as an simple object we know the desired shape of the level function, so we can test accuracy and speed of the approximations even without knowing the exact solution of the problem.

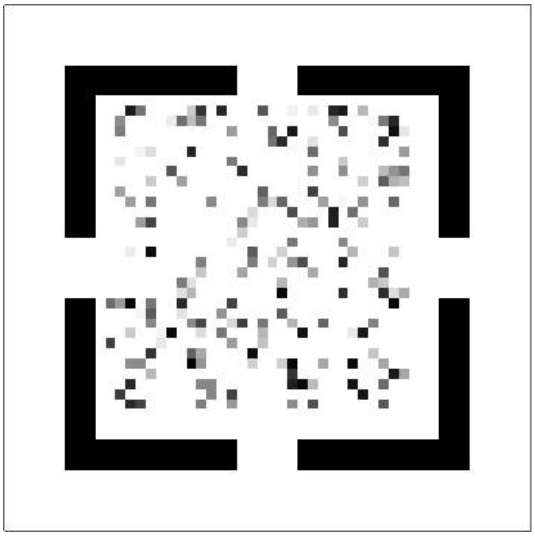


FIG. 4.1. *Object we want to segment.*

The approach we are presenting in this paper is based on the idea of the level set function. At the beginning of the segmentation process we construct initial level set function (Figure 4.2), which is developing in the time by the mean curvature and the constructed vector field tends the level set function to the border of the segmented object. Instead of creating developing curve to segment the object, we create the level set function and we monitor the development of the segmented area implicitly by looking on its isolines. This type of approach is robust against topological changes.

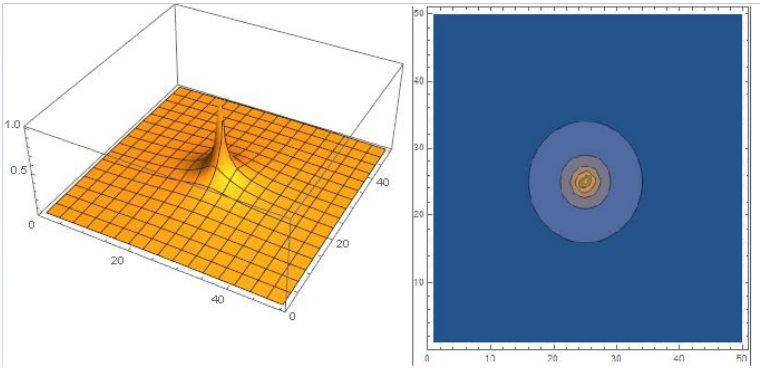


FIG. 4.2. *Initial level set function.*

4.1. Visual test. As the first comparison of the different choices of approximation of nonlinear smoothing term in (1) we chose the visual test.

We can take a look on the difference that made two different approximation on the initial image (see Figure 4.3). The difference is defined as model with **(APR2)** minus model with **(APR1)**. One can see that **(APR2)** is better in presmoothing of the noise in the image, but, on the other hand, the borders of the object are little bit more blurry.

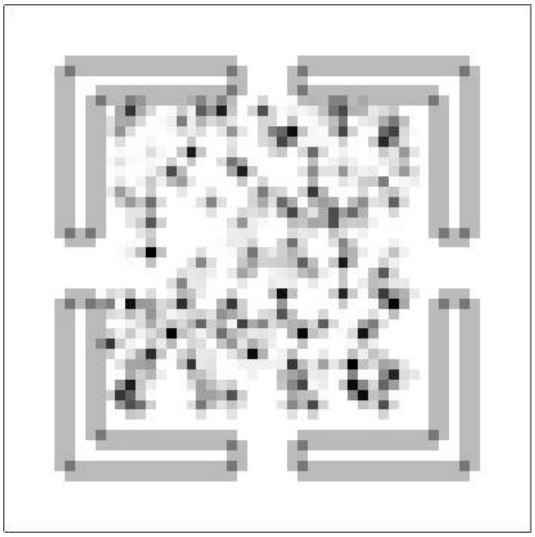
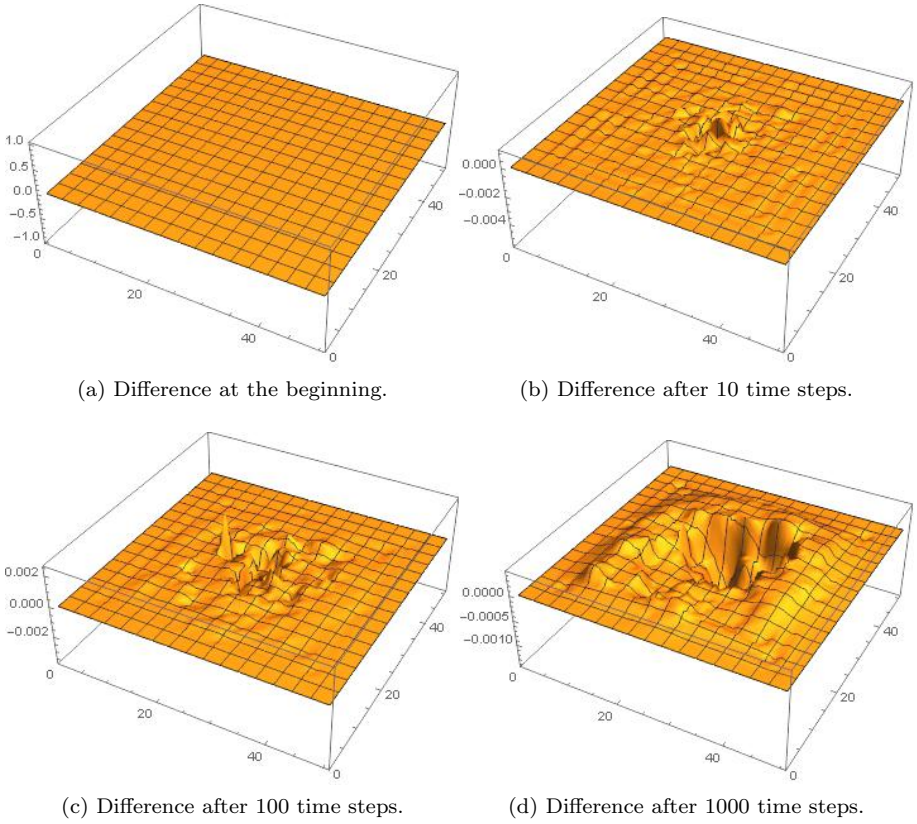


FIG. 4.3. *Difference for the initial image.*

This is the graphical impact of the choice of the approximation. Now take a look on the difference between level set functions in the various time steps. On the Figure 4.4 we can see that the difference in the time very slightly increase, but even after 1000 time steps, when the object is segmented the difference is still less than 0.001 in absolute numbers. So from graphical perspective it seems that the **(APR1)** is slightly better, but the difference is small. To make these initial observations more precise we do the numerical tests as well.



(a) Difference at the beginning.

(b) Difference after 10 time steps.

(c) Difference after 100 time steps.

(d) Difference after 1000 time steps.

FIG. 4.4. *Difference between level set functions.*

4.2. Numerical comparison. The second comparison of approximation of nonlinear smoothing term in (1) we are presenting in this paper are the absolute and relative L_1, L_2 and L_∞ norms of the difference between the segmentation level set functions:

TABLE 4.1
Absolute and relative norms for sensitivity constant $K = 1$.

Absolute difference after	1 step	10 steps	100 steps	1000 steps
L_1 norm	0.00612	0.13982	0.22316	0.29966
L_2 norm	0.00001	0.00004	0.00013	0.00009
L_∞ norm	0.00086	0.00244	0.00309	0.00098
Relative difference after	1 step	10 steps	100 steps	1000 steps
L_1 norm	0.00039	0.00091	0.00151	0.00098
L_2 norm	7.53e-08	2.78e-07	8.79e-07	7.20e-07
L_∞ norm	5.53e-06	1.59e-05	2.09e-05	7.44e-06

From these numbers we are able to conclude the same result as from the visual test - the difference between model with **(APR1)** and model with **(APR2)** is too small to make any relevant impact on the final result of segmentation (biggest relative error is less than 0.2 %).

There is one more parameter that can make an impact - constant K, the constant of sensitivity of the function g mentioned in (H6). In the first example above we set $K = 1$, so lets increase this value and check if it has a significant impact.

In the next table we list L_1, L_2 and L_∞ norms of the absolute and relative difference between the segmentation level set functions for sensitivity constant $K = 10$:

TABLE 4.2
Absolute and relative norms for sensitivity constant K = 10.

Absolute difference after	1 step	10 steps	100 steps	1000 steps
L_1 norm	0.03299	0.10126	0.18812	0.16514
L_2 norm	0.00001	0.00001	0.00008	0.00002
L_∞ norm	0.00082	0.00304	0.00158	0.00021
Relative difference after	1 step	10 steps	100 steps	1000 steps
L_1 norm	0.00021	0.00067	0.00133	0.00161
L_2 norm	5.22e-08	3.88e-07	5.88e-07	2.01e-07
L_∞ norm	5.29e-06	2.01e-05	1.11e-05	2.06e-06

Comparing these numbers with the ones from Table 4.1 one can see that the choice of the constant K do not play a big role in overall process of the segmentation when looking on the difference between segmantation level set functions.

5. Conclusion. In this paper we pay attention on the options of approximation of the nonlinear smoothing term in the image segmentation. We compared both approaches from theoretical and numerical perspective.

In the Section 3 we mention that model with **(APR2)** has better theoretical features, especially the stability of the scheme and convergence is unconditional compared to conditional stability and convergence of the semi-implicit shceme for model with **(APR1)**, here the time step and the space step have to be the same order.

Section 4 was dedicated to numerical comparison of both models. Overall result is that from numerical perspective is better the model with **(APR1)**, but the difference and impact of choice of the approximation is minimal and not significant.

Overall is seems more reasonable to use **(APR2)** as it is easier for implimentation, there is a proof of all needed theoretical aspects of the model and the difference in numerical computation is negligible.

For the future research we plan to study and evaluate the importance of the nonlinear smoothing term in the image segmentation overall.

REFERENCES

[1] Eymard R., Handlovičová A., Mikula K.: Study of a finite volume scheme for regularised mean curvature flow level set equation, IMA Journal on Numerical Analysis, Vol. 31, 813-846, 2011.
[2] Osher S., Sethian J. A.: Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations, J. Comput. Phys., 79(1):12-49, 1988.
[3] Mikula K., Sarti A., Sgallarri A.: Co-volume method for Riemannian mean curvature flow in subjective surfaces multiscale segmentation, Computing and Visualization in Science, Vol. 9, No. 1, 23-31, 2006.
[4] Mikula K., Sarti A., Sgallari F.: Co-volume level set method in subjective surface based medical image segmentation, in: Handbook of Medical Image Analysis: Segmentation and Registration Models (J.Suri et al., Eds.), Springer, New York, 583-626, 2005.
[5] Handlovičová A., Tibenský M.: Convergence of the numerical scheme for regularised Riemannian

mean curvature flow equation, submitted to Tatra Mountains Mathematical Publications, 2017.
[6] Mikula K., Ramarosy N.: Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing, Numerische Mathematik 89, No. 3, 561-590, 2001.
[7] Tibenský M.: Využitie metód založených na level set rovnici v spracovaní obrazu, Faculty of mathematics, physics and informatics, Comenius University, 2016.
[8] Droniou J., Nataraj N.: Improved L^2 estimate for gradient schemes, and super-convergence of the TPFA finite volume scheme, IMA Journal of Numerical Analysis 2017, 2016.

Proceedings of EQUADIFF 2017
pp. 237–246

STABILITY OF ALE SPACE-TIME DISCONTINUOUS GALERKIN METHOD

MILOSLAV VLASÁK*, MONIKA BALÁZSOVÁ†, AND MILOSLAV FEISTAUER‡

Abstract. We assume the heat equation in a time dependent domain, where the evolution of the domain is described by a given mapping. The problem is discretized by the discontinuous Galerkin (DG) method in space as well as in time with the aid of Arbitrary Lagrangian-Eulerian (ALE) method. The sketch of the proof of the stability of the method is shown.

Key words. ALE formulation, discontinuous Galerkin method, discrete characteristic function, stability

AMS subject classifications. 65M60, 65M99

1. Introduction.

Although many theoretical results are devoted to the numerical analysis of parabolic PDEs within a fixed domain, there are number of areas with many important applications of parabolic PDEs with time dependent domain. We can mention, for example, problems with moving boundaries, where the motion of the boundary is either prescribed or given by the PDE itself.

There are several approaches how to deal with problems in time dependent domains, e.g. fictitious domain method, see e.g. [21], or immersed boundary method, see e.g. [4]. A very popular technique is Arbitrary Lagrangian-Eulerian (ALE) method that is based on a one-to-one ALE mapping of the reference domain on the current one. ALE method is often applied in connection with conforming finite element method (FEM) in space and lower order time discretizations (backward Euler method, Crank-Nicolson method, BDF2) in time, see e.g. [18] or [19].

The class of discontinuous Galerkin methods seems to be one of the most promising candidates to construct high order accurate schemes for solving convection-diffusion problems, where narrow layers and steep gradients of the solution may appear. For a survey about DG space discretization, see [1], [10], [11]. The discontinuous Galerkin method could be applied for time discretization as well. For a survey about DG time discretization, see e.g. [23]. The discontinuous Galerkin method in space with BDF time discretization was applied with success to time dependent problems, see e.g. [7] or [22]. Moreover, in [8] space-time DG discretization was applied to the vibration of an airfoil problem and the results were compared with BDF time discretization. According to this comparison, DG time discretization seems to be more robust and accurate than BDF.

*Faculty of Mathematics and Physics, Charles University, Sokolovska 83, 186 75 Prague 8, Czech Republic (vlasak@karlin.mff.cuni.cz). The research of M. Vlasák was supported by grant 17-01747S of the Czech Science Foundation. M. Vlasák is a junior member of the university centre for mathematical modeling, applied analysis and computational mathematics (MathMAC).

†Faculty of Mathematics and Physics, Charles University, Sokolovska 83, 186 75 Prague 8, Czech Republic (balazsova@karlin.mff.cuni.cz). The research of M. Balázsová was supported by Charles University, project GA UK no. 127615.

‡Faculty of Mathematics and Physics, Charles University, Sokolovska 83, 186 75 Prague 8, Czech Republic (feist@karlin.mff.cuni.cz). The research of M. Feistauer was supported by grant 17-01747S of the Czech Science Foundation.

Numerical analysis of stability and a priori error estimates of time dependent problems with divergence free domain velocity and discretized by the conforming FEM in space and by DG in time could be found in [5] and [6]. Finally, the stability analysis of space-time DG discretization of nonlinear convection-diffusion problems is studied in [2] for lower degree polynomial approximations in time and in [3] for general polynomial degree.

2. Continuous problem. Let $T > 0$. We consider the following initial–
–boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= f \quad \text{in } \Omega_t \times (0,T), \\ u &= 0 \quad \text{in } \partial\Omega_t \times (0,T), \\ u &= u^0 \quad \text{in } \Omega_0, \end{aligned} \tag{2.1}$$

where $\Omega_t \subset R^d$ $(d = 1, 2, 3)$ is a bounded polyhedral time dependent domain with a Lipschitz continuous boundary $\partial\Omega_t$. We assume that the initial condition $u^0 \in L^2(\Omega_0)$ and the right-hand side $f \in L^2(0, T, L^2(\Omega_t))$. We denote by $(.,.)_t$ and $\|.\|_t$ the $L^2(\Omega_t)$ scalar product and norm, respectively.

The evolution of the domain Ω_t in time is described by a given regular one-to-one ALE mapping

$$\mathcal{A} : \overline{\Omega}_0 \times [0,T] \to \overline{\Omega}_t, \tag{2.2}$$

where $\overline{\Omega}_0$ or $\overline{\Omega}_t$ are closures of Ω_0 or Ω_t, respectively. For the purpose of the proof of the stability we introduce following regularity assumptions on the ALE mapping $\mathcal{A}$:

$$\mathcal{A} \in W^{1,\infty}(0, T, W^{1,\infty}(\Omega_0)), \quad \mathcal{A}^{-1} \in W^{1,\infty}(0, T, W^{1,\infty}(\Omega_t)). \tag{2.3}$$

Moreover, we denote the Jacobi matrix of $\mathcal{A}$ by $B = \frac{d\mathcal{A}}{dX}$, the corresponding determinant by $J = \det(B)$ and the *domain velocity* by $\omega = \frac{\partial \mathcal{A}}{\partial t} \circ \mathcal{A}^{-1}$. From the regularity assumptions (2.3) it is possible to show that B, B^{-1}, J, J^{-1}, ω and $\nabla \cdot \omega$ are bounded, i.e. there exists a constant $C_{\mathcal{A}} > 0$ such that

$$\begin{aligned} &\max(\|B\|_{L^\infty(0,T,L^\infty(\Omega_0))}, \|B^{-1}\|_{L^\infty(0,T,L^\infty(\Omega_0))}, \|J\|_{L^\infty(0,T,L^\infty(\Omega_0))} \\ &\|J^{-1}\|_{L^\infty(0,T,L^\infty(\Omega_0))}, \|\omega\|_{L^\infty(0,T,L^\infty(\Omega_t))}, \|\nabla \cdot \omega\|_{L^\infty(0,T,L^\infty(\Omega_t))}) \le C_{\mathcal{A}}. \end{aligned} \tag{2.4}$$

Problem (2.1) is usually transformed into the *ALE formulation*. To this end, we introduce ALE derivative

$$D_t u = \frac{\partial u}{\partial t} + \omega \cdot \nabla u. \tag{2.5}$$

Now we introduce the ALE formulation equivalent to problem (2.1)

$$\begin{aligned} D_t u - \Delta u - \omega \cdot \nabla u &= f \quad \text{in } \Omega_t \times (0,T), \\ u &= 0 \quad \text{in } \partial\Omega_t \times (0,T), \\ u &= u^0 \quad \text{in } \Omega_0. \end{aligned} \tag{2.6}$$

3. Discretization. In this section, we describe the interior penalty discontinuous Galerkin discretization in space variables together with the discontinuous Galerkin time discretization in the ALE framework.

We consider a space partition $\mathcal{T}_{h,0}$ consisting of a finite number of closed, d -dimensional simplices K with mutually disjoint interiors and covering $\overline{\Omega}_0$. We assume conforming properties, i.e. neighbouring elements share an entire edge or face. We set $h_K = \mathrm{diam}(K)$ and $h = \max_K h_K$. We assume that the mesh is quasi-uniform, i.e. there exists a constant $C_Q > 0$ such that $h_K \le C_Q h_{\bar{K}}$ for all neighbouring elements K and $\bar{K}$. By ρ_K we denote the radius of the largest d-dimensional ball inscribed into K. We assume shape regularity of elements, i.e. $h_K/\rho_K \le C$ for all $K \in \mathcal{T}_h$, where the constant does not depend on $\mathcal{T}_{h,0}$ for $h \in (0, h_0)$. By $\Gamma_{h,0}$ we denote the set of all edges of $\mathcal{T}_{h,0}$. We define a unit normal vector n to arbitrary edge from $\Gamma_{h,0}$. For inner edges the direction is arbitrary, for outer edges we assume that n is the unit outer normal vector.

Since the domain Ω_0 evolves into $\Omega_t = \mathcal{A}(\Omega_0, t)$, we define similarly the evolution of the mesh $\mathcal{T}_{h,t} = \mathcal{A}(\mathcal{T}_{h,0}, t)$, the evolution of the edges $\Gamma_{h,t} = \mathcal{A}(\Gamma_{h,0}, t)$.

We introduce the space for the semidiscrete solution on Ω_0

$$V_h = \{v \in L^2(\Omega_0) : v|_K \in P^p(K)\}, \tag{3.1}$$

where $P^p(K)$ denotes the space of polynomials up to the degree $p \ge 1$ on K. Functions from the space V_h are discontinuous across the edges of $\mathcal{T}_{h,0}$. For this reason we define one-sided limits

$$v_L(x) = \lim_{s\to 0+} v(x - ns), \quad v_R(x) = \lim_{s\to 0+} v(x + ns), \tag{3.2}$$

jumps and mean values

$$[v] = v_L - v_R, \quad \langle v \rangle = \frac{v_L + v_R}{2}. \tag{3.3}$$

For outer edges we define

$$[v] = \langle v \rangle = v_L = \lim_{s\to 0+} v(x - ns). \tag{3.4}$$

In order to discretize problem (2.6) in time, we consider a time partition $0 = t_0 < t_1 < \ldots < t_r = T$ with time intervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = t_m - t_{m-1}$ and $\tau = \max_{m=1,\ldots,r} \tau_m$. We define the solution space

$$V_h^\tau = \{v \in L^2(0, T, L^2(\Omega_t)) : (v \circ \mathcal{A})|_{I_m} \in P^q(I_m, V_h)\}. \tag{3.5}$$

For a function $v \in V_h^\tau$ we define the one–sided limits

$$v_\pm^m = v(t_m \pm) = \lim_{t\to t_m \pm} v(t) \tag{3.6}$$

and the jumps

$$\{v\}_m = v_+^m - v_-^m, \quad m \ge 1 \quad \text{and} \quad \{v\}_0 = v_+^0 - u^0. \tag{3.7}$$

We approximate the diffusion term by the *discontinuous Galerkin interior penalty* form

$$\begin{aligned} a_{h,t}(u, v) = &\sum_{K\in\mathcal{T}_{h,t}} \int_K \nabla u \cdot \nabla v dx - \sum_{e\in\Gamma_{h,t}} \int_e (\langle \nabla u \rangle \cdot n[v] + \theta \langle \nabla v \rangle \cdot n[u]) dS \\ &+ \sum_{e\in\Gamma_{h,t}} \int_e \sigma [u][v] dS. \end{aligned} \tag{3.8}$$

The choice of parameter $\theta = 1, 0, -1$ corresponds to SIPG, IIPG and NIPG formulation, respectively. Parameter σ is defined on the inner edges between elements K and $\bar{K}$ by

$$\sigma = \frac{C_W}{\frac{h_K + h_{\bar{K}}}{2}} \tag{3.9}$$

and on the boundary edges by

$$\sigma = \frac{C_W}{h_K}, \tag{3.10}$$

where the constant $C_W > 0$ needs to be chosen large enough to guarantee ellipticity of $a_{h,t}$. Lower bounds for C_W will be briefly discussed later. For more informations about different variants of discontinuous Galerkin method and their corresponding formulations approximating $(-\Delta u, v)_t$ see e.g. [1].

Now, we are able to formulate the fully discrete space-time discontinuous Galerkin scheme:

DEFINITION 3.1. *We say that a function $U \in V_h^\tau$ is the discrete solution of problem (2.6) obtained by space-time discontinuous Galerkin method, if the following conditions are satisfied*

$$\int_{I_m} (D_t U, v)_t + a_{h,t}(U, v) - (\omega \cdot \nabla U, v)_t dt + (\{U\}_{m-1}, v_+^{m-1})_{t_{m-1}} = \int_{I_m} (f, v)_t dt \quad \forall m = 1, \dots, r, \, \forall v \in V_h^\tau. \tag{3.11}$$

The time discretization in (3.11) can be viewed as a generalization of some specific classical one–step methods for parabolic problems. It is possible to show that setting $q = 0$, i.e. piecewise constant approximation in time, is equivalent (up to suitable quadrature of the right–hand side) to backward Euler method in time and discontinuous Galerkin method in space. Similarly, the higher polynomial degree approximations in time lead to methods that are equivalent (up to suitable quadrature of the right–hand side) to Radau IIA Runge-Kutta methods. For details about the relations between Galerkin methods and Runge-Kutta methods see e.g. [15] and [20]. For the descriptions of Radau IIA Runge-Kutta methods see e.g. [12] or [16] and [17].

4. Stability. The aim of this section is to show that the numerical scheme (3.11) is stable, i.e. the approximate solution obtained from (3.11) can be bounded in terms of the data f and u^0 of the problem (2.1).

An important auxiliary tool for the analysis of problems in time-dependent domains is the *Reynolds transport formula*:

$$\frac{d}{dt} \int_{\Omega_t} v(x,t) dx = \int_{\Omega_t} \frac{\partial v}{\partial t}(x,t) + \nabla \cdot (\omega v)(x,t) dx = \int_{\Omega_t} D_t v(x,t) + \nabla \cdot \omega(x,t) v(x,t) dx. \tag{4.1}$$

For the purpose of the forthcoming estimates we define discontinuous Galerkin energy norm

$$\|u\|_{DG,t}^2 = \sum_{K \in \mathcal{T}_{h,t}} \|\nabla u\|_{L^2(K)}^2 + \sum_{e \in \Gamma_{h,t}} \|\sigma^{1/2}[u]\|_{L^2(e)}^2. \tag{4.2}$$

Using this norm we can summarize the properties of $a_{h,t}$ in following lemma.

LEMMA 4.1. *Let $U, v \in V_h^\tau$. Then there exists a constant $C_a > 0$ such that*

$$a_{h,t}(U,v) \le C_a \|U\|_{DG,t} \, \|v\|_{DG,t}. \tag{4.3}$$

Moreover, let the constant C_W satisfy

$$\begin{aligned} &C_W > 0, && \theta = -1, \quad \textit{NIPG}, \\ &C_W \ge \frac{1}{2} C_M (C_I + 1)(C_Q + 1), && \theta = 0, \quad \textit{IIPG}, \\ &C_W \ge C_M (C_I + 1)(C_Q + 1), && \theta = 1, \quad \textit{SIPG}, \end{aligned} \tag{4.4}$$

where constant C_M and C_I come from the trace inequality and the inverse inequality, respectively, see [11]. Then

$$a_{h,t}(U,U) \ge \frac{1}{2} \|U\|^2_{DG,t}. \tag{4.5}$$

Proof. The ideas of the proof are well described in e.g. [11]. The generalization to the problems in the time dependent domains can be found in [2]. □

We need the estimate of the ALE derivative term.

LEMMA 4.2. *Let $U \in V_h^\tau$. Then*

$$\begin{aligned} &\int_{I_m} (D_t U, U)_t dt + (\{U\}_{m-1}, U_+^{m-1})_{t_{m-1}} \\ &\ge \frac{1}{2} \|U_-^m\|^2_{t_m} - \frac{1}{2} \|U_-^{m-1}\|^2_{t_{m-1}} - \frac{C_{\mathcal{A}}}{2} \int_{I_m} \|U\|^2_t dt. \end{aligned} \tag{4.6}$$

Proof. At first, we will study relation (4.6) elementwise for each element $K \in \mathcal{T}_{h,0}$. Let us denote $K_t = \mathcal{A}(K,t)$. Applying Reynolds transport formula with $v = U^2$ we get

$$\begin{aligned} &\int_{I_m} \int_{K_t} U \cdot D_t U dx dt + \int_{K_{t_{m-1}}} \{U\}_{m-1} U_+^{m-1} dx \\ &= \frac{1}{2} \int_{I_m} \int_{K_t} D_t U^2 dx dt + \int_{K_{t_{m-1}}} \{U\}_{m-1} U_+^{m-1} dx \\ &= \frac{1}{2} \int_{I_m} \frac{d}{dt} \int_{K_t} U^2 dx dt - \frac{1}{2} \int_{I_m} \int_{K_t} (\nabla \cdot \omega) U^2 dx dt + \int_{K_{t_{m-1}}} \{U\}_{m-1} U_+^{m-1} dx \\ &= \frac{1}{2} \|U_-^m\|^2_{L^2(K_{t_m})} - \frac{1}{2} \|U_+^{m-1}\|^2_{L^2(K_{t_{m-1}})} + \|U_+^{m-1}\|^2_{L^2(K_{t_{m-1}})} \\ &\quad - \int_{K_{t_{m-1}}} U_-^{m-1} U_+^{m-1} dx - \frac{1}{2} \int_{I_m} \int_{K_t} (\nabla \cdot \omega) U^2 dx dt \\ &= \frac{1}{2} \|U_-^m\|^2_{L^2(K_{t_m})} - \frac{1}{2} \|U_-^{m-1}\|^2_{L^2(K_{t_{m-1}})} + \frac{1}{2} \|\{U\}_{m-1}\|^2_{L^2(K_{t_{m-1}})} \\ &\quad - \frac{1}{2} \int_{I_m} \int_{K_t} (\nabla \cdot \omega) U^2 dx dt \\ &\ge \frac{1}{2} \|U_-^m\|^2_{L^2(K_{t_m})} - \frac{1}{2} \|U_-^{m-1}\|^2_{L^2(K_{t_{m-1}})} - \frac{C_{\mathcal{A}}}{2} \int_{I_m} \int_{K_t} U^2 dx dt. \end{aligned} \tag{4.7}$$

The lemma is proved by summing this relation over all $K_t \in \mathcal{T}_{h,t}$. □

Setting $v = U$ in (3.11) we get the basic identity

$$\int_{I_m} (D_t U, U)_t + a_{h,t}(U,U) - (\omega \cdot \nabla U, U)_t dt + (\{U\}_{m-1}, U_+^{m-1})_{t_{m-1}} = \int_{I_m} (f, U)_t dt. \tag{4.8}$$

Since

$$\int_{I_m} (\omega \cdot \nabla U, U)_t dt \le C_{\mathcal{A}} \int_{I_m} \|U\|_{DG,t} \|U\|_t dt \le C_{\mathcal{A}}^2 \int_{I_m} \|U\|_t^2 dt + \frac{1}{4} \int_{I_m} \|U\|_{DG,t}^2 dt, \tag{4.9}$$

applying Lemma 4.1 and Lemma 4.2 we get

$$\begin{aligned} &\frac{1}{2}\|U_-^m\|_{t_m}^2 - \frac{1}{2}\|U_-^{m-1}\|_{t_{m-1}}^2 + \frac{1}{2}\int_{I_m} \|U\|_{DG,t}^2 dt \\ &\le \|f\|_{L^2(I_m,L^2(\Omega_t))} \|U\|_{L^2(I_m,L^2(\Omega_t))} + C_{\mathcal{A}}^2 \int_{I_m} \|U\|_t^2 dt + \frac{1}{4}\int_{I_m} \|U\|_{DG,t}^2 dt \\ &\quad + \frac{C_{\mathcal{A}}}{2} \int_{I_m} \|U\|_t^2 dt \\ &\le \|f\|_{L^2(I_m,L^2(\Omega_t))}^2 + \frac{1}{4}\int_{I_m} \|U\|_{DG,t}^2 dt + \tau_m C_1 \sup_{t \in I_m} \|U\|_t^2, \end{aligned} \tag{4.10}$$

where the constant $C_1 = 1/4 + C_{\mathcal{A}}/2 + C_{\mathcal{A}}^2$.

To be able to get rid of the last supremum term, we need to derive a technique for estimating the values of the discrete solution inside of intervals I_m.

4.1. Discrete characteristic function. The concept of the discrete characteristic function comes from [9]. As we have seen in (4.10), application of the test function $v = U$ naturally leads to the nodal estimate. Setting $v = \chi_{(t_{m-1},s)}U$, where $\chi_{(t_{m-1},s)}$ is characteristic function of the interval (t_{m-1}, s) for $s \in [t_{m-1}, t_m]$, will lead to a similar estimate for $\|U(s)\|_s$ instead of $\|U_-^m\|_{t_m}$. Unfortunately, it is not possible to do it, since $\chi_{(t_{m-1},s)}U \notin V_h^\tau$. The idea of the discrete characteristic function is based on the construction of $U_s \in V_h^\tau$ for given $U \in V_h^\tau$ and $s \in [t_{m-1}, t_m]$ such that U_s will preserve similar properties to the classical characteristic function. For applications of the discrete characteristic function see, e.g. [11] or [24].

We will use a notation $\tilde{v} = v \circ \mathcal{A}$ for transformation of functions from the evolving space-time cylinder to the reference space-time cylinder. From the assumptions on the ALE mapping $\mathcal{A}$ and according to the definition of space V_h^τ it is possible to see that this transformation is bijection between V_h^τ and $\tilde{V}_h^\tau$, where

$$\tilde{V}_h^\tau = \{v \in L^2(0,T,L^2(\Omega_0)) : v|_{K \times I_m} \in P^q(I_m, P^p(K))\}, \tag{4.11}$$

i.e. $\tilde{V}_h^\tau$ represents the space of classical piecewise polynomial functions.

We define the discrete characteristic function for time dependent domains in three steps. At first, the given function $U \in V_h^\tau$ is transformed onto the reference domain,

i.e. $\tilde{U} = U \circ \mathcal{A} \in \tilde{V}_h^\tau$. Second step is the construction of discrete characteristic function in fixed domains, i.e. $\tilde{U}_s \in \tilde{V}_h^\tau$ such that

$$\tilde{U}_{s+}^{m-1} = \tilde{U}_{+}^{m-1}, \tag{4.12}$$
$$\int_{I_m} \left(\tilde{U}_s, \frac{\partial v}{\partial t}\right)_0 dt = \int_{t_{m-1}}^{s} \left(\tilde{U}, \frac{\partial v}{\partial t}\right)_0 dt \quad \forall v \in \tilde{V}_h^\tau.$$

The last step is the transformation back to the current domain, i.e. $U_s = \tilde{U}_s \circ \mathcal{A}^{-1} \in V_h^\tau$.

Now, we want to show a similar relation to the relation from Lemma 4.2 that will also describe the *contraction* property of the discrete characteristic function.

LEMMA 4.3. *Let $U \in V_h^\tau$ and $U_s \in V_h^\tau$ be its discrete characteristic function associated with $s \in I_m$. Then there exists a constant $C_D > 0$ depending only on the polynomial degree q and on the regularity of the ALE mapping (2.3) such that*

$$\int_{I_m} (D_t U, U_s)_t dt + (\{U\}_{m-1}, U_{s+}^{m-1})_{t_{m-1}} \tag{4.13}$$
$$\geq \frac{1}{2} \sup_{I_m} \|U(t)\|_t^2 - \frac{1}{2} \|U_-^{m-1}\|_{t_{m-1}}^2 - C_D \tau_m \sup_{t \in I_m} \|U\|_t^2$$

Proof. Since the proof is long and technical, it is skipped in this paper. The proof will be contained in [3]. ∎

Using Lemma 4.3, it is possible to deal with the ALE derivative term. For all the other terms we need to show that the process of creating the discrete characteristic function is stable with a constant independent of the parameter $s \in I_m$.

LEMMA 4.4. *Let $U \in V_h^\tau$ and $U_s \in V_h^\tau$ be its discrete characteristic function associated with $s \in I_m$. Then there exists a constant $C_{ST} > 0$ depending only on the polynomial degree q and on the regularity of ALE mapping (2.3) such that*

$$\int_{I_m} \|U_s(t)\|_t^2 dt \leq C_{ST} \int_{I_m} \|U(t)\|_t^2 dt, \tag{4.14}$$

$$\int_{I_m} \|U_s(t)\|_{DG,t}^2 dt \leq C_{ST} \int_{I_m} \|U(t)\|_{DG,t}^2 dt. \tag{4.15}$$

Proof. Since the proof is long and technical, it is skipped in this paper. The proof will be contained in [3]. ∎

4.2. Main result. Now, we are ready to formulate the main result.

THEOREM 4.5. *Let the parameter C_W satisfy (4.4) and let $U \in V_h^\tau$ be an approximate solution obtained by scheme (3.11). Then there exist constants $C > 0$ and $C^* > 0$ such that $\tau \leq C^*$ implies*

$$\sup_{I_m} \|U\|_t^2 \leq C(\|f\|_{L^2(0,T,L^2(\Omega_t))}^2 + \|u^0\|_0^2). \tag{4.16}$$

Proof. Setting $v = U_s$ in the left-hand side of (3.11), where $s \in [t_{m-1}, t_m]$ such that $\|U(s)\|_s = \sup_{t \in I_m} \|U\|_t$, and using Lemma 4.1, Lemma 4.3 and Lemma 4.4 we

get

$$
\begin{aligned}
(4.17)\quad & \int_{I_m} (D_t U, U_s)_t + a_{h,t}(U, U_s)_t - (\omega \cdot \nabla U, U_s)_t dt + (\{U\}_{m-1}, U_+^{m-1})_{t_{m-1}} \\
& \geq \frac{1}{2}\|U(s)\|_s^2 - \frac{1}{2}\|U_-^{m-1}\|_{t_{m-1}}^2 - C_D \tau_m \sup_{t \in I_m} \|U\|_t^2 \\
& \quad - \int_{I_m} C_a \|U\|_{DG,t} \|U_s\|_{DG,t} dt - C_{\mathcal{A}} \int_{I_m} \|U\|_{DG,t} \|U_s\|_t dt \\
& \geq \frac{1}{2} \sup_{I_m} \|U\|_t^2 - \frac{1}{2} \sup_{I_{m-1}} \|U\|_t^2 - C_D \tau_m \sup_{t \in I_m} \|U\|_t^2 - \frac{C_a}{2} \int_{I_m} \|U\|_{DG,t}^2 dt \\
& \quad - \frac{C_a C_{ST}}{2} \int_{I_m} \|U\|_{DG,t}^2 dt - \frac{1}{2} \int_{I_m} \|U\|_{DG,t}^2 dt - \frac{C_{\mathcal{A}}^2 C_{ST}}{2} \int_{I_m} \|U\|_t^2 dt,
\end{aligned}
$$

where we use the notation $\sup_{I_0} \|U\|_t^2 = \|u^0\|_0^2$. Similarly, setting $v = U_s$ in the right-hand side of (3.11) we get

$$
(4.18)\qquad \int_{I_m} (f, U_s)_t dt \leq \frac{1}{2}\|f\|_{L^2(I_m, L^2(\Omega_t))}^2 + \frac{C_{ST}}{2} \int_{I_m} \|U\|_t^2.
$$

Using these relations we get

$$
\begin{aligned}
(4.19)\qquad & \frac{1}{2} \sup_{I_m} \|U\|_t^2 - \frac{1}{2} \sup_{I_{m-1}} \|U\|_t^2 \leq \frac{1}{2}\|f\|_{L^2(I_m, L^2(\Omega_t))}^2 \\
& \qquad + C_2 \tau_m \sup_{t \in I_m} \|U\|_t^2 + C_3 \int_{I_m} \|U\|_{DG,t}^2 dt,
\end{aligned}
$$

where $C_2 = C_D + (C_{\mathcal{A}}^2 + 1) C_{ST}/2$ and $C_3 = (1 + C_a + C_a C_{ST})/2$.

Multiplying (4.10) by $4C_3$ and summing with (4.19) we get

$$
\begin{aligned}
(4.20)\qquad & \frac{1}{2}\left(4C_3\|U_-^m\|_{t_m}^2 + \sup_{I_m} \|U\|_t^2\right) - \frac{1}{2}\left(4C_3\|U_-^{m-1}\|_{t_m}^2 + \sup_{I_{m-1}} \|U\|_t^2\right) \\
& \leq \frac{8C_3 + 1}{2}\|f\|_{L^2(I_m, L^2(\Omega_t))}^2 + (4C_1C_3 + C_2)\tau_m \sup_{t \in I_m} \|U\|_t^2.
\end{aligned}
$$

Setting $C^* = 8C_1C_3 + 2C_2$ we get we get $(4C_1C_3 + C_2)\tau_m < 1/2$ and the statement of the theorem follows from the application of the discrete Gronwall lemma. □

5. Conclusion. We presented a higher order method for the heat equation in a time dependent domain based on the space-time discontinuous Galerkin method. For this problem, the idea of the proof of the unconditional stability for any polynomial degree is shown. There are several items for the future work.

- The extension of the discontinuous Galerkin discretization and the stability analysis to nonlinear problems.
- Deriving a priori error estimates.
- Investigating other suitable higher order time discretizations for problems with a time dependent domain, e.g. continuous Galerkin method, DIRK, etc.
- The numerical analysis of coupled problems, where the ALE mapping depends on the solution of the problem.

Acknowledgments. We are grateful to Z. Vlasáková for stimulating suggestions in the analysis of the discrete characteristic functions.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. *Unified analysis of discontinuous Galerkin methods for elliptic problems.* SIAM J. Numer. Anal., 39(5):1749–1779, 2002.

[2] M. Balázsová, and M. Feistauer. *On the stability of the space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains.* Appl. Math., 60:501–526, 2015.

[3] M. Balázsová, M. Feistauer, and M. Vlasák. *Stability of the ALE space-time discontinuous Galerkin method for nonlinear convection-diffusion problems in time-dependent domains.* (in preparation).

[4] D. Boffi, L. Gastaldi, and, L. Heltai. *Numerical stability of the finite element immersed boundary method.* Math. Models Methods Appl. Sci., 17:1479–1505, 2007.

[5] A. Bonito, I. Kyza, and, R. H. Nochetto. *Time-discrete higher-order ALE formulations: Stability.* SIAM J. Numer. Anal., 51(1):577–604,2013

[6] A. Bonito, I. Kyza, and, R. H. Nochetto. *Time-discrete higher order ALE formulations: a priori error analysis.* Numer. Math., 125:225–257,2013

[7] J. Česenek, M. Feistauer, J. Horáček, V. Kučera, and J. Prokopová. *Simulation of compressible viscous flow in time-dependent domains.* Appl. Math. Comput., 219:7139–7150,2013

[8] J. Česenek, M. Feistauer, and A. Kosík. *DGFEM for the analysis of airfoil vibrations induced by compressible flow.* ZAMM Z. Angew. Math. Mech., 93 No. 6-7:387–402,2013

[9] K. Chrysafinos, and N. J. Walkington. *Error estimates for the discontinuous Galerkin methods for parabolic equations.* SIAM J. Numer. Anal., 44:349–366,2006

[10] B. Cockburn, G. E. Karniadakis. and C.-W. Shu. *Discontinuous Galerkin methods.* In Lecture Notes in Computational Science and Engineering 11. Springer, Berlin, 2000.

[11] V. Dolejší and M. Feistauer. *Discontinuous Galerkin method, Analysis and applications to compressible flow.* Cham: Springer, 2015.

[12] B. L. Ehle. *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems.* Research report CSRR 2010, Dept. AACS, Univ. of Waterloo, Ontario, Canada, 1969.

[13] L. Formaggia, and F. Nobile. *A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements.* East-West J. Numer. Math., 7(2):105–131,1999

[14] L. Gastaldi. *A priori error estimates for the Arbitrary Lagrangian Eulerian formulation with finite elements.* East-West J. Numer. Math., 9(2):123–156,2001

[15] A. Guillo, and J. L. Soulé. *La résolution numérique des problemes différentiels aux conditions initiales par des méthodes de collocation.* R.A.I.R.O., R-3:17–44,1969

[16] E. Hairer, S. P. Norsett, and G. Wanner. *Solving ordinary differential equations I, Nonstiff problems.* Number 8 in Springer Series in Computational Mathematics. Springer Verlag, 2000.

[17] E. Hairer and G. Wanner. *Solving ordinary differential equations II, Stiff and differential-algebraic problems.* Springer Verlag, 2002.

[18] C. W.Hirt, A. A. Amsdem, and J. L, Cook. *An arbitrary Lagrangian-Eulerian computing method for all flow speeds.* J. Comput. Phys., 135(2):198–216, 1997.

[19] T. J. R. Hughes, W. K. Liu, and T. K, Zimmermann. *Lagrangian-Eulerian finite element formulation for incompressible viscous flows.* Comput. Methods Appl. Mech. Eng., 29(3):329–349, 1981.

[20] B. L. Hulme. *One step piecewise polynomial Galerkin methods for initial value problems.* Math. Comp., 26:415–424,1972

[21] K. Khadra, P. Angot, S. Parneix, and J. -P. Caltagirone. *Fictitious domain approach for numerical modelling of Navier-Stokes equations.* Int. J. Numer. Methods Fluids., 34(8):651–684, 2000.

[22] A. Kosík, M. Feistauer, M. Hadrava, and J. Horáček. *Numerical simulation of the interaction between a nonlinear elastic structure and compressible flow by the discontinuous Galerkin method.* Appl. Math. Comput., 267:382–396,2015

[23] V. Thomeé. *Galerkin finite element methods for parabolic problems. 2nd revised and expanded ed.* Springer, Berlin, 2006.

[24] M. Vlasák, V. Dolejší, and J. Hájek. *A Priori Error Estimates of an Extrapolated Space-Time Discontinuous Galerkin Method for Nonlinear Convection-Diffusion Problems.* Numer. Methods Partial Differ. Equations, 27(6):1453–1482,2011

Proceedings of EQUADIFF 2017
pp. 247–254

UPPER HAUSDORFF DIMENSION ESTIMATES FOR INVARIANT SETS OF EVOLUTIONARY SYSTEMS ON HILBERT MANIFOLDS

AMINA KRUCK AND VOLKER REITMANN

Abstract. We prove a generalization of the Douady-Oesterlé theorem on the upper bound of the Hausdorff dimension of an invariant set of a smooth map on an infinite dimensional manifold. It is assumed that the linearization of this map is a noncompact linear operator. A similar estimate is given for the Hausdorff dimension of an invariant set of a dynamical system generated by a differential equation on a Hilbert manifold.

Key words. Hilbert manifold, Hausdorff dimension, singular value

AMS subject classifications. 35B40, 35K57

1. Basic notation of manifold theory. Let us shortly introduce some definitions and properties for manifolds over a Hilbert space ([1, 8]). Suppose $\mathbb{H}$ is a Hilbert space and $\mathcal{M}$ is a set. A *chart* on $\mathcal{M}$ is a bijection $x : \mathcal{D}(x) \subset \mathcal{M} \to \mathcal{R}(x) \subset \mathbb{H}$, where $\mathcal{R}(x)$ is an open set. An *atlas* A of class $C^k (k \geq 1)$ on $\mathcal{M}$ is a set of charts, such that:
(AT1) $\cup_{x\in A}\mathcal{D}(x) = \mathcal{M}$;
(AT2) For arbitrary $x, y \in A$, such that $\mathcal{D}(y) \cap \mathcal{D}(x) \neq \emptyset$, the set $x(\mathcal{D}(x) \cap \mathcal{D}(y))$ is an open subset in $\mathbb{H}$;
(AT3) For arbitrary $x, y \in A$ the map $y \circ x^{-1} : x(\mathcal{D}(x) \cap \mathcal{D}(y)) \to y(\mathcal{D}(x) \cap \mathcal{D}(y))$ is a C^k diffeomorphism.

A pair $(\mathcal{M}, A)$ where $\mathcal{M}$ is a set and A is a C^k-atlas on $\mathcal{M}$, is called C^k *-manifold over the Hilbert space* $\mathbb{H}$.

Let x and y be two arbitrary charts on $\mathcal{M}$ around the point $u \in \mathcal{M}$. Let $\xi, \eta \in \mathbb{H}$ be arbitrary. Introduce the equivalence relation

$$(u, x, \xi) \sim (u, y, \eta) \Leftrightarrow \eta = (y \circ x^{-1})'(x(u))\xi.$$

The equivalence class

$$[u, x, \xi] = \{(u, y, \eta) | u \in \mathcal{D}(x) \cap \mathcal{D}(y), (u, y, \eta) \sim (u, x, \xi)\},$$

is called *tangent vector* at u. The *tangent space* of $\mathcal{M}$ at u is the set $T_u\mathcal{M}$ of all equivalence classes $[u, x, \xi]$ such that x is a chart, $u \in \mathcal{D}(x)$ and $\xi \in \mathbb{H}$. It is equipped with a vector space structure on $T_u\mathcal{M}$ given by:

$$\begin{aligned} [u, x, \xi] + [u, x, \eta] &= [u, x, \xi + \eta], \forall \xi \in \mathbb{H}, \eta \in \mathbb{H} \\ \lambda[u, x, \xi] &= [u, x, \lambda\xi], \qquad \forall \lambda \in \mathbb{R}, \xi \in \mathbb{H}. \end{aligned}$$

The *tangent bundel* $T\mathcal{M}$ of $\mathcal{M}$ is defined by $T\mathcal{M} = \cup_{u\in\mathcal{M}} T_u\mathcal{M}$.

Suppose that $\mathcal{M}$ is a C^k-manifold over the Hilbert space $\mathbb{H}$. The map $\varphi : \mathcal{U} \subset \mathcal{M} \to \mathcal{M}$ is said to be C^r *-differentiable* $(r \leq k)$ at $u \in \mathcal{M}$ if there are charts x around u and y around $\varphi(u)$ such that the map $y \circ \varphi \circ x^{-1}$ is C^r-differentiable in $x(u)$ in the sense of Fréchet.

The *differential* of φ at $u \in \mathcal{U}$ is the linear map $d_u\varphi : T_u\mathcal{M} \to T_{\varphi(u)}\mathcal{M}$, given by

$$d_u\varphi([u, x, \xi]) = [\varphi(u), y, (y \circ \varphi \circ x^{-1})'(x(u))\xi], \tag{1.1}$$

where x, y are charts around u and $\varphi(u)$, respectively, and $\xi \in \mathbb{H}$ is arbitrary.

Let a *Riemannian metric* of class C^{k-1} be defined on the connected C^k-manifold $\mathcal{M}(k \geq 2)$ over the Hilbert space $\mathbb{H}$. Suppose that at every point $u \in \mathcal{M}$ and for every chart x around u there is given a symmetric positive definite operator $G_x : \mathbb{H} \to \mathbb{H}$ with the following properties

(RM1) The map $G_x : \mathcal{D}(x) \to \mathcal{L}(\mathbb{H})$ is $\mathcal{C}^k$-smooth.

(RM2) $[(y \circ x^{-1})'(x(u))]^* G_y(u)[(y \circ x^{-1})'(x(u))] = G_x(u)$ for any two charts x, y around u.

Let $(\mathcal{M}, g)$ be a Riemannian C^r-manifold $(r \geq 3)$ over the Hilbert space $\mathbb{H}$. For any $u \in \mathcal{M}$ and any $v \in T_u\mathcal{M}$ there exists a unique geodesic $\varphi(\cdot, u, v)$ with $\varphi(0, u, v) = u$, $\dot{\varphi}(0, u, v) = v$. Then $(t, u, v) \mapsto \varphi(t, u, v)$ is a C^{r-2}-map.

DEFINITION 1.1. *The map $v \mapsto \exp_u v = \varphi(1, u, v)$ is called exponential map of class C^{r-2} around $0 \in T_u\mathcal{M}$.*

Let $\mathcal{V}$ be a sufficiently small neighborhood of $0 \in T_u\mathcal{M}$. Then the map $\exp_u : \mathcal{V} \to \exp_u \mathcal{V}$ is a C^{r-2} - diffeomorphism.

It follows for any $u \in \mathcal{M}$ and any sufficiently small number $\varepsilon > 0$ the map $\exp_u$ is a C^{r-2} -diffeomorphism on $\mathcal{B}_\varepsilon(0_u) \subset T_u\mathcal{M}$.

For any $v \in \mathcal{B}_\varepsilon(0_u)$ the map $t \mapsto c(t) = \exp_u(t, v)$ with $t \in [0, 1]$ is a geodesic on $\mathcal{M}$.

Let us define a dynamical system and an associated global attractor on the Riemannian manifold ([1, 8]). Let $(\mathcal{M}, \rho)$ be the metric space generated on the Riemannian manifold $(\mathcal{M}, G)$ and let $\{\varphi^t\}_{t \in \mathcal{J}}$ be a family of maps $\varphi^t : \mathcal{M} \to \mathcal{M}$, where $\mathcal{J} \in \{\mathbb{R}, \mathbb{R}_+, \mathbb{Z}, \mathbb{Z}_+\}$. The pair $(\{\varphi^t\}_{t \in \mathcal{J}}, (\mathcal{M}, \rho))$ is called a *dynamical system* on the metric space $(\mathcal{M}, \rho)$ if the following holds:

1. $\varphi^0 = \mathrm{id}_{\mathcal{M}}$;
2. $\varphi^{t+s} = \varphi^t \circ \varphi^s$ for all $s, t \in \mathcal{J}$;
3. $\varphi^{(\cdot)}(\cdot) : \mathcal{J} \times \mathcal{M} \to \mathcal{M}$ is smooth if $\mathcal{J} \in \{\mathbb{R}, \mathbb{R}_+\}$. The family $\varphi^t : \mathcal{M} \to \mathcal{M}$ of maps with $t \in \mathcal{J}$ is smooth if $\mathcal{J} \in \{\mathbb{Z}, \mathbb{Z}_+\}$

Let $(\{\varphi^t\}_{t \in \mathcal{J}}, (\mathcal{M}, \rho))$ be a dynamical system. A set $\mathcal{A} \subset \mathcal{M}$ is called a *global $\mathcal{B}$-attractor* for the dynamical system if the following conditions are satisfied:

(CM1) $\mathcal{A}$ is compact;

(CM2) $\mathcal{A}$ is an invariant set in the sense that $\varphi^t(\mathcal{A}) = \mathcal{A}, \forall t \in \mathcal{J}$;

(CM3) $\mathcal{A}$ attracts any bounded set $\mathcal{B} \subset \mathcal{M}$ under $\{\varphi^t\}_{t \in \mathcal{J}}$, i.e.

$$\mathrm{dist}(\varphi^t(\mathcal{B}), \mathcal{A}) \to 0 \qquad \text{for} \quad t \to \infty \tag{1.2}$$

$$\text{where} \quad \mathrm{dist}(\mathcal{Z}_1, \mathcal{Z}_2) = \sup_{u \in \mathcal{Z}_1} \inf_{v \in \mathcal{Z}_2} \rho(u, v) \tag{1.3}$$

for any nonempty subsets $\mathcal{Z}_1, \mathcal{Z}_2 \subset \mathcal{M}$ is the Hausdorff semidistance.

2. Hausdorff dimension and singular values. In the following we introduce some basic definitions and propositions of singular values for noncompact linear operators. Consider the linear not compact operator $T : \mathbb{K} \to \mathbb{K}'$, where $(\mathbb{K}, (\cdot, \cdot)_{\mathbb{K}})$ and $(\mathbb{K}', (\cdot, \cdot)_{\mathbb{K}'})$ are Hilbert spaces. (The case when $\mathbb{K} = \mathbb{K}'$ is considered in [[10]].) The *adjoint operator* $T^{[*]} : \mathbb{K}' \to \mathbb{K}$, is defined by the relation $(T\xi, \eta)_{\mathbb{K}'} = (\xi, T^{[*]}\eta)_{\mathbb{K}}$, $\forall \xi \in \mathbb{K}, \forall \eta \in \mathbb{K}'$.

The *singular values* of T, denoted by $\alpha_i(T)$, are given by

$$\alpha_k(T) = \sup_{\substack{\mathbb{L} \subset \mathbb{K} \\ \dim \mathbb{L} = k}} \inf_{\substack{\xi \in \mathbb{L} \\ |\xi|_{\mathbb{K}} = 1}} |T\xi|_{\mathbb{K}'}, \qquad k = 1, 2, \ldots. \tag{2.1}$$

Let $T^{\wedge k}:\mathbb{K}^{\wedge k}\to\mathbb{K}'^{\wedge k}$ and let consider $\omega_k(T)=\alpha(T^{\wedge k})$. The function

$$\omega_d(T)=\begin{cases}\omega_{d_0}^{1-s}(T)\cdot\omega_{d_0+1}^{s}(T), & d>0\\ 1, & d=0\end{cases}$$

is called *the singular value function* of T. Here $d\geq 0$ is written in the form $d=d_0+s$, $d_0\in\mathbb{N}_0$, $s\in(0,1]$.

Let $\{\xi_i\}_{i\in\mathcal{I}}$ be an orthonormal basis of $\mathbb{K}$ such that ξ_i is an eigenvector of $T^{[*]}T$ corresponding to the eigenvalue $\alpha_i(T)$, $i\in\mathcal{I}$. Then there exists an orthonormal basis $\{\eta_i\}_{i\in\mathcal{I}}$ in $\mathbb{K}'$ with $\eta_i=\frac{1}{\alpha_i}T\xi_i$ for any $i\in\mathcal{I}$ and $\alpha_i>0$. The image of the unit ball $B_1(0)\subset\mathbb{K}$ under the map T is the set

$$\left\{\sum_{i\in\mathcal{I},\alpha_i(T)\neq 0}c_i\eta_i\in\mathbb{K}'|\ \sum_{i\in\mathcal{I},\alpha_i(T)\neq 0}\left(\frac{c_i}{\alpha_i(T)}\right)^2\leq 1\right\}.$$

The operator $\tilde{T}=T^{[*]}T$ is positive, self-adjoint, and continuous but no longer compact. We introduce the sequence of numbers $\beta_n(\tilde{T})$, $n\geq 1$, defined by

$$\beta_n(\tilde{T})=\inf_{\substack{\mathbb{L}\subset\mathbb{K}\\ \dim\mathbb{L}=k}}\sup_{\substack{\xi\in\mathbb{L}\\ |\xi|_{\mathbb{K}}=1}}(\tilde{T}\xi,\xi)_{\mathbb{K}}. \tag{2.2}$$

The sequence $\{\beta_n(\tilde{T})\}$ is nonincreasing and we can easily see that the definition of $\beta_n(\tilde{T})$ is unchanged if we replace the infimum in (2.2) by the infimum for $\mathbb{L}\subset\mathbb{K}$. If $\tilde{T}$ is compact then, according to the well known min-max principle $\beta_n(\tilde{T})$ would be the eigenvalues of $\tilde{T}$.

We set

$$\beta_\infty(\tilde{T})=\lim_{n\to\infty}\beta_n(\tilde{T})=\inf_{n\geq 1}\beta_n(\tilde{T}). \tag{2.3}$$

The sequence is stationary at some stage:

$$\beta_1(\tilde{T})\geq\ldots\geq\beta_{n_0}(\tilde{T})>\beta_{n_0+1}(\tilde{T})=\beta_m(\tilde{T})=\beta_\infty(\tilde{T}),\qquad\forall m\geq n_0+1 \tag{2.4}$$

or

$$\beta_m(\tilde{T})>\beta_\infty(\tilde{T}),\quad\forall m\in N. \tag{2.5}$$

In the first case it follows from the above result that $\beta_1,\ldots\beta_{n_0}$, are eigenvalues of $\tilde{T}$, while in the second case each β_m is an eigenvalue of $\tilde{T}$. In both cases we decompose $\mathbb{K}$ into the direct sum $\mathbb{K}_v\oplus\mathbb{K}_v^\perp$, where $\mathbb{K}_v$ is the space spanned by the eigenvectors of $\tilde{T}, e_i, i\in I$, which we suppose orthonormalized ($I=(1,...,n_o)$ when (2.3) occurs, $I=\mathbb{N}$ when (2.4) holds). Of course, it may happen that $\mathbb{K}_v=\{O\}$ or $\mathbb{K}_v=\mathbb{K}$.

Let $\mathbb{K}=\mathbb{K}_v\oplus\mathbb{K}_v^\perp$ denote the decomposition of $\mathbb{K}$, where $\mathbb{K}_v$ and $\mathbb{K}_v^\perp$ are orthogonal. In the same way let us introduce the decomposition $\mathbb{K}'=\mathbb{K}'_v\oplus\mathbb{K}'_v{}^\perp$. Let $\{\xi_i\}_{i\in I}$ be an orthonormal basis of $\mathbb{K}_v$ such that ξ_i is an eigenvector of $T^{[*]}T$ corresponding to the eigenvalue $\alpha_i(T)$, $i\in I$. Then there exists an orthonormal basis $\{\eta_i\}_{i\in I}$ in $\mathbb{K}'$ with $\eta_i=\frac{1}{\alpha_i}T\xi_i$ for any $i\in I$ and $\alpha_i\neq 0$. We observe that the vectors $Te_i, i\in I$ are orthogonal, i. e.

$$(Te_i,Te_j)_{\mathbb{K}'}=(T^{[*]}Te_i,e_j)_{\mathbb{K}}=\beta_i(e_i,e_j)_{\mathbb{K}}=\beta_i\delta_{ij}\quad\forall i,j\in I, \tag{2.6}$$

where $\delta_{ij} = (e_i, e_j), \forall i, j \in I$.

The image of the unit ball $B_1(0) \subset \mathbb{K}$ under the map T is included in the sum of the ellipsoid $\sum_{i\in I} \frac{1}{\alpha_i^2} \left(\xi, \frac{Te_i}{\alpha_i}\right)^2 \leq 1$ of $\mathbb{K}'_v$ and of the ball of ${\mathbb{K}'_v}^{\perp}$ centered at 0 of radius $\alpha_\infty(T)$.

The next proposition is a generalization of a result of [10]

PROPOSITION 2.1. *Let $\mathbb{K}$ be a Hilbert space and B its unit ball. Let $T : \mathbb{K} \to \mathbb{K}'$ be a linear continuous operator and, if T is not compact, let $\mathbb{K}_v$ be defined as above. Then $T(B)$ is included in an ellipsoid $\mathcal{E}$:*
(i) If T is not compact, but $\mathbb{K}'_v = \mathbb{K}'$, the axes of $\mathcal{E}$ are directed along the vectors Te_i and their length is $\alpha_i(T)$, the e_i being the eigenvectors of $T^{[*]}T$.
(ii) If T is not compact and $\mathbb{K}'_v \neq \mathbb{K}'$, $\mathcal{E}$ is the product of the ball centered at 0 of radius α_∞ in ${\mathbb{K}'_v}^{\perp}$, and of the ellipsoid of $\mathbb{K}'_v$ whose axes are directed along the vectors Te_i with lengths $\alpha_i(T)$, the e_i being the eigenvectors of T spanning $\mathbb{K}'_v$.

Let $\mathcal{E}$ be an ellipsoid in the Hilbert space $\mathbb{H}'$ and let $a_1(\mathcal{E}) \geq a_2(\mathcal{E}) \geq \dots$ denote the lengths of the half-axes. For any $j \in \mathbb{N}_0$ we define

$$\omega_j(\mathcal{E}) = \begin{cases} a_1(\mathcal{E}) \cdot \ldots \cdot a_j(\mathcal{E})), & j \in \mathbb{N} \\ 1, & j = 0 \end{cases}.$$

For any $d > 0$ of the form $d = d_0 + s$ with $d_0 \in \mathbb{N}_0$ and $s \in (0, 1]$ we define

$$\omega_d(\mathcal{E}) = \omega_{d_0}^{1-s}(\mathcal{E}) \cdot \omega_{d_0}^{s}(\mathcal{E}).$$

Let $(\mathcal{M}, G)$ be a Riemannian manifold over the Hilbert space $\mathbb{H}$ and $\mathcal{K} \subset \mathcal{M}$ be a subset.

For arbitrary real numbers $\epsilon > 0$ and $d \geq 0$ we consider the *d-dimensional Hausdorff outer premeasure* at level ϵ of $\mathcal{K}$ given by

$$\mu_{\mathrm{H}}(\mathcal{K}, d, \epsilon) := \inf \sum_i r_i^d, \tag{2.7}$$

where the infimum is taken over all countable covers of $\mathcal{K}$ by balls $\mathcal{B}_{r_i}(u_i) = \{v \in \mathcal{M} | \rho(u_i, v) \leq r_i\}$ of radius $r_i \leq \epsilon$ and outer $u_i \in \mathcal{M}$. For fixed d and $\mathcal{K}$ the function $\mu_{\mathrm{H}}(\mathcal{K}, d, \epsilon)$ is monotone decreasing in $\mathcal{E}$.

Hence the limit

$$\mu_{\mathrm{H}}(\mathcal{K}, d) = \lim_{\epsilon \to 0+0} \mu_{\mathrm{H}}(\mathcal{K}, d, \epsilon) \tag{2.8}$$

exists and is called *d-dimensional Hausdorff outer measure* of $\mathcal{K}$.

For every subset $\mathcal{K} \subset \mathcal{M}$ there exists a critical number d^* with

$$\mu_{\mathrm{H}}(\mathcal{K}, d) = \begin{cases} \infty & \text{for any } 0 \leq d < d^*, \\ 0 & \text{for any } d > d^*. \end{cases} \tag{2.9}$$

This critical number can be characterized as

$$d^* = \sup\{d \geq 0 |\ \mu(\mathcal{K}, d) = \infty\}. \tag{2.10}$$

It is called *Hausdorff dimension* of $\mathcal{K}$ and denoted by $\dim_{\mathrm{H}}\mathcal{K}$.

Introduce *the global Lyapunov exponents* $\nu_1^u \geq \nu_2^u \geq \dots$ by

$$\nu_1^u + \nu_2^u + \ldots + \nu_m^u = \lim_{t\to\infty} \frac{1}{t} \log \max_{p\in\mathcal{K}} \omega_m(d_p\varphi^t), \quad m = 1, 2, \ldots.$$

The *upper Lyapunov dimension* of φ^t on $\mathcal{K}$ with respect to the global Lyapunov exponents is

$$\dim_L^u(\varphi^t,\mathcal{K}) \le N + \frac{\nu_1^u + \cdots \nu_N^u}{\nu_{N+1}^u},$$

where $N \ge 0$ denotes the smallest number satisfying $\nu_1^u + \nu_2^u + \cdots + \nu_N^u + \nu_{N+1}^u < 0$

3. Hausdorff dimension bounds for invariant sets of maps on Hilbert manifolds. Let $(\mathcal{M}, G)$ be a Riemannian manifold, let $\mathcal{U} \subset \mathcal{M}$ be an open subset and let us consider the map $\varphi : \mathcal{U} \to \mathcal{M}$ of class C^1. The tangent map of φ at a point $u \in \mathcal{U}$ is denoted by $d_u\varphi : T_u\mathcal{M} \to T_{\varphi(u)}\mathcal{M}$.

Let $u \in \mathcal{U}$ be an arbitrary point and consider charts x and x' at u and $\varphi(u)$, respectively. We introduce the operators $G_x(u) : \mathbb{H} \to \mathbb{H}$ and $G'_{x'}(\varphi(u))$ that realizes the metric fundamental tensor G in the canonical bases of $T_u\mathcal{M}$ and $T_{\varphi(u)}\mathcal{M}$, respectively. The tangent map of φ at u written in coordinates of the charts x and x' is given by the operator $\Phi = D(x' \circ \varphi \circ x^{-1})(x(u))$. The singular values of the tangent map $d_u\varphi : T_u\mathcal{M} \to T_{\varphi(u)}\mathcal{M}$ coincide with the singular values of the operator $\sqrt{G'}\Phi\sqrt{G^{-1}}$.

Let $\mathcal{K} \subset \mathcal{U}$ is a compact set and the tangent map $d_u\varphi$ be uniformly differentiable in the sense of Fréchet on the open set $\mathcal{U}$.

Let us consider the exponential map $\exp_u : T_u\mathcal{M} \to \mathcal{M}$.

By τ_v^u we denote the isometry between $T_u\mathcal{M}$ and $T_v\mathcal{M}$ defined by parallel transport along the geodesic for points lying sufficiently near to each other.

Let us fix a finite cover with balls $\mathcal{B}(u_i, r_i)_i$ of radius $r_i \le \varepsilon$ of $\mathcal{K}$. The Taylor formula for differentiable maps provides that for every $v \in \mathcal{B}(u_i, r_i)$

$$\begin{gathered} ||\exp^{-1}_{\varphi(u_i)} \varphi(v) - d_{u_i}\varphi(\exp^{-1}_{u_i}(v))|| \le \\ \sup_{w\in\mathcal{B}(u_i,r_i)} ||\tau^{\varphi(u_i)}_{\varphi(w)} d_w\varphi\tau^w_{u_i} - d_{u_i}\varphi|| \cdot ||\exp^{-1}_{u_i}(w)||. \end{gathered} \tag{3.1}$$

THEOREM 3.1. *Let $d > 0$ be a real number and $\mathcal{K} \subset \mathcal{U}$ a compact set which is negatively invariant with respect to φ, i.e. $\varphi(\mathcal{K}) \supset \mathcal{K}$. If the inequality*

$$\sup_{u\in\mathcal{K}} \omega_d(d_u\varphi) < 1 \tag{3.2}$$

holds, then $\dim_{\mathrm{H}} \mathcal{K} < d$.

In difference to the paper [7] we consider here the case when the linearization of the map φ may be a noncompact linear operator.

COROLLARY 3.2. *Let $\mathcal{K} \subset \mathcal{U} \subset \mathcal{M}$ be a compact set satisfying $\varphi(\mathcal{K}) \supset \mathcal{K}$. If for some continuous function $\kappa : \mathcal{U} \to \mathbb{R}_+$ and for some number $d > 0$ the inequality*

$$\sup_{u\in\mathcal{K}} \left(\frac{\kappa(\varphi(u))}{\kappa(u)} \omega_d(d_u\varphi) \right) < 1 \tag{3.3}$$

holds, then $\dim_{\mathrm{H}} \mathcal{K} < d$.

Let us describe the main ideas which are used in the proof of Theorem 3.1. Consider the exponential map

$$\exp_u : T_u\mathcal{M} \to \mathcal{M}, \tag{3.4}$$

where $u \in \mathcal{M}$ is an arbitrary point. Then the set $\exp_u(\mathcal{E})$ is the image of an ellipsoid $\mathcal{E}$ in the tangent space $T_u\mathcal{M}$ centered at 0 under the map $\exp_u$. Let $\mathcal{K} \subset \mathcal{U}$ be a compact set, let $\varepsilon > 0$ be a sufficiently small number and let us fix a number $d > 0$. The *outer ellipsoid premeasure* at level ε and of order d of $\mathcal{K}$ is given by

$$\tilde{\mu}_H(\mathcal{K}, d, \varepsilon) = \inf \left\{ \sum_i \omega_d(\mathcal{E}_i) \right\}, \tag{3.5}$$

where the infimum is taken over all finite covers $\cup_i \exp_{u_i}(\mathcal{E}_i) \subset \mathcal{K}$, where $u_i \in \mathcal{M}$, $\mathcal{E}_i \subset T_{u_i}\mathcal{M}$ are ellipsoids satisfying $\omega_d(\mathcal{E}_i)^{1/d} \le \varepsilon$.

The following two lemmas for the compact case of the differentional are proved in [1]. The proof for the noncompact case can be done using Proposition 2.1. The use of the two lemmas is an essential part in the proof of Theorem 3.1.

LEMMA 3.3. *For an arbitrary number $d > 0$, $d = d_0 + s$, $s \in (0, 1]$, $d_0 \in \mathbb{N}_0$ we define the numbers $\lambda = \sqrt{d_0 + 1}$ and $C_d \ge 2^{d_0}(d_0 + 1)^{d/2}$. Then for a compact set $\mathcal{K} \subset \mathcal{U}$ and for every sufficiently small $\varepsilon > 0$ the inequality*

$$\mu_H(\mathcal{K}, d, \varepsilon) \ge \tilde{\mu}_H(\mathcal{K}, d, \varepsilon) \ge C_d^{-1} \mu_H(\mathcal{K}, d, \lambda\varepsilon) \qquad \textit{holds.} \tag{3.6}$$

LEMMA 3.4. *Let $\mathcal{K} \subset \mathcal{U}$ be a compact set and consider a map $\varphi : \mathcal{U} \to \mathcal{M}$ of class C^1. For a number $d > 0$, we assume that $\sup_{u \in \mathcal{K}} \omega_d(d_u\varphi) \le k$. Then, for every $l > k$ there exists a number $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0]$*

$$\mu_H(\varphi(\mathcal{K}), d, \lambda l^{1/d}\varepsilon) \le C_d l \mu_H(\mathcal{K}, d, \varepsilon) \tag{3.7}$$

holds, where $\lambda = \sqrt{d_0 + 1}$, $C_d \ge 2^{d_0}(d_0 + 1)^{d/2}$, $d = d_0 + s$, $s \in (0, 1]$, $d_0 \in \mathbb{N}_0$.

4. Hausdorff dimension bounds for invariant sets of vector fields on Hilbert manifolds. Let $(\mathcal{M}, G)$ be a Riemannian manifold, let $\mathcal{U} \subset \mathcal{M}$ be an open subset and $\mathcal{I}_1 \subset \mathbb{R}$ be an open interval with 0. We consider a time-dependent vector field $F : \mathcal{I}_1 \times \mathcal{U} \to T\mathcal{U}$ of class C^1 and the corresponding differential equation

$$\dot{u} = F(t, u). \tag{4.1}$$

Suppose, that for a point $(t, u) \in \mathcal{I}_1 \times \mathcal{U}$ the *covariant derivative* of the vector field F is $\nabla F(t, u) : T_u\mathcal{M} \to T_u\mathcal{M}$ and ∇F is a compact operator. The case when ∇F is noncompact can be also considered with the help of Section 3.

Let $\mathcal{D} \subset \mathcal{U}$ be an open set and $\mathcal{I} \subset \mathcal{I}_1$ be an open interval such that the solution $\varphi(\cdot, u)$ with $\varphi(0, u) = u$, $u \in \mathcal{D}$ of equation (15) exists everywhere on $\mathcal{I}$.

For every $t \in \mathcal{I}$ there exists the operator $\varphi^t : \mathcal{D} \to \mathcal{U}$ such that $\varphi^t(u) = \varphi(t, u)$.

Since the vector field F is continuously differentiable, the same holds for the operator $\{\varphi^t\}_{t \in \mathcal{I}}$. For an arbitrary point $u \in \mathcal{D}$, the tangent map $d_u\varphi^t$ solves the variation equation

$$y' = \nabla F(t, \varphi^t(u))y \tag{4.2}$$

with initial condition $d_u\varphi^t|_{t=0} = \mathrm{id}_{T_u\mathcal{M}}$.

Here the *absolute derivative* y' is taken along the integral curve $t \mapsto \varphi^t(u)$ in the direction of the vector field F.

Let us denote the eigenvalues of the symmetric part of the covariant derivative ∇F, i.e., of the operator

$$S(t,u) = \frac{1}{2}[\nabla F(t,u) + \nabla F(t,u)^{[*]}], \tag{4.3}$$

by $\lambda_i(t,u)$, $i = 1, 2, \ldots$ and order them with respect to its size and multiplicity, i.e., $\lambda_1(t,u) \geq \lambda_2(t,u) \geq \ldots$.

Let us introduce on $\mathcal{U}$ a new metric tensor $\tilde{g}_{|u} = \kappa^2(u) g_{|u}$ by means of a function $\kappa : \mathcal{U} \to \mathbb{R}_+$ of class C^1. Let $u \in \mathcal{U}$ be a fixed point and consider the chart x around u. Let $V : \mathcal{U} \to \mathbb{R}$ be a differentiable function and the map $\dot{V} : \mathcal{I} \times \mathcal{U} \to \mathbb{R}$ be defined by $\dot{V}(t,u) = \langle d_u V, F(t,u) \rangle$. The symmetric part of the covariant derivative $\tilde{\nabla} F(t,u)$ at $u \in \mathcal{U}$ with respect to the new metric is given by

$$\frac{1}{2}[G^{-1}\Phi^T G + \Phi] + \frac{\dot{\kappa}}{\kappa}\mathrm{Id}, \tag{4.4}$$

where $\Phi = D(\tilde{x} \circ \varphi \circ x^{-1})(x(u))$ and the operator G represents $g_{|u}$.

If

$$\kappa(u) = e^{\frac{V(u)}{d}} \quad (u \in \mathcal{U}) \tag{4.5}$$

then $\dot{\kappa}(u) = \kappa(u)\frac{\dot{V}(u)}{d}$ implies that the eigenvalues $\tilde{\lambda}_i$ of (4.4) are related to the eigenvalues with respect to the original metric g by $\tilde{\lambda}_i = \lambda_i + \frac{\dot{V}}{d}, i = 1, 2, \ldots$.

The next theorems are corollaries of Theorem 3.1.

THEOREM 4.1. *Let $d > 0$, be a real number written in the form $d = d_0 + s$ with $d_0 \in \mathbb{N}_0$, $s \in (0,1]$ and let $\mathcal{K} \subset \mathcal{D}$ be a compact set satisfying $\varphi^\tau(\mathcal{K}) \supset \mathcal{K}$ for a certain $\tau \in \mathcal{I} \cap \mathbb{R}_+$. If the condition*

$$\sup_{u \in \mathcal{K}} \int_0^\tau [\lambda_1(t, \varphi^t(u)) + \lambda_2(t, \varphi^t(u)) + \ldots + \lambda_{d_0}(t, \varphi^t(u)) + s\lambda_{d_0+1}(t, \varphi^t(u))]dt < 0$$

holds, then $\dim_{\mathrm{H}} \mathcal{K} \leq d$.

THEOREM 4.2. *Let $\mathcal{K} \subset \mathcal{D}$ be a compact set such that $\varphi^\tau(\mathcal{K}) \supset \mathcal{K}$ is true for some $\tau \in \mathcal{I} \cap \mathbb{R}_+$. Let $V : \mathcal{U} \to \mathbb{R}$ be a differentiable function and denote by $\lambda_1(t,u) \geq \lambda_2(t,u) \geq \ldots$ the eigenvalues of $S(t,u)$. If for a real number $d > 0$ $d = d_0 + s$ with $d_0 \in \mathbb{N}_0$ and $s \in (0,1]$ the condition*

$$\begin{aligned} \sup_{u \in \mathcal{K}} \int_0^\tau [\lambda_1(t, \varphi^t(u)) + \lambda_2(t, \varphi^t(u)) + \ldots \\ + \lambda_{d_0}(t, \varphi^t(u)) + s\lambda_{d_0+1}(t, \varphi^t(u)) + \dot{V}(t, \varphi^t(u))]dt < 0 \end{aligned} \tag{4.6}$$

holds, then $\dim_{\mathrm{H}} \mathcal{K} \leq d$.

The application of the Theorem 4.1 and Theorem 4.2 for the compact case to the sine-Gordon equation given on the cylinder was considered in the paper [7]. The non-compact version of these theorems can be applied to estimate the Hausdorff dimension of an attractor for the Ginzburg-Landau equation [3] using a nontrivial metric tensor instead of the Lyapunov function used in this paper. Thus it is possible to calculate the Lyapunov dimension $\dim_L^u(\varphi^t, \mathcal{K})$, introduced in Section 2, for this equation.

REFERENCES

[1] V. A. Boichenko, G. A. Leonov and V. Reitmann, *Dimension Theory for Ordinary Differential Equations*. Wiesbaden:Vieweg-Teubner Verlag, 2005.

[2] I. D. Chueshov, *Introduction to the Theory of Infinite-Dimensional Dissipative Systems*. ACTA. Kharkov, 1999 (in Russian). English translation: Acta, Kharkov (2002) (see http://www.emis.de/monographs/Chueshov).

[3] C. R. Doering, J. D. Gibbon, D. D. Holm and B. Nicolaenko *Exact Lyapunov Dimension of the Universal Attractor for the Complex Ginzburg-Landau Equation*. Phys. Rev. Lett. 59, Iss. 26-28 (1987) pp. 2911–2914

[4] A. Douady, J. Oesterle, *Dimension de Hausdorff des attracteurs*. Comptes Rendus de l'Academie des Sciences Paris Serie A. (1980). 290. pp. 1135-1138

[5] M. Ghidaglia, R. Temam, *Attractors for damped nonlinear hyperbolic equations*. J. Math. Pures Appl., 66, (1987), pp. 273–319.

[6] M. A. Henon, *A two-dimensional mapping with a strange attractor*. Commun. Math. Phys. (1976). Vol. 50, 2. pp. 69–77.

[7] A. V. Kruck, A. E. Malykh and V. Reitmann, *Upper Hausdorff dimension estimates and stratification for invariant sets of evolutionary systems on Hilbert manifolds*. Differential equations, 2017 (to appear).

[8] S. Lang, *Differential and Riemannian Manifolds* . Springer, New York , 1995

[9] G. A. Leonov, V. Reitmann, V. B. Smirnova, *Non-local Methods for Pendulum-like Feedback Systems*. Teubner-Texte zur Mathematik, Bd. 132, B.G. Teubner Stuttgart- Leipzig, 1992.

[10] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. New York-Berlin: Springer-Verlag, 1988.

Proceedings of EQUADIFF 2017
pp. 255–264

GAUSSIAN CURVATURE BASED TANGENTIAL REDISTRIBUTION OF POINTS ON EVOLVING SURFACES*

MATEJ MEDĽA† AND KAROL MIKULA ‡

Abstract. There exist two main methods for computing a surface evolution, level-set method and Lagrangian method. Redistribution of points is a crucial element in a Lagrangian approach. In this paper we present a point redistribution that compress quads in the areas with a high Gaussian curvature. Numerical method is presented for a mean curvature flow of a surface approximated by quads.

Key words. surface evolution, point redistribution, finite volume method, mean curvature flow

AMS subject classifications. 53C44, 65M08, 65M50

1. Introduction. An important part in computing a surface evolution is a redistribution of points on a surface. An improper distribution of points could lead to an unstable numerical method.

There are several papers dedicated to this problem. One of the approaches is using the so-called Laplacian smoothing method [1]. A method that we build on is controlling the so-called local area density [4]. In a paper [5] a method for a redistribution of points on a curve by a curvature was presented. We generalize this method for evolving surfaces. In a case of surfaces we redistribute points by Gaussian curvature, since saddle point can have a zero mean curvature but it has non-zero Gaussian curvature.

2. Surface evolution models. Let us have an open parametric surface $E = \{\mathbf{x}(t,u,v) | t \in [0,T),\ (u,v) \in \Omega = [0,1] \times [0,1]\}$ evolving in a time t by the following partial differential equation

$$\frac{\partial \mathbf{x}}{\partial t}(t,u,v) = \beta(t,u,v)\mathbf{N}(t,u,v) + \mathbf{V_T}(t,u,v),\ t \in (0,T),\ (u,v) \in \Omega \setminus \partial\Omega, \tag{2.1}$$

where $\mathbf{N}(t,u,v)$ is a unit normal vector and $T > 0$. The vector $\mathbf{V_T}$ represents the evolution in a tangential direction along the surface. An evolution in the tangential direction does not change the image of the surface. In a quad (quadrilateral) approximation of a surface it only change the size and the shape of the quads.

In our numerical experiments we are focusing only on two special cases

$$\beta(t,u,v)\mathbf{N}(t,u,v) = \Delta_{\mathbf{x}}\mathbf{x}(t,u,v), \tag{2.2}$$

$$\beta(t,u,v)\mathbf{N}(t,u,v) = \Delta_{\mathbf{x}}\mathbf{x}(t,u,v) + \mathbf{N}(t,u,v), \tag{2.3}$$

where $\Delta_{\mathbf{x}}\mathbf{x}(t,u,v)$ is a Laplace-Beltrami operator applied on the position vector of the parametrized surface E. This is known to be the mean curvature vector of the

*This work was supported by Grant No.: APVV-15-0522 and VEGA 1/0608/15.

†Department of Mathematics, Slovak University of Technology in Bratislava, Slovakia medla@math.sk.

‡Department of Mathematics, Slovak University of Technology in Bratislava, Slovakia mikula@math.sk.

surface. We want to emphasize that $\Delta_{\mathbf{x}}\mathbf{x}$ is normal to the surface E and does not depend on the parametrization $\mathbf{x}$. It depends only on the shape of E.

We want to have a surface evolution that does not change the boundary curve but can change the distribution of points along the boundary. For this reason we have the following boundary conditions

$$\frac{\partial \mathbf{x}}{\partial t}(t,u,v) = \mathbf{V_T}(t,u,v) \quad t \in (0,T),\ (u,v) \in \partial\Omega, \tag{2.4}$$

where the vector $\mathbf{V_T}$ lies in a tangent direction of a boundary curve. Let us also have the initial condition

$$\mathbf{x}(0,u,v) = \mathbf{x}_0(u,v),\ (u,v) \in \Omega \setminus \partial\Omega. \tag{2.5}$$

3. The tangential redistribution. The variable that we want to control by the tangential redistribution is the local area density

$$g(t,u,v) = ||\partial_u \mathbf{x}(t,u,v) \times \partial_v \mathbf{x}(t,u,v)||. \tag{3.1}$$

It can be understood as the area of the parallelogram with sides $\partial_u \mathbf{x}(t,u,v)$ and $\partial_v \mathbf{x}(t,u,v)$. In a quad approximation of a surface the area density g is proportional to the area of the quads.

In the rest of this section we derive a formula for $\mathbf{V_T}$ in the equation (2.1) that provide us a desired area density.

3.1. Change of the area density in time. For the derivative of the area density it applies [4]

$$\partial_t g = g\Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N} + g\nabla_{\mathbf{x}} \cdot \mathbf{V_T}. \tag{3.2}$$

If we want the area density to converge to a prescribed local area density $c(t,u,v)$, one of the possibilities is to find an area density that satisfies the following ODE

$$\partial_t \left(\frac{g}{A}\right) = \left(\frac{c}{A} - \frac{g}{A}\right)\omega, \tag{3.3}$$

where A is the area of the surface and ω is a parameter controlling the rate at which g converges to c.

By rearranging the equation (3.3) and by substituting the equation (3.2) into it we get

$$\nabla_{\mathbf{x}} \cdot \mathbf{V_T} = \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N} - \frac{1}{A}\iint_E \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N}\, \mathrm{d}\mathbf{x} + \left(\frac{c}{g} - 1\right)\omega. \tag{3.4}$$

The equation (3.4) does not have a unique solution. By taking a vector field $\mathbf{V_T}$ that is a gradient of some potential φ and a Neumann boundary condition we obtain a PDR that has an infinity many solutions that differs only by a constant. By giving a Dirichlet boundary condition in one arbitrary point we ensure uniqueness of the solution. Since we are only interested in the gradient of φ it does not matter in which point we prescribe the Dirichlet BC. The equation for the potential with the boundary conditions is

$$\begin{aligned}\nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}}\varphi(\cdot,u,v) = \Delta_{\mathbf{x}}\varphi(\cdot,u,v) = \\ \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N} - \frac{1}{A}\iint_E \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N}\, \mathrm{d}\mathbf{x} + \left(\frac{c}{g} - 1\right)\omega, \quad (u,v) \in \Omega \setminus \partial\Omega,\end{aligned} \tag{3.5}$$

$$\nabla_{\mathbf{x}}\varphi(\cdot,u,v) \cdot \mathbf{n}\,(\cdot,u,v) = 0, \quad (u,v) \in \partial\Omega \setminus \{(0,0)\}, \tag{3.6}$$

$$\varphi(\cdot,u,v) = 0, \quad (u,v) = (0,0). \tag{3.7}$$

A Neumann boundary condition provides a tangential vector field which has a zero projection to the normal of the boundary. This ensures that points on the boundary are moving only in the direction of the tangential vector of the boundary curve.

Then equations (2.1)-(2.4) acquires the form

$$\frac{\partial \mathbf{x}}{\partial t}(t,u,v) = \Delta_{\mathbf{x}}\mathbf{x}(t,u,v) + \nabla_{\mathbf{x}}\varphi(t,u,v), \quad t \in (0,T),\ (u,v) \in \Omega \setminus \partial\Omega \tag{3.8}$$

$$\frac{\partial \mathbf{x}}{\partial t}(t,u,v) = \nabla_{\mathbf{x}}\varphi(t,u,v), \quad t \in (0,T),\ (u,v) \in \partial\Omega \tag{3.9}$$

4. Choice of the function c. The choice of the function c is crucial for the distribution of points on the surface. An appropriate choice of c can provide a quad approximation of the surface with large quads in the areas with a small Gaussian curvature $G(t,u,v)$ and vice versa. There are two properties that the function c has to satisfy,

$$c(t,u,v) > 0, \quad \iint_{\Omega} c(t,u,v)\ \mathrm{dudv} = A. \tag{4.1}$$

The first property has to be satisfied since the size of the quad cannot be negative. Numerical interpretation of the second property is that the sum of the quad sizes has to be equal to the area of the surface.

If we choose c to be inverse proportional to the Gaussian curvature, we obtain a surface approximation with smaller quads in areas of high Gaussian curvature. There are multiple options for how to choose this dependence. First let us define an auxiliary function $\hat{c}$ that has the form

$$\hat{c}(t,u,v) = \left(p \min\left(|G(t,u,v)|/\widetilde{G}, 1\right) + 1\right)^{-1}, \tag{4.2}$$

where $\widetilde{G}$ is a chosen value that is restricting the maximal value of a function nad p is the chosen parameter. Then the function c is the function $\hat{c}$ normalized

$$c(t,u,v) = A\frac{\hat{c}(t,u,v)}{\iint_{\Omega} \hat{c}(t,u,v)\ \mathrm{dudv}}. \tag{4.3}$$

This normalization ensures that the second property (4.1) is fulfilled.

5. Surface and PDEs approximation. Let us divide the surface E in the t-th time step into quadrilaterals. Let us denote the vertices of the quads of the surface $\mathbf{x}_{t,i},\ i \in \{1,...,N\}$. Let us have a Q_i quads that has a vertex $\mathbf{x}_{t,i}$. Then let us denote 4 vertices of the q-th quad of $\mathbf{x}_{t,i}$ by $\mathbf{x}_{t,i}^{q,j},\ j \in \{0,1,2,3\},\ q \in Q_i$. The vertex $\mathbf{x}_{t,i}^{q,0} = \mathbf{x}_{t,i}$ and other vertices are numbered in an anticlockwise direction. For the vertex $\mathbf{x}_{t,i}^{q,3}$ holds $\mathbf{x}_{t,i}^{q,3} = \mathbf{x}_{t,i}^{q+1,1}$, where $q+1$ is as a $\mathrm{mod}(q+1, Q_i)$. Let us have a function $k(i,q,j)$ that takes the local indexes of a vertex and return its global index. For a better understanding see Fig. 7.1.

Let us interpolate values of $\mathbf{x}$ on quads using a bilinear interpolation

$$\mathbf{x}_{t,i}^{q}(\phi,\rho) = (1-\phi)(1-\rho)\mathbf{x}_{t,i}^{q,0} + \phi(1-\rho)\mathbf{x}_{t,i}^{q,1} + (1-\phi)\rho\mathbf{x}_{t,i}^{q,3} + \phi\rho\mathbf{x}_{i}^{q,2}. \tag{5.1}$$

Every function defined on the surface E is also approximated using bilinear interpolation

$$f_{t,i}^{q}(\phi,\rho) = (1-\phi)(1-\rho)f_{t,i}^{q,0} + \phi(1-\rho)f_{t,i}^{q,1} + (1-\phi)\rho f_{t,i}^{q,3} + \phi\rho f_{t,i}^{q,2}, \tag{5.2}$$

where $f_{t,i}^{q,j}$ is the value of the function f in the vertex $\mathbf{x}_{t,i}^{q,j}$.

Let us have a finite volume $V_{t,i}$ composed of quads defined by the centers of the original quads and the centers of their edges. Let us denote the edges of $V_{t,i}$ on the q-th quad by

$$e_{t,i}^{q,1} = \{\mathbf{x}_{t,i}^{q}(1/2,\rho); \rho \in (0,1/2)\}, \quad e_{t,i}^{q,3} = \{\mathbf{x}_{t,i}^{q}(\phi,1/2); \phi \in (0,1/2)\} \tag{5.3}$$

At last let us integrate the equation (3.5) over the finite volume

$$\begin{aligned} \iint_{V_{t,i}} \Delta_{\mathbf{x}}\varphi \,\mathrm{d}\mathbf{x} &= \iint_{V_{t,i}} \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N} \,\mathrm{d}\mathbf{x} \\ &- \iint_{V_{t,i}} \frac{1}{A} \iint_{E} \Delta_{\mathbf{x}}\mathbf{x} \cdot \beta\mathbf{N} \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x} + \iint_{V_{t,i}} \left(\frac{c}{g} - 1\right) \omega \,\mathrm{d}\mathbf{x} \end{aligned} \tag{5.4}$$

and also the equation (3.8)

$$\iint_{V_{t,i}} \frac{\partial \mathbf{x}}{\partial t} \,\mathrm{d}\mathbf{x} = \iint_{V_{t,i}} \Delta_{\mathbf{x}}\mathbf{x} \,\mathrm{d}\mathbf{x} + \iint_{V_{t,i}} \nabla_{\mathbf{x}}\varphi \,\mathrm{d}\mathbf{x} \tag{5.5}$$

For the boundary condition (3.9) it holds that the derivative of a potential φ on the right side in the normal direction is zero. That means yhe direction of the gradient of φ is the tangential direction to the boundary curve. For this reason we have 1D finite volumes on the boundary. They are defined by the points $\left(\mathbf{x}_{t,i}^{1,0} + \mathbf{x}_{t,i}\right)/2$, $\mathbf{x}_{t,i}$, $\left(\mathbf{x}_{t,i}^{Q_i,3} + \mathbf{x}_{t,i}\right)/2$ and after integrating (3.9) on this finite volume we get

$$\int_{V_{t,i}} \frac{\partial \mathbf{x}}{\partial t} \,\mathrm{d}\mathbf{x} = \int_{V_{t,i}} \nabla_{\mathbf{x}}\varphi \,\mathrm{d}\mathbf{x}. \tag{5.6}$$

The integral equations (5.4)-(5.6) form a basis for the finite volume method which leads for (5.5)-(5.6) to a system of equations in a matrix form

$$T_t^+ \mathbf{X}_{t+1} + T_t^- \mathbf{X}_t = B_t \mathbf{X}_{t+1} + A_t \mathbf{X}_{t+1}, \tag{5.7}$$

$$\mathbf{X}_{t+1} = [\mathbf{x}_{t+1,1}, \mathbf{x}_{t+1,2}, \ldots, \mathbf{x}_{t+1,N}]^T, \quad \mathbf{X}_t = [\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \ldots, \mathbf{x}_{t,N}]^T \quad . \tag{5.8}$$

The matrices T_{t+1}^+, T_t^- are related to the time derivative, the matrix B_t is related to the evolution in the normal direction and the matrix A_t is related to the evolution in the tangential direction.

For the equation (5.4), it leads to a system of equations

$$D_t \Phi_t = \mathbf{b}_t, \quad \Phi_t = [\varphi_{t,1}, \varphi_{t,2}, \ldots, \varphi_{t,N}]^T. \tag{5.9}$$

The matrix D_t is related to the Laplace-Beltrami operator and $\mathbf{b}_t$ is related to the right hand side of the equation (5.4).

6. The computational algorithm. The algorithm to numerically solve the equations (2.1)-(2.4) (or (3.8)-(3.9)) is as follows.

Let us have a known initial condition $\mathbf{X}_0$ and a number of time steps M.

For$(t = 0; t < M; t++)$

- compute the matrices T_t^+, T_t^-, B_t

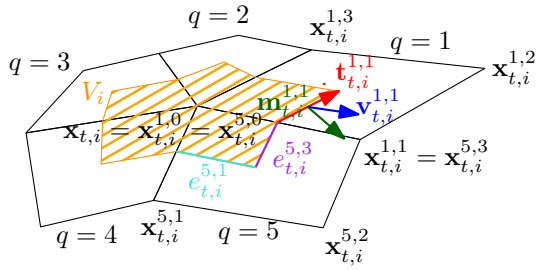


FIG. 7.1. *A sketch of the finite volume V_i composed of five quads. Local notation for quad number 1 and 5 are labeled. For the quad 1, vectors $\mathbf{m}_{t,i}^{1,1}$, $\mathbf{t}_{t,i}^{1,1}$, $\mathbf{v}_{t,i}^{1,1}$ are labeled. For the quad 5, edges $e_{t,i}^{5,1}$, $e_{t,i}^{5,3}$ are labeled.*

- use these matrices to explicitly compute $(\beta\mathbf{N})_{t,i}$ that is used in $\mathbf{b}_t$
- compute the matrix D_t and $\mathbf{b}_t$
- find Φ_t by solving $D_t\Phi_t = \mathbf{b}_t$
- use Φ_t to compute the matrix A_t
- find $\mathbf{X}_{t+1}$ by solving (5.7)

7. The finite volume method. In this section we present the coefficients of the matrices from the previous section derived by the finite volume method. A detailed derivation of the coefficients can be found in a forthcoming paper [2].

7.1. The approximation of the time derivative. Let us assume a constant time derivative on the finite volume and let us approximate the time derivative by a finite difference. Then the first integral in the equation (5.5) becomes

$$m(V_{t,i})\frac{\mathbf{x}_{t+1,i}-\mathbf{x}_{t,i}}{\tau}, \tag{7.1}$$

where

$$\begin{aligned} m(V_{t,i}) = \sum_{q=1}^{Q_i} &\left\| \frac{\mathbf{x}_{t,i}^{q,1}-\mathbf{x}_{t,i}^{q,0}}{2} \times \left(\frac{\mathbf{x}_{t,i}^{q,0}+\mathbf{x}_{t,i}^{q,1}+\mathbf{x}_{t,i}^{q,2}+\mathbf{x}_{t,i}^{q,3}}{4} - \mathbf{x}_{t,i}^{q,0} \right) \right\| \Big/ 2 \\ &+ \left\| \frac{\mathbf{x}_{t,i}^{q,3}-\mathbf{x}_{t,i}^{q,0}}{2} \times \left(\frac{\mathbf{x}_{t,i}^{q,0}+\mathbf{x}_{t,i}^{q,1}+\mathbf{x}_{t,i}^{q,2}+\mathbf{x}_{t,i}^{q,3}}{4} - \mathbf{x}_{t,i}^{q,0} \right) \right\| \Big/ 2 \end{aligned} \tag{7.2}$$

and τ is the time step. Then the only non-zero coefficients of the matrices T_t^+ and T_t^- are the diagonal coefficients

$$T_{t,i,i}^+ = m(V_{t,i})/\tau, \quad T_{t,i,i}^- = -m(V_{t,i})/\tau. \tag{7.3}$$

7.2. The finite volume approximation of the Laplace-Beltrami operator. We are applying Laplace-Beltrami operator to the vector function $\mathbf{x}(t,u,v)$ and to a scalar function $\varphi(t,u,v)$. Using a bilinear approximation we can derive the following vectors on the edges $e_{t,i}^{q,1}$, $e_{t,i}^{q,3}$ (see Fig. 7.1)

$$\begin{aligned} \mathbf{t}_{t,i}^{q,1} &= -\frac{1}{2}\mathbf{x}_{t,i}^{q,0} - \frac{1}{2}\mathbf{x}_{t,i}^{q,1} + \frac{1}{2}\mathbf{x}_{t,i}^{q,3} + \frac{1}{2}\mathbf{x}_{t,i}^{q,2}, \\ \mathbf{t}_{t,i}^{q,3} &= -\frac{1}{2}\mathbf{x}_{t,i}^{q,0} + \frac{1}{2}\mathbf{x}_{t,i}^{q,1} - \frac{1}{2}\mathbf{x}_{t,i}^{q,3} + \frac{1}{2}\mathbf{x}_{t,i}^{q,2}. \end{aligned} \tag{7.4}$$

$$\begin{aligned}\mathbf{v}_{t,i}^{q,1} &= -\frac{3}{4}\mathbf{x}_{t,i}^{q,0} + \frac{3}{4}\mathbf{x}_{t,i}^{q,1} - \frac{1}{4}\mathbf{x}_{t,i}^{q,3} + \frac{1}{4}\mathbf{x}_{t,i}^{q,2},\\ \mathbf{v}_{t,i}^{q,3} &= -\frac{3}{4}\mathbf{x}_{t,i}^{q,0} - \frac{1}{4}\mathbf{x}_{t,i}^{q,1} + \frac{3}{4}\mathbf{x}_{t,i}^{q,3} + \frac{1}{4}\mathbf{x}_{t,i}^{q,2},\end{aligned} \tag{7.5}$$

$$\mathbf{m}_{t,i}^{q,1} = \mathbf{v}_{t,i}^{q,1} - \frac{\mathbf{v}_{t,i}^{q,1}\cdot\mathbf{t}_{t,i}^{q,1}}{\mathbf{t}_{t,i}^{q,1}\cdot\mathbf{t}_{t,i}^{q,1}}\mathbf{t}_{t,i}^{q,1}, \quad \mathbf{m}_{t,i}^{q,3} = \mathbf{v}_{t,i}^{q,3} - \frac{\mathbf{v}_{t,i}^{q,3}\cdot\mathbf{t}_{t,i}^{q,3}}{\mathbf{t}_{t,i}^{q,3}\cdot\mathbf{t}_{t,i}^{q,3}}\mathbf{t}_{t,i}^{q,3}. \tag{7.6}$$

If $f_{t,i}^{q,j}$ is one of the coordinates of $\mathbf{x}_{t,i}^{q,j}$ then the q-th quad contributes to the coefficients $B_{t,i,k(i,q,j)}$ by the values

$$\begin{aligned}B_{t,i,k(i,q,0)} &+= \frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(-\frac{3}{4}+\frac{1}{2}a_i^{q,1}\right) + \frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(-\frac{3}{4}+\frac{1}{2}a_i^{q,3}\right),\\ B_{t,i,k(i,q,1)} &+= \frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(\frac{3}{4}+\frac{1}{2}a_i^{q,1}\right) + \frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(-\frac{1}{4}-\frac{1}{2}a_i^{q,3}\right),\\ B_{t,i,k(i,q,2)} &+= \frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(\frac{1}{4}-\frac{1}{2}a_i^{q,1}\right) + \frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(\frac{1}{4}-\frac{1}{2}a_i^{q,3}\right),\\ B_{t,i,k(i,q,3)} &+= \frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(-\frac{1}{4}-\frac{1}{2}a_i^{q,1}\right) + \frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(\frac{3}{4}+\frac{1}{2}a_i^{q,3}\right).\end{aligned} \tag{7.7}$$

where

$$a_{t,i}^{q,1} = \frac{\mathbf{v}_{t,i}^{q,1}\cdot\mathbf{t}_{t,i}^{q,1}}{\mathbf{t}_{t,i}^{q,1}\cdot\mathbf{t}_{t,i}^{q,1}}, \quad a_{t,i}^{q,3} = \frac{\mathbf{v}_{t,i}^{q,3}\cdot\mathbf{t}_{t,i}^{q,3}}{\mathbf{t}_{t,i}^{q,3}\cdot\mathbf{t}_{t,i}^{q,3}}. \tag{7.8}$$

If $f_{t,i}^{q,j} = \varphi_{t,i}^{q,j}$ then the q-th quad contributes to the coefficients $\Phi_{t,i,k(i,q,j)}$ by the same values.

7.3. The approximation of the right hand side in the equation (5.4). Let us approximate the first integral in the equation (5.4) by assuming $\beta\mathbf{N}$ and $\left(\frac{c}{g}-1\right)\omega$ are constant on a finite volume. Let us denote this constant value on the finite volume $V_{t,i}$ by $(\beta\mathbf{N})_{t,i}$ and $\left(\frac{c_{t,i}}{g_{t,i}}-1\right)\omega$. This vector can be approximated by explicitly computing the movement in the normal direction

$$(\beta\mathbf{N})_{t,i} = \left((B_{t,i}-T_{t,i}^{-})\cdot\mathbf{X}_t/T_{t,i,i}^{+} - \mathbf{x}_{t,i}\right)/\tau. \tag{7.9}$$

Then the right hand side has the form

$$\begin{aligned}\mathbf{b}_{t,i} &= (\beta\mathbf{N})_{t,i}\cdot\iint_{V_{t,i}}\Delta_{\mathbf{x}}\mathbf{x}\,\mathrm{d}\mathbf{x}\\ &\quad - m(V_{t,i})\frac{1}{A}(\beta\mathbf{N})_{t,i}\cdot\sum_{j=1}^{N}\iint_{V_j}\Delta_{\mathbf{x}}\mathbf{x}\,\mathrm{d}\mathbf{x} + m(V_{t,i})\left(\frac{c_{t,i}}{g_{t,i}}-1\right)\omega\end{aligned} \tag{7.10}$$

and $\iint_{V_{t,i}}\Delta_{\mathbf{x}}\mathbf{x}\,\mathrm{d}\mathbf{x}$ is approximated as in section (7.2).

The local area density $g_{t,i}$ in (7.10) is dependent on the parametrization $\mathbf{x}(u,v)$. In the numerical approximation of the surface we do not have any so we take such

$\mathbf{x}(u,v)$ that projects a rectangle dudv with size 1 onto the quarter of quad. So $g_{t,i}$ is approximated by

$$g_{t,i} = m(V_{t,i})/Q_i \tag{7.11}$$

and $c_{t,i}$ is approximated by

$$c_{t,i} = 1\Big/\left(p\min\left(|G_{t,i}|/\widetilde{G},1\right)+1\right)\Big/\sum_{j=1}^{N} Q_j\Big/\left(p\min\left(|G_{t,j}|/\widetilde{G},1\right)+1\right). \tag{7.12}$$

Finally the Gaussian curvature is approximated by [3]

$$\begin{aligned} G_{t,i} = \frac{4}{m(V_{t,i})}\Bigg(2\pi - \sum_{q=1}^{Q_i} &\arccos\left(\frac{(\mathbf{x}_{t,i}^{q,1}-\mathbf{x}_{t,i}^{q,0})\cdot(\mathbf{x}_{t,i}^{q,2}-\mathbf{x}_{t,i}^{q,0})}{||\mathbf{x}_{t,i}^{q,1}-\mathbf{x}_{t,i}^{q,0}||\,||\mathbf{x}_{t,i}^{q,2}-\mathbf{x}_{t,i}^{q,0}||}\right)\\ &+\arccos\left(\frac{(\mathbf{x}_{t,i}^{q,3}-\mathbf{x}_{t,i}^{q,0})\cdot(\mathbf{x}_{t,i}^{q,2}-\mathbf{x}_{t,i}^{q,0})}{||\mathbf{x}_{t,i}^{q,3}-\mathbf{x}_{t,i}^{q,0}||\,||\mathbf{x}_{t,i}^{q,2}-\mathbf{x}_{t,i}^{q,0}||}\right)\\ &+\arccos\left(\frac{(\mathbf{x}_{t,i}^{q,3}-\mathbf{x}_{t,i}^{q,0})\cdot(\mathbf{x}_{t,i}^{q,1}-\mathbf{x}_{t,i}^{q,0})}{||\mathbf{x}_{t,i}^{q,3}-\mathbf{x}_{t,i}^{q,0}||\,||\mathbf{x}_{t,i}^{q,1}-\mathbf{x}_{t,i}^{q,0}||}\right)\Bigg). \end{aligned} \tag{7.13}$$

7.4. The finite volume approximation of the surface gradient. In this section we approximate the integral of the surface gradient in the equation (5.5). Let us approximate the function φ using a bilinear interpolation. Let us define

$$\begin{aligned} \phi_{t,i}^{q,1} &= \frac{3}{8}\varphi_{t,i}^{q,0}+\frac{3}{8}\varphi_{t,i}^{q,1}+\frac{1}{8}\varphi_{t,i}^{q,2}+\frac{1}{8}\varphi_{t,i}^{q,3}-\tilde{\varphi}_{t,i},\\ \phi_{t,i}^{q,3} &= \frac{3}{8}\varphi_{t,i}^{q,0}+\frac{1}{8}\varphi_{t,i}^{q,1}+\frac{1}{8}\varphi_{t,i}^{q,2}+\frac{3}{8}\varphi_{t,i}^{q,3}-\tilde{\varphi}_{t,i}, \end{aligned} \tag{7.14}$$

where

$$\tilde{\varphi}_{t,i} = \frac{1}{Q_i}\sum_{q=1}^{Q_i}\frac{1}{8}\varphi_{t,i}^{q,0}+\frac{2}{8}\varphi_{t,i}^{q,1}+\frac{1}{8}\varphi_{t,i}^{q,2}+\frac{2}{8}\varphi_{t,i}^{q,3}. \tag{7.15}$$

Then the q-th quad contributes to the coefficients $A_{t,i,k(i,q,j)}$ by the values

$$\begin{aligned} A_{t,i,k(i,q,0)} &+= \phi_{t,i}^{q,1}\frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(-\frac{3}{4}+\frac{1}{2}a_i^{q,1}\right)+\phi_{t,i}^{q,3}\frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(-\frac{3}{4}+\frac{1}{2}a_i^{q,3}\right),\\ A_{t,i,k(i,q,1)} &+= \phi_{t,i}^{q,1}\frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(\frac{3}{4}+\frac{1}{2}a_i^{q,1}\right)+\phi_{t,i}^{q,3}\frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(-\frac{1}{4}-\frac{1}{2}a_i^{q,3}\right),\\ A_{t,i,k(i,q,2)} &+= \phi_{t,i}^{q,1}\frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(\frac{1}{4}-\frac{1}{2}a_i^{q,1}\right)+\phi_{t,i}^{q,3}\frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(\frac{1}{4}-\frac{1}{2}a_i^{q,3}\right),\\ A_{t,i,k(i,q,3)} &+= \phi_{t,i}^{q,1}\frac{m(e_{t,i}^{q,1})}{||\mathbf{m}_{t,i}^{q,1}||}\left(-\frac{1}{4}-\frac{1}{2}a_{t,i}^{q,1}\right)+\phi_{t,i}^{q,3}\frac{m(e_{t,i}^{q,3})}{||\mathbf{m}_{t,i}^{q,3}||}\left(\frac{3}{4}+\frac{1}{2}a_{t,i}^{q,3}\right). \end{aligned} \tag{7.16}$$

For the special case of the boundary finite volumes (5.6) we have the coefficients

$$A_{t,i,k(i,1,1)} = \left(\frac{\varphi_{t,i}^{1,1}+\varphi_{t,i}}{2}-\frac{\varphi_{t,i}^{1,1}+2\varphi_{t,i}+\varphi_{t,i}^{Q_i,3}}{4}\right)\frac{1}{||\mathbf{x}_{t,i}^{1,1}-\mathbf{x}_{t,i}||},$$

$$A_{t,i,k(i,Q_i,3)} = \left(\frac{\varphi_{t,i}^{Q_i,3} + \varphi_{t,i}}{2} - \frac{\varphi_{t,i}^{1,1} + 2\varphi_{t,i} + \varphi_{t,i}^{Q_i,3}}{4} \right) \frac{1}{||\mathbf{x}_{t,i}^{Q_i,3} - \mathbf{x}_{t,i}||}, \tag{7.17}$$

$$A_{t,i,i} = -A_{t,i,k(i,1,1)} - A_{t,i,k(i,Q_i,3)}.$$

8. Numerical experiments. In this section we present three numerical experiments. In first two experiments we present mean curvature flow (2.2) of an open surface with redistribution of points by the Gaussian curvature. In the last experiment we present an evolution of a closed surface by (2.3). A value of interest is the difference between the area density g and the desired area density c. This norm is numerically computed as

$$error_t = \sqrt{\frac{\sum_{i=1}^{N}(g_{t,i} - c_{t,i})^2}{\sum_{i=1}^{N}(g_{t,i})^2}}. \tag{8.1}$$

For all the experiments we used the time step $\tau = 0.1$ and the following parameters, $\omega = 1$, $p = 10$, $\widetilde{G} = 2.4$.

The first experiment has the initial condition in the shape of a cylinder with radius 1 and height 1 with 1225 points; and 250 time steps were computed. The surface in time steps 0, 5, 10, 250 can be seen on figure 8.2. A decreasing $error_t$ for this evolution can be seen on figure 8.1, top. In time 0, there is constant g on the surface and also a constant Gaussian curvature, hence $error_0 = 0$. After some time, points with a higher Gaussian curvature occur near the boundary. The redistribution responds to this and decreases the area of the corresponding quads. After that, the highest Gaussian curvature points move to the center of the cylinder. Then the surface acquires a steady state and the local area density does not change in time although it is not constant on the surface.

The second experiment has an initial condition in the shape of a hyperbolic paraboloid $z = x^2 - y^2$ on a domain $(x, y) \in (-1, 1) \times (-1, 1)$ with 400 points and 250 time steps were computed. The surface in time steps 0, 20, 40, 60 can be seen on figure 8.3. A decreasing $error_t$ for this evolution can be seen on figure 8.1, left. In the beginning the points with a high Gaussian curvature are in the middle of the surface. Thus the quads start to accumulate in this area. After some time, the mean curvature evolution causes a decrease of the Gaussian curvature in this area. This results in an enlarging of quads.

Special case of an evolving surface is presented in the last experiment. The surface is closed, therefore there are no boundary conditions (2.4), (3.6). The initial condition is a dumbbell like surface with 2168 points and 500 time steps were computed. The surface in time steps 0, 20, 60, 300 can be seen on figure 8.4. A decreasing $error_t$ for this evolution can be seen on figure 8.1, right. In the beginning there are points with a higher curvature in the corners and on the edges of the surface. Then the surface starts to smooth out. In the steady state there is a constant Gaussian curvature on the surface, hence constant c and g.

9. Conclusion. We presented a method for a redistribution of points by Gaussian curvature. We have shown 3 experiments presenting the performance of this method. We checked that the local area density converges to the prescribed local are density resulting in refinement of the surface approximation in areas of high Gaussian curvature. The method can be generalized to triangular meshes and a mean curvature dependent redistribution, which can be an objective of our further research.

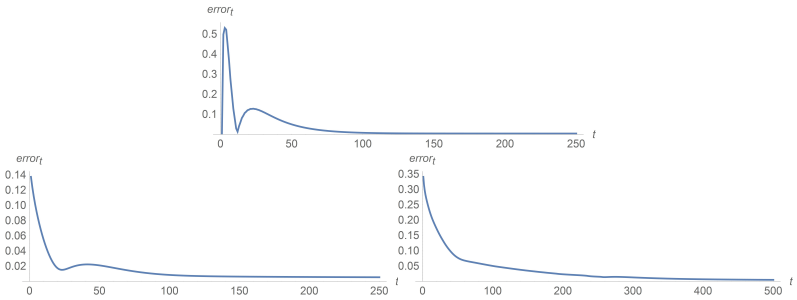


FIG. 8.1. *A graph of error$_t$ for the experiments. Top: the evolving cylinder. Bottom left: the evolving hyperbolic paraboloid. Bottom right: the evolving dumbbell like surface.*

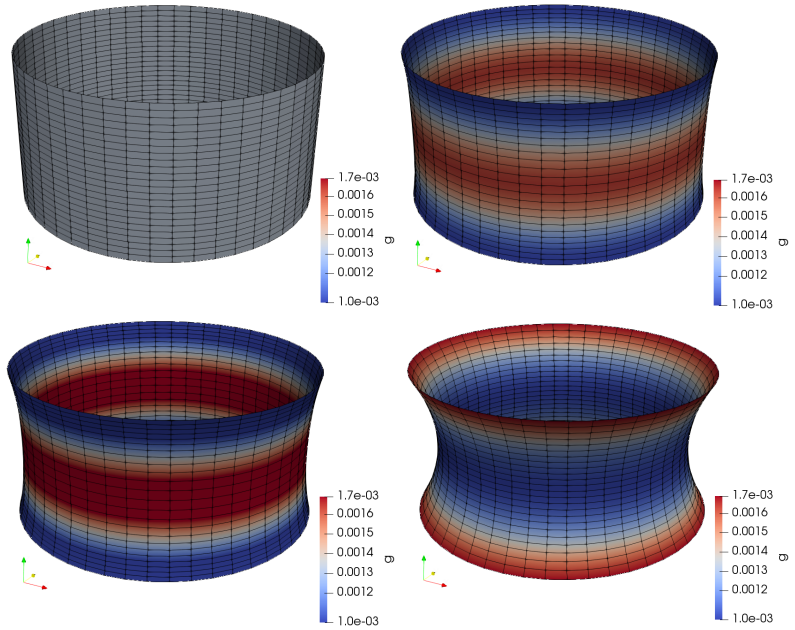


FIG. 8.2. *An evolving surface at time steps 0, 5, 10, 250.*

REFERENCES

[1] L. A. FREITAG, *On combining Laplacian and optimization-based mesh smoothing techniques*, American Society of Mechanical Engineers, Applied Mechanics Division, AMD, (1999)

[2] M. HÚSKA, M. MEDĽA, K. MIKULA, S. MORIGI, *Surface quadrangulation*, in preparation

[3] D. LIU, G. XU, *Angle Deficit Approximation of Gaussian Curvature and Its Convergence over Quadrilateral meshes*, In Computer-Aided Design, Volume 39, Issue 6, 2007, Pages 506-517, ISSN 0010-4485, https://doi.org/10.1016/j.cad.2007.01.007.

[4] K. MIKULA, M. REMEŠÍKOVÁ, P. SARKOCI, D. ŠEVČOVIČ, *Manifold evolution with tangential redistribution of points*, SIAM J. Scientific Computing, Vol. 36, No.4 (2014), pp. A1384-A1414

[5] D. ŠEVČOVIČ, S. YAZAKI, *Evolution of plane curves with a curvature adjusted tangential velocity*, Japan J. Indust. Appl. Math., 28(3) (2011), 413-442

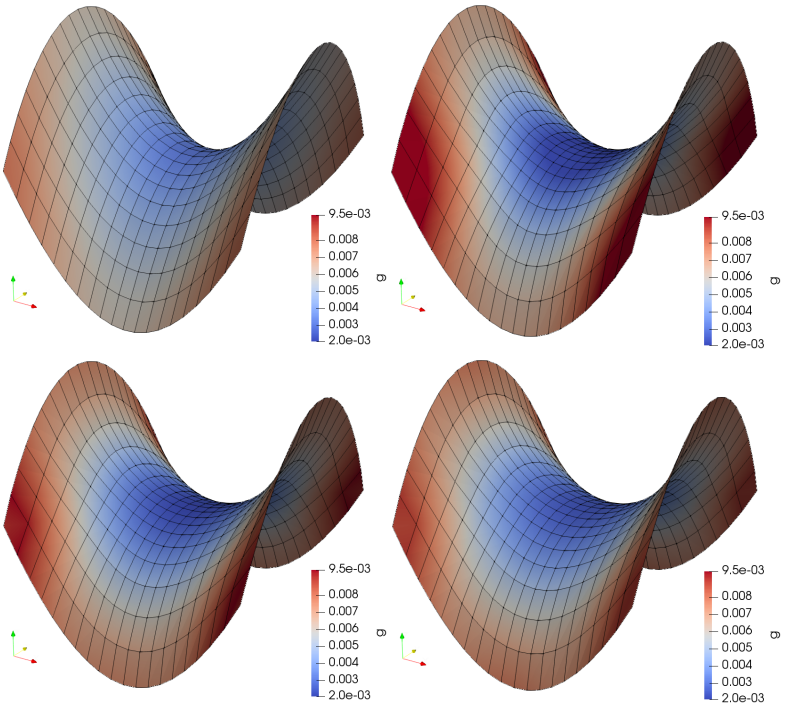


FIG. 8.3. *An evolving paraboloid at time steps 0, 20, 40, 60.*

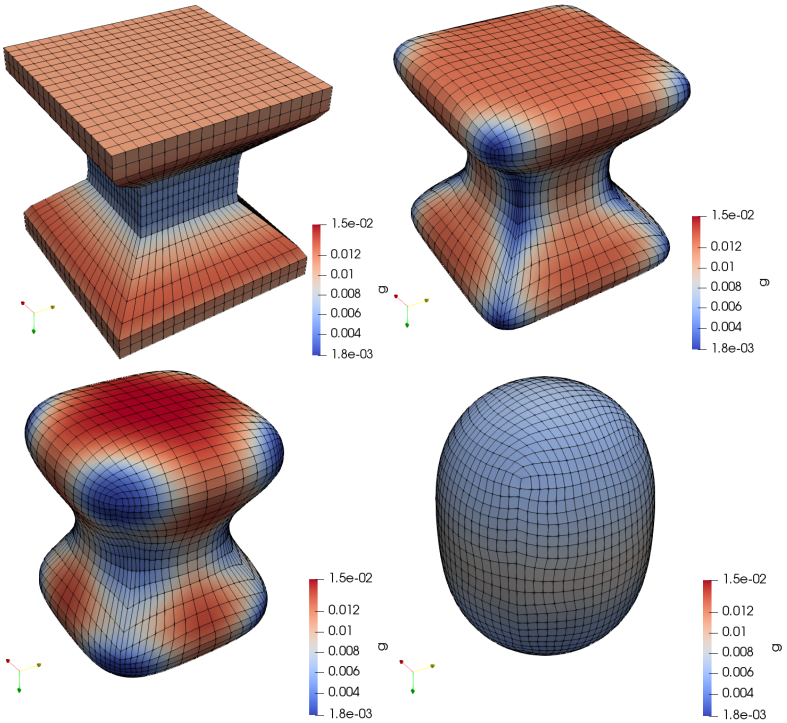


FIG. 8.4. *An evolving dumbbell like surface at time steps 0, 20, 60, 300.*

Proceedings of EQUADIFF 2017
pp. 265–274

COMPUTATIONAL DESIGN OPTIMIZATION OF LOW-ENERGY BUILDINGS *

JIŘÍ VALA †

Abstract. European directives and related national technical standards force the substantial reduction of energy consumption of all types of buildings. This can be done thanks to the massive insulation and the improvement of quality of building enclosures, using the simple evaluation assuming the one-dimensional stationary heat conduction. However, recent applications of advanced materials, structures and technologies force the proper physical, mathematical and computational analysis coming from the thermodynamic principles.

This paper shows the non-expensive evaluation of energy consumption of buildings with controlled indoor temperature, decomposing a building, considered as a thermal system, into particular subsystems and elements, coupled by interface thermal fluxes. We come to a rather large parabolic system of partial differential equations, containing the nonlinearities i) from the surface Stefan - Boltzmann radiation and ii) from the heating control; this can be handled using some properties of semilinear systems. The Fourier multiplicative decomposition together with the finite element technique enables us to derive a sparse system of ordinary differential equations, appropriate for the input of climatic data (temperature, beam and diffuse solar radiation). For the approximate solutions the spectral analysis is helpful; all nonlinearities can be overcome thanks to quasi-Newton iterations.

All above sketched simulations have been implemented in MATLAB. An example shows the validation of this approach, utilizing the time series of measured energy consumption from the real family house in Ostrov u Macochy (30 km northern from Brno). Additional procedures for the support of design of low-energy buildings come namely from the Nelder - Mead optimization algorithm.

Key words. Low-energy buildings, heat transfer, computational modelling, optimization techniques, MATLAB software tools.

AMS subject classifications. 35K05, 35K20, 65K10, 65M60, 65M70, 80A20.

1. Introduction. Knowledge of the position of Sun on the sky, used for natural winter heating and summer shading, dates back to the antique architecture and to the manuscripts by Aischylos and Socrates. However, the modern history of solar, low-energy and similar houses starts from the global economical crisis in the 30ties of 20th century, with the MIT "solar houses" (Massachusetts Institute of Technology, USA), coupling the new trends in architecture and civil engineering with the technological progress oriented to the reduction of energy requirements of buildings, namely of the cost of artificial heating. The actual European concept of passive house, forced by the directive [29] and national technical standards, is connected with the project CEPHEUS (Cost Efficient Passive House as European Standard, 1998–2001), whose ideas are explained in [8] in all details. All energy gains rely on the massive insulation of the building enclosure, together with available technological equipments (heat pumps, air recuperation, etc.) and certain exploitation of solar benefits; this is reflected by the rather simple software tool [9].

The approach of [8] does not handle the thermal accumulation and available climatic data properly, namely in the case of buildings with carefully controlled interior

*This work was supported by the project LO1408 AdMaS UP (Advanced Materials, Structures and Technologies), Ministry of Education, Youth and Sports of the Czech Republic, National Sustainability Programme I).

†Institute of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Brno University of Technology, 602 00 Brno, Veveří 331/95, Czech Republic.

temperature in their particular zones and rooms, as in the freezing and cooling plants where the substantial effect of decrease of energy consumption thanks to their optimal design can be expected. Moreover, the inhabitants of family houses or block of flats frequently prefer quite other criteria of well-being than the minimization of energy cost, as reviewed in [4], to suppress (often intuitively) the "sick building syndrom", occurring just in advanced structures minimizing the heat loss without proper ventilation. Also some new experimental research outputs like [25] do not coincide with traditional simplified calculation results. Software simulation packages for building energy performance developed in the last 2 decades, introduced in [5], involve much more physical processes than [9]; however, their complicated "black box" structure with extensive direct computations is not very friendly to the design optimization aims of architects and civil engineers.

In this paper we shall introduce a computational model of a building as a thermal system, whose basic ideas come from [21] and [23]. The decomposition of a building to building parts, as walls, roof, floor, ceilings, etc., as subsystems, with their own interior structure, containing particular constructive, insulation and other layers, as included subsystems, up to particular elements, incorporating selected physical processes with necessary geometric and material characteristics, enables us to obtain a compromise between model complexity and practically reliable, robust and inexpensive computations, supporting the above mentioned optimization of various types. The modular structure of the corresponding software in MATLAB respects such system approach in our practical implementation. Unlike [12], referring to [15] and [24], based on the analogy with the analysis of LC-electric circuits, coupling the finite difference approach with the Euler or similar time integration scheme, we shall work with the finite element technique, the Fourier multiplicative decomposition and the spectral properties of solutions, following some results of [13] (for direct computations) and [14] (for optimization algorithms).

2. Physical and mathematical fundamentals. We shall demonstrate the approach sketched above on the rather simple case of non-stationary heat conduction in the isotropic materials (at least macroscopically, not homogeneous in general), driven by boundary heat transfer from external environment, as studied in [6], including such interface transfer between adjacent subsystems, up to the level of particular elements, occupying a domain Ω in the 3-dimensional Euclidean space R^3. To avoid technical difficulties, we assume certain regularity of Ω, sufficient for the validity of standard Sobolev embedding and trace theorems in the sense of [20], p. 15; for possible generalizations see [18], pp. 69, 160, 512. The development of similar considerations with slightly stronger results in the Euclidean spaces of lower dimensions R^1 and R^2 are left to the patient reader. The following notations hold literally for constructive, insulation, etc. elements of buildings, whereas their modification for empty rooms (representing a majority of volume of a building) needs to set zero values of thermal conductivity; potential generalizations will be mentioned later.

2.1. A simple model problem. Let R^3 be supplied by some Cartesian coordinate system $x = (x_1, x_2, x_3)$. Let the boundary $\partial\Omega$ of Ω in R^3 having a local vector of outward unit normal $\nu(x) = (\nu_1(x), \nu_2(x), \nu_3(x))$ almost everywhere. The usual notation for the Hamilton operators $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$ will be used. Moreover, let us consider a time interval $J = [0, T]$ with some real positive T (the limit passage $T \to \infty$ is not prohibited); the upper dot symbol is reserved for partial derivatives with respect to the time $t \in J$. The standard notation of Lebesgue, Sobolev, Bochner, etc. (abstract) function spaces will be utilized in the following considerations, following

[20], pp. 10, 22.

Let us introduce 2 basic material characteristics on Ω: the thermal conductivity $\lambda(x)$ (for the insulation ability) and the thermal capacity $\kappa(x)$ (for the accumulation ability, related to unit volume here). It is natural to suppose that λ and κ are functions from $L^\infty(\Omega)$ (for homogeneous materials only constants), a. e. with values greater than certain positive constant. The weak formulation of a heat transfer equation, using the temperature $\vartheta(x,t)$ on $\Omega \times J$ as the reference variable and working with some volume sources $f(x,t,\vartheta(x,t))$ on $\Omega \times J$ and surface sources $g(x,t,\vartheta(x,t))$ on $\partial\Omega \times J$, reads

$$(v, \kappa\dot{\vartheta}) + (\nabla v, \lambda\nabla\vartheta) = (v,f) + \langle v,g\rangle \quad \text{on } J \tag{2.1}$$

where $(.,.)$ denotes scalar products (for any fixed t) both in $L^2(\Omega)$ and in $L^2(\Omega)^3$, $\langle .,.\rangle$ those in $L^2(\partial\Omega)$, v is an arbitrary test function from V and ϑ must be contained in $L^2(J,V)$, with certain $\dot{\vartheta}$ in $L^2(J,H)$; here we set $H = L^2(\Omega)$, V will be specified later due to the particular choice of f and g, crucial for the implementation of the model. The Cauchy initial condition

$$\vartheta(.,0) = \vartheta_0 \tag{2.2}$$

with a priori known $\vartheta_0 \in V$ then completes the problem definition.

Let us notice that, regardless of (2.2), the formal application of the Green-Ostrogradskij theorem (at least in the sense of distributions – cf. [28], p. 244), using the central dots for the scalar products in R^3, can convert (2.1) to its strong form

$$\dot{\varepsilon} + \nabla \cdot q = f\,, \quad \varepsilon = \kappa\vartheta\,, \quad q = -\lambda\nabla\vartheta \quad \text{on } \Omega \times J\,, \quad q \cdot \nu = g \quad \text{on } \partial\Omega \times J\,, \tag{2.3}$$

compatible with [1], pp. 5, 14: the 1st equation of (2.3) represents the principle of conservation of energy ε related to unit volume, due to some thermal flux q, the 2nd equation quantifies the thermal energy, the 3rd equation is the well-known empirical Fourier constitutive relation between thermal fluxes and temperature gradients, finally the 4th equation represents a general boundary (or interface) condition.

2.2. Fourier multiplicative decomposition. Following the approach of [3], p. 346, let us consider the temperature ϑ on $\Omega \times J$ in the form of multiplicative decomposition

$$\vartheta(x,t) = \phi_i(x)\theta_i(t) \tag{2.4}$$

for any $x \in \Omega$ and $t \in J$ where i denotes the Einstein summation index from $\{1,\dots,n\}$ for certain integer n, with the aim of the limit passage $n \to \infty$, and $\phi_1(x),\dots,\phi_n(x)$ represents a basis of some finite-dimensional approximation V_n of V. For simplicity let us assume $V_n \subset V$; possible "variational crimes" violating such assumptions can be handled by [27]. Consequently in (2.1) we are allowed to consider $v = \phi_j$ for arbitrary $j \in \{1,\dots,n\}$, i. e.

$$(\phi_j, \kappa\phi_i)\dot{\theta}_i + (\nabla\phi_j, \lambda\nabla\phi_i)\theta_i = (\phi_j, f) + \langle\phi_j, g\rangle \quad \text{on } J\,. \tag{2.5}$$

The least squares minimization of $(\theta_k\phi_k - \vartheta_0, \kappa(\theta_i\phi_i - \vartheta_0))$, referring to (2.2), involving also the Einstein summation over $k \in \{1,\dots,n\}$, yields

$$(\phi_j, \kappa\phi_i)\theta_i(0) = (\phi_j, \kappa\vartheta_0)\,. \tag{2.6}$$

The matrix form of (2.5), useful for an efficient software (e. g. MATLAB-based) implementation, is

$$M\dot{\theta} + K\theta = F \quad \text{on } J \tag{2.7}$$

where M and K are positive definite symmetric square matrices from $R^{n\times n}$, $\theta(t) = (\theta_1(t),\dots,\theta_n(t))^{\mathrm{T}}$ is a column vector from R^n for any fixed t, as well as $F(t)$, covering the whole right hand side of (2.5); however, its evaluation is not easy in general. (2.7) forms a system of ordinary differential equations, which should by analysed analytically. Due to practical reasons for m equidistant time steps (where environmental data

needed for the composition of F are measured usually) are introduced: $\theta^r = \theta(rh)$ with $r \in \{1, \dots, m\}$, m being en integer number, $h = T/m$; this is compatible with $\theta^0 = \theta(0)$ by (2.6). Also (2.6) can be rewritten as

$$K\theta^0 = \theta_\star \,, \tag{2.8}$$

with $\theta_\star$ (a column vector from R^n again) generated by the right hand side of (2.6).

Finite element approximations by [28], pp. 247, 427, work usually with some continuous functions ϕ_i ($i \in \{1, \dots, n\}$) with values from $[-1, 1]$ and small compact support, not orthogonal exactly, unlike classical Fourier analysis. The Lebesgue measure of supports of such functions on Ω is not greater than $c^{-1}n^{-3}$ and their Hausdorff measure on $\partial\Omega$ is not greater than $c^{-1}n^{-2}$ where c is a positive (sufficiently small) constant independent of n. Moreover, we shall consider the integer upper bound N for the number of functions ϕ_i supported on the same part of Ω or $\partial\Omega$ of non-zero relevant measure. It is reasonable to suppose that this choice guarantees also $cn^{-3}|a|^2 \le a \cdot Ma \le c^{-1}n^{-3}|a|^2$, $cn^{-1}|a|^2 \le a \cdot Ka \le c^{-1}n^{-1}|a|^2$, the last couple of inequalities also for K constructed with $\lambda = 1$ everywhere instead of the correct λ formally, for all $a \in R^n$ (considered as column vectors) where $|.|$ denotes the norm in R^n (not only in R^1); the central dots here are used for the scalar products also in R^n (similarly to those in R^3 by (2.3)).

2.3. Existence and uniqueness of solution. Let us start with the purely linear (not very realistic) case $f \in L^2(J, H)$, $g \in L^2(J, X)$ where $X = L^4(\partial\Omega)$, with f and g independent of ϑ; in this case we can take $V = W^{1,2}(\Omega)$. For any fixed $t \in J$ we can rewrite (2.7), supplied by θ^0 from (2.8), as in two different forms as

$$\int_0^t \theta'(\tau) \cdot M\theta'(\tau)\,\mathrm{d}\tau + \frac{1}{2}\theta(t) \cdot K\theta(t) = \frac{1}{2}\theta^0 \cdot K\theta^0 + \int_0^t \theta'(\tau) \cdot F(\tau)\,\mathrm{d}\tau \,, \tag{2.9}$$

with the prime symbol replacing the dot one for all time derivatives with respect to τ instead of t. Utilizing the above introduced estimates, (2.9) yields

$$\frac{c}{2n^3}\int_0^t |\theta'(\tau)|^2\,\mathrm{d}\tau + \frac{c}{2n}|\theta(t)|^2 \le \frac{1}{2cn}|\theta^0|^2 + \frac{c}{4n^3}\int_0^t |\theta'(\tau)|^2\,\mathrm{d}\tau + \frac{n^3}{c}\int_0^t |F(\tau)|^2\,\mathrm{d}\tau \,. \tag{2.10}$$

For the last additive term of (2.10) we have

$$\begin{aligned} \int_0^t |F(\tau)|^2\,\mathrm{d}\tau \ &\le \int_0^t\!\!\int_\Omega \phi_i(x) f(x,\tau) \cdot \phi_i(x) f(x,\tau)\,dx\,\mathrm{d}\tau \\ &+ \int_0^t\!\!\int_{\partial\Omega} \phi_i(x) g(x,\tau) \cdot \phi_i(x) g(x,\tau)\,ds(x)\,\mathrm{d}\tau \le \mu_f \|f\|^2_{L^2(J,H)} + \mu_g \|g\|^2_{L^2(J,X)} \,, \end{aligned} \tag{2.11}$$

utilizing the measures

$$\mu_f = N\left(\left(\frac{1}{cn^3}\right)^{1-1/2}\right)^2 = \frac{N}{cn^3} \,, \qquad \mu_g = N\left(\left(\frac{1}{cn^2}\right)^{1-1/4}\right)^2 = \frac{N}{c^{3/2}n^3} \,. \tag{2.12}$$

Combining (2.10), (2.11) and (2.12), we obtain the brief result

$$\int_0^t |\theta'(\tau)|^2\,\mathrm{d}\tau \le \mathcal{C}n^3 \,, \qquad |\theta(t)|^2 \le \mathcal{C}n \tag{2.13}$$

for some positive constant $\mathcal{C}$ independent of n. Thus, inserting (2.13) into (2.4), we get

$$\|\vartheta(.,t)\|^2_H \le \frac{N\mathcal{C}n^3}{cn^3} = \frac{N\mathcal{C}}{c} \,, \quad \int_0^t \|\nabla\vartheta(.,\tau)\|^2_{H^3}\,\mathrm{d}\tau \le \frac{N\mathcal{C}n}{cn} = \frac{N\mathcal{C}}{c} \,. \tag{2.14}$$

Let us notice that ϑ in (2.14) involves the dependence on n, inherited from (2.4), generating certain sequences $\vartheta^{(n)}$. Due to the reflexivity of both V and $L^2(J, H)$, the Eberlein - Shmul'yan theorem (as introduced in [7], p. 66) yields, up to subsequences,

the existence of a weak limit $\vartheta(.,t)$ of $\vartheta^{(n)}(t)$ in V for each $t \in J$, which is strong in H (because of the existence of compact embedding of H into V); simultaneously $\dot{\vartheta}$ is a weak limit of $\dot{\vartheta}^{(n)}$ in $L^2(J,H)$. Such ϑ can be then identified with the solution of (2.1) with (2.2).

Let $\bar{\vartheta}$ be the difference between 2 solutions of (2.1) with (2.2) and t an arbitrary time from J. Then the choice $v = \bar{\vartheta}(.,t)$ gives

$$\frac{1}{2}(\bar{\vartheta}(.,t), \kappa\bar{\vartheta}(.,t)) + \int_0^t (\nabla\bar{\vartheta}(.,\tau), \lambda\nabla\bar{\vartheta}(.,\tau))\,\mathrm{d}\tau = 0\,. \tag{2.15}$$

Thanks to the positive-valued κ and λ, from (2.15) we receive $\bar{\vartheta} = 0$ on J, which implies the uniqueness of ϑ satisfying (2.1) with (2.2).

Similar arguments can be repeated also for the limit case $\lambda \to 0$: this is important for the simplification of temperature development in empty rooms where no more detailed information is available, unlike constructive and insulation building parts. Consequently $\vartheta(.,t)$ is constant for any fixed $t \in J$.

2.4. Realistic classes of thermal sources. More realistic cases for the choice of f and g, needed in computational tools for thermal analysis of buildings, are:

i) $g = \beta(\vartheta_* - \vartheta)$ for the thermal transfer from external environment with some prescribed external temperature $\vartheta_* \in L^2(J,X)$ and some known a. e. positive transfer factor $\beta \in L^\infty(\partial\Omega)$, taking the rigid body – air convection into account, later used also for the thermal transfer between two neighbour domain through their interface analogously,

ii) $f = \alpha(\vartheta_* - \vartheta)$ for the obligatory ventilation by technical standards, similar to i), but applied to the above mentioned case of constant $\vartheta(.,t)$ for a fixed $t \in J$, with some known a. e. positive transfer factor $\alpha \in L^\infty(\Omega)$: such simplified "volumetric convection" is needed to include the heat exchange caused by various installed equipments (without deeper analysis of their performance) between rooms and external environment,

iii) g coming from the beam and diffuse components of solar radiation, occurring just on the building envelope (not on internal interfaces) evaluable from the climatic records of the so-called reference climatic year, due to the day and year quasi-cycles, the mutual position of Sun and Earth, the geographical location of our building object and on the slope and orientation of the relevant building surface, under certain astronomical simplifications presented (including numerous further references) in [13], with the resulting setting of $g \in L^2(J,X)$,

iv) $g = \sigma(\vartheta_*^4 - \vartheta^4)$ for the thermal radiation on the building envelope due to the physical Stefan - Boltzmann law and some known a. e. positive factor $\sigma \in L^\infty(\partial\Omega)$, interpretable as the Stefan - Boltzmann constant (exact for the ideal black body), modified by the empirical surface emissivity, which cannot be incorporated to i) properly because of the presence of ϑ^4,

v) f coming from the artificial heating (or air conditioning, too) in the case similar to ii), but with the requirement of the type $\vartheta \geq \vartheta_\diamond$ for some prescribed indoor temperature $\vartheta_\diamond \in V$ (depending on the room categories by technical standards) at least in the least square sense, due to the real maximal power of heating equipments and to their expected (summer, winter, etc.) different regimes – for more details see [13] again.

All such volume sources f and surface sources g are able to generate additive contributions to the right hand side of (2.1). However, it is useful to incorporate some their parts to the left hand side of (2.1).

Whereas i), ii) and iii) can be handled inside the theory of linear parabolic equations, iv) forces the redefinition of V and the inequalities in v) will be overcome using some facts from the control theory. In i) and ii) g and f force 2 new additive terms $\langle v, \beta\vartheta\rangle$ and $(v, \alpha\vartheta)$ on the left hand side of (2.1); $\beta\vartheta_*$ and $\alpha\vartheta_*$ can be then hidden in g and f on the right hand side as above. Consequently K in (2.7) is replaced by $K + K_f + K_g$ formally with some sparse positive symmetrical matrices K_f from ii) and K_g from i), even with certain regularizing effect. Due to the limited extent of this paper, the detailed analysis can be performed by the patient reader without substantial difficulties. Then iii) brings no new left hand side modification of (2.1) unlike i) and ii); its significance lies in practical long evaluations, accounting for all available environmental data: the temperature θ_*, needed in i) and ii), too, and both relevant components of solar radiation. The repeated application of such data leads to certain quasiperiodicity of the solution of (2.1), suppressing the effect of (2.2) for increasing time.

For iv) the rough heuristic approximation (acceptable for the usual range of temperature) $\vartheta^4 - \vartheta_*^4 = (\vartheta^2 + \vartheta_*^2)(\vartheta + \vartheta_*)(\vartheta - \vartheta_*) \approx 4\vartheta_*^3(\vartheta - \vartheta_*)$ highlights certain quasilinearity of the problem. Using the notation $\langle ., . \rangle$ also for the duality between $L^5(\partial\Omega)$ and $L^{5/4}(\partial\Omega)$, we are able to introduce $V = \{v \in W^{1,2}(\Omega) : v \in L^5(\partial\Omega)\}$ (in the sense of traces), supplied with the norm $\|v\|_{W^{1,2}(\Omega)} + \|v\|_{L^5(\partial\Omega)}$ by [20], pp. 64, 253 (which generates a reflexive Banach space again), and, motivated by i), to add $\langle v, \sigma|\vartheta|^3\vartheta\rangle$ to the left hand side and $\langle v, \sigma|\vartheta_*|^3\vartheta_*\rangle$ to the right hand side of (2.1). Consequently, in addition to the 2nd left-side additive term of (2.9), we have the contribution of the type $\frac{1}{5}|\theta(t)|^{3/2}\theta(t) \cdot S|\theta(t)|^{3/2}\theta(t)$, containing certain sparse positive symmetrical matrix S; the enrichment of the right side of (2.9) is evident. The existence and uniqueness of solution of (2.1) with (2.2) can be then verified as above.

To handle v), the best choice is seemingly to convert (2.1) to the form of a variational inequality. However, the above sketched technical specifications bring serious complications to the design of an efficient computational algorithm, thus another approach, avoiding general optimization strategies, based on the careful control of a heating equipment, is considered: $\vartheta \geq \vartheta_\diamond$ is satisfied in every time step just during the correct (a priori prescribed) heating season, thanks to the controlled heating source f in a corresponding room; the maximum value for the heating power is still considered if this is insufficient.

2.5. Building as a thermal system. All generalizations i) – v) are useful for the development of a model of thermal behaviour of buildings. Understanding Ω as a building element at the lowest (most detailed) level, we are able to compose substructrues at the finite number of levels, using the transfer conditions by i) and ii), up to the whole structure. If ϑ_* and consequently θ_* refer to the external environment, this contributes both to the matrix K in (2.7) (using the matrices K_f and K_g from the preceding discussion) and to the right hand side F. Usually such conditions are applied only in the case when some interface to the room is present, otherwise it is acceptable to take $\alpha \to \infty$, i. e. to force the continuity of temperature on the interface in the normal direction. Clearly iii) and iv) occur only on the external interfaces (building claddings). The existence and uniqueness considerations, handling all possible interface types, can be repeated without substantial difficulties.

Such computational model is open to various generalizations. In particular, let us remind that physical and mathematical homogenization approaches, trying to involve (even incomplete) information on material microstructure, lead to effective anisotropic material characteristics even in the case of composites with isotropic components,

due to their location, orientation, etc. (as typically in fibre concrete). Removing the isotropy assumption, we come to the direction-dependent material characteristics λ and κ on Ω and α, β and σ on $\partial\Omega$, generating certain square matrices from $L^\infty(\Omega)^{3\times 3}$ or $L^\infty(\partial\Omega)^{3\times 3}$ (using the notation from an introductory simple problem for brevity again). At least for the case that all such matrices are a. e. symmetrical and positive definite, the above sketched existence and uniqueness considerations can be repeated with slight technical modifications.

Even more general case with the material characteristics $\lambda(.,\vartheta)$, $\kappa(.,\vartheta)$ on Ω and $\alpha(.,\vartheta,\vartheta_*)$, $\beta(.,\vartheta,\vartheta_*)$ $\sigma(.,\vartheta,\vartheta_*)$ on $\partial\Omega$, important in building practice, can be handled as a quasilinear problem, using selected results on pseudomonotone or weakly continuous mappings by [20], p. 321. However, some additional growth assumptions are needed and all proofs become much more complicated, thus they are not presentable in this short conference paper.

Deeper generalizations cover both the 1st thermodynamic principle of conservation of mass, (linear and angular) momentum and energy (not only of thermal energy as above) and the 2nd thermodynamic principle, handling the irreversibility of some thermal processes, as [22], pp. 145 (for closed systems) and 231 (for open systems). Unfortunately, there is a lot of open questions in the mathematical analysis of corresponding systems of equations of evolutions and related inequalities, as well as in the suggestion of computational algorithms constructing some sequences of reasonable approximate solutions; this is still true even in the particular case of Navier - Stokes equations (cf. the "mysteriously difficult problem" of [20], p. 257).

Fortunately, some simplified approaches for the analysis of parallel physical processes, as heat and moisture transfer in porous media, are available: instead of ϑ we have the couple of unknown variables (ϑ, u) where u evaluates certain moisture content (related to the mass or volume unit), considering the conservation of mass (moisture in pores) and (thermal) energy. The Fick constitutive relation between u and some moisture flux η can be written in the similar way as the Fourier one between ϑ and q in (2.3); however, in the complete system of 2 equations of evolution we need (and must be able to identify in practice) additional material characteristics to handle the Dufour effect (time redistribution of ϑ depends not only on q, but also on η) and the Soret effect (time redistribution of u depends not only on η, but also on q). The proper mathematical and numerical analysis is based on generalization of the results sketched above to the system of 2 equations; practical computations must take the slow moisture transfer in comparison with the thermal one into account.

3. Computational modelling and optimization. Computational tools, at least for direct calculations, including those minimizing the energy consumption, can be based on (2.7) with (2.8). Since some sources are frequently prescribed by their time derivatives in practice, namely those by ii) and v), it is useful to consider the right hand side of (2.7) as $F(t) = \Phi(t) + \dot{\Psi}(t)$ for any $t \in J$, namely for $t \in \{h, 2h, \dots, mh\}$ where $h = T/m$; the reliable construction of the limit passage $m \to \infty$ depends on the environmental data by iii). To derive the semi-analytic formulae for the evaluation of θ in time, the spectral decomposition $MV\Lambda = KV$ with the generalized real diagonal eigenvalue matrix Λ and the matrix of eigenvectors V is then helpful.

3.1. Direct calculations with heating control. For the brevity, let us consider $\theta^1, \dots \theta^m$ instead of $\theta(h), \dots \theta(mh)$ (a priori unknown temperatures) and also $\Phi^1, \dots, \Phi^m$ and $\Psi^1, \dots, \Psi^m$ (characterizing all prescribed thermal sources) in the similar sense. For the beginning, let us neglect all nonlinear thermal sources by iv) and v). Applying the classical integral calculus, namely the method of variations of constants,

for any time step index $s \in \{1, \dots, m\}$ we come to the direct evaluation formula

$$\theta^s - V\exp(-\Lambda h)V^T M\theta^{s-1} = V\Lambda^{-1}V^T\Phi^s - V\Lambda^{-1}\exp(-\Lambda h)V^T\Phi^{s-1} + V(I-\exp(-\Lambda h))\left(\Lambda^{-1}V^T\frac{\Psi^s-\Psi^{s-1}}{h} - \Lambda^{-2}V^T\frac{\Phi^s-\Phi^{s-1}}{h}\right), \tag{3.1}$$

exact for any $\Phi(t)$ and $\Psi(t)$ with $t \in J$ considered as a Lagrangian linear spline using the nodes $\{0, h, 2h, \dots, T\}$. This holds for an arbitrary positive h, unlike the Euler explicit or implicit, Crank - Nicholson, etc. discretization schemes.

To adopt (3.1) to handle iv), at least for sufficiently small h, we can add some $|\theta|^{3/2}S|\theta|^{3/2}$ to K, inserting some reasonable estimate of θ, and apply the quasi-Newton iterations inside each s-th time step; the exploitation of the inexact Newton method is expected to reduce the number of algebraic operations. The same is true for v) where, using the least squares approach, some $\mathcal{G}$ must be added to $\dot\Psi$, to minimize (if possible and required, due to technical specifications) $|\theta-\theta_\diamond|^2$; this can be modified by some prescribed weights for particular rooms if needed. Since $\mathcal{G}$ is just a vector of constants $\mathcal{G}^s \in R^n$ for $(s-1)h < t \le sh$, the total consumption of energy for heating can be evaluated easily as

$$Q = h\sum_{s=1}^{m}\mathcal{G}^s\,. \tag{3.2}$$

Fortunately, both corrections iv) and v) can be unified in one iteration procedure; its details (together with the instructive example), distinguishing between 4 typical heating regimes, are discussed in [12].

The validation of this approach here works with the real living house and atelier in Ostrov u Macochy (30 km northern from Brno), presented (as an example of low-energy house from ecological materials) in [11], p. 146. This small experimental house, designed by architect M. Hudec, built from wood and straw balls, contains 2 floors and 4 rooms, whose 26 mutual interfaces, including those to external environment, are assumed to consist of finite numbers of homogeneous isotropic layers. The design temperature for all rooms is $\vartheta_\diamond = 20^{\rm o}$ C; $\theta_\diamond$ can be then set analogously to θ_0 in (2.8). The annual climatic records for $h = 1$ hour from the international airport Brno - Tuřany need improvements using the incomplete data from the (colder and wetter) Moravian Karst. The original software code implementing (3.1) and its iterative generalizations has been written in MATLAB. Certain type of optimization is built even in the seemingly direct computational algorithm, thanks to the least squares technique in v). The 1st block of results in the following table documents the process of validation of the algorithm; the comparative variable is Q by (3.2) everywhere.

3.2. Selection of design parameters. The work of architects and civil engineers is far from the optimization of one physically transparent goal function under some simple set of additional conditions: it contains aesthetic, artistic, ecological and other criteria, whose deterministic quantification would be very complicated or quite impossible. The resulting project is typically a result of discussion based on comparison of a finite number of variants, supported by some auxiliary calculations.

As an example, we consider the principal motivation by the economy of heating here, e. g. we are seeking for a sub-optimal (sufficiently small) value of Q, corresponding to some of the prepared variants. The 2nd block in the table demonstrates the effect of the installation of particular heating devices on every floor, or even in every room, instead of one central device, as well as the effect of 2 types of possible replacement of materials in walls. The computation works just with $h = 1$ hour, assuming $\vartheta_0 = 20^{\rm o}$ C everywhere, repeating the same climatic data for all considered years; it finishes in the case of quasi-periodicity of results, here after 3 years in all cases.

TABLE 3.1
Consumption of energy for heating by various methods including design optimization.

Q [MWh]	evaluation method
1.881	new software, correction for building location
1.419	new software, original climatic data from Brno - Tuřany
1.897	software Energie 2009 (related to Czech technical specifications)
1.710	qualified estimate from time series of user payments for energy
1.915	heating on both floors: 2 devices, total power preserved
1.900	heating in all rooms: 4 devices, total power preserved
1.849	partial replacement of glass garden frontage by non-transparent one
3.039	replacement of straw balls in walls by clay blocks
1.841	Nelder - Mead optimization, 1 parameter: vertical rotation 20.81°
1.769	Nelder - Mead optimization, 2 parameters: vertical rotation 21.37°, glass transparency factor 0.1

3.3. Nelder - Mead simplex algorithm. In the case of proper mathematical optimization, no simple numerical evaluation of gradients like [2] is available, which justifies the choice of the Nelder - Mead downhill simplex method, coming from [19] originally. In the formulation of [26] this method works, in general, with the 5-step algorithm, involving (after sorting simplex vertices) 1) reflection, 2) expansion, 3) outer contraction, 4) inner contraction and 5) shrinkage. Theoretical convergence results for this method are not quite satisfactory: namely by [16], assuming Q (in our notation) as a strictly convex function of 1 or 2 parameters with bounded level sets, the convergence is guaranteed just for 1 parameter, whereas for 2 parameters only the simplex diameter tends to 0 (but need not converge to any minimizer); [17] presents the computer-supported 25-page convergence proof for 2 parameters by contradiction, but only for the restricted algorithm with missing step 2). However, some unpleasant cases of total divergence or numerical stagnation of the algorithm, even for more parameters, can be overcome using some ad hoc adaptive strategies, following [10].

The 3rd block in the table shows the application of this method, making use of the MATLAB function *fminsearch* from the *optimization* toolbox (in addition to the above sketched software code for direct calculations) for 1 and 2 parameters with respect to their lower and upper bounds, included via simple penalty functions: the 1st parameter is the hypothetical vertical rotation of the house, the 2nd one is certain glass transparency factor. More practical considerations and recommendations of this type, including graphs, figures and further references, have been recently published in [14].

4. Conclusion. The computer-supported design of high-performance buildings, accenting their thermal behaviour, motivated by the development of new structures, materials and technologies, as well as by the requirements of sustainable environmental solutions for buildings, contributing to the health and well-being of their inhabitants, reflected by [29], brings new challenges also for physicists, mathematicians, hardware and software developers and other experts. Existing modelling and simulation tools, even those declared as multi-physical, frequently predict other results then those observed in situ; to identify all substantial sources of such differences is not easy.

The system approach, presented in this paper, can be helpful to meet the requirements of reliable and robust optimization with the work style of architects and civil engineers, as well as with investors' money, time and patience. Nevertheless, the need of deeper interdisciplinary discussion is evident.

REFERENCES

[1] A. Bermúdez de Castro, *Continuum Thermodynamics*, Birkhäuser, Basel, 2005.

[2] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, New York, 2006.

[3] R. Brigola, *Fourier-Analysis und Distributionen* (in German), Co-Verlag, Berlin, 2012.

[4] S. Carluci and L. Pagliano, A review of indices for the long-term evaluation of the general thermal comfort conditions in buildings, *Energy and Buildings*, 53 (2012), pp. 194–205.

[5] D. B. Crawley, Contrasting the capabilities of building energy performance simulation programs. *Building and Environment*, 43 (2008), pp. 661–673.

[6] M. G. Davies, *Building Heat Transfer*, J. Wiley & Sons, Chichester, 2004.

[7] P. Drábek and J. Milota, *Lectures on Nolinear Analysis*, University of West Bohemia, Pilsen, 2004.

[8] W. Feist, *Gestaltungsgrundlagen Passivhäuser* (in German), Das Beispiel, Darmstadt, 1999.

[9] W. Feist, R. Pfluger, B. Kaufmann, J. Schnieders and O. Kah, *Passive House Planning Package*, Passive House Institute, Darmstadt, 2004.

[10] F. Gao and L. Han, Implementing the Nelder-Mead simplex algorithm with adaptive parameters, *Computer Optimizations and Applications*, 99 (2010), pp. 111–222.

[11] M. Hudec, B. Johanisováand T. Mansbart, *Pasivní domy z přírodních materiálů* (in Czech), Grada, Prague, 2012.

[12] P. Jarošová, Computational approaches to the design of low-energy buildings, *Programs and Algorithms of Numerical Mathematics* in Dolní Maxov (Czech Republic), Proceedings, 2014, pp. 92–99, Institute of Mathematics AS CR, Prague, 2015.

[13] P. Jarošová and J. Vala, On a computational model of building thermal dynamic response, *Thermophysics* in Terchová (Slovak Republic), Proceedings, 2016, pp. 40011/1–6, American Institute of Physics, Melville (USA), 2016.

[14] P. Jarošová and J. Vala, Optimization approaches in the thermal system analysis of buildings, *ICNAAM (International Conference on Numerical Analysis and Applied Mathematics)* in Thessaloniki, Proceedings, 2017, 4 pp., American Institute of Physics, Melville (USA), 2018, accepted for publication.

[15] J. H. Kämpf and D. A. Robinson, A simplified thermal model to support analysis of urban resource flows, *Energy and Buildings*, 39 (2007), pp. 445–453.

[16] J. C. Lagarias, J. A. Reeds, M. H. Wrigth and P. E. Wrigth, Convergence properties of the Nelder-Mead simplex method in low dimensions, *SIAM Journal of Optimization*, 9 (1998), pp. 112–147.

[17] J. C. Lagarias, B. Poonen and M. H. Wrigth, Convergence of the restricted Nelder-Mead algorithm in two dimensions, SIAM Journal on Optimization, 22 (2012), pp. 501–532.

[18] V. G. Maz'ya, *Sobolev Spaces with Applications to Elliptic Partial Differential Equations*, Springer, Berlin, 2011.

[19] J. A. Nelder and R. Mead, Simplex method for function minimization, *Computer Journal*, 7 (1965), pp. 308–313.

[20] T. Roubíček, *Nonlinear Partial Differential Equations with Applications*, Birkhäuser, Basel, 2005.

[21] J. Řehánek, *Tepelná akumulace budov* (in Czech), ČKAIT, Prague, 2002.

[22] M. Shukuya, *Exergy – Theory and Applications in the Built Environment*, Springer, London, 2013.

[23] S. Šťastník and J. Vala, On the thermal stability in dwelling structures, *Building Research Journal*, 52 (2004), pp. 31–55.

[24] J. C. Underwood, An improved lumped parameter method for building thermal modelling, *Energy and Buildings*, 79 (2014), pp. 191–201.

[25] H. Viggers, M. Keall, K. Wickens and P. Howden-Chapman, Increased house size can cancel out the effect of improved insulation on overall heating energy requirements, *Energy Policy*, 107 (2017), pp. 248–257.

[26] M. H. Wright, Nelder, Mead, and the other simplex method, *Documenta Mathematica*, extra volume *Optimization Stories* (2012), pp. 271–276.

[27] A. Ženíšek, Finite element variational crimes in parabolic-elliptic problems, *Numerische Mathematik*, 55 (1989), pp. 343–376.

[28] A. Ženíšek, *Sobolev Spaces and Their Applications in the Finite Element Method*. Brno University of Technology, 2005.

[29] *Directive 2010/31/EU of the European Parliament and of the Council on the Energy Performance of Buildings*, Official Journal of the European Union, L 153/13 (2010).

Proceedings of EQUADIFF 2017
pp. 275–282

A GENERALIZATION OF THE KELLER–SEGEL SYSTEM TO HIGHER DIMENSIONS FROM A STRUCTURAL VIEWPOINT*

KENTAROU FUJIE† AND TAKASI SENBA‡

Abstract. We consider initial boundary problems of a two-chemical substances chemotaxis system. In the four-dimensional setting, it was shown that solutions exist globally in time and remain bounded if the total mass is less than $(8\pi)^2$, whereas the solution emanating from some initial data of large magnitude may blows up.

This result can be regarded as a generalization of the well-known 8π problem in the Keller–Segel system to higher dimensions. We will compare mathematical structures of the Keller–Segel system and our system and discuss the difference.

Key words. chemotaxis; global existence; Lyapunov functional; Adams' inequality

AMS subject classifications. 35B45, 35K45, 35Q92, 92C17

1. Problem.

Consider the following fully parabolic system:

$$\begin{cases} u_t = \Delta u - \chi \nabla \cdot (u \nabla v) & \text{in } \Omega \times (0,\infty), \\ \tau_1 v_t = \Delta v - v + w & \text{in } \Omega \times (0,\infty), \\ \tau_2 w_t = \Delta w - w + u & \text{in } \Omega \times (0,\infty), \end{cases} \tag{1.1}$$

in a bounded domain $\Omega \subset \mathbb{R}^n$ $(n \in \mathbb{N})$ with smooth boundary $\partial\Omega$, where the parameters τ_1, τ_2, and χ are positive. Suppose that the boundary condition:

$$\frac{\partial u}{\partial \nu} - \chi u \frac{\partial v}{\partial \nu} = v = w = 0 \qquad \text{on } \partial\Omega \times (0,\infty). \tag{1.2}$$

Moreover assume that

$$u(\cdot,0) = u_0, \quad v(\cdot,0) = v_0, \quad w(\cdot,0) = w_0 \qquad \text{in } \Omega, \tag{1.3}$$

where the initial data (u_0, v_0, w_0) satisfies

$$\begin{cases} u_0 \in C^0(\overline{\Omega}), \quad u_0 \geq 0 \quad \text{in } \overline{\Omega}, \\ v_0 \in C^2(\overline{\Omega}), \quad v_0 \geq 0 \quad \text{in } \overline{\Omega}, \\ w_0 \in C^2(\overline{\Omega}) \quad u_0 \geq 0 \quad \text{in } \overline{\Omega} \end{cases} \tag{1.4}$$

and the boundary condition

$$v_0 = w_0 = 0 \qquad \text{on } \partial\Omega \times (0,\infty). \tag{1.5}$$

* The first author is partially supported by Grant-in-Aid for Research Activity start-up (No. 17H07131), Japan Society for the Promotion of Science. The second author is partially supported by Grant-in-Aid for Scientific Research (C) (No. 22540200), Japan Society for the Promotion of Science.

†Faculty of Science Division 1, Tokyo University of Science, Tokyo, 162-8601, JAPAN (fujie@rs.tus.ac.jp).

‡Faculty of Science, Fukuoka University, Fukuoka, 814-0180, JAPAN (senba@fukuoka-u.ac.jp).

2. Background and motivation. In 1970 Keller and Segel ([17]) proposed a mathematical model describing a movement of cells, which is the following reaction-diffusion system:

$$\begin{cases} u_t = \Delta u - \chi \nabla \cdot (u \nabla v), \\ v_t = \Delta v - v + u. \end{cases} \tag{2.1}$$

Here functions u and v represent the population of cells and the density of a chemical substance, respectively. The term $-\chi \nabla \cdot (u \nabla v)$ represents the chemotaxis effect.

From a mathematical view point, the type of (2.1) has been studied well (see surveys [14, 12, 1]). Under suitable boundary conditions, smooth solutions of (2.1) conserve the total mass, i.e., $\|u(t)\|_{L^1(\Omega)} = \|u_0\|_{L^1(\Omega)}$ for all $t > 0$. Considering the simplified system of (2.1) such as

$$\begin{cases} u_t = \Delta u - \chi \nabla \cdot (u \nabla v), \\ v_t = \Delta v + u \end{cases}$$

in $\mathbb{R}^n$, we can confirm that the above system is invariant by the standard scaling $u_\lambda(x,t) = \lambda^2 u(\lambda x, \lambda^2 t)$ and $v_\lambda(x,t) = v(\lambda x, \lambda^2 t)$ with $\lambda > 0$ and

$$\|u_\lambda(\cdot,t)\|_{L^1(\mathbb{R}^n)} = \lambda^{2-n} \|u(\cdot,t)\|_{L^1(\mathbb{R}^n)} \qquad t > 0.$$

Hence in the above sense, the two-dimensional setting is the critical case. Moreover, in [10, 19], it is shown that the system (2.1) has the particular mathematical structure, the Lyapunov functional:

$$\frac{d}{dt}\mathcal{F}(u(t), v(t)) + \mathcal{D}(u(t), v(t)) = 0 \qquad \text{for all } t \in (0, T),$$

where

$$\mathcal{F}(u,v) = \int_\Omega (u \log u - \chi u v) + \frac{\chi}{2} \int_\Omega |\nabla v|^2 + \frac{\chi}{2} \int_\Omega v^2,$$

$$\mathcal{D}(u,v) = \int_\Omega u |\nabla (\log u - \chi v)|^2.$$

This Lyapunov functional is the key ingredient in the study of behaviors of solutions to the Keller–Segel system (2.1) ([19, 15, 25]). The Trudinger–Moser inequality ([5]): for all $\varepsilon > 0$ there exists some $C_\varepsilon > 0$ such that for all $u \in H^1(\Omega)$,

$$\log \left(\int_\Omega e^{|u(x)|} \, dx \right) \le \left(\frac{1}{2 \cdot 8\pi} + \varepsilon \right) \|\nabla u\|^2_{L^2(\Omega)} + C_\varepsilon \|u\|_{L^1(\Omega)},$$

plays a role of judgement of the balance of terms in the Lyapunov functional in the critical case $n = 2$. This combination implies "8π-problem", which seems to be one of the main topic in the study of the Keller–Segel system ([16, 3, 18, 2]). Precisely, in the two-dimensional and radially symmetric setting, the behavior of radial solutions to the Neumann problem of (2.1) is classified as follows:

- if $\|u_0\|_{L^1\Omega} < 8\pi/\chi$ then the solution exists globally and remains bounded ([19]).

- there exists some initial data with $\|u_0\|_{L^1\Omega} > 8\pi/\chi$ such that the corresponding solution blows up in finite [11, 13].

As to nonradial solutions, the critical constant changed to $4\pi/\chi$ ([19, 15]). Here the critical constants $8\pi/\chi$ and $4\pi/\chi$ come from the critical constants in the Trudinger–Moser inequality. As to the subcritical case, in [20] it was established that for all regular initial data the system (2.1) has global bounded solution in the one-dimensional setting. As to the supercritical case, that is, the higher dimensional case $n \geq 3$, solutions of (2.1) exist globally in time and converge to the constant steady state provided that $\|u_0\|_{L^{\frac{n}{2}}(\Omega)} + \|\nabla v_0\|_{L^n(\Omega)}$ is sufficiently small ([4]). Moreover there are many finite time blowup radial solutions with $\|u_0\|_{L^1(\Omega)} = m$ for all $m > 0$ ([25]).

Motivation. The motivation of this study is to give a generalization of the Keller–Segel system (2.1) to higher dimensions in the sense of a mathematical structure. Indeed, the system (1.1) has a similar structural properties as the Keller–Segel system. Smooth solutions of (1.1) conserve the total mass, i.e., $\|u(t)\|_{L^1(\Omega)} = \|u_0\|_{L^1(\Omega)}$ for all $t > 0$. We confirm that the simplified system of (1.1) such as

$$\begin{cases} u_t = \Delta u - \chi \nabla \cdot (u \nabla v), \\ \tau_1 v_t = \Delta v + w, \\ \tau_2 w_t = \Delta w + u \end{cases}$$

in $\mathbb{R}^n$ is invariant by the following standard scaling

$$\begin{cases} u_\lambda(x,t) & = \lambda^4 u(\lambda x, \lambda^2 t), \\ v_\lambda(x,t) & = v(\lambda x, \lambda^2 t), \\ w_\lambda(x,t) & = \lambda^2 w(\lambda x, \lambda^2 t) \qquad (\lambda > 0). \end{cases}$$

Moreover we have

$$\|u_\lambda(\cdot,t)\|_{L^1(\mathbb{R}^n)} = \lambda^{4-n} \|u(\cdot,t)\|_{L^1(\mathbb{R}^n)} \qquad t > 0.$$

Hence the four-dimensional setting is the critical case in the above sense. Moreover the system (1.1) has a Lyapunov functional, which seems to be a natural generalization of one of the Keller–Segel system (2.1):

$$\frac{d}{dt}\mathcal{F}(u(t), v(t)) + \mathcal{D}(u(t), v(t)) = 0 \qquad \text{for all } t \in (0, T),$$

where

$$\mathcal{F}(u,v) = \int_\Omega (u \log u - \chi u v) + \frac{\tau_1 \tau_2 \chi}{2} \int_\Omega |v_t|^2 + \frac{\chi}{2} \int_\Omega |(-\Delta + 1)v|^2,$$

$$\mathcal{D}(u,v) = \chi(\tau_1 + \tau_2) \int_\Omega \left(|\nabla v_t|^2 + |v_t|^2 \right) + \int_\Omega u |\nabla(\log u - \chi v)|^2.$$

Now, in the critical case $n = 4$, an Adams type inequality, which is a generalization of the Trudinger–Moser inequality to higher derivatives, plays a key role to decide the balance of the Lyapunov functional in the same way that the Trudinger–Moser inequality does in the study of the Keller–Segel system. Hence the system (1.1) has a generalized mathematical structure of the Keller–Segel system.

3. Main results. Our main results read as follows.

THEOREM 3.1 ([8]). *Let $n \leq 3$. Suppose that (u_0, v_0, w_0) satisfies* (1.4) *and* (1.5). *Then the problem* (1.1)–(1.2)–(1.3) *has a unique classical positive solution, which exists globally in time. Moreover the solution is uniformly bounded in time in the sense that*

$$\sup_{t\in[0,\infty)} \Big(\|u(t)\|_{L^\infty(\Omega)} + \|v(t)\|_{W^{2,\infty}(\Omega)} + \|w(t)\|_{W^{1,\infty}(\Omega)} \Big) < \infty.$$

REMARK 3.2. *This result corresponds to the study of the Keller–Segel system in the one-dimensional case. In [20] it is shown that for all regular initial data the Keller–Segel system* (2.1) *has global and bounded solution.*

THEOREM 3.3 ([8]). *Let $n = 4$. Suppose that the initial data (u_0, v_0, w_0) satisfies* (1.4), (1.5) *and*

$$\int_\Omega u_0 < \frac{(8\pi)^2}{\chi}.$$

Then the problem (1.1)–(1.2)–(1.3) *has a unique classical positive solution, which exists globally in time. Moreover the solution is uniformly bounded in time in the sense that*

$$\sup_{t\in[0,\infty)} \Big(\|u(t)\|_{L^\infty(\Omega)} + \|v(t)\|_{W^{2,\infty}(\Omega)} + \|w(t)\|_{W^{1,\infty}(\Omega)} \Big) < \infty.$$

REMARK 3.4. *As to the initial-boundary problem of the Keller–Segel system with the mixed boundary condition, nonradial solutions exist globally in time and remain bounded if $\|u_0\|_{L^1(\Omega)} < 8\pi/\chi$. Hence the above theorem is regarded as a generalization of the study of the Keller–Segel system.*

REMARK 3.5. *By the standard compactness methods, we can show asymptotic behavior of the globally bounded solutions in Theorem 3.1 and Theorem 3.3. Precisely, there exists some increasing sequence $T_k \in (0, \infty)$ such that $(u(T_k), v(T_k), w(T_k))$ converges to a solution of the stationary problem.*

We consider blowup solutions to (1.1)–(1.2)–(1.3). The following is the definition of blowup of solutions.

DEFINITION 3.6. *We say that a solution (u, v, w) to* (1.1) *blows up, if the solution satisfies*

$$\limsup_{t\nearrow T_{max}} (\|u(t)\|_{L^\infty(\Omega)} + \|v(t)\|_{L^\infty(\Omega)} + \|w(t)\|_{L^\infty(\Omega)}) = \infty,$$

where T_{max} is the maximal existence time of the classical solution (u, v, w).

THEOREM 3.7 ([9]). *Suppose $n = 4$, Ω be a convex bounded domain and $\Lambda \in ((8\pi)^2/\chi, \infty) \setminus \{(8\pi)^2/\chi\}\mathbb{N}$. Then there exist blowup solutions (u, v, w) to* (1.1)–(1.2)–(1.3) *satisfying $\|u(t)\|_{L^1(\Omega)} = \Lambda$.*

REMARK 3.8. *By Theorem 3.3 and Theorem 3.7, we established that the case where $n = 4$ and $\int_\Omega u_0 = (8\pi)^2/\chi$ is critical and that this case is corresponding to the case $n = 2$ and $\int_\Omega u_0 = 8\pi/\chi$ of the Keller–Segel system.*

4. Strategy and mathematical challenge. As compared with the Keller–Segel system, we should control the power balance between the terms $\int_\Omega u\log u + (\chi/2)\int_\Omega |(-\Delta+1)v|^2$ and $\chi\int_\Omega uv$. Instead of the Trudinger–Moser inequality, we will apply the Adams type inequality ([21, 24]): for all $\varepsilon > 0$ there exists some $C_\varepsilon > 0$ such that for all $u \in H^2(\Omega)$,

$$\log\left(\int_\Omega e^{|u(x)|}\,dx\right) \le \left(\frac{1}{2(8\pi)^2}+\varepsilon\right)\|(-\Delta+1)v\|^2_{L^2(\Omega)} + C_\varepsilon.$$

We remark that the critical constant of the Adams type inequality implies the constant $(8\pi)^2/\chi$. Invoking the smallness of the mass, we can combine these estimates and deduce the lower estimate for the Lyapunov functional.

The mathematical challenge is also in regularity estimates. After deriving the energy estimate from the lower estimate for the Lyapunov functional, we will proceed to deduce L^p estimate for u. We cannot adopt the approach in the study of the Keller–Segel system to our system (1.1) because the four-dimensional setting disturbs the relationships of exponents in the Sobolev inequality. Moreover the particular structure of (1.1), i.e., the system (1.1) consists of three parabolic equations, causes a difficulty. From this reason, we use the localizing method, which is introduced in [22, 23, 6, 7].

As to the blowup result, our method has the same spirit in [13, 15]. We first consider a blowing up sequence of stationary solutions. Stationary solutions (u, v, w) to (1.1)–(1.2)–(1.3) satisfy that

$$\begin{cases} 0 = \Delta u - \chi\nabla\cdot(u\nabla v) & \text{in } \Omega,\\ 0 = \Delta v - v + w & \text{in } \Omega,\\ 0 = \Delta w - w + u & \text{in } \Omega,\\ u \ge 0,\ v \ge 0,\ w \ge 0 & \text{in } \Omega,\\ \dfrac{\partial u}{\partial \nu} - \chi u\dfrac{\partial v}{\partial \nu} = v = w = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

Put $\Lambda = \|u\|_{L^1(\Omega)} \in (0,\infty)$. The system (4.1) can be rewritten as the following:

$$\begin{cases} (-\Delta+1)^2 v = \dfrac{\Lambda}{\int_\Omega e^{\chi v}}e^{\chi v} & \text{in } \Omega,\\ u = \dfrac{\Lambda}{\int_\Omega e^{\chi v}}e^{\chi v}, \quad w = -\Delta v + v & \text{in } \Omega,\\ v = \Delta v = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.2}$$

Here and henceforth, we say that (u, v, w, Λ) is a solution to (4.2), if the function (u, v, w) and the positive constant Λ satisfies (4.2). The following proposition plays a key role in our analysis. This claim is about a quantization property of solutions to (4.2).

PROPOSITION 4.1 ([9]). *Let $\Lambda > 0$. Suppose that solutions $\{(u_k, v_k, w_k, \Lambda)\}_k$ to (4.2) satisfy that $\lim_{k\to\infty}\|v_k\|_{L^\infty(\Omega)} = \infty$. Then $J = \Lambda\chi/(8\pi)^2$ is a positive integer and there is a set of points $\{Q(j)\}_{j=1}^J \subset \Omega$ satisfying that*

$$u_k \to \sum_{j=1}^J \frac{(8\pi)^2}{\chi}\delta_{Q(j)} \quad \text{in } \mathcal{M}(\overline{\Omega}) \quad \text{as } k\to\infty,$$

where $\delta_{Q(j)}$ is the delta function whose support is the point $Q(j)$ and $\mathcal{M}(\overline{\Omega})$ is a set of Radon measures on $\overline{\Omega}$.

For $\Lambda > 0$ put the set $\mathcal{S}(\Lambda)$ as

$$\big\{(u,v,w) \in C^2(\overline{\Omega}) : (u,v,w) \text{ is a stationary solution to (1.1)–(1.2)–(1.3) with } \|u\|_{L^1(\Omega)} = \Lambda\big\}.$$

The following lemma is an immediate consequence of Proposition 4.1.

LEMMA 4.2 ([9]). *For $\Lambda \in (0,\infty) \setminus \{(8\pi)^2/\chi\}\mathbb{N}$, there exists a constant $C > 0$ satisfying*

$$\sup\{\|(u,v,w)\|_{L^\infty(\Omega)} : (u,v,w) \in \mathcal{S}(\Lambda)\} \le C$$

and

$$F_*(\Lambda) := \inf\{\mathcal{F}(u,v,w) : (u,v,w) \in \mathcal{S}(\Lambda)\} \ge -C.$$

In order to find a blowup solution, we construct a triplet of nonnegative functions (u_0, v_0, w_0) satisfying

$$\mathcal{F}(u_0,v_0,w_0) < F_*(\Lambda) \qquad \text{for } \Lambda > (8\pi)^2/\chi \quad \text{with } \Lambda \notin \{(8\pi)^2/\chi\}\mathbb{N}.$$

5. Further comments and conjectures. Let us first give some comments on Neumann boundary case. Suppose that the following boundary conditions:

$$\frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu} = \frac{\partial w}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0,\infty) \tag{5.1}$$

and the initial data satisfies the boundary condition

$$\frac{\partial v_0}{\partial \nu} = \frac{\partial w_0}{\partial \nu} = 0 \qquad \text{on } \partial\Omega \times (0,\infty). \tag{5.2}$$

Moreover we assume the radial symmetry:

$$\Omega = B(R) = \{x \in \mathbb{R}^4 \,|\, |x| \le R\} \quad \text{with} \quad R > 0 \qquad \text{and}$$
$$(u_0, v_0, w_0): \text{ radial symmetry.}$$

THEOREM 5.1 ([8]). *Let $n = 4$, $\Omega = B(R) = \{x \in \mathbb{R}^4 \,|\, |x| \le R\}$ $(R > 0)$. Suppose that (u_0, v_0, w_0) is radially symmetric and satisfies* (1.4), (5.2) *and*

$$\int_\Omega u_0 < \frac{(8\pi)^2}{\chi}.$$

Then the problem (1.1)–(5.1)–(1.3) *has a unique classical positive solution, which exists globally in time. Moreover the solution is uniformly bounded in time in the sense that*

$$\sup_{t\in[0,\infty)} \Big(\|u(t)\|_{L^\infty(\Omega)} + \|v(t)\|_{W^{2,\infty}(\Omega)} + \|w(t)\|_{W^{1,\infty}(\Omega)}\Big) < \infty.$$

REMARK 5.2. *Comparing with the study of the two-dimensional Keller–Segel system, the critical constant is changed from* $8\pi/\chi$ *to* $(8\pi)^2/\chi$.

REMARK 5.3. *We used the assumption of radial symmetry to deduce an Adams type inequality in [8]. We conjecture that without this assumption the threshold constant seems to be* $(8\pi)^2/2\chi$.

As to blowup of solution, at least, the following questions have been left as an open (especially, the second one is related to the result [25]):

- does the blowup in Theorem 3.3 occur at finite time or infinite time?;
- does the solution blows up independently of the size of the initial data in the super critical case ($n \geq 5$)?

REFERENCES

[1] N. BELLOMO, A. BELLOUQUID, Y. TAO, M. WINKLER: *Toward a mathematical theory of Keller–Segel models of pattern formation in biological tissues.* Math. Models Methods Appl. Sci. **25** (2015), 1663–1763.

[2] P. BILER, G. KARCH, P. LAURENÇOT, T. NADZIEJA: *The 8π-problem for radially symmetric solutions of a chemotaxis model in a disc.* Topol. Methods Nonlinear Anal. **27** (2006), 133–147.

[3] P. BILER, T. NADZIEJA: *Existence and nonexistence of solutions for a model of gravitational interaction of particles. I.* Colloq. Math. **66** (1994), 319–334.

[4] X. CAO: *Global bounded solutions of the higher-dimensional Keller-Segel system under smallness conditions in optimal spaces.* Discrete Contin. Dyn. Syst. **35** (2015), 1891–1904.

[5] S.Y.A. CHANG, P. YANG: *Conformal deformation of metrics on* S^2. J. Differential Geom. **27** (1988), 259–296.

[6] K. FUJIE, T. SENBA: *Global existence and boundedness in a parabolic-elliptic Keller-Segel system with general sensitivity.* Discrete Contin. Dyn. Syst. Ser. B **21** (2016), 81–102.

[7] K. FUJIE, T. SENBA: *Global existence and boundedness of radial solutions to a two dimensional fully parabolic chemotaxis system with general sensitivity.* Nonlinearity **29** (2016), 2417–2450.

[8] K. FUJIE, T. SENBA: *Application of an Adams type inequality to a two-chemical substances chemotaxis system.* J. Differential Equations **263** (2017), 88–148.

[9] K. FUJIE, T. SENBA: *Blow-up of solutions to a two-chemical substances chemotaxis system in the critical dimension.* In preparation.

[10] H. GAJEWSKI, K. ZACHARIAS: *On a reaction-diffusion system modelling chemotaxis.* International Conference on Differential Equations, Vol. 1, 2 (Berlin, 1999), 1098–1103, World Sci. Publ., River Edge, NJ, 2000.

[11] M.A HERRERO, J.J.L. VELÁZQUEZ: *A blow-up mechanism for a chemotaxis model.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. **24** (1997), 663–683.

[12] T. HILLEN, K. PAINTER: *A user's guide to PDE models for chemotaxis.* J. Math. Biol. **58** (2009), 183–217.

[13] D. HORSTMANN: *On the existence of radially symmetric blow-up solutions for the Keller-Segel model.* J. Math. Biol. **44** (2002), 463–478.

[14] D. HORSTMANN: *From 1970 until present: the Keller-Segel model in chemotaxis and its consequences. I.* Jahresber. Deutsch. Math.-Verein. **105** (2003), 103–165.

[15] D. HORSTMANN, G. WANG: *Blow-up in a chemotaxis model without symmetry assumptions.* European J. Appl. Math. **12** (2001), 159–177.

[16] W. JÄGER, S. LUCKHAUS: *On explosions of solutions to a system of partial differential equations modelling chemotaxis.* Trans. Amer. Math. Soc. **329** (1992), 819–824.

[17] E.F. KELLER, L.A. SEGEL: *Initiation of slime mold aggregation viewed as an instability.* J. Theor. Biol. **26** (1970), 399–415.

[18] T. NAGAI: *Blow-up of radially symmetric solutions to a chemotaxis system.* Adv. Math. Sci. Appl. **5** (1995), 581–601.

[19] T. NAGAI, T. SENBA, K. YOSHIDA: *Application of the Trudinger-Moser inequality to a parabolic system of chemotaxis.* Funkc. Ekvacioj, Ser. Int. **40** (1997), 411–433.

[20] K. OSAKI, A. YAGI: *Finite dimensional attractor for one-dimensional Keller-Segel equations.* Funkcial. Ekvac. **44** (2001), 441–469.

[21] B. Ruf, F. Sani: *Sharp Adams-type inequalities in* $\mathbb{R}^n$. Trans. Amer. Math. Soc. **365** (2013), 645–670.
[22] Y. Sugiyama: *On* ε*-regularity theorem and asymptotic behaviors of solutions for Keller-Segel systems.* SIAM J. Math. Anal. **41** (2009), 1664–1692.
[23] T. Senba, T. Suzuki: *Chemotactic collapse in a parabolic-elliptic system of mathematical biology.* Adv. Differential Equations **6** (2001), 21–50.
[24] C. Tarsi: *Adams' inequality and limiting Sobolev embeddings into Zygmund spaces.* Potential Anal. **37** (2012), 353–385.
[25] M. Winkler: *Finite-time blow-up in the higher-dimensional parabolic-parabolic Keller-Segel system.* J. Math. Pures Appl. **100** (2013), 748–767.

Proceedings of EQUADIFF 2017
pp. 283–286

A NOTE ON THE UNIQUENESS AND STRUCTURE OF SOLUTIONS TO THE DIRICHLET PROBLEM FOR SOME ELLIPTIC SYSTEMS*

JANN-LONG CHERN†, SHOJI YOTSUTANI ‡, AND NICHIRO KAWANO §

Abstract. In this note, we consider some elliptic systems on a smooth domain of R^n. By using the maximum principle, we can get a more general and complete results of the identical property of positive solution pair, and thus classify the structure of all positive solutions depending on the nonlinearities easily.

Key words. elliptic system, uniqueness, solutions structure

AMS subject classifications. 35J57, 35B09, 35J91

1. Introduction. In this paper, we consider the smooth positive solutions of the following elliptic system

$$\begin{cases} \Delta u + \phi(x) u^p v^q = 0 \\ \Delta v + \phi(x) u^q v^p = 0 \end{cases} \tag{1.1}$$

in Ω with boundary condition

$$(u, v) = (0, 0) \text{ on } \partial\Omega, \tag{1.2}$$

where $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}, n \geq 3, p > 0, q > 0, \Omega \subset \mathbf{R}^n$ is a domain with the maximum principle holds true, and ϕ is a positive and continuous function in Ω. If $n = 3, \phi \equiv 1$ and $(p, q) = (2, 3)$, system (1.1) arises from the stationary Schrödinger system with critical exponent for Bose-Einstein condensate. We refer the readers to [5], [8], [11], [12], and the references therein. If $\phi = \frac{1}{1+|x|^2}$ then system (1.1) is called a Matukuma-type system. Recently, if $\phi \equiv 1, 1 \leq p, q \leq \frac{n+2}{n-2}$ and $p + q = \frac{n+2}{n-2}$, i.e., the critical exponent case, Li-Ma[10] used the Hardy-Littlewood-Sobolev inequality to prove that any $L^{\frac{2n}{n-2}}(R^n) \times L^{\frac{2n}{n-2}}(R^n)$ positive solution (u, v) to system (1.1) is radial symmetric. Furthermore, they also showed that any $L^{\frac{2n}{n-2}}(R^n) \times L^{\frac{2n}{n-2}}(R^n)$ radial symmetric solution (u, v) is unique and $u \equiv v$. In this note, we consider the general case, $p > 0, q > 0$, and, by using the maximum principle to get a more general and complete result.

Our first theorem is the following.

THEOREM 1.1. *Let $\phi > 0$ in Ω, and $p, q > 0$. Then if $q \geq p$ then any positive smooth solution (u, v) of (1.1)-(1.2) satisfies $u \equiv v$.*

*Work by the first author was partially supported by the Ministry of Science and Technology(MOST) of Taiwan; Grant No.:104-2115-M-008-010-MY3.

†Departmentof Mathematics, National Central Univeristy, Chung-Li 32001,Taiwan(chern@math.ncu.edu.tw).

‡Department of Applied Mathematics and Informatics, Ryukoku University Seta, Otsu, 520-2194, JAPAN (shoji@math.ryukoku.ac.jp).

§Department of Education and Culture, University of Miyazaki, JAPAN.

By Theorem 1.1 we easily obtain the following results about the symmetry, existence and uniqueness results.

COROLLARY 1.2. *Let $\phi \equiv 1$. Then the following properties are valid.*

(i) *If $0 < p \le q \le \frac{n+2}{n-2}, 1 \le p+q < \frac{n+2}{n-2}$ and Ω is bounded, then (1.1)-(1.2) possesses one and only one positive solution (u, v) and $u \equiv v$ in Ω. In addition, if Ω is symmetric with respect to some $x_i -$ axis, then (u, v) is symmetric with respect to $x_i -$ axis.*

(ii) *If $0 < p \le q,\ p+q = \frac{n+2}{n-2}$ and $\Omega = \mathbf{R}^n$, then every positive solution (u, v) of (1.1) satisfies $u \equiv v$, and it is radial symmetric with one parameter family of functions*

$$\phi_\lambda(x) = \Big(\frac{\lambda\sqrt{n(n-2)}}{\lambda^2 + |x - x_0|^2}\Big)^{\frac{n-2}{2}},$$

where $\lambda > 0$ is a parameter and $x_0 \in \mathbf{R}^n$.

(iii) *If $p + q \ge \frac{n+2}{n-2}$ and $\Omega \ne \mathbf{R}^n$ is a star-shape domain, then (1.1)-(1.2) does not have any positive solution.*

Remark 1.3

(A) If $q < p$ and $\phi \equiv 1$, then equations (1.1)-(1.2) may possess infinite many positive solutions. The details can be found in [2]. For example, if $q = p-1,\ 1 < 2p-1 < \frac{n+2}{n-2}$ and Ω is bounded, then for any $\lambda > 0$, $(\lambda v, v)$ is a positive solution of equations (1.1)-(1.2), where v is the positive solutionof the following equation

$$\begin{cases} \Delta u + \lambda^{p-1} u^{2p-1} = 0 \ \text{ in } \ \Omega \\ v = 0 \ \text{ on } \partial\Omega. \end{cases} \tag{1.3}$$

(B) We note that Corollary 1.2-(ii) was proved by Li-Ma [10] if $1 \le p < q \le \frac{n+2}{n-2}$, $p+q = \frac{n+2}{n-2}$ and (u, v) is in $L^{\frac{2n}{n-2}}(R^n) \times L^{\frac{2n}{n-2}}(R^n)$. In this case, system (1.1) will reduce to one equation. Then, if Ω is a ball or R^n, and under the respective condition in the parts (i) and (ii) of Corollary 1.2, by using the method of moving plane, we can finally get that all positive solutions of (1.1)-(1.2) are radially symmetry. By the way, from the Pohozaev identity, we can also get the non-existence result of part (iii) in Corollary 1.2.

(C) By using the same ideas and proofs, some classes of systems, e.g., Schrödinger-type system, Matukuma-type system, etc, have also their respective results of Theorem 1.1 and Corollary 1.2. The details can be found in [2]. For example, let $\phi(x) = \frac{1}{1+|x|^2}$, we can also consider the following Matukuma-type system

$$\begin{cases} \Delta u + \frac{1}{1+|x|^2} u^p v^q = 0 \ \text{ in } \ R^3 \\ \Delta v + \frac{1}{1+|x|^2} u^q v^p = 0 \ \text{ in } \ R^3. \end{cases} \tag{1.4}$$

Then, from Theorem 1.1 and by using Theorems 1-2 in [13], we easily obtain the following results

THEOREM 1.3. *Suppose $0 < p \le q$ and $1 < p + q < 5$. Then the following statements are valid.*

(i) *Every positive entire solution (u, v) of equation (1.4) satisfies $u \equiv v$ in R^3, and it is radially symmetric about the origin with $u'(r) < 0\ \forall r > 0$.*

(ii) *Let* $TM(u) = \frac{1}{4\pi}\int_{R^3} \frac{u^p}{1+|x|^2}\, dx$ *be the total mass of* u. *Then equation (1.4) has an unique positive entire solution with finite total mass, and has infinitely many positive entire solutions with infinite total mass.*

In Section 2, based on the maximum principle, we can get the proof of Theorem 1.1. Applying Theorem 1.1 and by using the well-known results of Yamabe problem, we also easily get Corollary 1.2.

2. Identical Property and Proof of Main Results. We prove Theorem 1.1 and Corollary 1.2 in this section.

Proof of Theorem 1.1. Let (u, v) be a positive solution of (1.1)-(1.2), and let $w = u - v$. Then, by (1.1), w satisfies

$$\Delta w = \phi(x)(-u^p v^q + u^q v^p) \ \text{ in } \ \Omega, \ w = 0 \ \text{ on } \ \partial\Omega. \tag{2.1}$$

We divide the proof into the following steps.

Step 1. If $q > p > 0$ then we want to show $w \equiv 0$, i.e., $u \equiv v$, in Ω.

First, we prove $v(x) \geq u(x)\ \forall x \in \Omega$. Suppose not, then there exists some $x_0 \in \Omega$ such that $w(x_0) = u(x_0) - v(x_0) > 0$. Then there exists some $x_1 \in \text{ int}(\Omega)$ such that

$$w(x_1) = \max_{x\in\Omega} w(x) > 0 \ \text{ and } \ \Delta w(x_1) \leq 0. \tag{2.2}$$

By $\phi > 0$ and (2.1)-(2.2), we easily obtain

$$0 \geq \Delta w(x_1) = -\phi(x_1)(u^p(x_1)v^q(x_1)(1 - (\frac{u(x_1)}{v(x_1)})^{q-p})) > 0.$$

This contradiction shows $v(x) \geq u(x)\ \forall x \in \Omega$.

Now, suppose $v \not\equiv u$ in Ω. Then by (1.1)-(1.2), we easily deduce that

$$0 = \int_\Omega (v\Delta u - u\Delta v)dx = \int_\Omega -\phi(x)u^p v^{q+1}(1 - (\frac{u}{v})^{q+1-p})dx < 0.$$

This contradiction proves $u \equiv v$ if $q > p > 0$.

Step 2. If $p = q > 0$. then by (2.1) we easily obtain

$$\Delta w = \phi(x)(u^p v^p - u^p v^p) = 0 \ \text{ in } \ \Omega \ \text{ and } \ w = 0 \ \text{ on } \ \partial\Omega.$$

This shows $u \equiv v$ in Ω.

By **Steps 1 and 2** we complete the proof of Theorem 1.1. q.e.d.

Now we are in a position to prove Corollary 1.2.

Proof of Corollary 1.2. Let $\phi \equiv 1$ and (u, v) be a positive solution of (1.1). By our main result, Theorem 1.1, we obtain that $u \equiv v$, and then system (1.1) reduces to the following one equation

$$\begin{aligned} &\Delta u + u^{p+q} = 0 \ \text{ in } \ \Omega \\ &u = 0 \ \text{ on } \ \partial\Omega. \end{aligned} \tag{2.3}$$

Then by the well-known Yamabe problem and the prescribing scalar curvature problem, e.g., see [6], [1], [4], [3], [9] and the references therein, we easily obtain the results (i)-(iii) of Corollary 1.2. We complete the proof. q.e.d

Acknowledgments. Work by the first author was partially supported by the Ministry of Science and Technology(MOST) of Taiwan; Grant No.:104-2115-M-008-010-MY3.

REFERENCES

[1] L. A. Caffarelli, B. Gidas and J. Spruck, *Asymptotic symmetry and local behavior of semilinear elliptic equations with critical Sobolev growth*, Comm. Pure Appl. Math. 42 (1989), no. 3, 271–297.

[2] Jann-Long Chern, Nichiro Kawano and Shoji Yotsutani, *Approximations and Analysis of Positive Solutions for Some Elliptic Systems*, Preprint, 2017.

[3] W. Chen and C. Li, *Classification of solutions of some nonlinear elliptic equations*, Duke Math. J. 63 (1991), 615–622.

[4] C.-C. Chen and C.-S. Lin, *Uniqueness of the ground state solutions of* $\Delta u + f(u) = 0$ *in* $\mathbf{R}^n, n \geq 3$, Comm. Partial Diff. Eqns 16 (1991), 1549-1572.

[5] D.G. de Figueiredo and O. Lopes, *Solitary waves for some nonlinear Schrödinger systems*, Ann. I. H. Poincaré Anal. Nonlinéaire 25 (2008), 149-161.

[6] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys. 68 (1979), 209-243.

[7] D. Gilbarg and N.S. Trudinger, *Elliptic partial differential equations of second order*, Grundlehren der Mathematischen Wissenschaften. vol. 224, Springer-Verlag, Berlin, 1983.

[8] T. Kanna and M. Lakshmanan, *Effect of phase shift in shape changing collision of solitons in coupled nonlinear Schrodinger equations*, Topical issue on geometry, integrability and nonlinearity in condensed matter physics. Eur. Phys. J. B Condens. Matter Phys. 29 (2002), no. 2, 249–254.

[9] C. Li, *Local asymptotic symmetry of singular solutions to nonlinear elliptic equations*, Invent. Math. 123 (1996), no. 2, 221–231.

[10] C. Li and L. Ma, *Uniqueness of positive bound states to Schrodinger systems with critical exponents*, SIAM J. Math. Anal. 40 (2008), no. 3, 1049–1057.

[11] T. Lin and J. Wei, *Ground state of N coupled nonlinear Schrodinger equations in* $\mathbf{R}^n n$, $n \leq 3$, Comm. Math. Phys. 255 (2005), 629–653.

[12] T. Lin and J. Wei, *Spikes in two coupled nonlinear Schrodinger equations*, Ann. Inst. H. Poincare Anal. Non Lineaire 22 (2005), 403–439.

[13] Y. Li, *On the positive solutions of the Matukuma equation*, Duke Math, J. 70 (1993), no. 3, 575-589.

Proceedings of EQUADIFF 2017
pp. 287–294

CLASSICAL AND GENERALIZED JACOBI POLYNOMIALS ORTHOGONAL WITH DIFFERENT WEIGHT FUNCTIONS AND DIFFERENTIAL EQUATIONS SATISFIED BY THESE POLYNOMIALS *

MARIANA MARČOKOVÁ† AND VLADIMÍR GULDAN‡

Abstract. In this contribution we deal with classical Jacobi polynomials orthogonal with respect to different weight functions, their special cases - classical Legendre polynomials and generalized brothers of them. We derive expressions of generalized Legendre polynomials and generalized ultraspherical polynomials by means of classical Jacobi polynomials.

Key words. orthogonal polynomial, weight function, classical Jacobi polynomial, classical Legendre polynomial, generalized orthogonal polynomial, differential equation

AMS subject classifications. 33C45, 42C05

1. Introduction. This paper presents relations of generalized Legendre polynomials of a certain type to classical Jacobi polynomials with some different weight functions. Also generalization to ultraspherical polynomials is given. Further, we deal with influence to Jacobi polynomials, when their weight function is multiplied by even function. In the conclusion we derive the differential equations satisfied by the introduced generalized Legendre polynomials. The motivation for such investigation was obtained when studying the book [2] dealing with physical geodesy and the papers [3], [5], [7], [10], and [11] using Legendre polynomials in applications.

1.1. Definition and basic properties of orthogonal polynomials. We recall the definition and the basic properties of orthogonal polynomials that can be found in the basic literature on orthogonal polynomials (cf. [1], [4], [8], and [9]).

DEFINITION 1.1. *Let $(a,b) \subset R$ be a finite or infinite interval. A function $v(x)$ is called the weight function if at this interval it fulfills the following conditions:*

(i) *$v(x)$ is nonnegative at (a,b), i.e.*

$$v(x) \geq 0,$$

(ii) *$v(x)$ is integrable at (a,b), i.e.*

$$0 < \int_a^b v(x)dx < \infty$$

and

(iii) *for every $n = 0, 1, 2, \ldots$*

$$0 < \int_a^b |x|^n v(x)dx < \infty.$$

*This work was supported by the Grant No.:14-49-00079-P of Russian Science Foundation.

†Faculty of Civil Engineering, University of Žilina, Univerzitná 1, Žilina, Slovak Republic and Moscow Power Engineering Institute, Moscow, Russia (mariana.marcokova@gmail.com).

‡Faculty of Mechanical Engineering, University of Žilina, Univerzitná 1, Žilina, Slovak Republic (vladimir.guldan@fstroj.uniza.sk).

DEFINITION 1.2. *Let $\{P_n(x)\}_{n=0}^{\infty}$ be a system of polynomials, where every polynomial $P_n(x)$ has the degree n. If for all polynomials of this system*

$$\int_a^b P_n(x)P_m(x)v(x)dx = 0,\ n \neq m,$$

then the polynomials $\{P_n(x)\}_{n=0}^{\infty}$ are called orthogonal in (a,b) with respect to the weight function $v(x)$. If moreover

$$||P_n(x)||_{v(x)} = \left[\int_a^b P_n^2(x)v(x)dx\right]^{\frac{1}{2}} = 1$$

for every $n = 0, 1, 2, \ldots$, then the polynomials are called orthonormal in (a,b).

So the condition of the orthonormality of the system $\{P_n(x)\}_{n=0}^{\infty}$ has the form

$$\int_a^b P_n(x)P_m(x)v(x)dx = \delta_{nm},$$

where δ_{nm} is Kronecker delta.

THEOREM 1.3. *For every weight function $v(x)$ there exists one and only one system of polynomials $\{P_n(x)\}_{n=0}^{\infty}$ orthonormal in (a,b), where*

$$P_n(x) = \sum_{k=0}^{n} a_k^{(n)} x^{n-k}\ ,\ a_0^{(n)} > 0.$$

THEOREM 1.4. *A polynomial $P_n(x)$ is orthogonal in (a,b) with respect to the weight function $v(x)$, if and only if for arbitrary polynomial $S_m(x)$ of the degree $m < n$ the following condition is fulfilled*

$$\int_a^b P_n(x)S_m(x)v(x)dx = 0.$$

THEOREM 1.5. *If the interval of orthogonality is symmetric according to the origin of coordinate system and weight function $v(x)$ is even function, then every orthogonal polynomial $P_n(x)$ fulfils the equality*

$$P_n(-x) = (-1)^n P_n(x).$$

1.2. Classical Jacobi polynomials, classical Legendre polynomials and differential equations satisfied by them. It is well-known that Jacobi polynomials $\{P_n(x;\alpha,\beta)\}_{n=0}^{\infty}$ are orthogonal in the interval $I=(-1,1)$ with respect to the weight function

$$J(x)=(1-x)^{\alpha}(1+x)^{\beta},\ x\in(-1,1), \tag{1.1}$$

where $\alpha>-1,\beta>-1$. Very important special case of Jacobi polynomials are classical Legendre polynomials $\{P_n(x;0,0)\}_{n=0}^{\infty}$, for which $\alpha=\beta=0$ in the weight function $J(x)$. In the next we denote them by $\{P_n(x)\}_{n=0}^{\infty}$. As it is seen the Legendre classical polynomials $\{P_n(x)\}_{n=0}^{\infty}$ are orthogonal in $I=(-1,1)$ with respect to the weight function $L(x)=1$. If $\alpha=\beta$, then polynomials $\{P_n(x;\alpha,\alpha)\}_{n=0}^{\infty}$ are called ultraspherical polynomials.

Classical orthogonal polynomials are solutions of the second order linear homogeneous differential equations of the form (cf. e.g. [4], [8], and [9]):

$$a(x)y_n''(x)+b(x)y_n'(x)+\lambda_n y_n(x)=0,$$

where $a(x)$ is a polynomial of the degree at most 2 , $b(x)$ is a polynomial of the degree 1 and λ_n does not depend of x . For the classical Jacobi polynomials this equation has the form

$$(1-x^2)y_n''(x)+[\beta-\alpha-(\alpha+\beta+2)x]\,y_n'(x)+n(n+\alpha+\beta+1)y_n(x)=0,$$

which in the case of the classical Legendre polynomials is reduced to the equation

$$(1-x^2)y_n''(x)-2xy_n'(x)+n(n+1)y_n(x)=0. \tag{1.2}$$

2. Generalized Legendre polynomials of a certain type and classical Jacobi polynomials with different weight functions. In [6] we introduced the system of polynomials $\{Q_n(x)\}_{n=0}^{\infty}$ which are the polynomials orthonormal in I with respect to the weight function

$$Q(x)=\left(x^2\right)^{\gamma},$$

where $\gamma>0$ and $Q_n(+\infty)>0$. It is clear that these polynomials are generalization of the classical Legendre polynomials, which can be obtained by substituting $\gamma=0$ in the weight function $Q(x)$.

Further in [6] we introduced two classes of orthonormal polynomials:

1. polynomials $\{P_n(x;0,\gamma-\frac{1}{2})\}_{n=0}^{\infty}$ orthonormal in I with respect to the weight function

$$J_1(x)=(1+x)^{\gamma-\frac{1}{2}}$$

and

2. polynomials $\{P_n(x;0,\gamma)\}_{n=0}^{\infty}$ orthonormal in I with respect to the weight function

$$J_2(x)=(1+x)^{\gamma}.$$

In both these cases we have classical Jacobi polynomials orthogonal with the weight function (1.1) for $\alpha = 0$, $\beta = \gamma - \frac{1}{2}$ and $\alpha = 0$, $\beta = \gamma$, respectively. In the next theorem we proved relations between them and the polynomials $\{Q_n(x)\}_{n=0}^{\infty}$ (cf. [6]). Here we give this theorem with its proof because it is essential for our further investigation.

THEOREM 2.1. *In the notations introduced in the previous sections we have*

$$Q_{2n}(x) = 2^{\frac{\gamma}{2}-\frac{1}{4}} P_n\left(2x^2-1; 0, \gamma-\frac{1}{2}\right) \tag{2.1}$$

and

$$Q_{2n+1}(x) = 2^{\frac{\gamma}{2}} x P_n\left(2x^2-1; 0, \gamma\right). \tag{2.2}$$

Proof. According to the Theorem 1.5, the function $Q_{2n}(x)$ is even function. Putting $t = x^2$ we denote $W_n(t) = Q_{2n}(x)$. The orthogonality of the polynomials $\{Q_n(x)\}_{n=0}^{\infty}$ for $r = 0, 1, \ldots, n-1$ and $n > 0$ yields

$$0 = \int_0^1 x^{2r} Q_{2n}(x) x^{2\gamma} dx = \frac{1}{2}\int_0^1 t^r W_n(t) t^{\gamma-\frac{1}{2}} dt =$$

$$= \frac{1}{2^2}\int_{-1}^1 \left(\frac{\tau+1}{2}\right)^r W_n\left(\frac{\tau+1}{2}\right)\left(\frac{\tau+1}{2}\right)^{\gamma-\frac{1}{2}} d\tau =$$

$$= \frac{1}{2^{\gamma+\frac{3}{2}}}\int_{-1}^1 \left(\frac{\tau+1}{2}\right)^r W_n\left(\frac{\tau+1}{2}\right)(\tau+1)^{\gamma-\frac{1}{2}} d\tau\ .$$

From that it is clear that the polynomials $W_n\left(\frac{x+1}{2}\right)$ are orthogonal in I with respect to the weight function $J_1(x)$. According to the Theorem 1.3, taking into account the uniqueness of these polynomials, we have

$$W_n\left(\frac{x+1}{2}\right) = k\, P_n\left(x; 0, \gamma-\frac{1}{2}\right),$$

where $k > 0$ in consequence of the fact that $P_n\left(\infty; 0, \gamma-\frac{1}{2}\right) > 0$ and $W_n(+\infty) > 0$.

From the orthonormality of the polynomials $W_n(t)$ we derive

$$\frac{1}{2} = \int_0^1 W_n^2(t)\, t^{\gamma-\frac{1}{2}} dt = k^2 \int_{-1}^1 P_n^2\left(\tau; 0, \gamma-\frac{1}{2}\right)\left(\frac{\tau+1}{2}\right)^{\gamma-\frac{1}{2}} \frac{1}{2}\, d\tau =$$

$$= \frac{1}{2^{\gamma+\frac{1}{2}}} k^2 \int_{-1}^1 P_n^2\left(\tau; 0, \gamma-\frac{1}{2}\right)(\tau+1)^{\gamma-\frac{1}{2}} d\tau$$

from where we have $k = 2^{\frac{\gamma}{2}-\frac{1}{4}}$ and the relation (2.1), i.e.

$$Q_{2n}(x) = 2^{\frac{\gamma}{2}-\frac{1}{4}} P_n\left(2t-1; 0, \gamma-\frac{1}{2}\right),\ t = x^2.$$

Now we prove the relation (2.2). Putting $t = x^2$ we have

$$\overline{W}_n(t) = x^{-1} Q_{2n+1}(x),$$

where $\overline{W}_n(t)$ is the polynomial of the degree n and $Q_{2n+1}(x)$ is odd function. For $r = = 0, 1, \ldots, n-1$ and $n > 0$ the orthogonality of the polynomials $\{Q_n(x)\}_{n=0}^{\infty}$ yields

$$0 = \int_0^1 x^{2r+1} Q_{2n+1}(x)\, x^{2\gamma} dx = \frac{1}{2} \int_0^1 t^r\, \overline{W}_n(t)\, t^{\gamma+\frac{1}{2}} dt =$$

$$= \frac{1}{2^2} \int_{-1}^1 \left(\frac{\tau+1}{2}\right)^r \overline{W}_n\left(\frac{\tau+1}{2}\right) \left(\frac{\tau+1}{2}\right)^{\gamma+\frac{1}{2}} d\tau =$$

$$= \frac{1}{2^{\gamma+\frac{5}{2}}} \int_{-1}^1 \left(\frac{\tau+1}{2}\right)^r \left(\frac{\tau+1}{2}\right)^{\frac{1}{2}} \overline{W}_n\left(\frac{\tau+1}{2}\right) (\tau+1)^{\gamma} d\tau\ .$$

From there

$$\left(\frac{x+1}{2}\right)^{\frac{1}{2}} \overline{W}_n\left(\frac{x+1}{2}\right) = \overline{k}\, P_n(x; 0, \gamma)\ ,$$

where $\overline{k} > 0$ and from the orthonormality of the polynomials $t^{\frac{1}{2}}\, \overline{W}_n(t)$ we derive

$$\frac{1}{2} = \int_0^1 x^{-2}\, Q_{2n+1}^2(x)\, x^{2\gamma} dx = \int_0^1 t\, \overline{W}_n^2(t)\, t^{\gamma} dt =$$

$$= \frac{1}{2} \int_{-1}^1 \left(\frac{\tau+1}{2}\right) \overline{W}_n^2 \left(\frac{\tau+1}{2}\right) \left(\frac{\tau+1}{2}\right)^{\gamma} d\tau = \frac{1}{2^{\gamma+1}} \overline{k}^{\,2} \int_{-1}^1 P_n^2(\tau; 0, \gamma)(\tau+1)^{\gamma} d\tau\ .$$

Finally we get $\overline{k} = 2^{\frac{\gamma}{2}}$ and the relation (2.2) of the theorem. □

3. Generalized ultraspherical polynomials and their relation to certain classical Jacobi polynomials. In the next theorem we generalize the relations derived in the Theorem 2.1 for generalized ultraspherical polynomials taking into account polynomials orthonormal in I with respect to the weight function

$$\widetilde{Q}(x) = \left(1-x^2\right)^{\alpha} \left(x^2\right)^{\gamma}$$

instead of the weight function $Q(x) = \left(x^2\right)^{\gamma}$.

THEOREM 3.1. *Let* $\{\widetilde{Q}_n(x)\}_{n=0}^{\infty}$ *be the polynomials orthonormal in* $I = (-1,1)$ *with the weight function*

$$\widetilde{Q}(x) = \left(1-x^2\right)^{\alpha}\left(x^2\right)^{\gamma},$$

where $\alpha > -1, \gamma > 0$ *and* $\widetilde{Q}_n(+\infty) > 0$. *Let* $\{P_n\left(x;\alpha,\gamma-\frac{1}{2}\right)\}_{n=0}^{\infty}$ *be the polynomials orthonormal in* I *with the weight function*

$$\widetilde{J}_1(x) = (1-x)^{\alpha}(1+x)^{\gamma-\frac{1}{2}}$$

and $\{P_n(x;\alpha,\gamma)\}_{n=0}^{\infty}$ *be the polynomials orthonormal in* I *with the weight function*

$$\widetilde{J}_2(x) = (1-x)^{\alpha}(1+x)^{\gamma}.$$

Then

$$\widetilde{Q}_{2n}(x) = 2^{\frac{\alpha+\gamma}{2}-\frac{1}{4}} P_n\left(2x^2-1;\alpha,\gamma-\frac{1}{2}\right) \tag{3.1}$$

and

$$\widetilde{Q}_{2n+1}(x) = 2^{\frac{\alpha+\gamma}{2}} x P_n\left(2x^2-1;\alpha,\gamma\right). \tag{3.2}$$

Proof. Similarly to the proof of the Theorem 2.1 we put the appropriate substitutions to the integrals proving the orthonormality of the polynomials $\{\widetilde{Q}_n(x)\}_{n=0}^{\infty}$. In all the integrals the term $\left(\frac{1-\tau}{2}\right)^{\alpha}$ will appear and after some algebra and integration the term $2^{\frac{\alpha}{2}}$ will appear in the relations (3.1) and (3.2). □

4. Even multiple of the weight function of Jacobi polynomials. The result of the following theorem is the analogy of the well-known relation for classical Jacobi polynomials.

THEOREM 4.1. *Let* $\{Q_n(x;\alpha,\beta,\gamma)\}_{n=0}^{\infty}$ *be the polynomials orthonormal in the interval* $I = (-1,1)$ *with the weight function*

$$Q(x;\alpha,\beta,\gamma) = (1-x)^{\alpha}(1+x)^{\beta}\left(x^2\right)^{\gamma}$$

where $\alpha > -1, \beta > -1, \gamma > 0$. *Then*

$$Q_n(-x;\alpha,\beta,\gamma) = (-1)^n Q_n(x;\beta,\alpha,\gamma). \tag{4.1}$$

Proof. According to the orthogonality criterion (Theorem 1.4) the necessary and sufficient condition of the orthogonality of the polynomials $\{Q_n(x;\alpha,\beta,\gamma)\}_{n=0}^{\infty}$ has the form

$$\int_{-1}^{1}(1-x)^{\alpha}(1+x)^{\beta}\left(x^2\right)^{\gamma} Q_n(x;\alpha,\beta,\gamma)F_m(x)dx = 0,$$

where $F_m(x)$ is an arbitrary polynomial of the degree $m = 0,1,\ldots,n-1$. Substituting $x = -t$ this condition will obtain the form

$$\int_{-1}^{1}(1+t)^{\alpha}(1-t)^{\beta}(t^2)^{\gamma} Q_n(-t;\alpha,\beta,\gamma)F_m(-t)dt = 0.$$

Because $F_m(t)$ is an arbitrary polynomial of the degree m, then also $F_m(-t)$ is an arbitrary polynomial of the degree m. So, in the consequence of the same theorem, the polynomial $Q_n(-t;\alpha,\beta,\gamma)$ is also orthogonal, but with the weight $Q(t;\beta,\alpha,\gamma)$ and it may differ from the orthogonal polynomial $Q_n(t;\beta,\alpha,\gamma)$ only by constant multiple. So

$$Q_n(-t;\alpha,\beta,\gamma) \equiv c\, Q_n(t;\beta,\alpha,\gamma).$$

Because the polynomials are orthonormal, it yields $|c| = 1$. Comparing the coefficients at the highest powers of these two polynomials, we get $|c| = (-1)^n$ and the relation (4.1). □

It is obvious that the result of this theorem can be generalized for polynomials orthogonal in I with the weight function $J(x)h(x)$, where the factor $h(x)$ is an even function on I.

THEOREM 4.2. *Let $\{\widetilde{P}_n(x;\alpha,\beta)\}_{n=0}^{\infty}$ be the polynomials orthonormal in the interval $I = (-1,1)$ with the weight function*

$$\widetilde{J}(x;\alpha,\beta) = (1-x)^{\alpha}(1+x)^{\beta}h(x),$$

where $\alpha > -1, \beta > -1, h(x) \geq 0$ in I and $h(x)$ is an even function in I. Then

$$\widetilde{P}_n(x;\alpha,\beta) = (-1)^n \widetilde{P}_n(x;\beta,\alpha).$$

Proof. Similar to the proof of the Theorem 4.1. □

5. Consequences of differential equations with generalized Legendre polynomials solutions. Differentiating both sides of (2.1) according to x, then expressing $P_n\left(2x^2-1;0,\gamma-\frac{1}{2}\right)$, $P_n'\left(2x^2-1;0,\gamma-\frac{1}{2}\right)$, and $P_n''\left(2x^2-1;0,\gamma-\frac{1}{2}\right)$ by means of polynomials $Q_{2n}(x)$, $Q_{2n}'(x)$, and $Q_{2n}''(x)$, then substituting the derivatives P_n, P_n', and P_n'' into the differential equation with these Jacobi polynomials solutions, we have the following equation:

$$(1-x^2)Q_{2n}''(x) - \frac{-2\gamma + 2x^2 + 2\gamma x^2}{x}Q_{2n}'(x) = -2n(2n+2\gamma+1)Q_{2n}(x).$$

For $\gamma = 0$ it reduces to the equation

$$(1-x^2)Q_{2n}''(x) - 2xQ_{2n}'(x) = -2n(2n+1)Q_{2n}(x).$$

Comparing it with (1.2) we observe the last equation to be the differential equation for the $(2n)$-th degree Legendre polynomial.

By the similar way from (2.2) we derive the following equation:

$$(1-x^2)Q_{2n+1}''(x) + \frac{-1+2\gamma-x^2-2\gamma x^2}{x}Q_{2n+1}'(x) + \frac{1-2\gamma+x^2+2\gamma x^2}{x^2}Q_{2n+1}(x) =$$

$$= -4n(n+\gamma+1)Q_{2n+1}(x).$$

For $\gamma = \frac{1}{2}$ it reduces to the equation

$$(1-x^2)Q_{2n+1}''(x) - 2xQ_{2n+1}'(x) + 2Q_{2n+1}(x) = -4n\left(n+\frac{3}{2}\right)Q_{2n+1}(x).$$

The last equation is the differential equation for the $(2n+1)$-st degree Legendre polynomial.

In such a way we have proved the following theorem:

THEOREM 5.1.

1. *The Jacobi polynomial* $P_n\left(2x^2-1;0,-\frac{1}{2}\right)$ *of the argument* $2x^2-1$ *orthogonal with respect to the weight function* $(1+x)^{-\frac{1}{2}}$ *is the Legendre polynomial of the argument* x *and of the degree* $2n$.

2. *The polynomial* $xP_n\left(2x^2-1;0,\frac{1}{2}\right)$, *where* $P_n\left(2x^2-1;0,\frac{1}{2}\right)$ *is the Jacobi polynomial of the argument* $2x^2-1$, *orthogonal with respect to the weight* $(1+x)^{\frac{1}{2}}$, *is the Legendre polynomial of the argument* x *and of the degree* $2n+1$.

REFERENCES

[1] L. C. ANDREWS, *Special Functions of Mathematics for Engineers*, Second ed., McGraw-Hill, Inc., 1992.

[2] B. HOFMAN-WELLENHOF AND H. MORITZ, *Physical Geodesy*, Second ed., Springer, 2006.

[3] D. JANECKI AND K. STEPIEN, *Legendre polynomials used for the approximation of cylindrical surfaces*, Communications - Scientific Letters of the University of Žilina, 4 (2005), 59–61.

[4] J. KOROUS, *Special Parts of Mathematics. Orthogonal Functions and Orthogonal Polynomials*, SNTL, Prague, 1958 (in Czech).

[5] P. W. LIVERMORE, CH. A. JONES, AND S. J. WORLAND, *Spectral radial basis functions for full sphere computations*, J. Comput. Phys., 227 (2007), 1209–1224.

[6] M. MARČOKOVÁ AND V. GULDAN, *Jacobi Polynomials and Some Related Functions*, in Mathematical Methods in Engineering, N. M. F. Ferreira and J. A. T. Machado, eds., Springer, 2014, pp. 219–227.

[7] S. P. MIREVSKI, L. BOYADIJEV, AND R. SCHERER, *On the Riemann-Liouville fractional calculus, g-Jacobi functions and F-Gauss functions*, Appl. Math. Comput., 187(1) (2007), 315–325.

[8] P. K. SUJETIN, *Classical orthogonal polynomials*, Nauka, Moskva, 1979 (in Russian).

[9] G. SZEGÖ, *Orthogonal polynomials*, Nauka, Moskva, 1969 (in Russian).

[10] R. TENZER, *Geopotential model of Earth - approximation of Earth shape, geopotential model testing methods*, Communications - Scientific Letters of the University of Žilina, 4 (2001), 50–58.

[11] R. TENZER, *Methodology for testing of geopotential model*, Studies of University in Žilina, Civil Engineering series, 25 (2002), 113–116.

Proceedings of EQUADIFF 2017
pp. 295–304

STOCHASTIC MODULATION EQUATIONS ON UNBOUNDED DOMAINS*

LUIGI A. BIANCHI† AND DIRK BLÖMKER‡

Abstract. We study the impact of small additive space-time white noise on nonlinear stochastic partial differential equations (SPDEs) on unbounded domains close to a bifurcation, where an infinite band of eigenvalues changes stability due to the unboundedness of the underlying domain. Thus we expect not only a slow motion in time, but also a slow spatial modulation of the dominant modes, and we rely on the approximation via modulation or amplitude equations, which acts as a replacement for the lack of random invariant manifolds on extended domains.

One technical problem for establishing error estimates in the stochastic case rises from the spatially translation invariant nature of space-time white noise on unbounded domains, which implies that at any time the error is always very large somewhere far out in space. Thus we have to work in weighted spaces that allow for growth at infinity.

As a first example we study the stochastic one-dimensional Swift-Hohenberg equation on the whole real line [1, 2]. In this setting, because of the weak regularity of solutions, the standard methods for deterministic modulation equations fail, and we need to develop new tools to treat the approximation. Using energy estimates we are only able to show that solutions of the Ginzburg-Landau equation are Hölder continuous in spaces with a very weak weight, which provides just enough regularity to proceed with the error estimates.

Key words. modulation equations, amplitude equations, convolution operator, regularity, Rayleigh-Bénard, Swift-Hohenberg, Ginzburg-Landau

AMS subject classifications. 60H15,60H10

1. Experiments.

A celebrated model in pattern formation is the Rayleigh-Bénard convection, an experimental phenomenon where a fluid between two plates is heated from below and kept at a constant temperature from above. Here the full description would be a 3D-Navier-Stokes equation coupled to the heat equation, a mathematical model that is yet too complicated for our analytical tools. In this article we review the results of [1] and [2] and thus we consider the simpler Swift-Hohenberg model [8] that is used as a reduced model for the convective instability.

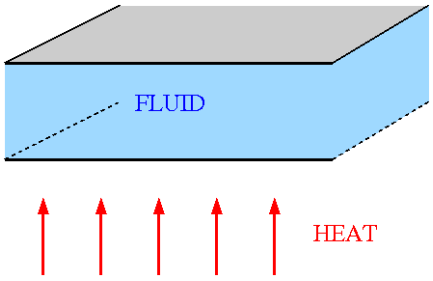


FIGURE 1.1. *Rayleigh-Bénard convection*

*This work was supported by Grant No.: DFG BL535-9/2.
† FB Mathematik und Statistik, Universität Konstanz, 78457 Konstanz
luigi-amedeo.bianchi@uni-konstanz.de
‡Institut für Mathematik, Universität Augsburg, 86135 Augsburg
dirk.bloemker@math.uni-augsburg.de

1.1. Convective instability. The convective instability is the first bifurcation in the Rayleigh-Bénard problem. Below a critical temperature T_c the fluid is at rest and no pattern is formed. The heat is just transported by conduction through the system.

Above the critical temperature T_c convection rolls start to form. Hot fluid is going up and cold fluid is going down, and they cannot do that in the same place, so we have areas where the motion is upwards and other areas where it is downwards. In a view from above, a striped pattern starts to show up.

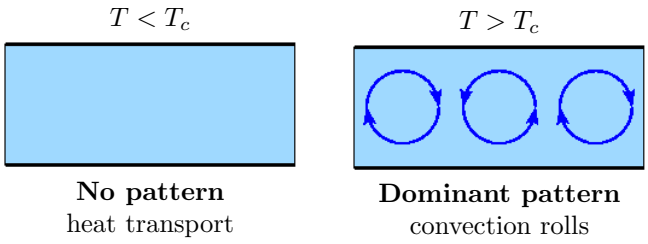


FIGURE 1.2. *Bifurcation at the convective instability, the figure shows a cut through the fluid with the plates above and below.*

1.2. Pattern formation below criticality. Very close to the critical point, stochastic effects were observed first in electro-convection (see Rehberg et al. [18]) and much later in Rayleigh-Bénard convection (see Oh, Ahlers et al. [16, 17]). In both experiments, pattern formation slightly below the critical threshold (i.e., a critical temperature T_c in Rayleigh Bénard) was observed. Nevertheless the distance to bifurcation had to be of the order of the noise's strength, which made it extremely difficult to observe in experiments, as the source of noise in Rayleigh-Bénard are thermal fluctuations. Similar observations in numerical experiments and using formal center manifold approximations were done by Hutt et al. [10, 9].

OBSERVATION FROM EXPERIMENTS, [18]:

Below threshold (but close)	**Well above threshold**
trivial solution is not stable	convection rolls are stable
pattern is slowly modulated	pattern is almost periodic

2. Introduction. The typical setting in the following presentation of our results shows complicated systems given for example by (stochastic) partial differential equations. Near a change of stability (or bifurcation) of the trivial solution, we have a natural separation of time-scales. The (Fourier) modes similar to the bifurcating pattern move on a slow time-scale given by the distance from bifurcation, while the other modes move and disappear on an order one time-scale.

The typical results we are aiming at are the approximation of the full dynamics by means of the *amplitude* of the bifurcating pattern, which is given by a (stochastic) differential equation. On unbounded domains a full band of eigenfunctions changes stability. In order to take this into account the amplitude of the dominating pattern is slowly *modulated* in space.

This approximation by modulation (or amplitude) equations is well established in the physics literature, but only on a formal level. From a mathematical point of view, the deterministic problems are well studied. Starting from the first publications [4, 11, 14, 13] there is a rich literature, featuring also recent contributions, for example [20, 6], just to name two.

Let us point out that the celebrated center manifold reduction, which works well for deterministic PDEs on bounded domains, is not available for PDEs on unbounded domains. Moreover, it is not useful in the stochastic setting: because of the inherent non-autonomy of the system due to noise, the manifold itself would move through the whole phase space, and thus any reduction to the manifold does not reduce the complexity of the dynamics at all.

We finally give an outline of the paper. In Section 3 we state the setting of the Swift-Hohenberg equation and discuss the spectrum of the linearized operator together with modulated pattern. We then briefly recall the main results on large domains in Section 4, while in Section 5 we state in detail the results available on unbounded domains. In the final two sections we give a remark on pattern formation below criticality and provide an outlook on several possible extensions of the result.

3. Swift-Hohenberg. For our results we consider for simplicity only a toy problem given by the Swift-Hohenberg equation. It can be derived via heuristic reduction from the Rayleigh-Bénard problem close to the convective instability, as was originally shown by Swift & Hohenberg [8]. See also [19] for a more rigorous approach. The equation is given as:

$$\partial_t u = -(1+\partial_x^2)^2 u + \nu\varepsilon^2 u - u^3 + \varepsilon^{3/2}\xi \,, \tag{SH}$$

where we assume

- $u(t,x) \in \mathbb{R}, \quad t > 0,\ x \in \mathbb{R}$
- periodic boundary conditions – or – unbounded domain
- $\xi = \partial_t W$ Gaussian space-time white noise.

Thus in the sense of generalized processes the mean of the noise is zero and it is uncorrelated in space and time:

$$\mathbb{E}\,\xi(t,x) = 0 \,, \qquad \mathbb{E}\,\xi(t,x)\xi(s,y) = \delta(t-s)\delta(x-y) \,.$$

As a mathematical model, the noise is given as a derivative of a standard cylindrical Wiener process $\{W(t)\}_{t\geq 0}$ in $L^2(\mathbb{R})$, meaning that

$$W(t) = \sum_k \beta_k(t) e_k$$

where $\{e_k\}_k$ is any orthonormal basis in $L^2(\mathbb{R})$ and $\{\beta_k\}_k$ is a sequence of i.i.d. real-valued Brownian motions.

3.1. Eigenvalues – Spectral gap. In our example of the Swift-Hohenberg operator we can calculate all eigenvalues of the linearized operator explicitly:

$$\mathcal{L} = -(1+\partial_x^2)^2 \quad \text{and thus} \quad \mathcal{L}\mathrm{e}^{ikx} \;=\; \lambda(k)\,\mathrm{e}^{ikx} \,,$$

subject to periodic boundary conditions on an interval or on the whole real line. Obviously,

$$\lambda(k) = -(1-k^2)^2 \,.$$

In Figure 3.1 we plotted the eigenvalues of $\mathcal{L}$ for those $k \in \mathbb{R}$ which lead to admissible eigenfunctions that satisfy the boundary conditions. We see that the spectral gap between the largest two eigenvalues shrinks as the domain gets larger: on an interval of length $\mathcal{O}(\varepsilon^{-1})$ already many eigenvalues are $\mathcal{O}(\varepsilon^2)$ away from the largest eigenvalue 0.

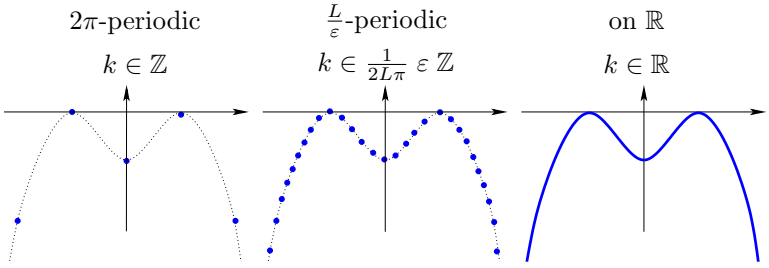


FIGURE 3.1. *Band of Eigenvalues for the example* $\mathcal{L} = -(1+\partial_x^2)^2$ *on bounded, large, and unbounded domains. We plot the wave-number* k *against the eigenvalue* $\lambda(k) = -(1-k^2)^2$ *for the corresponding eigenfunction* $\mathrm{e}^{\mathrm{i}kx}$.

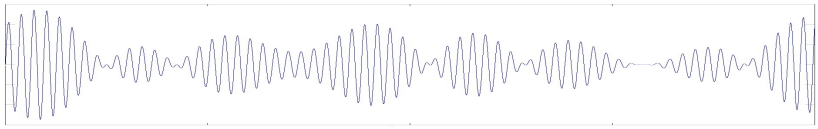
3.2. Modulated pattern. As many eigenvalues are close to the change of stability, we need to understand how many eigenfunctions with wave-number around $k = \pm 1$ influence the pattern.

Let us compare a 2π-periodic pattern

$$u(x) = \varepsilon A \mathrm{e}^{ix} + \text{c.c.} \quad \text{with} \quad A \in \mathbb{C}$$

with a modulated pattern

$$u(x) = \varepsilon A(\varepsilon x)\mathrm{e}^{ix} + \text{c.c.} \quad \text{with} \quad A : \mathbb{R} \mapsto \mathbb{C}$$



If we consider the amplitude A in polar coordinates, then its absolute value $|A|$ determines the size of the modulated pattern, while the angle is a phase shift of the pattern. Both move slowly in space here.

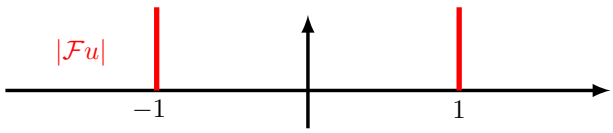
We can calculate that

$$u(x) = \varepsilon A(\varepsilon x)\mathrm{e}^{ix} + \text{c.c.}$$

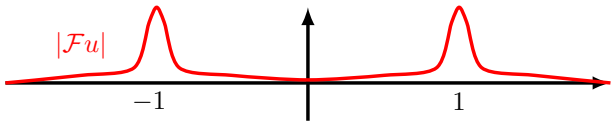
has Fourier transform

$$\mathcal{F}u(k) = \mathcal{F}A\,(\,(k-1)/\varepsilon\,) + \overline{\mathcal{F}A}\,(\,(k+1)/\varepsilon\,)\ .$$

For 2π-periodic pattern the function is in the span of e^{ix} and e^{-ix}. Thus the Fourier transform is only a Dirac at wave-numbers $k \in \{-1, 1\}$.



For the slow modulation of a 2π-periodic pattern, the Fourier transform widens up but it is still concentrated around $k \in \{-1, 1\}$. A whole band of infinitely many Fourier modes defines the structure of the solution.



4. Large domains. Here we present the results of Blömker, Hairer & Pavliotis [3] without stating the technical details.

THEOREM 4.1 (Approximation [3]). *Consider a $2L/\varepsilon$-periodic solution u of (SH) If $u(0,x) = \varepsilon A(0,\varepsilon x)\cdot \mathrm{e}^{ix} + c.c. + \mathcal{O}(\varepsilon^2)$ is a modulated wave with admissible initial condition $A(0,\cdot) = \mathcal{O}(1)$, then*

$$\forall t \in [0, T_0\varepsilon^{-2}] \qquad u(t,x) = \varepsilon A(\varepsilon^2 t, \varepsilon x)\cdot \mathrm{e}^{ix} + c.c. + \mathcal{O}\big(\varepsilon^{2-}\big)\ ,$$

where the amplitude $A(T,X) \in \mathbb{C}$ solves (GL).

The amplitude equation is a stochastic Ginzburg-Landau equation:

$$\partial_T A = (4\partial_X^2 + \nu)A - 3|A|^2 A + \eta \tag{GL}$$

with

- $2L$-periodic solutions
- $\mathbb{C}$-valued space-time white noise $\eta = \partial_T \mathcal{W}$

The complex-valued standard cylindrical Wiener-process $\mathcal{W}$ arises from rescaling the discrete Fourier transform of the real-valued Wiener process W for Fourier-modes with wave-number k close to 1. See also Section 5.2.

Let us remark that even in the case of ξ in (SH) colored and regular in space, the amplitude equation (GL) has space-time white noise, due to rescaling in space and time.

The estimates in Theorem 4.1 above are given in C^0-norms and the initial condition $A(0)$ is called *admissible* if it splits into a more regular H^1-part, and a Gaussian part, which we can bound in C^0. This is a quite natural assumption for SPDEs using the standard transformation with the stochastic convolution.

5. Unbounded domains. The key technical problem for deriving an approximation result via amplitude equations for (SH) on unbounded domains is the regularity of solutions. All previous results require too much regularity that we do not have in the stochastic setting. The theory for deterministic PDEs always uses uniform bounds in space on derivatives of the amplitude A. While the pioneering works [4, 11], which needed a uniform bound on the fourth derivative, were much improved since then, all results still need a uniform bound.

The previously stated Theorem 4.1 on large but still bounded domains needs a split condition in space for a more regular H^1-part and a Gaussian part only in C^0. Nevertheless, solutions are always uniformly bounded in space.

In two papers Klepel, Mohammed & Blömker [15, 12] discussed the case of spatially constant noise. Also in this setting they need too much regularity, as the solution of the amplitude equation (GL) has to be $H^{1/2+}$ in space and thus it is uniformly bounded.

We formulate the regularity that we expect for the amplitude A as a theorem:

THEOREM 5.1 (Lack of regularity). *With space-time white noise on the whole real line and with sufficiently smooth initial conditions the amplitude A solving (GL) is*

- *γ-Hölder-continuous in space and time only with $\gamma < 1/2$,*

- *unbounded in space, i.e.* $\|A(T,\cdot)\|_\infty = \infty$ *for all* $T > 0$.

To address the lack of regularity we can on one hand consider mild solutions, that take care of the problems with differentiability. On the other hand, we need weighted Hölder spaces which are defined by the norm (for some small $\kappa > 0$)

$$\|u\|_{C^{0,\alpha}_\kappa} = \sup_{L>1}\{L^{-\kappa}\|u\|_{C^{0,\alpha}([-L,L])}\} .$$

5.1. Mild formulation. Recall the Swift-Hohenberg equation:

$$\partial_t u = \underbrace{\mathcal{L}u + \nu\varepsilon^2 u}_{=:\mathcal{L}_\nu u} - u^3 + \varepsilon^{3/2}\partial_t W \tag{SH}$$

Its *mild solution* (see [5]), also called variation of constants formula, is

$$u(t) = \mathrm{e}^{t\mathcal{L}_\nu}u(0) - \int_0^t \mathrm{e}^{(t-s)\mathcal{L}_\nu}u^3(s)\,ds + \varepsilon^{3/2}W_{\mathcal{L}_\nu}(t)$$

with the *stochastic convolution* given by

$$W_{\mathcal{L}_\nu}(t) = \int_0^t \mathrm{e}^{(t-s)\mathcal{L}_\nu}dW(s) .$$

Remark 1. *Results for existence and uniqueness of mild solutions are usually straightforward using fixed-point theorems. Unfortunately this is not the case in the weighted spaces we are considering. The nonlinearity is unbounded and the semigroup is only regularizing in terms of differentiability but not in terms of weights. Thus the right-hand-side of the fixed-point equation is not a self-mapping.*

So the existence and uniqueness is first established for weak solutions via an approximation with large but bounded domains, and then one can show that weak solutions are sufficiently regular to be also mild. We will go not into details here, for those see [2].

5.2. Results for the linearized equation. This is the key stochastic result from Bianchi & Blömker [1]. It is one of the essential building blocks to prove a result for the residuum of the nonlinear equation.

Theorem 5.2 (Approximation). *Given the Wiener process* W *from (SH), there is a complex-valued Wiener process* $\mathcal{W}$ *for (GL) such that for any* $\kappa > 0$ *with probability almost* 1

$$\sup_{[0,\frac{T_0}{\varepsilon^2}]}\left\|\varepsilon^{\frac{3}{2}}W_{\mathcal{L}_\nu}(t,x) - \left[\varepsilon\mathcal{W}_{4\partial_x^2+\nu}(\varepsilon^2 t,\varepsilon x)\cdot \mathrm{e}^{ix} + c.c.\right]\right\|_{C^0_\kappa} \le C\varepsilon^{\frac{3}{2}-}$$

Definition 5.3. *We say that an* ε*-dependent event* $\mathcal{A}_\varepsilon$ *has* probability almost 1*, if for all* $p \ge 1$ *there is a constant* $C_p > 0$ *such that* $\mathbb{P}(\mathcal{A}_\varepsilon) \ge 1 - C_p\varepsilon^p$.

Let us remark that, in order to control the cubic term in the nonlinear result afterwards, we use the weighted supremum norm and not weaker (and actually much simpler) weighted L^2-norm.

Proof. We provide here only a brief sketch of the proof, for all the technical details see [1]. We rescale the stochastic convolution to the slow time ($T = \varepsilon^2 t$) and large space ($X = \varepsilon x$). Then we split into Fourier-modes larger than 0 and the

complex conjugate corresponding to Fourier-modes smaller than 0. This defines the complex valued Wiener process, as there is one canonical process $\mathcal{W}$ such that we can summarize the difference as a single stochastic integral w.r.t. $\mathcal{W}$:

$$\varepsilon^{1/2} W_{\mathcal{L}}(T\varepsilon^{-2}, X\varepsilon^{-1}) - [\mathcal{W}_{4\partial_x^2}(T,X)\cdot \mathrm{e}^{iX/\varepsilon} + \text{c.c.}] = \int_0^T \mathcal{H}_\tau d\mathcal{W}(\tau)\cdot \mathrm{e}^{iX/\varepsilon} + \text{c.c.}$$

with a convolution operator $\mathcal{H}_\tau u = H_\tau \star u$ that mainly contains rescaled differences of the semigroups.

We use a technical estimate that allows to bound $\int_0^T \mathcal{H}_\tau d\mathcal{W}(\tau)$ in weighted Hölder spaces with small exponent and small weight in terms of bounds on the Fourier-transform $\hat{H}_\tau$ in spaces with slightly more regularity than $L^2([0,T_0]\times\mathbb{R})$.

The remaining and lengthy part of the proof shows the bounds for the norm of $\hat{H}_\tau$ in different areas of the Fourier-space. □

5.3. Nonlinear result. The full nonlinear result for (SH) and (GL) was treated in Bianchi, Blömker & Schneider [2]. It contains of two steps: first we bound the residual of the Swift-Hohenberg equation, and then via standard energy-type estimates we establish the approximation result.

5.3.1. Residual. Let A be a solution of (GL) with some conditions on $A(0,\cdot)$. It basically has to be in any $W^{1,p}_\rho$, $p>1$ for an integrable weight ρ.

Definition 5.4 (Approximation). *For A from above, we define the approximation*

$$u_A(t,x) = \varepsilon A(\varepsilon^2 t, \varepsilon x)\mathrm{e}^{ix} + c.c.$$

The key step towards an approximation result is to bound the residual for u_A.

Definition 5.5 (Residual). *For u_A from above we define*

$$Res(t) = u(t) - \mathrm{e}^{t\mathcal{L}_\nu} u_A(0) + \int_0^t \mathrm{e}^{(t-s)\mathcal{L}_\nu} u_A^3(s)\, ds - \varepsilon^{3/2} W_{\mathcal{L}_\nu}(t)$$

We can prove the following result:

Theorem 5.6 (Residual). *For every small $\kappa>0$ with probability almost* 1

$$\sup_{[0,T_0\varepsilon^{-2}]} \| Res \|_{C^0_\kappa} \le C\varepsilon^{3/2-}.$$

The proof can be found in [2, Theorem 5.9]. Its main strategy is as follows:

- use suitable exchange Lemmas to replace Swift-Hohenberg semigroups by Ginzburg-Landau semigroups,
- take advantage of Theorem 5.2 for stochastic convolution $W_{\mathcal{L}_\nu}$,
- notice that all terms of order $\mathcal{O}(\varepsilon)$ cancel due to (GL).

The key problem is that for the exchange Lemmas some regularity (or Gaussianity) is needed to estimate:

$$\mathrm{e}^{t\mathcal{L}}[D(\varepsilon x)\mathrm{e}^{ix}] \approx [\mathrm{e}^{4T\partial_X^2} D](\varepsilon x)\cdot \mathrm{e}^{ix} \quad \text{and} \quad \mathrm{e}^{t\mathcal{L}}[D(\varepsilon x)\mathrm{e}^{3ix}] \approx 0$$

If D is very smooth the proofs are straightforward, but here $D \in \{A^3, A|A|^2\}$ thus we only have Hölder-regularity.

5.3.2. Approximation. For a solution u of (SH) and the approximation u_A we define

$$R = u - u_A - Res$$

which solves

$$\partial_t R = \mathcal{L}_\nu R - (R + u_A + Res)^3 - {u_A}^3.$$

Use standard energy estimates in a weighted L^2-norm (for $\rho > 1$)

$$\|R\|^2_{L^2_{\rho,\varepsilon}} = \int_{\mathbb{R}} (1 + \varepsilon^2 x^2)^{-\rho/2} |R(x)|^2 dx$$

we obtain the following result.

THEOREM 5.7. *With probability almost* 1

$$\sup_{[0,T_0\varepsilon^{-2}]} \|u - u_A\|_{L^2_{\rho,\varepsilon}} \leq C\|u(0) - u_A(0)\|_{L^2_{\rho,\varepsilon}} + C\varepsilon^{1-}.$$

Details of the proof can be found in [2, Theorem 6.3].

6. A comment on pattern formation below criticality. Using amplitude equations, the question of pattern formation has a simple answer. Let us consider (SH) below the bifurcation, but sufficiently close. To be more precise, if σ is the noise-strength, then the distance from bifurcation should be $\mathcal{O}(\sigma^{4/3})$. In such scaling the effective dynamic is described by the amplitude equation, which is independent of σ. Thus the amplitude A is always $\mathcal{O}(1)$ and hence the pattern is visible.

7. Outlook. Let us conclude by commenting on some possible extensions of the results above.

7.1. Other types of noise in (SH). In the result presented here we only treat space-time white noise in both equations. But we could try more regular noise to overcome regularity barriers.

For colored, spatially smooth and translation invariant noise, it seems straightforward that in the approximation result (GL) still has space-time white noise, due to the rescaling both in space and time. Thus it does not help with the regularity.

If we consider trace class noise in $L^2(\mathbb{R})$ then we impose a decay-condition at infinity for (SH). But in that case, due to the spatial rescaling, we expect point-forcing in (GL).

In order to have noise that does not change under the space-time rescaling, one could try to consider algebraic decay of correlations. Here we expect a similar algebraic decay of correlations also for the noise in (GL). However, these types of noise seem to yield poor regularity of solutions, too.

7.2. Quadratic non-linearities. A more accurate Swift-Hohenberg model of the real Rayleigh-Bénard convection has a quadratic nonlinearity. In that setting the analysis is much more involved, as one has much more complicated interaction of Fourier-modes. But it is known from the deterministic results that even in the Rayleigh-Bénard phenomenon the amplitude equation is of Ginzburg-Landau type. Consequently, we expect a similar result to hold in the stochastic case, too.

7.3. Higher-dimensional models. Considering higher dimensional models is a difficult problem, as already in 2D the Ginzburg-Landau equation is no longer well-defined. See Hairer, Ryser & Weber [7] for a result on Allen-Cahn, which should generalize to (GL).

Consider for example Swift-Hohenberg in $\mathbb{R}^2$

$$\partial_t u = -(1+\Delta)^2 u + 4\partial_y^2 u + \nu\varepsilon^2 u - u^3 + \varepsilon\partial_t W \tag{2D-SH}$$

subject to space-time white noise or even smoother spatially colored noise. This formally has the amplitude equation

$$\partial_T A = -4\Delta A + \nu A - 3A|A|^2 + \partial_T \mathcal{W} \tag{2D-GL}$$

also with space-time white noise, which is no longer well-defined, as noted in the aforementioned [7]. Nevertheless, results like these are used in the applied literature.

Here in the spirit of [7], we can consider a smaller strength of the noise to obtain a meaningful limit. In that case the amplitude equation has no longer an additive noise, but additional deterministic terms should appear due to the presence of noise in (SH-2D) and averaging effects in the nonlinearity.

REFERENCES

[1] L. A. Bianchi and D. Blömker. Modulation equation for SPDEs in unbounded domains with space–time white noise—linear theory. *Stochastic Processes and their Applications*, 126(10):3171–3201, (2016).

[2] L. A. Bianchi, D. Blömker, and G. Schneider. Modulation equation and SPDEs on unbounded domains. Preprint, arXiv, (2017).

[3] D. Blömker, M. Hairer, and G. A. Pavliotis. Modulation equations: Stochastic bifurcation in large domains. *Commun. Math. Physics.*, 258(2):479–512, (2005).

[4] P. Collet and J.-P. Eckmann. The time dependent amplitude equation for the Swift-Hohenberg problem. *Comm. Math. Phys.*, 132(1):139–153, (1990).

[5] G. Da Prato and J. Zabczyk. Stochastic equations in infinite dimensions, 2nd Edition, vol. 152 of Encyclopedia of of Mathematics and its Applications. Cambridge University Press, Cambridge, 2014.

[6] W.-P. Düll, K. S. Kashani, G. Schneider, and D. Zimmermann. Attractivity of the Ginzburg-Landau mode distribution for a pattern forming system with marginally stable long modes. *J. Differ. Equations*, 261(1):319–339, (2016).

[7] M. Hairer, M. D. Ryser, and H. Weber. Triviality of the 2D stochastic Allen-Cahn equation. *Electron. J. Probab.* 17, Paper No. 39, 14 p. (2012).

[8] P. C. Hohenberg and J. B. Swift. Effects of additive noise at the onset of Rayleigh-Bénard convection. *Physical Review A,* 46:4773–4785, (1992).

[9] A. Hutt. Additive noise may change the stability of nonlinear systems. *Europhys. Lett.* 84, 34003:1–4, (2008).

[10] A. Hutt, A. Longtin, and L. Schimansky-Geier. Additive global noise delays Turing bifurcations. *Physical Review Letters*, 98, 230601, (2007).

[11] P. Kirrmann, G. Schneider, and A. Mielke. The validity of modulation equations for extended systems with cubic nonlinearities. *Proc. R. Soc. Edinb., Sect. A* 122(1-2):85–91, (1992).

[12] K. Klepel, D. Blömker, and W. W. Mohammed. Amplitude equation for the generalized Swift Hohenberg equation with noise. *Z. Angew. Math. Phys.* 65(6):1107–1126, (2014).

[13] I. Melbourne. Derivation of the time-dependent Ginzburg-Landau equation on the line. *J. Nonlinear Sci.*, 8(1):1–15, (1998).

[14] A. Mielke and G. Schneider. Attractors for modulation equations on unbounded domains – existence and comparison. *Nonlinearity*, 8(5):743–768, (1995).

[15] W. W. Mohammed, D. Blömker, and K. Klepel. Modulation equation for stochastic Swift–Hohenberg equation. *SIAM Journal on Mathematical Analysis*, 45(1):14–30, (2013).

[16] J. Oh, G. Ahlers. Thermal-Noise Effect on the Transition to Rayleigh-Bénard Convection, *Phys. Rev. Lett.* 91, 094501, (2003).

[17] J. Oh, J. Ortiz de Zarate, J. Sengers, G. Ahlers. Dynamics of fluctuations in a fluid below the onset of Rayleigh-Benard convection, *Phys. Rev.* E 69 , 021106, (2004).

[18] I. Rehberg, S. Rasenat, J. M. de la Torre, H. R. Brand. Thermally induced hydrodynamic fluctuations below the onset of electroconvection *Physical Review Letters* 67(5):596–599, (1991)

[19] A. Roberts. Planform evolution in convection–an embedded centre manifold. *J. Austral. Math. Soc.* B., 34(2), 174–198, (1992).

[20] G. Schneider and H. Uecker. The amplitude equations for the first instability of electroconvection in nematic liquid crystals in the case of two unbounded space directions. *Nonlinearity*, 20(6):1361–1386, (2007).

Proceedings of EQUADIFF 2017
pp. 305–314

AN EFFICIENT LINEAR NUMERICAL SCHEME FOR THE STEFAN PROBLEM, THE POROUS MEDIUM EQUATION AND NONLINEAR CROSS-DIFFUSION SYSTEMS

MOTLATSI MOLATI* AND HIDEKI MURAKAWA†

Abstract. This paper deals with nonlinear diffusion problems which include the Stefan problem, the porous medium equation and cross-diffusion systems. We provide a linear scheme for these nonlinear diffusion problems. The proposed numerical scheme has many advantages. Namely, the implementation is very easy and the ensuing linear algebraic systems are symmetric, which show low computational cost. Moreover, this scheme has the accuracy comparable to that of the well-studied nonlinear schemes and make it possible to realize the much faster computation rather than the nonlinear schemes with the same level of accuracy. In this paper, numerical experiments are carried out to demonstrate efficiency of the proposed scheme.

Key words. Stefan problem, Porous medium equation, Cross-diffusion system, Degenerate convection-reaction-diffusion equation, Linear scheme, Error estimate, Numerical method

AMS subject classifications. 35K55, 65M12, 80A22, 92D25

1. Introduction. In this paper, we propose an efficient linear scheme for the following nonlinear diffusion problem: Find $\boldsymbol{z} = (z_1, \dots, z_M) : \overline{\Omega} \times [0, T) \to \mathbb{R}^M$ $(M \in \mathbb{N})$ such that

$$\begin{cases} \dfrac{\partial \boldsymbol{z}}{\partial t} = \Delta \boldsymbol{\beta}(\boldsymbol{z}) + \boldsymbol{f}(\boldsymbol{z}) & \text{in} \quad Q := \Omega \times (0, T), \\ \boldsymbol{\beta}(\boldsymbol{z}) = \boldsymbol{0} & \text{on} \quad \partial\Omega \times (0, T), \\ \boldsymbol{z}(\cdot, 0) = \boldsymbol{z}^0 & \text{in} \quad \Omega. \end{cases} \tag{1.1}$$

Here, $\Omega \subset \mathbb{R}^d$ $(d \in \mathbb{N})$ is a bounded domain with smooth boundary $\partial\Omega$, T is a positive constant, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)$, $\boldsymbol{f} = (f_1, \dots, f_M)$: $\mathbb{R}^M \to \mathbb{R}^M$ and $\boldsymbol{z}^0 = (z_1^0, \dots, z_M^0) \in L^2(\Omega)^M$ are given functions. Let $(\beta_i)_j$ denote the derivative of the ith component of $\boldsymbol{\beta}$ with respect to the jth variable. If there is a point s where $(\beta_i)_i(s) = 0$ for some i, then the diffusion vanishes at that point. In this case, (1.1) is called a degenerate parabolic system. This type of problem with $M = 1$ includes the Stefan problem and the porous medium equation, and such problems have been widely studied for a long time. In Problem (1.1), the diffusivity β_i of the ith component depends not only on the ith variable but also on the jth $(j \neq i)$ variables in general. This mixture of diffusion terms is called cross-diffusion. This type of problems appears in many fields of applications. A typical example is called the Shigesada-Kawasaki-Teramoto cross-diffusion system [9].

In this paper, we propose an efficient numerical scheme to approximate the solutions of Problem (1.1). Our scheme has many advantages, e.g., it is very easy-to-implement and stable, computational costs are low, the discretization matrices are symmetric, and the accuracy is comparable to that of the widely studied nonlinear

*Department of Mathematics and Computer Science, National University of Lesotho, P.O. Roma 180, Lesotho (m.molati@nul.ls).

†Faculty of Mathematics, Kyushu University, 744 Motooka, Nishiku, Fukuoka 819-0395, Japan(murakawa@math.kyushu-u.ac.jp) (Corresponding author).

schemes. The contents of this paper are as follows. In the next section, we give a brief introduction of numerical schemes for (1.1) which are covered in the literature. In Section 3, we propose an efficient linear scheme, and give a brief summary of theoretical results. In Section 4, the numerical experiments are carried out. The numerical results illustrate the efficiency of the proposed scheme. Concluding remarks are made in the final section of the paper.

2. Numerical schemes. Before proposing our scheme, let us summarize numerical schemes in the literature (see references in [8]). We discuss discrete-time approximations. They are simpler than fully discrete numerical schemes but play a crucial role in developing numerical methods. Put $\tau = T/N_T$ ($N_T \in \mathbb{N}$) be the time step size. Let $\boldsymbol{Z}^0$ and $\boldsymbol{Z}^n (n = 1, \dots, N_T)$denote the approximations of the initial function $\boldsymbol{z}^0$ and the solution $\boldsymbol{z}(\cdot, \tau n)$ at time $t = \tau n$, respectively. When we consider the 'equation' (1.1), that is, the case where $M = 1$, we do not use boldfaced variables and omit the subscript for the component. A lot of numerical schemes have been developed and analyzed for equation (1.1). Many researchers have considered nonlinear schemes of the following type:

$$\begin{cases} \dfrac{\beta_\varepsilon^{-1}(U^n) - \beta_\varepsilon^{-1}(U^{n-1})}{\tau} = \Delta U^n + f(\beta_\varepsilon^{-1}(U^n)) & \text{in} \quad \Omega, \\ U^n = 0 & \text{on} \quad \partial\Omega, \\ Z^n := \beta_\varepsilon^{-1}(U^n) & \text{in} \quad \Omega. \end{cases} \tag{2.1}$$

Here, the auxiliary functions U^n represent approximations to $\beta(z(\cdot, \tau n))$, and β_ε is a smooth and strictly increasing function which regularizes the non-smooth and non-strictly increasing function β. Nonlinear schemes of type (2.1) show better accuracy in practice. For solving the corresponding nonlinear algebraic systems arising from fully implicit schemes, some iterative methods such as the Newton method have to be used to linearize the schemes. Therefore, it requires much time for numerical computation. Incidentally, nonlinear schemes of type (2.3) stated below are also employed for the degenerate parabolic equations. However, the algebraic systems arising in (2.3) are non-symmetric, while those in (2.1) are symmetric. Thus, schemes of type (2.1) are more convenient to handle than those of type (2.3), especially, in multi-dimensional case.

Berger, Brezis and Rogers [2] proposed the following linear scheme for the degenerate parabolic equation:

$$\begin{cases} \mu U^n - \tau \Delta U^n = \mu \beta(Z^{n-1}) + \tau f(Z^{n-1}) & \text{in} \quad \Omega, \\ U^n = 0 & \text{on} \quad \partial\Omega, \\ Z^n := Z^{n-1} + \mu(U^n - \beta(Z^{n-1})) & \text{in} \quad \Omega. \end{cases} \tag{2.2}$$

Here, μ is a given positive constant. This is quite simple in that the scheme amounts to solving linear elliptic equations in U^n and then to performing explicit corrections for Z^n. After discretizing this scheme in space, we obtain an easy-to-implement numerical method. Implementation and calculation time are almost the same as the implicit method for the linear heat equation requires. However, the accuracy is low compared with the nonlinear scheme because the nonlinear diffusion is approximated by the linear diffusion with a constant diffusion coefficient.

The history of numerical analysis for the cross-diffusion systems is not long, and the list of references is very short compared to the one for the degenerate parabolic

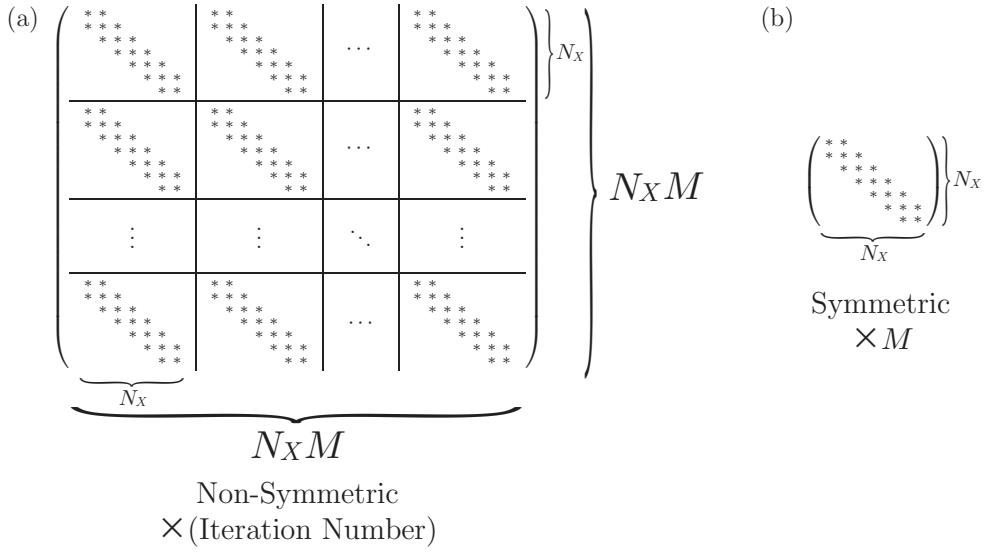


FIG. 2.1. (a) *Type of matrices arising in the nonlinear scheme* (2.3) *in one space dimension.* (b) *Type of matrices arising in the linear schemes* (2.4) *and* (3.1) *in one space dimension. Here,* N_X *and* M *denote the numbers of spatial mesh points and components in* (1.1)*, respectively.*

equations. Most researchers have treated the following type of fully implicit nonlinear schemes.

$$\begin{cases} \dfrac{\boldsymbol{Z}^n - \boldsymbol{Z}^{n-1}}{\tau} = \Delta\boldsymbol{\beta}(\boldsymbol{Z}^n) + \boldsymbol{f}(\boldsymbol{Z}^n) & \text{in} \quad \Omega, \\ \boldsymbol{\beta}(\boldsymbol{Z}^n) = \boldsymbol{0} & \text{on} \quad \partial\Omega. \end{cases} \tag{2.3}$$

The matrices generated by the discretization in space are large, sparse and non-symmetric even in one space dimension (FIG 2.1(a)). The implementation is complicated and the computational costs are high. In multi-component case and/or in multi-dimensional space, this drawback becomes even bigger.

In references [5, 6, 7], the author proposed and analyzed the following linear scheme for the cross-diffusion system (1.1):

$$\begin{cases} \mu\boldsymbol{U}^n - \tau\Delta\boldsymbol{U}^n = \mu\boldsymbol{\beta}(\boldsymbol{Z}^{n-1}) + \tau\boldsymbol{f}(\boldsymbol{Z}^{n-1}) & \text{in} \quad \Omega, \\ \boldsymbol{U}^n = \boldsymbol{0} & \text{on} \quad \partial\Omega, \\ \boldsymbol{Z}^n := \boldsymbol{Z}^{n-1} + \mu(\boldsymbol{U}^n - \boldsymbol{\beta}(\boldsymbol{Z}^{n-1})) & \text{in} \quad \Omega. \end{cases} \tag{2.4}$$

This scheme is regarded as an extension of (2.2) to the system. Likewise, the scheme amounts to solving M independent linear elliptic equations in $\boldsymbol{U}^n$ and updating $\boldsymbol{Z}^n$ explicitly. The boundary condition becomes quite simple. The difficulty of implementation is almost the same as appears in the implicit method for the linear heat equation. The computational cost is less than M times the computational cost of the linear heat equation, because the ensuing linear algebraic system keeps the same matrix for all time steps and for all $i \in \{1, \ldots, M\}$. The type of matrices is shown in FIG 2.1(b).

3. Proposed linear scheme. Taking the advantages and disadvantages of the nonlinear and the linear schemes into consideration, we modify the linear scheme

(2.4), and then, propose an efficient linear scheme for Problem (1.1).

We rewrite equation (1.1) with $M = 1$ and the linear scheme (2.2) formally as follows:

$$\begin{cases} \dfrac{1}{\beta'(z)}\dfrac{\partial \beta(z)}{\partial t} = \Delta \beta(z) + f(z), \\ \dfrac{\partial z}{\partial t} = \dfrac{1}{\beta'(z)}\dfrac{\partial \beta(z)}{\partial t}, \end{cases} \qquad \begin{cases} \mu \dfrac{U^n - \beta(Z^{n-1})}{\tau} = \Delta U^n + f(Z^{n-1}), \\ \dfrac{Z^n - Z^{n-1}}{\tau} = \mu \dfrac{U^n - \beta(Z^{n-1})}{\tau}. \end{cases}$$

By comparing these expressions, the parameter μ can be regarded as an approximation to $1/\beta'(z)$. In practice, we usually choose $\mu = L_\beta^{-1}$, where L_β is the Lipschitz constant of β. But the accuracy of the numerical solutions is low, because of the rough choice of μ. So, it is expected that a good approximation μ to $1/\beta'(z)$ gives the numerical solution with high accuracy, for example, $\mu \approx 1/\beta'(Z^{n-1})$. Along this idea, Murakawa [8] proposed the following scheme for Problem (1.1).

$$\begin{cases} \mu_i^n U_i^n - \tau \Delta U_i^n = \mu_i^n \beta_i(\boldsymbol{Z}^{n-1}) + \tau f_i(\boldsymbol{Z}^{n-1}) & \text{in} \quad \Omega, \\ U_i^n = 0 & \text{on} \quad \partial\Omega, \quad (i = 1, \dots, M). \\ Z_i^n = Z_i^{n-1} + \mu_i^n (U_i^n - \beta_i(\boldsymbol{Z}^{n-1})) & \text{in} \quad \Omega \end{cases} \tag{3.1}$$

Here, $\mu_i^n = \mu_i^n(x)$ $(i = 1, \dots, M)$ are given functions. Thus, we just change μ from a constant to functions. This minor change makes the scheme more accurate. The difficulty of implementation and computational costs do not greatly differ from those of (2.2) and (2.4).

The shape of matrices arising in the scheme (3.1) is the same as in the implicit scheme for the linear heat equation (FIG 2.1(b)). Since the matrices are symmetric, we can employ efficient solver such as conjugate gradient method. On the other hand, the matrices arising in the scheme (2.3) (FIG 2.1(a)) are large, sparse and non-symmetric even in one space dimension. Moreover, computational costs are high.

Rates of convergence of (3.1) with respect to τ were derived theoretically in [8]. Since there is some difference between the handling of the degenerate-diffusion and that of the cross-diffusion from mathematical points of view, it is difficult to treat degenerate cross-diffusion systems in general settings. Therefore, we deal with each case separately. The results can be summarized as follows. Let $\boldsymbol{z}$ be the weak solution of (1.1), and $\boldsymbol{U}$, $\boldsymbol{Z}$ be piecewise constant interpolations in time of a solution of (3.1). We define the global error E by

$$\begin{aligned} E :=& \|\boldsymbol{\beta}(\boldsymbol{z}) - \boldsymbol{U}\|_{L^2(Q)^M} + \left\| \int_0^t (\boldsymbol{\beta}(\boldsymbol{z}) - \boldsymbol{U}) \right\|_{L^\infty(0,T;H^1(\Omega))^M} \\ &+ \|\boldsymbol{z} - \boldsymbol{Z}\|_{L^\infty(0,T;H^{-1}(\Omega))^M}. \end{aligned}$$

Then, the following orders were derived under some assumptions.

- For degenerate parabolic systems (without cross-diffusion),

$$\boldsymbol{z}^0 \in L^2(\Omega)^M \implies E = O(\tau^{1/4}), \tag{3.2}$$

$$\boldsymbol{z}^0 \in L^\infty(\Omega)^M,\ \Delta \boldsymbol{\beta}(\boldsymbol{z}^0) \in L^1(\Omega)^M \implies E = O(\tau^{1/2}). \tag{3.3}$$

- For (non-degenerate) cross-diffusion systems,

$$\boldsymbol{z}^0 \in L^2(\Omega)^M \implies E + \|\boldsymbol{z} - \boldsymbol{Z}\|_{L^2(Q)^M} = O(\tau^{1/2}), \tag{3.4}$$

$$\boldsymbol{z}^0 \in H_0^1(\Omega)^M \implies E + \|\boldsymbol{z} - \boldsymbol{Z}\|_{L^2(Q)^M} = O(\tau). \tag{3.5}$$

The orders (3.3)–(3.5) are sharp on account of the global regularity in time. These optimal error estimates (3.2)–(3.5) are the same as in the case where μ is a constant, and were obtained by Magenes, Nochetto and Verdi [3] for the degenerate parabolic equations and by Murakawa [6] for the cross-diffusion systems. However, actual errors in numerical computation become significantly smaller if we choose $\mu_i^n(x)$ suitably.

4. Numerical experiments. In this section, we carry out numerical experiments in one space dimension in order to demonstrate the performance of our scheme. Both the nonlinear and the linear schemes are tried, and these schemes are discretized in space by the standard finite difference method with a uniform mesh. All experiments were performed on a Laptop equipped with Intel Core(TM) i7-3667U CPU using a single thread. The C sources are complied by the GCC compiler with option -O3.

We calculate the discrete relative $L^2(Q)^M$ error $E_{\beta(z)}$, namely,

$$E_{\beta(z)} = \left(\sum_{\substack{0\le j\le N_X \\ 1\le n\le N_T}} \left| \boldsymbol{U}^{j,n} - \boldsymbol{\beta}(\boldsymbol{z}(x_j, n\tau)) \right|^2 \Big/ \sum_{\substack{0\le j\le N_X \\ 1\le n\le N_T}} |\boldsymbol{\beta}(\boldsymbol{z}(x_j, n\tau))|^2 \right)^{1/2}.$$

Here, $N_X + 1$ is the number of mesh points and x_j $(0 \le j \le N_X)$ imply the spatial grid points.

4.1. The porous medium equation. We deal with the following porous medium equation, that describes the isentropic flow through a porous medium.

$$\frac{\partial z}{\partial t} = \Delta z^m \quad \text{in } \Omega \times (0, T], \tag{4.1}$$

where $m > 1$, $\Omega = (-L, L) = (-8, 8)$ and $T = 10$. With appropriately chosen initial and boundary data, this problem has the following exact solution, which was derived by Barenblatt [1]:

$$z(x,t) = \frac{1}{(t+1)^{m+1}} \left[1 - \frac{(m-1)x^2}{2m(m+1)(t+1)^{\frac{2}{m+1}}} \right]_+^{\frac{1}{m-1}}.$$

Here, $[\cdot]_+$ implies the positive part. For the nonlinear scheme, the following approximate inverse function with $\varepsilon = 10^{-4}$ is used:

$$\beta_\varepsilon^{-1}(u) = \begin{cases} u^{\frac{1}{m}} & \text{if } u \ge \varepsilon^{\frac{m}{m-1}}, \\ \dfrac{1}{\varepsilon} u & \text{otherwise.} \end{cases}$$

For the linear schemes, we set $\mu = 1/m$ in the fixed μ case, and choose μ^n as follows in the case where μ^n are functions:

$$\mu^n = 1/(10^{-3} + \beta'(Z^{n-1})).$$

The spatial mesh size is fixed as $h = 2L/N_X = 2^{-10}$ and we inquire into rates of convergence with respect to the time step size τ. We consider the case where $m = 16$. The Barenblatt solution is shown in FIG 4.1(a). FIG 4.1(b) illustrates errors versus time step size with $\tau = 2^{-4}, 2^{-5}, \ldots, 2^{-9}$. The errors in the proposed linear

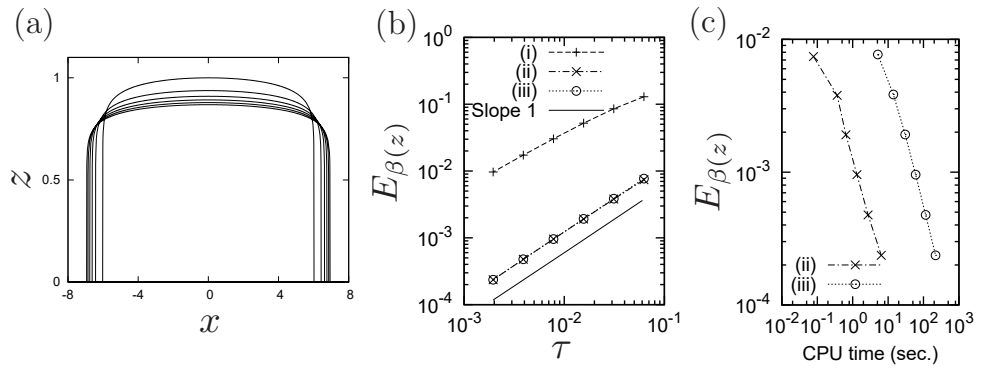


FIG. 4.1. (a) *The Barenblatt solution of the porous medium equation* (4.1) *with* $m = 16$ *at* $t = 0, 2, \ldots, 10$. (b), (c) *Numerical results for the porous medium equation* (4.1), *where* (i) *represents the linear scheme* (2.2), (ii) *represents the linear scheme* (3.1), (iii) *represents the nonlinear scheme* (2.1).

scheme (3.1) and in the nonlinear scheme (2.1) are almost the same, and are quite smaller than those in the linear scheme (2.2) with fixed μ. The errors are along a straight line having slope 1, which implies that the numerical rate of convergence with respect to τ is of order 1 for each scheme. This is much better than the theoretical result (3.3). The proposed linear and the nonlinear schemes are compared in terms of CPU time. The results are shown in FIG 4.1(c). The proposed linear scheme is about 50 times faster than the nonlinear scheme to achieve the same level of accuracy in this experiment. These results indicate that the proposed linear scheme is superior in speed to the nonlinear scheme even though the linear scheme is very easy-to-implement and it is computationally less costly. These advantages become even more when we deal with higher dimensional and/or multi-component problems.

4.2. The Shigesada-Kawasaki-Teramoto cross-diffusion system. We deal with the following cross-diffusion system that was proposed by Shigesada, Kawasaki and Teramoto [9] to understand temporal and spatial behaviours of two animal species under the influence of the population pressure due to intra- and interspecific interferences:

$$\begin{cases} \dfrac{\partial z_1}{\partial t} = \Delta\left[(a_{10} + a_{11}z_1 + a_{12}z_2)z_1\right] + (c_{10} - c_{11}z_1 - c_{12}z_2)z_1 + f_1(x,t), \\ \dfrac{\partial z_2}{\partial t} = \Delta\left[(a_{20} + a_{21}z_1 + a_{22}z_2)z_2\right] + (c_{20} - c_{21}z_1 - c_{22}z_2)z_2 + f_2(x,t). \end{cases} \tag{4.2}$$

Here, we set $a_{10} = 1$, $a_{20} = 1/(3c_{12})$, $c_{10} = 1$, $c_{11} = 1$, $c_{12} = 2.5$, $c_{20} = 1$,

$c_{21} = 2 + 5c_{20}/3 - c_{20}c_{12}$, $c_{22} = 1$, and

$$\begin{aligned}
f_1(x,t) =& \frac{1}{32}\mathrm{sech}\left(\frac{1}{4}(t+\sqrt{2}x)\right)^4 \left(-4a_{11} + 2a_{12} + (2a_{11} - 5a_{12})\tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right.\\
&\left. + \cosh\left(\frac{1}{2}(t+\sqrt{2}x)\right)\left(2a_{11} - a_{12} + (2a_{11} + a_{12})\tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right)\right),\\
f_2(x,t) =& \frac{1}{64}\mathrm{sech}\left(\frac{1}{4}(t+\sqrt{2}x)\right)^4 \left(-1 + \tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right)\\
&\times \left(-7a_{21} + 10a_{22} + 3(a_{21} - 2a_{22})\tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right.\\
&\left. + \cosh\left(\frac{1}{2}(t+\sqrt{2}x)\right)\left(5a_{21} - 4a_{22} + (3a_{21} + 4a_{22})\tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right)\right).
\end{aligned}$$

This problem has the following exact solution:

$$\begin{aligned}
z_1(x,t) &= \frac{1}{2}\left(1 + \tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right),\\
z_2(x,t) &= \frac{1}{4}\left(1 - \tanh\left(\frac{1}{4}(t+\sqrt{2}x)\right)\right)^2.
\end{aligned} \tag{4.3}$$

The functions f_1 and f_2 are determined so that (z_1, z_2) defined in (4.3) is a solution of system (4.2).

We carry out numerical experiments with $a_{11} = 0$, $a_{12} = 10$, $a_{21} = 10$, $a_{22} = 0$ in space $\Omega = (0, 10)$ and in time interval $(-10, -5)$. The spatial mesh size is fixed as $h = 2^{-8}$. The initial and the Dirichlet boundary data are given by the exact solution. The solution is shown in FIG 4.2(a). Looking at the shapes of matrices arising in the

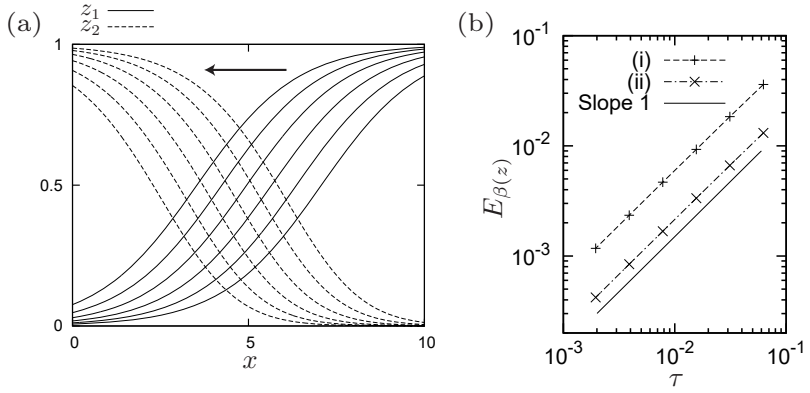


FIG. 4.2. (a) *The solution of* (4.2) *at* $t = -10, -9, \ldots, -5$. (b) *Numerical results for* (4.2), *where* (i) *and* (ii) *represent the linear schemes* (2.4) *and* (3.1), *respectively.*

schemes, which are shown in FIG 2.1, it is easy to imagine that the linear scheme (3.1) is superior than the nonlinear scheme (2.3) in terms of simplicity of implementation and computational costs. We treat only the linear schemes (2.4) and (3.1) in the case where μ is fixed as $\mu = 0.1$ and in the case where μ_i^n are functions, respectively. Using $\boldsymbol{Z}^{n-1}$, we define μ_i^n as follows:

$$\mu_i^n(x) = 1/(\beta_i)_i(\boldsymbol{Z}^{n-1}(x)).$$

In the fixed μ case, if we choose μ_i larger than 0.1, then the numerical solutions become unstable. FIG 4.2(b) shows the numerical results with $\tau = 2^{-4}, 2^{-5}, \ldots, 2^{-9}$. Numerical convergence rate with respect to τ is observed to be of order 1, which corresponds to the theoretical result. The proposed scheme (3.1) shows higher accuracy compared to the fixed μ case. The difference (about three times difference) is not so large in this experiment. This difference becomes considerably large in the problem of which solution shows the profile with sharp peaks (see Section 5.4 in [8]).

4.3. A degenerate convection-reaction-diffusion equation. We deal with the following degenerate convection-reaction-diffusion equation in one space dimension:

$$\frac{\partial z}{\partial t} = \frac{\partial}{\partial x^2}\beta(z) - \frac{\partial}{\partial x}(b_1 z - b_2\beta(z)) - c(z - \beta(z)), \tag{4.4}$$

where $b_1, b_2, c \in \mathbb{R}$. The function β is defined as follows.

$$\beta(s) := [\![s]\!]^m := \begin{cases} s^m & \text{if } s \geq 0, \\ -(-s)^m & \text{if } s < 0. \end{cases}$$

This problem has the following exact solution [4].

$$z(x,t) = k_1 \exp\left(-ct - \frac{b_2(x - b_1 t)}{2m}\right) \left[\!\!\left[\cos\left(\frac{1}{2}(x - b_1 t - k_2)\sqrt{4c - b_2^2}\right) \right]\!\!\right]^{1/m},$$

where k_1 and k_2 are arbitrary constants. Since $\cos(\sqrt{-1}x) = \cosh x$, the value in the bracket on the right hand side can be determined for arbitrary parameters.

The linear scheme (3.1) can be applied to this problem because (4.4) is linear in z and $\beta(z)$. Therefore, we have the following linear scheme for (4.4).

$$\begin{cases} \mu^n \dfrac{U^n - \beta(Z^{n-1})}{\tau} = \dfrac{\partial}{\partial x^2}U^n - \dfrac{\partial}{\partial x}(b_1 Z^n - b_2 U^n) - c(Z^n - U^n), \\ Z^n = Z^{n-1} + \mu^n(U^n - \beta(Z^{n-1})). \end{cases}$$

Substituting the second equation into the first one, we have

$$\begin{cases} ((1+\tau c)\mu^n - \tau c)U^n - \tau \dfrac{\partial}{\partial x^2}U^n + \tau \dfrac{\partial}{\partial x}((b_1\mu^n - b_2)U^n) \\ \qquad = (1+\tau c)\mu^n \beta(Z^{n-1}) - \tau c Z^{n-1} - \tau b_1 \dfrac{\partial}{\partial x}(Z^{n-1} - \mu^n \beta(Z^{n-1})), \\ Z^n = Z^{n-1} + \mu^n(U^n - \beta(Z^{n-1})). \end{cases} \tag{4.5}$$

This scheme, which consists of solving the linear problem in U^n and explicit correction for Z^n, is quite simpler than nonlinear schemes.

We carry out numerical simulations for (4.4). We set $m = 10$, $b_1 = 2mc/b_2$, $b_2 = c = k_2 = 1$, $k_1 = 3\pi/4$, $\Omega = (0, L) = (0, 10)$, $(0, T) = (0, 0.05)$. The exact solution is presented in FIG 4.3(a). Because of the appearance of the convection term, we set $\tau = h/(4L)$ and used the standard upwind technique. We deal with the linear schemes with fixed μ and with varying $\mu^n(x)$. These parameters are chosen to be the same as used in Subsection 4.1. FIG 4.3(b) shows the numerical results with $h = 2^{-6}, 2^{-7}, \ldots, 2^{-10}$, which demonstrates numerical convergence of both linear schemes. The numerical rates of convergence with respect to h (and/or τ) is slightly less than 1. The scheme with varying μ shows higher accuracy than the linear scheme with fixed μ.

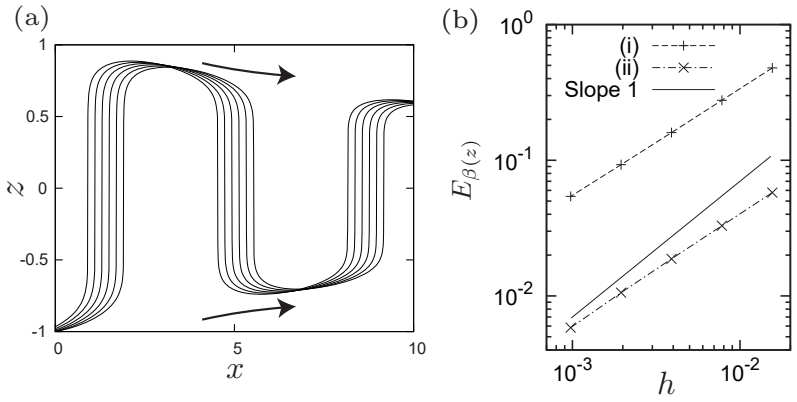


FIG. 4.3. (a) *The solution of* (4.4) *at* $t = 0, 0.01, \ldots, 0.05$. (b) *Numerical results for* (4.4)*, where* (i) *represents the linear scheme* (4.5) *with fixed* μ*,* (ii) *represents the linear scheme* (4.5) *with varying* $\mu^n(x)$*.*

5. Conclusion. The linear scheme with simple implementation has been proposed instead of the widely used nonlinear schemes for the nonlinear diffusion problem (1.1). The motivation is based on the fact that the proposed linear scheme (3.1), which is an improvement of the linear scheme (2.4), retains the same accuracy as obtained from the nonlinear schemes (2.1) and (2.3) with less difficulty of the implementation. The difficulty is much the same as for the linear heat equation, whereas the advantages are many. For instance, the type of linear algebraic systems in (3.1) is the same as in the implicit method for the linear heat equation. Moreover, it is easy to set the parameters appropriately and the computational costs are low. These advantages and those mentioned earlier work as well even for multi-dimensional and multi-component systems. On the other hand, in general, the nonlinear schemes are complicated to implement and require high computational costs. Taking account of accuracy, efficiency, stability and computational cost into consideration, we proposed the linear scheme with simple implementation, of which advantages are proved in complicated problem such as (4.4).

Acknowledgments. This work was partially supported by JSPS KAKENHI Grant nos. 26287025, 15H03635 and 17K05368, and JST CREST Grant No. JPMJCR14D3. MM acknowledges the Matsumae International Foundation for financial support and the Faculty of Mathematics at Kyushu University for hosting him during the research fellowship.

REFERENCES

[1] G.I. BARENBLATT, *On some unsteady motion of a liquid or a gas in a porous medium*, Prikl. Math. Meh., 16 (1952), pp. 67–78.

[2] A.E. BERGER, H. BREZIS AND J.C.W. ROGERS, *A numerical method for solving the problem* $u_t - \Delta f(u) = 0$, R.A.I.R.O. Anal. Numér., 13 (1979), pp. 297–312.

[3] E. MAGENES, R.H. NOCHETTO AND C. VERDI, *Energy error estimates for a linear scheme to approximate nonlinear parabolic problems*, Math. Mod. Numer. Anal., 21 (1987), pp. 655–678.

[4] M. MOLATI AND H. MURAKAWA, *Exact solutions of nonlinear diffusion-convection-reaction equation: A Lie symmetry analysis approach*, preprint.

[5] H. Murakawa, *A linear scheme to approximate nonlinear cross-diffusion systems*, Math. Mod. Numer. Anal., 45 (2011), pp. 1141–1161.
[6] H. Murakawa, *Error estimates for discrete-time approximations of nonlinear cross-diffusion systems*, SIAM J. Numer. Anal., 52(2) (2014), pp. 955–974.
[7] H. Murakawa, *A linear finite volume method for nonlinear cross-diffusion systems*, Numer. Math., 136(1) (2017), pp. 1–26.
[8] H. Murakawa, *An efficient linear scheme to approximate nonlinear diffusion problems*, to appear in Jpn. J. Ind. Appl. Math., DOI: 10.1007/s13160-017-0279-3.
[9] N. Shigesada, K. Kawasaki and E. Teramoto, *Spatial segregation of interacting species*, J. Theor. Biol., 79 (1979), pp. 83–99.

Proceedings of EQUADIFF 2017
pp. 315–324

CONTINUOUS DEPENDENCE FOR BV-ENTROPY SOLUTIONS TO STRONGLY DEGENERATE PARABOLIC EQUATIONS WITH VARIABLE COEFFICIENTS *

HIROSHI WATANABE†

Abstract. We consider the Cauchy problem for degenerate parabolic equations with variable coefficients. The equation has nonlinear convective term and degenerate diffusion term which depends on the spatial and time variables. In this paper, we prove the continuous dependence for entropy solutions in the space BV to the problem not only initial function but also all coefficients.

Key words. strongly degenerate parabolic, continuous dependence, BV-entropy solution

AMS subject classifications. 35K65, 35K55, 35L65

1. Introduction. Let $0 < T < \infty$ and $N \in \mathbb{N}$ be constants. We consider the Cauchy problem for a degenerate parabolic equation of the form

$$\begin{cases} \partial_t u + \nabla \cdot A(x,t,u) + B(x,t,u) = \Delta \beta(x,t,u), \quad (x,t) \in \mathbb{R}^N_T := \mathbb{R}^N \times (0,T), \\ u(x,0) = u_0(x), \quad u_0 \in L^\infty(\mathbb{R}^N) \cap BV(\mathbb{R}^N). \end{cases} \tag{P}$$

Here $\partial_t := \partial/\partial t$, $\nabla := (\partial/\partial x_1, \dots, \partial/\partial x_N)$ and $\Delta := \sum_{i=1}^N \partial^2/\partial x_i^2$ are the spatial nabla and the laplacian in $\mathbb{R}^N$, respectively. $A(x,t,\xi) = (A^1, ..., A^N)(x,t,\xi)$ is an $\mathbb{R}^N$-valued function on $\mathbb{R}^N \times [0,T] \times \mathbb{R}$ and $B(x,t,\xi)$ and $\beta(x,t,\xi)$ are $\mathbb{R}$-valued functions on $\mathbb{R}^N \times [0,T] \times \mathbb{R}$. The function $\beta(x,t,\xi)$ is supposed to be monotone nondecreasing and locally Lipschitz continuous with respect to ξ for any $(x,t) \in \mathbb{R}^N_T$. Therefore, the set of points ξ where $\partial_\xi \beta(x,t,\xi) = 0$ may have a positive measure for any $(x,t) \in \mathbb{R}^N_T$. In this sense, we say that the equation posed as (P):

$$\partial_t u + \nabla \cdot A(x,t,u) + B(x,t,u) = \Delta \beta(x,t,u) \tag{1.1}$$

is a *strongly degenerate parabolic equation*. The equation (1.1) can be applied to several mathematical models; hyperbolic conservation laws, porous medium, Stefan problem, filtration problem, sedimentation process, traffic flow, and so on. Moreover, (1.1) is regarded as a linear combination of the time dependent conservation laws (quasilinear hyperbolic equation) and the porous medium equation (nonlinear degenerate parabolic equation). Thus, (1.1) has both properties of hyperbolic equations and those of parabolic equations. In particular, up to the assumptions on β, "parabolicity" of (1.1) and "hyperbolicity" of it are not necessarily comparable.

Our mathematical treatment of the equation (1.1) is L^1-framework. More specifically, we consider (1.1) in the space $L^1(\mathbb{R}^N)$ and construct solutions to (1.1) in the space $L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$. Here, solutions to (1.1) should be defined in generalized sense. To ensure the existence and uniqueness of it, it is necessitate to consider distributional solutions satisfying a special condition. This framework was first treated by Vol'pert-Hudjaev [9]. In fact, it is well known that the Fréchet-Kolmogorov compactness theorem in the space BV and the Kružkov's doubling variable method [8]

*This work was supported by Grant-in-Aid for Scientific Research (C) (No. 17K05294) JSPS

†Division of Mathematical Sciences, Faculty of Science and Technology, Oita University 700 Dannoharu, Oita, Japan, 870-1192 (hwatanabe@oita-u.ac.jp).

are available. Under the above framework, the existence and uniqueness of entropy solutions to (1.1) are given ([3, 4, 5, 7, 10, 11, 12, 13]). Here, entropy solutions are weak solutions satisfying an entropy inequality which is derived by Kružkov [8]. In particular, Watanabe [11] proved the existence and uniqueness of entropy solutions to (P) in the space BV.

In this paper, we prove the continuous dependence of the BV-entropy solution to (P) not only initial data but also coefficients A, B and β. Feature of the present paper is to consider the equation (1.1) with variable coefficients. In particular, the equation with variable diffusion coefficients is treated in few literatures. For example, Chen-Karlsen [4] considered the equation with a separation variable type convective term and a quasi-linear type diffusion term. Notice that, these coefficients do not depend on time variable. In this article, we consider the time dependent nonlinear type convection $\nabla \cdot A(x,t,u)$ and diffusion $\Delta\beta(x,t,u)$. To prove desired estimate, we modify the choice of entropy triplet and the calculation in [4].

Throughout this paper, we employ the following notations and terminologies. For $1 \le p \le \infty$, the Lebesgue space of real-valued Lebesgue-measurable functions on $\mathbb{R}^N$ equipped with the usual norm $||\cdot||_p$ is denoted by $L^p(\mathbb{R}^N)$. The space of functions of bounded variation in $\mathbb{R}^N$ is denoted by $BV(\mathbb{R}^N)$ and the total variation on $\mathbb{R}^N$ is denoted by $TV(\cdot)$ (cf. [2, 6, 14]). The space $C_0^\infty(\mathbb{R}^N)^+$ is the class of nonnegative valued $C_0^\infty(\mathbb{R}^N)$-functions. The function $\mathrm{sgn}(\xi)$ means the usual *signum* function.

2. Assumptions and the main result. In this section, we present some assumptions and the main result. Before that, we write the nabla of the function $A(x,t,u)$ and the laplacian of the function $\beta(x,t,u)$ as follows:

$$\nabla \cdot A(x,t,u) = [\nabla \cdot A](x,t,u) + [\partial_\xi A](x,t,u) \cdot \nabla u$$

and

$$\begin{aligned}\Delta\beta(x,t,u) &= \nabla \cdot ([\nabla\beta](x,t,u) + [\partial_\xi\beta](x,t,u) \cdot \nabla u)\\ &= [\Delta\beta](x,t,u) + 2[\partial_\xi \nabla\beta](x,t,u) \cdot \nabla u + [\partial_\xi^2\beta](x,t,u)|\nabla u|^2 + [\partial_\xi\beta](x,t,u)\Delta u\end{aligned}$$

for $(x,t) \in \mathbb{R}_T^N$ and some regular function u. These are based on the chain rule formulas in [1] (see also [2, Theorem 3.99], [14, Theorem 2.1.11]).

Throughout this paper, we impose the following assumptions on the functions A, B, β and u_0. Here, we write $\partial_{x_i} := \partial/\partial_{x_i}$ for $i = 1, \dots, N$, $\partial_{x_{N+1}} := \partial_t$, $\widehat{\nabla} := (\partial_{x_1}, \partial_{x_2}, \dots, \partial_{x_N}, \partial_{x_{N+1}})$ and $\mathcal{U} := [-U, U]$ for any $U > 0$. For any $U > 0$, the following conditions hold:

{A0} $u_0(x) \in L^\infty(\mathbb{R}^N) \cap BV(\mathbb{R}^N)$;

{A1} $\begin{cases} A \in L^1(\mathbb{R}_T^N \times \mathcal{U})^N \cap L^\infty(\mathbb{R}_T^N \times \mathcal{U})^N \cap L^\infty(\mathcal{U}; L^2(\mathbb{R}_T^N))^N, \\ \partial_\xi A \in L^1(\mathbb{R}_T^N \times \mathcal{U})^N \cap L^\infty(\mathbb{R}_T^N \times \mathcal{U})^N, \quad \nabla \cdot A,\ \partial_\xi \nabla \cdot A \in L^1(\mathbb{R}_T^N \times \mathcal{U}); \end{cases}$

{A2} $\begin{cases} B \in L^1(\mathbb{R}_T^N \times \mathcal{U}) \cap L^\infty(\mathbb{R}_T^N \times \mathcal{U}) \cap L^\infty(\mathcal{U}; L^1(\mathbb{R}_T^N)), \\ |\widehat{\nabla} B| \in L^\infty(\mathcal{U}; L^1(\mathbb{R}_T^N)), \quad \partial_\xi B \in L^\infty(\mathbb{R}_T^N \times \mathcal{U}); \end{cases}$

{A3} $\begin{cases} \beta \in L^1(\mathbb{R}_T^N \times \mathcal{U}) \cap L^\infty(\mathbb{R}_T^N \times \mathcal{U}), \\ \nabla\beta \in L^1(\mathbb{R}_T^N \times \mathcal{U})^N \cap L^\infty(\mathcal{U}; L^2(\mathbb{R}_T^N))^N,\ \partial_t\beta \in L^\infty(\mathcal{U}; L^1(\mathbb{R}_T^N)), \\ \partial_\xi\beta \in L^\infty(\mathbb{R}_T^N \times \mathcal{U}),\ \partial_\xi\nabla\beta \in L^1(\mathbb{R}_T^N \times \mathcal{U})^N,\ \Delta\beta,\ \partial_\xi\Delta\beta \in L^1(\mathbb{R}_T^N \times \mathcal{U}); \end{cases}$

{A4} $B(x,t,0) = 0$ and $\nabla\beta(x,t,0) - A(x,t,0) = \vec{0}$ for $(x,t) \in \mathbb{R}_T^N$;

{A5} Let $\Psi(x,t,\xi) := \nabla \cdot A(x,t,\xi) - \Delta\beta(x,t,\xi) + B(x,t,\xi)$. Then, there exist positive constants c_0, c_1 such that

$$\sup_{(x,t)\in\mathbb{R}_T^N} |\Psi(x,t,0)| \le c_0, \qquad \sup_{(x,t,\xi)\in\mathbb{R}_T^N\times\mathbb{R}} (-\partial_\xi \Psi(x,t,\xi)) \le c_1;$$

{A6} For $i = 1, 2, \dots, N+1$,
$$\begin{cases} \partial_\xi \partial_{x_i}(\nabla\beta - A) \in L^\infty(\mathbb{R}^N_T \times \mathcal{U})^N, \\ \nabla\cdot(\nabla\beta - A),\ |\widehat{\nabla}(\Delta\beta - \nabla\cdot A)| \in L^\infty(\mathcal{U}; L^1(\mathbb{R}^N_T)); \end{cases}$$

{A7} For $(x,t,\xi) \in \mathbb{R}^N_T \times \mathcal{U}$ and $\lambda = (\lambda_1, \dots, \lambda_{N+1}) \in \mathbb{R}^{N+1}$, there exists a constant $\kappa > 0$ such that
$$\textstyle\sum_{i,j=1}^{N+1} (\partial_\xi \beta(x,t,\xi)\lambda_i\lambda_j - \kappa(\partial_{x_i}\partial_\xi\beta(x,t,\xi)\lambda_j)^2) \geq 0.$$

The conditions $\{A1\}$-$\{A3\}$ are regularity assumptions for the functions A, B and β with respect to x, t and ξ. {A4}-{A6} are used to prove L^∞, L^1 and BV-estimates for approximate solutions. {A6} is also interpreted the regularity assumptions for the flux $A(x,t,\xi) - \nabla\beta(x,t,\xi)$ to (1.1). The condition {A7} fulfills to get a BV-estimate with respect to x and t for approximate solutions to (P).

REMARK 1. *By the assumption $\{A7\}$, it is deduced that*
$$\partial_\xi \beta(x,t,\xi) \geq 0 \quad \textit{for } (x,t,\xi) \in \mathbb{R}^N_T \times \mathcal{U}. \tag{2.1}$$
More specifically, $\beta(x,t,\xi)$ degenerate on nondegenerate intervals with respect to ξ. In particular, if $\beta(x,t,\xi) \equiv \beta(\xi)$, then $\{A7\}$ is equivalent to (2.1).

We also impose an additional regularity assumption to prove the uniqueness of BV-entropy solution: for any $i, j = 1, \dots, N$,

{A8} $$\begin{cases} \partial_\xi\partial_{x_i}A^j,\ \partial_\xi\partial_{x_i}\partial_{x_j}\beta \in L^\infty(\mathbb{R}^N_T \times \mathcal{U}),\quad \sqrt{\partial_\xi\beta},\ \partial_{x_i}\sqrt{\partial_\xi\beta} \in L^1(\mathbb{R}^N_T \times \mathcal{U}), \\ \partial_{x_i}\partial_\xi\beta,\ \partial_{x_i}\sqrt{\partial_\xi\beta},\ \partial_{x_i}\partial_{x_j}\sqrt{\partial_\xi\beta} \in L^\infty(\mathbb{R}^N_T \times \mathcal{U}). \end{cases}$$

Next, we introduce generalized solutions to (P). Usually, the weak solution is interpreted as a generalized solution to equations with divergence form. Then, the existence and uniqueness of it may be shown. However, we can not prove the uniqueness of weak solutions to (P) in general. Because, discontinuities break out from the nonlinear convective term $\nabla\cdot A(x,t,u)$ and the uniqueness of weak solutions are possibly broken because of it. Therefore, we formulate the weak solution satisfying a special condition. It is called by the name entropy solution. To define it, we state the concept of entropy:

DEFINITION 2.1. *Let $\eta(\xi) \in C^2(\mathbb{R})$ and $q(x,t,\xi), r(x,t,\xi) \in L^1(\mathbb{R}^N_T \times \mathbb{R})^N \cap L^\infty(\mathbb{R}^N_T \times \mathbb{R})^N$ satisfying $q(x,t,\cdot), r(x,t,\cdot) \in C^2(\mathbb{R})^N$ for $(x,t) \in \mathbb{R}^N_T$. A triplet (η, q, r) is entropy triplet to (P) if it satisfies*
$$\partial_\xi q(x,t,\xi) = \eta'(\xi)\partial_\xi A(x,t,\xi), \quad \partial_\xi r(x,t,\xi) = \eta'(\xi)\partial_\xi\nabla\beta(x,t,\xi)$$
for a.e. $(x,t,\xi) \in \mathbb{R}^N_T \times \mathbb{R}$. Then, η is called entropy and (q,r) is called entropy flux.

DEFINITION 2.2. *Let $u_0 \in L^\infty(\mathbb{R}^N) \cap BV(\mathbb{R}^N)$. A function $u \in L^\infty(\mathbb{R}^N_T) \cap BV(\mathbb{R}^N_T)$ is called an BV-entropy solution to (P), if it satisfies the two conditions below:*

(I) *$u \in C([0,T]; L^1(\mathbb{R}^N))$ and $L^1\text{-}\lim_{t\downarrow 0} u(t) = u_0$;*

(II) *$\beta(x,t,u) \in L^2(0,T; H^1(\mathbb{R}^N))$, and for all $\varphi \in C_0^\infty(\mathbb{R}^N_T)^+$,*
$$\begin{aligned} \int_{\mathbb{R}^N_T} \{&\eta(u)\partial_t\varphi + (q(x,t,u) - r(x,t,u))\cdot\nabla\varphi + ([\nabla\cdot q](x,t,u) - [\nabla\cdot r](x,t,u))\varphi \\ &-\eta'(u)(\nabla\beta(x,t,u) - [\nabla\beta](x,t,u))\cdot\nabla\varphi \\ &-\eta'(u)([\nabla\cdot A](x,t,u) - [\Delta\beta](x,t,u) + B(x,t,u))\varphi\}dxdt \\ \geq \int_{\mathbb{R}^N_T} &\eta''(u)|\sqrt{\partial_\xi\beta(x,t,u)}Du|^2\varphi dxdt. \end{aligned}$$

The existence and uniqueness of the BV-entropy solution to (P) is given by Watanabe [11] as follows:

THEOREM 2.3 (Watanabe [11]). *We assume the conditions $\{A0\}$-$\{A7\}$. Then, the following statements hold:*

(I) There exists a BV-entropy solution u to (P). Moreover, if we take $U > 0$ satisfying $(||u_0||_{L^\infty(\mathbb{R}^N)} + c_0T)e^{c_1T} < U$ for the positive constants c_0 and c_1 in $\{A5\}$, then it follows that $u(x,t) \in \mathcal{U}$ for $(x,t) \in \mathbb{R}^N_T$. Additionally, there exist positive constants C_0 and C_1 which depend on T such that $TV(u(\cdot,t)) \le e^{C_0t}(TV(u_0) + C_1)$ for $t \in (0,T)$;

(II) We additionally impose the assumption $\{A8\}$. Let u, v be a pair of BV-entropy solutions to (P) with initial values u_0 and v_0, respectively. Then, there exist a positive constant C_2 which depend on T such that

$$\int_{\mathbb{R}^N} |u(x,t) - v(x,t)|dx \le e^{(\alpha+C_2)t} \int_{\mathbb{R}^N} |u_0(x) - v_0(x)|dx,$$

where $\alpha := ||\partial_\xi B||_{L^\infty(\mathbb{R}^N_T \times \mathcal{U})}$ for $t \in (0,T)$. In particular, for each initial value u_0, a BV-entropy solution is uniquely determined.

In the above result, we give the assumptions {A0}-{A7} to prove the existence of BV-entropy solutions. In this paper, we prove the continuous dependence of the BV-entropy solution to the function u_0, A, B and β under the additional assumption: for any $i, j = 1, \dots, N$,

$$\textbf{\{A8\}'} \quad \begin{cases} \sqrt{\partial_\xi \beta},\ \partial_{x_i}\sqrt{\partial_\xi \beta} \in L^1(\mathbb{R}^N_T \times \mathcal{U}), \quad \partial_{x_i}\partial_\xi \beta,\ \partial_{x_i}\sqrt{\partial_\xi \beta} \in L^\infty(\mathbb{R}^N_T \times \mathcal{U}), \\ \partial_{x_i} B,\ \nabla \cdot A,\ \partial_{x_j}\partial_{x_i} A^i,\ \Delta\beta,\ \partial_\xi \partial_{x_i}\partial_{x_j}\beta,\ \partial_{x_j}\partial^2_{x_i}\beta \in L^\infty(L^1), \end{cases}$$

where $L^\infty(L^1) := L^\infty((0,T) \times \mathcal{U}; L^1(\mathbb{R}^N))$. Notice that, since we consider nonlinear type coefficients, we need stronger regularity assumption than separation variable and quasilinear diffusion case [4].

THEOREM 2.4. *Let u_i be the BV-entropy solution to (P) with initial functions $u_{i,0}$ and coefficients A_i, B_i, β_i satisfying the assumptions $\{A0\}$-$\{A7\}$ and $\{A8\}'$ for $i = 1, 2$, respectively. For any $t \in (0,T)$, the following inequality holds:*

$$\begin{aligned} ||u_1(x,t) - u_2(x,t)||_{L^1(\mathbb{R}^N)} &\le e^{\alpha' t}||u_{1,0} - u_{2,0}||_{L^1(\mathbb{R}^N)} + \tfrac{e^{\alpha' t}-1}{\alpha'}\{||B_1 - B_2||_{L^\infty(L^1)} \\ &+ |||[\nabla_x \cdot A_1] - [\nabla_x \cdot A_2]|||_{L^\infty(L^1)} + |||[\Delta_x \beta_1] - [\Delta_x \beta_2]|||_{L^\infty(L^1)} \\ &+ e^{C_0 t}(TV(u_0) + C_1)(|||[\partial_\xi \nabla_x \beta_1] - [\partial_\xi \nabla_x \beta_2]|||_{(L^\infty)^N} + |||[\partial_\xi A_1] - [\partial_\xi A_2]|||_{(L^\infty)^N} \\ &\qquad + 2\left|\left|\nabla_x \sqrt{[\partial_\xi \beta_2]}\right|\right|_{(L^\infty)^N} \left|\left|\sqrt{[\partial_\xi \beta_1]} - \sqrt{[\partial_\xi \beta_2]}\right|\right|_{L^\infty})\} \\ &+ \widehat{C}\sqrt{t}e^{\alpha' t}\left|\left|\sqrt{[\partial_\xi \beta_1]} - \sqrt{[\partial_\xi \beta_2]}\right|\right|_{L^\infty} \end{aligned}$$

for some positive constants $\widehat{C}$ and $\alpha' := \max_{i=1,2}\{||\partial_\xi B_i||_{L^\infty}\}$. Here, $TV(u_0) := \max_{i=1,2}\{TV(u_{i,0})\}$, $L^\infty := L^\infty(\mathbb{R}^N \times (0,T) \times \mathcal{U})$, $L^\infty(L^1) := L^\infty((0,T) \times \mathcal{U}; L^1(\mathbb{R}^N))$ and C_0, C_1 are positive constants in Theorem 2.3 (I).

3. Proof of Main Theorem. *Step 0.* Let $\varphi \in C_0^\infty(\mathbb{R}^N_T)^+$. In addition, we introduce a symmetric function $\theta \in C_0^\infty(\mathbb{R})^+$ satisfying $\int_{\mathbb{R}} \theta(t)dt = 1$ and $\mathrm{supp}[\theta(t)] \subset \{|t| \le 1\}$. Similarly, we use a spherically symmetric function $\omega \in C_0^\infty(\mathbb{R}^N)^+$ satisfying $\int_{\mathbb{R}^N} \omega(x)dx = 1$ and $\mathrm{supp}[\omega(x)] \subset \{|x| \le 1\}$. Let $\delta_0, \delta > 0$ and define $\theta_{\delta_0}(t) = (1/\delta_0)\theta(t/\delta_0)$ and $\omega_\delta(x) = (1/\delta^N)\omega(x/\delta)$. These are smooth functions on $\mathbb{R}$ and $\mathbb{R}^N$, respectively, and satisfy

$$\lim_{\delta_0 \downarrow 0} \int_0^T \theta_{\delta_0}(t)\varphi(x,t)dt = \varphi(x,0), \quad \lim_{\delta \downarrow 0} \int_{\mathbb{R}^N} \omega_\delta(x)\varphi(x,t)dx = \varphi(0,t)$$

for $(x,t) \in \mathbb{R}^N_T$. Moreover, let $\nu, \tau \in (0,T)$ with $\nu < \tau$. For any $\alpha_0 > 0$, we define

$$\varphi_{\alpha_0}(t) := H_{\alpha_0}(t-\nu) - H_{\alpha_0}(t-\tau), \quad H_{\alpha_0}(t) := \int_{-\infty}^{t} \theta_{\alpha_0}(\sigma) d\sigma.$$

Then, we now employ the test function $\phi_\delta^{\delta_0,\alpha_0}$ defined by

$$\phi_\delta^{\delta_0,\alpha_0}(x,y,t,s) := \varphi_{\alpha_0}(t)\omega_\delta(x-y)\theta_{\delta_0}(t-s) \tag{3.1}$$

for $0 < \alpha_0 < \min(\nu, T-\tau)$ and $(x,t,y,s) \in (\mathbb{R}^N_T)^2$. Then, the following property holds

$$\lim_{\delta \downarrow 0} \int_{(\mathbb{R}^N)^2} |x-y| \left|\omega_\delta\left(\tfrac{x-y}{2}\right)\right| dxdy = 0$$

and there exists a constant $C > 0$ such that

$$\lim_{\delta \downarrow 0} \int_{\mathbb{R}^N_T \times \mathbb{R}^N} |x-y| \left|\partial_{x_i}\omega_\delta\left(\tfrac{x-y}{2}\right)\right| \varphi\left(\tfrac{x+y}{2}, t\right) dxdydt \le C \int_{\mathbb{R}^N_T} \varphi(x,t) dxdt,$$
$$\lim_{\delta \downarrow 0} \int_{\mathbb{R}^N_T \times \mathbb{R}^N} |x-y|^2 \left|\partial_{x_i}\partial_{x_j}\omega_\delta\left(\tfrac{x-y}{2}\right)\right| \varphi\left(\tfrac{x+y}{2}, t\right) dxdydt \le C \int_{\mathbb{R}^N_T} \varphi(x,t) dxdt$$

for $1 \le i, j \le N$. Moreover, it follows that

$$(\partial_t + \partial_s)\phi_\delta^{\delta_0,\alpha_0} = (\theta_{\alpha_0}(t-\nu) - \theta_{\alpha_0}(t-\tau))\theta_{\delta_0}(t-s)\omega_\delta(x-y), \quad (\nabla_x + \nabla_y)\phi_\delta^{\delta_0,\alpha_0} = 0.$$

In this section, the proof of Theorem 2.4 is presented. Hereafter, we give the entropy triplet in the following concrete form:

$$\begin{aligned} \eta(u) = \eta_\rho(u) &:= \int_k^u \operatorname{sgn}_\rho(\xi - k) d\xi, \\ q(x,t,u) = q_\rho(x,t,u) &:= \int_k^u \operatorname{sgn}_\rho(\xi - k)[\partial_\xi A](x,t,\xi) d\xi, \\ r(x,t,u) = r_\rho(x,t,u) &:= \int_k^u \operatorname{sgn}_\rho(\xi - k)[\partial_\xi \nabla\beta](x,t,\xi) d\xi \end{aligned} \tag{3.2}$$

for $k \in \mathbb{R}$. Here, we use the approximated signum function $\operatorname{sgn}_\rho(\xi)$ for $\rho > 0$ by $\operatorname{sgn}_\rho(\xi) = \operatorname{sgn}(\xi)$ for $|\xi| \ge \rho$ and $\operatorname{sgn}_\rho(\xi) = \sin\left(\frac{\pi}{2\rho}\xi\right)$ for $|\xi| < \rho$. Then, it can be seen that

$$\begin{aligned} \eta_\rho(u) \to |u-k|, \quad q_\rho(x,t,u) \to \operatorname{sgn}(u-k)(A(x,t,u) - A(x,t,k)), \\ r_\rho(x,t,u) \to \operatorname{sgn}(u-k)([\nabla\beta](x,t,u) - [\nabla\beta](x,t,k)) \end{aligned} \tag{3.3}$$

as $\rho \to 0$. Moreover, we set

$$\begin{aligned} [\nabla \cdot q_\rho](x,t,u) &:= \int_k^u \operatorname{sgn}_\rho(\xi - k)[\partial_\xi \nabla \cdot A](x,t,\xi) d\xi, \\ [\nabla \cdot r_\rho](x,t,u) &:= \int_k^u \operatorname{sgn}_\rho(\xi - k)[\partial_\xi \Delta\beta](x,t,\xi) d\xi. \end{aligned} \tag{3.4}$$

Then, it can be also seen that

$$\begin{aligned} [\nabla \cdot q_\rho](x,t,u) \to \operatorname{sgn}(u-k)([\nabla \cdot A](x,t,u) - [\nabla \cdot A](x,t,k)), \\ [\nabla \cdot r_\rho](x,t,u) \to \operatorname{sgn}(u-k)([\Delta\beta](x,t,u) - [\Delta\beta](x,t,k)) \end{aligned} \tag{3.5}$$

as $\rho \to 0$. Then, the entropy inequality in the definition of BV-entropy solutions implies that

$$\begin{aligned}
&\int_{\mathbb{R}_T^N} \{\eta_\rho(u)\partial_t\varphi + (q_\rho(x,t,u) - r_\rho(x,t,u)) \cdot \nabla\varphi + ([\nabla \cdot q_\rho](x,t,u) - [\nabla \cdot r_\rho](x,t,u))\varphi \\
&\qquad - \mathrm{sgn}_\rho(u-k)(\nabla\beta(x,t,u) - [\nabla\beta](x,t,u)) \cdot \nabla\varphi \\
&\qquad - \mathrm{sgn}_\rho(u-k)([\nabla \cdot A](x,t,u) - [\Delta\beta](x,t,u) + B(x,t,u))\varphi\} dxdt \\
&\quad \geq \int_{\mathbb{R}_T^N} \mathrm{sgn}'_\rho(u-k)|\sqrt{[\partial_\xi\beta](x,t,u)}Du|^2\varphi dxdt.
\end{aligned}$$

Step 1. Let u_i be the BV-entropy solution to (P) with $u_{i,0}$, A_i, B_i, β_i satisfying {A0}-{A7} and {A8}$'$. We put $k = u_2(y,s)$ and $\varphi = \phi_\delta^{\delta_0,\alpha_0}(x,y,t,s)$ (see (3.1)) in the definition of BV-entropy solution u_1. Integrating the inequality on $\mathbb{R}_T^N$ with respect to (y,s), then it follows that

$$\begin{aligned}
&\int_{(\mathbb{R}_T^N)^2} \{\eta_\rho(u_1)\partial_t\phi_\delta^{\delta_0,\alpha_0} + (q_{\rho,1}(x,t,u_1) - r_{\rho,1}(x,t,u_1)) \cdot \nabla_x\phi_\delta^{\delta_0,\alpha_0} \\
&\quad + ([\nabla_x \cdot q_{\rho,1}](x,t,u_1) - [\nabla_x \cdot r_{\rho,1}](x,t,u_1))\phi_\delta^{\delta_0,\alpha_0} \\
&\quad - \mathrm{sgn}_\rho(u_1(x,t) - u_2(y,s))((\nabla_x\beta_1(x,t,u_1) - [\nabla_x\beta_1](x,t,u_1)) \cdot \nabla_x\phi_\delta^{\delta_0,\alpha_0} \\
&\qquad\quad - ([\nabla_x \cdot A_1](x,t,u_1) - [\Delta_x\beta_1](x,t,u_1) + B_1(x,t,u_1))\phi_\delta^{\delta_0,\alpha_0})\} d\mathbf{x}d\mathbf{y} \qquad (3.6) \\
&\geq \int_{(\mathbb{R}_T^N)^2} \mathrm{sgn}'_\rho(u_1(x,t) - u_2(y,s))|\sqrt{[\partial_\xi\beta_1](x,t,u_1)}D_x u_1|^2\phi_\delta^{\delta_0,\alpha_0} d\mathbf{x}d\mathbf{y}.
\end{aligned}$$

Here, we write that $d\mathbf{x} = dxdt$ and $d\mathbf{y} = dyds$. Moreover, $q_{\rho,1}$, $[\nabla_x \cdot q_{\rho,1}]$, $r_{\rho,1}$ and $[\nabla_x \cdot r_{\rho,1}]$ are defined in (3.2) and (3.4) using A_1 and β_1, respectively. Similarly, we define the inequality (3.6)$'$ using the definition of another BV-entropy solution u_2. Moreover, we then set (EI) := (3.6) + (3.6)$'$ in what follows. The desired result is obtained by combining the estimates for (EI). In fact, using the same way in [11, Section 4] (see also [4, Section 4]), (EI) implies that

$$\begin{aligned}
&\int_{(\mathbb{R}_T^N)^2} \mathrm{sgn}(u_1-u_2)\{(u_1-u_2)(\partial_t+\partial_s)\phi_\delta^{\delta_0,\alpha_0} + (B_1(x,t,u_1) - B_2(y,s,u_2))\phi_\delta^{\delta_0,\alpha_0} \\
&\qquad + ((A_1(x,t,u_1) - A_1(x,t,u_2)) + (A_2(y,s,u_2) - A_2(y,s,u_1))) \cdot \nabla_x\phi_\delta^{\delta_0,\alpha_0} \\
&\qquad - ([\nabla_x \cdot A_1](x,t,u_2) - [\nabla_y \cdot A_2](y,s,u_1))\phi_\delta^{\delta_0,\alpha_0} \\
&\qquad - ([\nabla_x\beta_1](x,t,u_1) - [\nabla_x\beta_1](x,t,u_2)) \cdot (\nabla_x\omega_\delta)\varphi_{\alpha_0}\theta_{\delta_0} \\
&\qquad + ([\nabla_y\beta_2](y,s,u_1) - [\nabla_y\beta_2](y,s,u_2)) \cdot (\nabla_x\omega_\delta)\varphi_{\alpha_0}\theta_{\delta_0} \\
&\qquad - ([\Delta_y\beta_2](y,s,u_1) - [\Delta_x\beta_1](x,t,u_2))\phi_\delta^{\delta_0,\alpha_0}\} d\mathbf{x}d\mathbf{y} \\
&\geq \int_{(\mathbb{R}_T^N)^2} \left(\int_{u_2}^{u_1} \mathrm{sgn}(\xi-u_2)\varepsilon(x,t,y,s,\xi)d\xi\right) \varphi_{\alpha_0}\theta_{\delta_0}(\nabla_x \cdot \nabla_y\omega_\delta) d\mathbf{x}d\mathbf{y} \\
&\quad + \int_{(\mathbb{R}_T^N)^2} \left(\int_{u_2}^{u_1} \mathrm{sgn}(\xi-u_2)\nabla_y\varepsilon(x,t,y,s,\xi)d\xi\right) \cdot (\nabla_x\omega_\delta)\varphi_{\alpha_0}\theta_{\delta_0} d\mathbf{x}d\mathbf{y} \\
&\quad + \int_{(\mathbb{R}_T^N))^2} \left(\int_{u_2}^{u_1} \mathrm{sgn}(\xi-u_2)\nabla_x\varepsilon(x,t,y,s,\xi)d\xi\right) \cdot (\nabla_y\omega_\delta)\varphi_{\alpha_0}\theta_{\delta_0} d\mathbf{x}d\mathbf{y} \\
&\quad + \int_{(\mathbb{R}_T^N)^2} \left(\int_{u_2}^{u_1} \mathrm{sgn}(\xi-u_2)\nabla_x \cdot \nabla_y\varepsilon(x,t,y,s,\xi)d\xi\right) \varphi_{\alpha_0}\theta_{\delta_0}\omega_\delta d\mathbf{x}d\mathbf{y}
\end{aligned} \qquad (3.7)$$

by (3.2)-(3.5) and the properties of $\phi_\delta^{\delta_0,\alpha_0}$. Here, we set:

$$\varepsilon(x,t,y,s,\xi) := [\partial_\xi\beta_1](x,t,\xi) - 2\sqrt{[\partial_\xi\beta_1](x,t,\xi)}\sqrt{[\partial_\xi\beta_2](y,s,\xi)} + [\partial_\xi\beta_2](y,s,\xi).$$

To derive (3.7), we make the following terms:

$$\begin{aligned}
&\{-(\nabla_x\beta_1(x,t,u_1) - [\nabla_x\beta_1](x,t,u_1)) \\
&\quad +(\nabla_y\beta_2(y,s,u_2) - [\nabla_y\beta_2](y,s,u_2))\}\cdot(\nabla_x+\nabla_y)\phi_\delta^{\delta_0,\alpha_0} \\
&+(\nabla_x\beta_1(x,t,u_1) - [\nabla_x\beta_1](x,t,u_1))\cdot\nabla_y\phi_\delta^{\delta_0,\alpha_0} \\
&-(\nabla_y\beta_2(y,s,u_2) - [\nabla_y\beta_2](y,s,u_2))\cdot\nabla_x\phi_\delta^{\delta_0,\alpha_0}.
\end{aligned} \tag{3.8}$$

After that, we move the last two terms in (3.8) to the right-hand side in (EI) and use the notation $\varepsilon(x,t,y,s,\xi)$. Detailed calculation is referred to [4, 11].

Step 2. We investigate the diffusion terms in (3.7). First, the right-hand side of (3.7) is equal to

$$\begin{aligned}
&-\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\varepsilon(x,t,y,s,u_1)\varphi_{\alpha_0}(t)\theta_{\delta_0}(t-s)\nabla_y\omega_\rho\cdot Du_1 dtd\mathbf{y} \\
&-\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\varphi_{\alpha_0}(t)\theta_{\delta_0}(t-s)\omega_\rho\nabla_y\varepsilon(x,t,y,s,u_1)\cdot Du_1 dtd\mathbf{y} =: R_\delta^{\delta_0,\alpha_0},
\end{aligned}$$

using the Gauss divergence theorem. Therefore, it is deduced that

$$\begin{aligned}
\lim_{\alpha_0\to0}\lim_{\delta_0\to0} R_\delta^{\delta_0,\alpha_0} &\geq -\int_\nu^\tau\int_{(\mathbb{R}^N)^2}|\varepsilon(x,t,y,s,u_1)||\nabla_y\omega_\delta(x-y)||Du_1|dydt \\
&\quad -\int_\nu^\tau\int_{(\mathbb{R}^N)^2}|\omega_\delta(x-y)||\nabla_y\varepsilon(x,t,y,s,u_1)||Du_1|dydt =: R_\delta^1 + R_\delta^2.
\end{aligned}$$

By $\int_{\mathbb{R}^N}|\nabla_y\omega_\delta(x-y)|dy \leq \frac{C}{\delta}$ for some constant $C>0$ independent of δ, we have

$$\begin{aligned}
R_\delta^1 &\geq -2\int_\nu^\tau\int_{(\mathbb{R}^N)^2}\{\Big(\sqrt{[\partial_\xi\beta_1](x,t,u_1)} - \sqrt{[\partial_\xi\beta_2](x,t,u_1)}\Big)^2 \\
&\qquad +\Big(\sqrt{[\partial_\xi\beta_2](x,t,u_1)} - \sqrt{[\partial_\xi\beta_2](y,t,u_1)}\Big)^2\}|\nabla_y\omega_\delta(x-y)||Du_1|dydt \\
&\geq -2\{\tfrac{C(\tau-\nu)}{\delta}\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))\Big|\Big|\sqrt{[\partial_\xi\beta_1]} - \sqrt{[\partial_\xi\beta_2]}\Big|\Big|^2_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})} \\
&\qquad +C\delta(\tau-\nu)\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))\textstyle\sum_{i=1}^N\Big|\Big|\partial_{x_i}\sqrt{[\partial_\xi\beta_2]}\Big|\Big|^2_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}\}.
\end{aligned}$$

Here, we set $\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)} \equiv \mathbb{R}^N\times(\nu,\tau)\times\mathcal{U}$. In addition, it follows that

$$\begin{aligned}
R_\delta^2 &\geq -2\int_\nu^\tau\int_{(\mathbb{R}^N)^2}\Big|\nabla_y\sqrt{[\partial_\xi\beta_2](y,s,u_1)}\Big| \\
&\qquad\times\Big|\sqrt{[\partial_\xi\beta_1](x,t,u_1)} - \sqrt{[\partial_\xi\beta_2](y,t,u_1)}\Big||\omega_\delta(x-y)||Du_1|dydt \\
&\geq -2(\tau-\nu)\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))\Big|\Big|\nabla_y\sqrt{[\partial_\xi\beta_2]}\Big|\Big|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N} \\
&\quad\times(\Big|\Big|\sqrt{[\partial_\xi\beta_1]} - \sqrt{[\partial_\xi\beta_2]}\Big|\Big|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})} + \delta\textstyle\sum_{i=1}^N\Big|\Big|\partial_{x_i}\sqrt{[\partial_\xi\beta_2]}\Big|\Big|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}).
\end{aligned}$$

On the other hand, the diffusion terms of the right-hand side in (3.7) are calculated as follows:

$$\begin{aligned}
&\int_{(\mathbb{R}_T^N)^2} \operatorname{sgn}(u_1-u_2)(\{-([\nabla_x\beta_1](x,t,u_1)-[\nabla_x\beta_1](x,t,u_2))\\
&\qquad +([\nabla_y\beta_2](y,s,u_1)-[\nabla_y\beta_2](y,s,u_2))\}\cdot\nabla_x\omega_\delta(x-y)\varphi_{\alpha_0}(t)\theta_{\delta_0}(t-s)\\
&\qquad -([\Delta_y\beta_2](y,s,u_1)-[\Delta_x\beta_1](x,t,u_2))\phi_\delta^{\delta_0,\alpha_0})d\mathbf{x}d\mathbf{y}\\
&=\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\phi_\delta^{\delta_0,\alpha_0}([\partial_\xi\nabla_x\beta_1](x,t,u_1)-[\partial_\xi\nabla_y\beta_2](y,s,u_1))\cdot Du_1dtd\mathbf{y}\\
&\quad-\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\{([\Delta_y\beta_2](y,s,u_1)-[\Delta_y\beta_1](y,s,u_1))\\
&\qquad +([\Delta_y\beta_1](y,s,u_1)-[\Delta_x\beta_1](x,t,u_1))\}\phi_\delta^{\delta_0,\alpha_0}d\mathbf{x}d\mathbf{y}=:L_{\delta,\beta}^{\delta_0,\alpha_0}.
\end{aligned}$$

Then, we can see that

$$\begin{aligned}
\lim_{\alpha_0\to0}\lim_{\delta_0\to0}L_{\delta,\beta}^{\delta_0,\alpha_0}&\le(\tau-\nu)|||[\Delta_y\beta_2]-[\Delta_y\beta_1]|||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&\quad+\delta(\tau-\nu)\textstyle\sum_{i,j=1}^N|||[\partial_{x_j}\partial_{x_i}^2\beta_1]|||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&\quad+(\tau-\nu)\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))|||[\partial_\xi\nabla_x\beta_1]-[\partial_\xi\nabla_x\beta_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}\\
&\quad+\delta(\tau-\nu)\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))\textstyle\sum_{i,j=1}^N|||[\partial_{x_i}\partial_\xi\partial_{x_j}\beta_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}.
\end{aligned}$$

Step 3. We investigate the convection terms in (3.7) as follows:

$$\begin{aligned}
&\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\{((A_1(x,t,u_1)-A_1(x,t,u_2))\\
&\qquad+(A_2(y,s,u_2)-A_2(y,s,u_1)))\cdot\nabla_x\phi_\delta^{\delta_0,\alpha_0}\\
&\qquad-([\nabla_x\cdot A_1](x,t,u_2)-[\nabla_yA_2](y,s,u_1))\phi_\delta^{\delta_0,\alpha_0}\}d\mathbf{x}d\mathbf{y}\\
&=-\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)([\partial_\xi A_1](x,t,u_1)-[\partial_\xi A_2](y,s,u_1))\phi_\delta^{\delta_0,\alpha_0}Du_1dtd\mathbf{y}\\
&\quad+\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)\{([\nabla_y\cdot A_2](y,s,u_1)-[\nabla_y\cdot A_1](y,s,u_1))\\
&\qquad+([\nabla_y\cdot A_1](y,s,u_1)-[\nabla_x\cdot A_1](x,t,u_1))\}\phi_\delta^{\delta_0,\alpha_0}d\mathbf{x}d\mathbf{y}=:L_{\delta,A}^{\delta_0,\alpha_0}.
\end{aligned}$$

Hence, we deduce that

$$\begin{aligned}
\lim_{\alpha_0\to0}\lim_{\delta_0\to0}L_{\delta,A}^{\delta_0,\alpha_0}&\le(\tau-\nu)\sup_{t\in(\nu,\tau)}TV(u_1(\cdot,t))|||[\partial_\xi A_1]-[\partial_\xi A_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}\\
&\quad+(\tau-\nu)|||[\nabla_y\cdot A_2]-[\nabla_y\cdot A_1]|||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&\quad+\delta(\tau-\nu)\textstyle\sum_{i,j=1}^N||\partial_{x_j}\partial_{x_i}A_1^i||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}.
\end{aligned}$$

On the other hand, it can be seen that

$$\begin{aligned}
&\int_{(\mathbb{R}_T^N)^2}|u_1-u_2|(\partial_t+\partial_s)\phi_\delta^{\delta_0,\alpha_0}d\mathbf{x}d\mathbf{y}\\
&=\int_{(\mathbb{R}_T^N)^2}(|u_1(x,t)-u_1(y,t)|+|u_1(y,t)-u_2(y,t)|+|u_2(y,t)-u_2(y,s)|)\\
&\qquad\times(\theta_{\alpha_0}(t-\nu)-\theta_{\alpha_0}(t-\tau))\theta_{\delta_0}(t-s)\omega_\delta(x-y)d\mathbf{x}d\mathbf{y}=:L_\delta^{\delta_0,\alpha_0}.
\end{aligned}$$

Then, it is deduced that

$$\begin{aligned}
&\lim_{\alpha_0\to 0}\lim_{\delta_0\to 0} L_\delta^{\delta_0,\alpha_0}\\
&=\int_{(\mathbb{R}^N)^2}(|u_1(x,\nu)-u_1(y,\nu)|-|u_1(x,\tau)-u_1(y,\tau)|)\omega_\delta(x-y)dxdy\\
&\quad+\int_{(\mathbb{R}^N)^2}(|u_1(y,\nu)-u_2(y,\nu)|-|u_1(y,\tau)-u_2(y,\tau)|)\omega_\delta(x-y)dxdy\\
&\le 2\delta\sup_{t\in(\nu,\tau)} TV(u_1(\cdot,t))+||u_1(y,\nu)-u_2(y,\nu)||_{L^1(\mathbb{R}^N)}\\
&\quad-||u_1(y,\tau)-u_2(y,\tau)||_{L^1(\mathbb{R}^N)}.
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
&\lim_{\alpha_0\to 0}\lim_{\delta_0\to 0}\int_{(\mathbb{R}_T^N)^2}\operatorname{sgn}(u_1-u_2)(B_1(x,t,u_1)-B_2(y,s,u_2))\phi_\delta^{\delta_0,\alpha_0}d\mathbf{x}d\mathbf{y}\\
&\le(\tau-\nu)||B_1-B_2||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}+\delta(\tau-\nu)\textstyle\sum_{i=1}^N||\partial_{x_i}B_2||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&\quad+||\partial_\xi B_2||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}\int_\nu^\tau\int_{\mathbb{R}^N}|u_1-u_2|d\mathbf{x}.
\end{aligned}$$

Step 4. According to the above estimates and Theorem 2.3 (I), we see that

$$\begin{aligned}
&||u_1(y,\tau)-u_2(y,\tau)||_{L^1(\mathbb{R}^N)}\le||u_1(y,\nu)-u_2(y,\nu)||_{L^1(\mathbb{R}^N)}+(\alpha_1^{\nu,\tau}+\alpha_2^{\nu,\tau}\delta)(\tau-\nu)\\
&\quad+||\partial_\xi B_2||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}\int_\nu^\tau\int_{\mathbb{R}^N}|u_1-u_2|d\mathbf{x}+2\delta e^{C_0\tau}(TV(u_{1,0})+C_1)\\
&\quad+\tfrac{C(\tau-\nu)}{\delta}e^{C_0\tau}(TV(u_{1,0})+C_1)\left|\left|\sqrt{[\partial_\xi\beta_1]}-\sqrt{[\partial_\xi\beta_2]}\right|\right|^2_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})},
\end{aligned}$$

where $\alpha_1^{\nu,\tau}$ and $\alpha_2^{\nu,\tau}$ are constants depending ν and τ which are defined as follows

$$\begin{aligned}
\alpha_1^{\nu,\tau}:=&||B_1-B_2||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}+|||[\nabla_y\cdot A_1]-[\nabla_y\cdot A_2]|||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&+|||[\Delta_y\beta_1]-[\Delta_y\beta_2]|||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}+e^{C_0\tau}(TV(u_{1,0})+C_1)\\
&\times\{|||[\partial_\xi\nabla_x\beta_1]-[\partial_\xi\nabla_x\beta_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}+|||[\partial_\xi A_1]-[\partial_\xi A_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}\\
&+2\left|\left|\nabla_y\sqrt{[\partial_\xi\beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}\left|\left|\sqrt{[\partial_\xi\beta_1]}-\sqrt{[\partial_\xi\beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}\}
\end{aligned}$$

and

$$\begin{aligned}
\alpha_2^{\nu,\tau}:=&\textstyle\sum_{i=1}^N||\partial_{x_i}B_2||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}+\sum_{i,j=1}^N||\partial_{x_j}\partial_{x_i}A_1^i||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}\\
&+\textstyle\sum_{i,j=1}^N||\partial_{x_j}\partial_{x_i}^2\beta_1||_{L^\infty((\nu,\tau)\times\mathcal{U};L^1(\mathbb{R}^N))}+e^{C_0\tau}(TV(u_{1,0})+C_1)\\
&\times\{\textstyle\sum_{i,j=1}^N|||[\partial_{x_i}\partial_\xi\partial_{x_j}\beta_2]|||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}+\sum_{i=1}^N\left|\left|\partial_{x_i}\sqrt{[\partial_\xi\beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}\\
&\times(2\left|\left|\nabla_y\sqrt{[\partial_\xi\beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})^N}+C\textstyle\sum_{i=1}^N\left|\left|\partial_{x_i}\sqrt{[\partial_\xi\beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})})\}.
\end{aligned}$$

Here, we set $\delta = \sqrt{\tau-\nu}\left|\left|\sqrt{[\partial_\xi \beta_1]} - \sqrt{[\partial_\xi \beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}$. Then, we obtain

$$
\begin{aligned}
&||u_1(y,\tau) - u_2(y,\tau)||_{L^1(\mathbb{R}^N)} \leq ||u_1(y,\nu) - u_2(y,\nu)||_{L^1(\mathbb{R}^N)} \\
&\quad + ||\partial_\xi B_2||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})} \int_\nu^\tau \int_{\mathbb{R}^N} |u_1 - u_2| dxdt + \alpha_1^{\nu,\tau}(\tau-\nu) + \sqrt{\tau-\nu} \\
&\quad \times (\alpha_2^{\nu,\tau}(\tau-\nu) + (C+2)e^{C_0\tau}(TV(u_{1,0}) + C_1)) \left|\left|\sqrt{[\partial_\xi \beta_1]} - \sqrt{[\partial_\xi \beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(\nu,\tau)})}.
\end{aligned}
$$

Consequently, we conclude that

$$
\begin{aligned}
||u_1(y,t) - u_2(y,t)||_{L^1(\mathbb{R}^N)} \leq e^{\alpha' t}||u_1(y,0) - u_2(y,0)||_{L^1(\mathbb{R}^N)} + \tfrac{(e^{\alpha' t}-1)}{\alpha'}\alpha_1^{0,T} \\
+ \widehat{C}\sqrt{t}e^{\alpha' t}\left|\left|\sqrt{[\partial_\xi \beta_1]} - \sqrt{[\partial_\xi \beta_2]}\right|\right|_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(0,T)})},
\end{aligned}
$$

where $\alpha' := \max_{i=1,2}\{||\partial_\xi B_i||_{L^\infty(\mathbb{R}^{N,\mathcal{U}}_{(0,T)})}\}$ and $\widehat{C} := \alpha_2^{0,T}T + (C+2)e^{C_0 T}(TV(u_{1,0}) + C_1)$.

Acknowledgments. The author would like to express my hearty gratitude to the anonymous referee for the valuable comments and suggestions.

REFERENCES

[1] L. Ambrosio, G. Crasta, V. De Cicco and G. De Philippis, *A nonautonomous chain rule in $W^{1,p}$ and BV*, Manuscripta Math. **140** (2013) no.3, 461–480.

[2] L. Ambrosio, N. Fusco and D. Pallara, "Functions of Bounded Variation and Free Discontinuity Problems", Oxford Science Publications, (2000).

[3] J. Carrillo, *Entropy solutions for nonlinear degenerate problems*, Arch. Rational. Anal., **147** (1999), 269-361.

[4] G. Q. Chen and K. H. Karlsen, *Quasilinear anisotropic degenerate parabolic equations with time-space dependent diffusion coefficients*, Commun. Pure Appl. Anal., **4** (2005), 241–266.

[5] G. Q. Chen and B. Perthame, *Well-posedness for non-isotropic degenerate parabolic-hyperbolic equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **20**(4) (2003), 645–668.

[6] L. C. Evans and R. Gariepy, "Measure theory and fine properties of functions", Studies in Advanced Math., CRC Press, London, (1992).

[7] K. H. Karlsen, N. H. Risebro, *On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients*, Discrete Contin. Dyn., **9**(5) (2003), 1081–1104.

[8] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, **10** (1970), 217-243.

[9] A. I. Vol'pert and S. I. Hudjaev, Cauchy's problem for degenerate second order quasi-linear parabolic equations, Math. USSR-Sb., **7** (1969), 365-387.

[10] H. Watanabe, *Entropy solutions to strongly degenerate parabolic equations with zero-flux boundary conditions*, Adv. Math. Sci. Appl., **23** (2013), no.1, 209–234.

[11] H. Watanabe, *Strongly degenerate parabolic equations with variable coefficients*, Adv. Math. Sci. Appl., **26** (2017), 143–173.

[12] H. Watanabe and S. Oharu, *BV-entropy solutions to strongly degenerate parabolic equations*, Adv. Differential Equations **15**(7-8) (2010), 757–800.

[13] H. Watanabe and S. Oharu, *Finite-difference approximations to a class of strongly degenerate parabolic equations*, Adv. Math. Sci. Appl., **20** (2010), no.2, 319–347.

[14] W. P. Ziemer, "Weakly differentiable functions", Springer-Verlag, New York, (1989).

Proceedings of EQUADIFF 2017
pp. 325–330

NUMERICAL STUDY ON THE BLOW-UP RATE TO A QUASILINEAR PARABOLIC EQUATION *

KOICHI ANADA†, TETSUYA ISHIWATA‡, AND TAKEO USHIJIMA§

Abstract. In this paper, we consider the blow-up solutions for a quasilinear parabolic partial differential equation $u_t = u^2(u_{xx}+u)$. We numerically investigate the blow-up rates of these solutions by using a numerical method which is recently proposed by the authors [3].

Key words. blow-up rate, type II blow-up, numerical estimate, scale invariance, rescaling algorithm, curvature flow

AMS subject classifications. 35B44, 35K59, 65M99

1. Introduction. In this paper we consider the following quasilinear parabolic partial differential equation:

$$u_t = u^2(u_{xx} + u), \quad x \in (-a, a) \subset \mathbb{R}, t > 0. \tag{1.1}$$

This equation describes the motion of curves by their curvature ([2, 5, 6]). For this equation it was shown that there exist finite time blow-up solutions of so-called Type II under the following initial and boundary conditions

$$\begin{cases} u(t, \pm a) = 0, & t > 0, \\ u(0, x) = u_0(x), & x \in [-a, a] \end{cases} \tag{1.2}$$

where $a > \pi/2$ ([1, 8]). Here, we call Type I the blow-up solutions with the blow-up rate $(T-t)^{-1/2}$ which is determined by the spatially uniform blow-up solution of the equation (1.1). The non Type I blow-up solutions are called Type II. In [2] Anada & Ishiwata proved that there exists a solution which blows up in a finite time T with the blow-up rate

$$\left(\frac{1}{(T-t)} \log \log \frac{1}{(T-t)} \right)^{\frac{1}{2}}. \tag{1.3}$$

In [2], they posed several assumptions on the initial function u_0:

(K1) $u_0(x) > 0, \quad x \in (-a, a)$,
(K2) there exists $A > 0$ such that $(u_0(x))^2 + (u_{0x}(x))^2 < A^2, \quad x \in (-a, a)$,
(K3) $u_0(-x) = u_0(x)$,
(K4) $(u_{0x})(x) < 0, \quad x \in (0, a)$,
(K5) $Z\left(\frac{d}{dx}\left[u_0(u_{0x} + u_0)\right], (-a, a)\right) \leq 3$.

*This work was supported by KAKENHI (No.15H03632, No.15K13461, No.16H03953)

†Waseda University Senior High School, 3-31-1 Kamishakujii, Nerima-ku, Tokyo 177-0044, Japan, anada-koichi@waseda.jp

‡Department of Mathematical Sciences, Shibaura Institute of Technology, 307 Fukasaku, Minuma-ku, Saitama 337-8570, Japan, tisiwata@shibaura-it.ac.jp

§Deparment of Mathematics, Faculty of Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba 278-8510, Japan, ushijima_takeo@ma.noda.tus.ac.jp

Here, $Z(f(x), I)$ is the zero number of f on the interval $I \subset \mathbb{R}$, namely, the number of zeros of f in the interval I and $u_{0x} = \frac{d}{dx}u_0$. To our best knowledge, it is known nothing about the blow-up rates of the Type II blow-up solutions whose initial data do not satisfy the assumptions (K1)–(K5).

There are many evolution equations which has the scaling invariance:

(*) there exist α, β such that if $u(t,\cdot)$ solves the equation then for any $\lambda > 0$, $u^\lambda(t,\cdot) = \lambda^\alpha u(\lambda^\beta t, \cdot)$ solves the same equation.

The equation (1.1) satisfies the scaling invariance (*) in the case of $\beta = 2\alpha$. Recently, we proposed a numerical method to estimate the blow-up rates of the blow-up solutions for evolution equations which possess this scaling invariance ([3]). In this paper, we numerically investigate the blow-up rates of the solutions of the problem (1.1) and (1.2). For this purpose, we adopt our numerical method to this problem.

The organization of the paper is as follows: in section 2, we explain our numerical method, in section 3 we exhibit several numerical examples, in the last section we will give a conclusion at this moment and several remarks.

2. Algorithm using the scaling invariance. We consider the problems which satisfy the following conditions:

- The solution u blows up in a finite time, say T, namely, there exists $T < \infty$ such that $\lim_{t\to T} \|u(t)\| = \infty$.
- The equation has a scaling invariance (*)

Here and hereafter, we use the notation $u(t)$ considering the solution u of the problem as a function from time interval to a normed space X with a norm $\|\cdot\|$. For example, the equation

$$\frac{du}{dt} = u^p \tag{2.1}$$

with a positive initial data satisfies the assumptions above. In fact, the equation possesses blow-up solutions and for a solution $u(t)$ of this equation and for any $\lambda > 0$ if we set

$$u^\lambda(t) = \lambda^\alpha u(\lambda^\beta t), \quad \alpha = \frac{\beta}{p-1}$$

then u^λ is also a solution of (2.1). For this problem $X = (\mathbb{R}, |\cdot|)$. As we already noted, our problem (1.1) also satisfies the assumptions above. For (1.1) we choose $X = L^\infty(-a, a)$.

2.1. Rescaling algorithm. We explain our proposed method. First, we fix constants $M > 0$ and λ $(0 < \lambda < 1)$. Second, by using the scaling invariance (*) we repeat to rescale the solution and construct $\{t_m\}$ and $\{\tau_m\}$ as follows:

$$t_0 = 0, \quad t_m = \min\{\ t \mid\ (\lambda)^{-\alpha(m-1)} M = \|u(t)\|\}\ (m = 1, 2, 3, \dots),$$

$$\lambda^{\beta(m-1)}\tau_m = (t_m - t_{m-1}). \tag{2.2}$$

Then we have

$$T = \sum_{k=1}^{\infty} (\lambda^\beta)^{k-1}\tau_k, \quad (T - t_m) = \lambda^{\beta(m-1)} \sum_{l=1}^{\infty} \lambda^{\beta l}\tau_{l+m},$$

$$(T-t_m)^{\frac{\alpha}{\beta}}\|u(t_m)\| = \left(\sum_{l=1}^{\infty}\lambda^{\beta l}\tau_{l+m}\right)^{\frac{\alpha}{\beta}} M =: (f(m))^{\frac{\alpha}{\beta}} M, \tag{2.3}$$

$$(T-t_m) = \lambda^{\beta(m-1)} f(m). \tag{2.4}$$

Using these equations we can prove a relation between the sequence $\{\tau_m\}$ and the blow-up rate of the solution (Theorem 2.1). Third, using appropriate numerical approximation for the problem, we numerically construct the sequence $\{\tau_m\}$ for finite ms and observe the behavior of it with respect to m. Such kind of numerical method is called rescaling algorithm which is originally proposed by Berger & Kohn [7] for the blow-up problem of Fujita type. At last, supposing the behavior of $\{\tau_m\}$ as $m \to \infty$ is same as the observed one we estimate the blow-up rate from the Theorem 2.1.

In [3], we examined the effectiveness of our method by applying our method to several examples where the blow-up rates of the solutions are theoretically known. For all of these examples, we could estimate the blow-up rates correctly, especially, we could estimate not only the blow-up rates of simple power type (Type I) but also the blow-up rates of complex forms with log or log log (Type II).

2.2. Relation between τ_m and the blow-up rate. We proved in [3] the following relation between the sequence $\{\tau_m\}$ and the blow-up rate:

THEOREM 2.1.

1. *if $\tau_m = O(1)$ then the blow-up rate is $O((T-t_m)^{-\frac{\alpha}{\beta}})$,*
2. *if $\tau_m = Cm^k + o(m^k)$ for some integer k then the blow-up rate is $O((T-t_m)^{-\frac{\alpha}{\beta}}(\log(T-t_m)^{-1})^{k\frac{\alpha}{\beta}})$,*
3. *if $\tau_m = C\log m + o(\log m)$ then the blow-up rate is $O((T-t_m)^{-\frac{\alpha}{\beta}}(\log\log(T-t_m)^{-1})^{\frac{\alpha}{\beta}})$,*
4. *if $\tau_m = O(e^{km})$ then the blow-up rate is $O((T-t_m)^{r(\lambda)})$, $\quad r(\lambda) = \dfrac{\alpha\log\frac{1}{\lambda}}{k-\beta\log\frac{1}{\lambda}}$.*

REMARK 2.1. *Although it is not complete classification for the behavior of $\{\tau_m\}$, it is still useful. We note that this relation holds for the exact solution of the problem and the numerically constructed $\{\tau_m\}$ inevitably contains errors. We need to observe the behavior of $\{\tau_m\}$ for finite m.*

3. Numerical experiments. Now we exhibit several results of numerical experiments where the initial condition does not satisfy at least one of (K1)–(K5). Here we note that we use the standard finite difference scheme for numerical solutions of (1.1) and we set the numerical parameters as $\lambda = 0.5$ and $M = 2$ for all examples below.

In the figures 3.1 and 3.2 we plot the results which breaks the condition (K4) and (K5). In these figures, the horizontal axis is the number of rescaling times m and the vertical axis is $\exp(\tau_m)/m^\alpha$ for several α. In the figure 3.1, we plot the case where $a = \pi$ and $u_0(x) = 0.5(\cos(x/2) + \sin^2 x)$ with $\alpha = 1.1, 1.4, 1.5, 1.6$. Here, we note that we do not need to determine the precise value of α for evaluating the behavior of τ_m. Indeed if we know that

$$\frac{\exp(\tau_m)}{m^\alpha} \sim 1$$

then we have

$$\tau_m \sim \log m^\alpha = \alpha\log m \sim \log m.$$

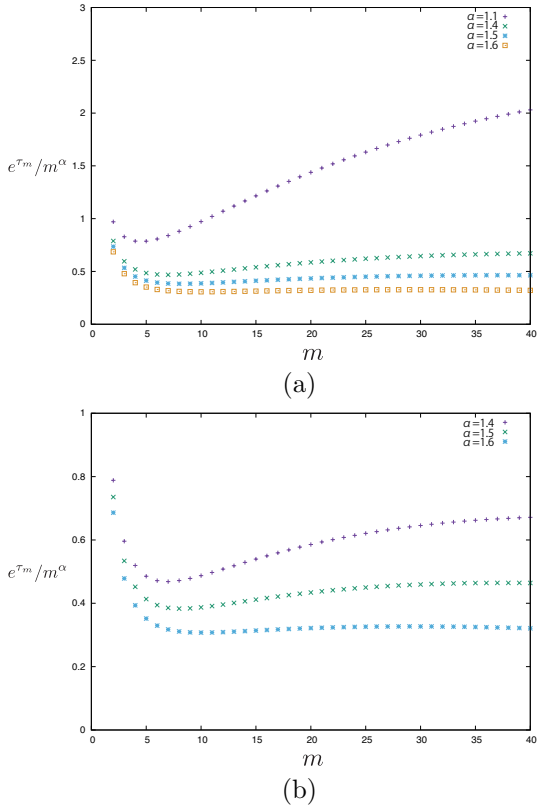


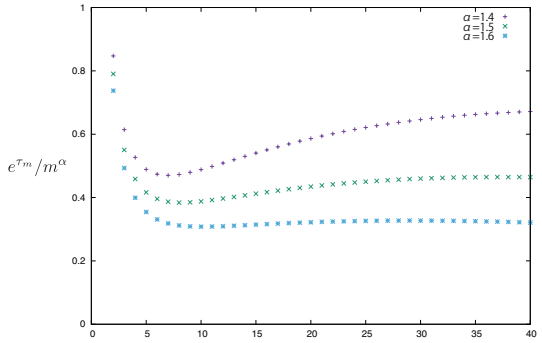
FIG. 3.1. *(a)* $\exp(\tau_m)/m^\alpha (\alpha = 1.1, 1.4, 1.5, 1.6)$*, (b) zoomed version of (a) for* $\alpha = 1.4, 1.5, 1.6$.

Here $A \sim B$ means $c_1 A \leq B \leq c_2 A$ for some $c_1, c_2 > 0$. Hence we can apply the third part of the Theorem 2.1 and we can conclude the blow-up rate is of log log type. From the figure, we can see that in the case where $\alpha = 1.1$, $\exp(\tau_m)/m^\alpha$ increases and from the figure 3.2 (b) we can suppose that $\exp(\tau_m)/m^\alpha$ is bounded from both bellow and above for some value of α between 1.4 and 1.6, namely we can suppose $\tau_m = O(\log m)$.

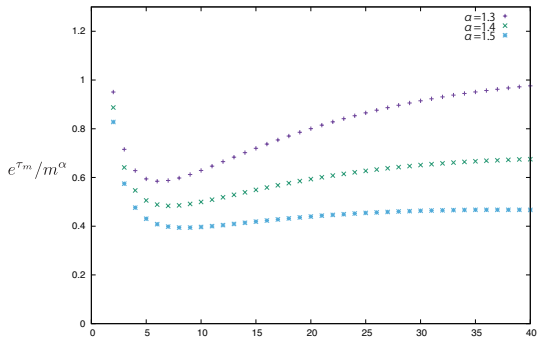
In the figures 3.2(a) and (b), we plot the case where $a = \pi, u_0(x) = 0.5(\cos(x/2) + \sin^2 2x)$, $a = \pi, u_0(x) = 0.5(\cos(x/2) + \sin^2 4x)$ and $a = 2\pi, u_0(x) = \sin\left(\pi\left(\frac{x}{2\pi}\right)^8\right) + 0.01\cos(x/4)$. The initial data (a) and (b) have much undulation than the previous one and initial data (c) has a peak near the boundary. From these figures, we can also observe that $\tau_m = O(\log m)$.

Thus, the blow-up rate of all these numerical solutions are estimated as (1.3).

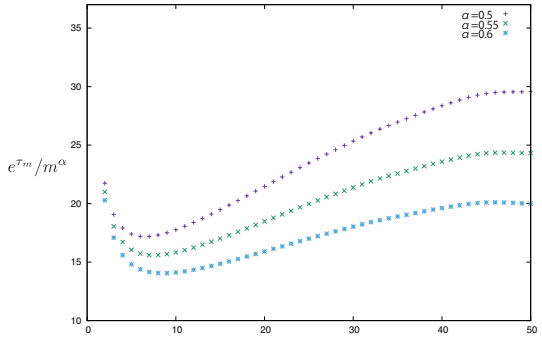
4. Conclusion and remarks. In this paper we numerically estimate the blow-up rate of type II solutions for the problem (1.1) and (1.2) as (1.3). Thus we conclude that the estimate (1.3) may be valid for wider class of initial data. On the other hand it is still open whether the other blow-up rates of type II blow-up solutions exist or not. Moreover, there are no information on the blow-up rate of type II blow-up solutions to $u_t = u^\delta(u_{xx} + u)$ for the case $\delta > 2$ and the higher dimensional problem: $u_t = u^\delta(\Delta u + u)$ with $\delta \geq 2$. These are challenging issues.



(a)



(b)



(c)

FIG. 3.2. *We plot* m *vs.* $\exp(\tau_m)/m^\alpha$. *(a)* $a = \pi, u_0(x) = 0.5(\cos(x/2) + \sin^2 2x)$ *(*$\alpha = 1.4, 1.5, 1.6$*), (b)* $a = \pi, u_0(x) = 0.5(\cos(x/2) + \sin^2 4x)$ *(*$\alpha = 1.3, 1.4, 1.5$*). (c)* $a = 2\pi, u_0(x) = \sin\left(\pi\left(\frac{x}{2\pi}\right)^8\right) + 0.01\cos(x/4)$ *(*$\alpha = 0.5, 0.55, 0.6$*).*

REFERENCES

[1] K. ANADA, I. FUKUDA AND M. TSUTSUMI, *Regional blow-up and decay of solutions to the Initial-Boundary value problem for* $u_t = uu_{xx} - \gamma(u_x)^2 + ku^2$, Funkcialaj Ekvacioj, 39 (1996), pp. 363–387.

[2] K. ANADA AND T. ISHIWATA, *Blow-up rates of solutions of initial-boundary value problems for a quasi-linear parabolic equation*, J. Differential Equations, 262 (2017), pp. 181–271.

[3] K. ANADA, T. ISHIWATA AND T. USHIJIMA, *A numerical method of estimating blow-up rates*

for nonlinear evolution equations by using rescaling algorithm, to appear in Japan J. Ind. Appl. Math.

[4] B. Andrews, *Singularities in crystalline curvature flows*, Asian J. Math., 6 (2002), pp. 101–122.

[5] S. B. Angenent, *On the formation of singularities in the curve shortening flow*, J. Diff. Geo., 33 (1991), pp. 601–633.

[6] S. B. Angenent and J. J. L. Velázquez, *Asymptotic shape of cusp singularities in curve shortening*, Duke Math. J., 77 (1995), pp. 71–110.

[7] M. Berger and R. V. Kohn, *A rescaling algorithm for the numerical calculation of blowing-up solutions*, Cmmm. Pure Appl. Math., 41 (1988), pp. 841–863.

[8] A. Friedman and B. McLeod, *Blow-up of solutions of nonlinear degenerate parabolic equations*, Arch. Rational Mech. Anal., 96 (1987), pp. 55–80.

[9] T. Ishiwata and S. Yazaki, *On the blow-up rate for fast blow-up solutions arising in an anisotropic crystalline motion*, J. Comput. Appl. Math., 159 (2003), pp. 55–64.

[10] P. A. Watterson, *Force-free magnetic evolution in the reversed-field pinch*, Thesis, Cambridge University (1985).

[11] M. Winkler, *Blow-up in a degenerate parabolic equation*, Indiana Univ. Math. J., 53 (2004), pp. 1415–1442.

Proceedings of EQUADIFF 2017
pp. 331–340

TWO METHODS FOR OPTICAL FLOW ESTIMATION*

PETER FROLKOVIČ† AND VIERA KLEINOVÁ.

Abstract. In this paper we describe two methods for optical flow estimation between two images. Both methods are based on the backward tracking of characteristics for advection equation and the difference is on the choice of advection vector field. We present numerical experiments on 2D data of cell nucleus.

Key words. optical flow, advection equation, level-set motion, characteristic curves

AMS subject classifications. 35L45, 65M25, 68U10

1. Introduction. Optical flow is an important topic in various fields including computer vision and image processing. It is a technique that is based on estimating a deformation between images of video sequence.

The most popular methods for optical flow estimation are so-called differential methods [1, 2, 3, 4, 5, 9]. These methods are based on spatial derivatives of images. We are interested in two approaches.

The first approach is based on the method created by Lucas and Kanade [5] where it is assumed that the optical flow is constant locally within some neighborhood of each pixel in images. This approach is extended to a nonlocal form by e.g. Horn and Schunck [3] where the optical flow is estimated globally over entire image.

The second approach is based on level-set formulation [6] and it is directly motivated by models described by Sapiro et al. [1] and Vemuri et al. [9]. The methods are appropriate especially to estimate a non-rigid deformation when the objects in images change their shape.

The main goal of this paper is to apply both approaches with the backward tracking of characteristic curves for related advection equation to estimate the optical flow together with their numerical implementation. Moreover we show the results obtained by Lucas-Kanade method and the method based on level-set motion and we suggest their combination for some type of images.

2. Formulation of optical flow. Let us represent two images $F(\mathbf{x}) \in \mathbf{R}$ and $G(\mathbf{x}) \in \mathbf{R}$ by functions of intensity on a domain $\Omega \subset \mathbf{R}^2$ for $\mathbf{x} \in \Omega$ and $\mathbf{x} = (x, y)$. The main goal of optical flow estimation is to find a deformation $\vec{U}(\mathbf{x})$ between $F(\mathbf{x})$ and $G(\mathbf{x})$ such that $F(\mathbf{x} - \vec{U}(\mathbf{x})) = G(\mathbf{x})$.

The basic idea of our approach is to search for a function $f = f(\mathbf{x}, t)$ that fulfils the advection equation

$$\partial_t f(\mathbf{x}, t) + \vec{u}(\mathbf{x}, t) \cdot \nabla f(\mathbf{x}, t) = 0, \qquad f(\mathbf{x}, 0) = F(\mathbf{x}), \tag{2.1}$$

for $t \in [0, T]$ where $T > 0$ and $\vec{u} = \vec{u}(\mathbf{x}, t) = (u(\mathbf{x}, t), v(\mathbf{x}, t))$ has to be specified such that $f(\mathbf{x}, T) = G(\mathbf{x})$.

*This work was supported by VEGA 1/0728/15. The second author would like to thank for financial contribution from the STU Grant scheme for Support of Young Researchers.

†Department of Mathematics and Descriptive Geometry, Slovak University of Technology, Radlinského 11, 810 05 Bratislava, Slovakia (peter.frolkovic@stuba.sk).

Once the advection equation (2.1) is solved, the characteristic curves $X(\mathbf{x},\tilde{t};t)$ generated by $\vec{u}$ can be used that are obtained as solutions of ordinary differential equations

$$\dot{X}(\mathbf{x},\tilde{t};t) = \vec{u}(X(\mathbf{x},\tilde{t};t),t)\,, \qquad X(\mathbf{x},\tilde{t};\tilde{t}) = \mathbf{x}\,, \tag{2.2}$$

for $\tilde{t} \in [0,T]$ and $\mathbf{x} \in \Omega$. The value $X(\mathbf{x},\tilde{t};t)$ is a position X of characteristic curve at time t such that the position at time $\tilde{t}$ is $\mathbf{x}$.

For the solution $f(\mathbf{x},t)$ of advection equation (2.1) we see that the time derivative of $f(X(\mathbf{x},\tilde{t};t),t)$ vanishes along $X(\mathbf{x},\tilde{t};t)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(X(\mathbf{x},\tilde{t};t),t) &= \partial_t f(X(\mathbf{x},\tilde{t};t),t) + \dot{X}(\mathbf{x},\tilde{t};t)\cdot\nabla f(X(\mathbf{x},\tilde{t};t),t)\\ &= \partial_t f(X(\mathbf{x},\tilde{t};t),t) + \vec{u}(X(\mathbf{x},\tilde{t};t),t)\cdot\nabla f(X(\mathbf{x},\tilde{t};t),t) = 0\,.\end{aligned}$$

We can conclude that $f(\mathbf{x},t)$ is constant along the characteristics.

In this paper we use the backward tracking of characteristics to compute the solution f of (2.1) by

$$f(\mathbf{x},\tilde{t}) = F(X(x,\tilde{t};0)) \tag{2.3}$$

for $\tilde{t} > 0$. Consequently, the deformation $\vec{U}(\mathbf{x})$ is defined for $\tilde{t} = T$ by

$$\vec{U}(\mathbf{x}) = \mathbf{x} - X(\mathbf{x},T;0)\,. \tag{2.4}$$

Next we describe two methods how to obtain the vector field $\vec{u}$.

2.1. Lucas-Kanade method. The method belongs to local methods and it solves the advection equation (2.1) for unknowns $\vec{u} = (u,v)$ separately for each $\mathbf{x} \in \Omega$ and $t = 0,1,\dots$. The solution is found starting with $t = 0$ by minimizing the function

$$H(u,v) = W_\sigma * (\partial_x f u + \partial_y f v + \partial_t f)^2\,, \tag{2.5}$$

where $*$ denotes the convolution. Here W_σ is a weight function and it is usually set to a Gaussian of a standard deviation σ [5]. The minimum of $H(u,v)$ is reached if $\partial_u H(u,v) = 0$ and $\partial_v H(u,v) = 0$

$$\begin{aligned} 0 &= W_\sigma * [2(\partial_x f u + \partial_y f v + \partial_t f)\partial_x f]\,,\\ 0 &= W_\sigma * [2(\partial_x f u + \partial_y f v + \partial_t f)\partial_y f]\,.\end{aligned}$$

The unknowns $(u(\mathbf{x},t), v(\mathbf{x},t))$ are obtained from the linear system in the form

$$\begin{pmatrix} W_\sigma * (\partial_x f)^2 & W_\sigma * (\partial_x f \partial_y f) \\ W_\sigma * (\partial_y f \partial_x f) & W_\sigma * (\partial_y f)^2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -W_\sigma * (\partial_x f \partial_t f) \\ -W_\sigma * (\partial_y f \partial_t f) \end{pmatrix}. \tag{2.6}$$

Once the vector field $\vec{u}$ is found for each $\mathbf{x} \in \Omega$, one can compute $f(\mathbf{x},t+1)$ using (2.3) and the method can return to (2.5) to compute $\vec{u}(\mathbf{x},t+1)$ and so on.

2.2. Method based on level-set motion. The method is motivated by Sapiro et al. [1] and Vemuri et al. [9]. In this case we consider $\vec{u}$ in the advection equation (2.1) of the form

$$\vec{u}(\mathbf{x},t) = \begin{cases} -S(\mathbf{x},t)\frac{\nabla f(\mathbf{x},t)}{|\nabla f(\mathbf{x},t)|} & |\nabla f(\mathbf{x},t)| \neq 0 \\ \vec{0} & |\nabla f(\mathbf{x},t)| = 0\,, \end{cases} \tag{2.7}$$

where $S(\mathbf{x},t)$ is a speed in normal direction, so the equation (2.1) can be rewritten in the form

$$\partial_t f(\mathbf{x},t) = S(\mathbf{x},t)|\nabla f(\mathbf{x},t)|\,, \qquad f(\mathbf{x},0) = F(\mathbf{x})\,. \tag{2.8}$$

The natural choice for the speed $S(\mathbf{x},t)$ is

$$S(\mathbf{x},t) = \alpha(\mathbf{x},t)(G(\mathbf{x}) - f(\mathbf{x},t))\,, \tag{2.9}$$

where $\alpha(\mathbf{x},t) > 0$ is a free parameter function that will be defined conveniently in numerical method later. Using (2.7) the advection equation (2.1), resp. (2.8), is solved directly for the unknown function f and $\vec{u}$ is determined from (2.7).

3. Numerical implementation. We assume 2D gray scale images with centers $\mathbf{x}_{ij} = (i,j)$ of pixels for $i = 0,\dots,I-1$ and $j = 0,\dots,J-1$. The distance between two centers $\mathbf{x}_{i+1j} - \mathbf{x}_{ij}$ and $\mathbf{x}_{ij+1} - \mathbf{x}_{ij}$ is 1 and the time points are chosen to be $t^n = n$ for $n = 0,1,\dots$. The images $F(\mathbf{x})$ and $G(\mathbf{x})$ for $\mathbf{x} \in \Omega$ are represented by bilinear interpolation of the discrete values of their intensities $F_{ij} = F(\mathbf{x}_{ij})$ and $G_{ij} = G(\mathbf{x}_{ij})$. The main goal of this paper is to approximate the deformation $\vec{U}(\mathbf{x})$ such that $G_{ij} \approx F(\mathbf{x}_{ij} - \vec{U}_{ij})$, where $\vec{U}_{ij} \approx \vec{U}(\mathbf{x}_{ij})$.

Once the vector field $\vec{u}$ is approximated by discrete values $\vec{u}_{ij}^n \approx \vec{u}(\mathbf{x}_{ij},t^n)$, see later, the characteristic curves are approximated by $X_{ij}^{n,m} \approx X(\mathbf{x}_{ij},t^n;t^m)$ at any time t^n by numerical approximation of (2.2) for $m = n-1,\dots,0$, namely

$$X_{ij}^{n,m} = X_{ij}^{n,m+1} - \vec{u}^m(X_{ij}^{n,m+1})\,, \qquad X_{ij}^{n,n} = \mathbf{x}_{ij}\,, \tag{3.1}$$

where $\vec{u}^m(\mathbf{x})$ is the bilinear interpolation of discrete values $\vec{u}_{ij}^m$.

Consequently, we can approximate $f_{ij}^n \approx f(\mathbf{x}_{ij},t^n)$ as in equation (2.3) by

$$f_{ij}^n = F(X_{ij}^{n,0}) \approx F(X(\mathbf{x}_{ij},t^n;0))\,, \tag{3.2}$$

for $n > 0$ and for $n = 0$ we set $f_{ij}^0 = F_{ij}$.

When $n = N$ the deformation $\vec{U}(\mathbf{x})$ between $F(\mathbf{x})$ and $G(\mathbf{x})$ is given by

$$\vec{U}_{ij} = \mathbf{x}_{ij} - X_{ij}^{N,0}. \tag{3.3}$$

The stopping time t^n for some $n = N$ is determined by estimating the distance between f_{ij}^n and G_{ij}.

3.1. Numerical implementation of Lucas-Kanade method. The numerical approximation of the linear system (2.6) is obtained by solving

$$\begin{pmatrix} W_\sigma * (\partial_x f_{ij}^n)^2 & W_\sigma * (\partial_x f_{ij}^n \partial_y f_{ij}^n) \\ W_\sigma * (\partial_y f_{ij}^n \partial_x f_{ij}^n) & W_\sigma * (\partial_y f_{ij}^n)^2 \end{pmatrix} \begin{pmatrix} u_{ij}^n \\ v_{ij}^n \end{pmatrix} = \\ = \begin{pmatrix} -W_\sigma * (\partial_x f_{ij}^n \partial_t f_{ij}^n) \\ -W_\sigma * (\partial_y f_{ij}^n \partial_x f_{ij}^n) \end{pmatrix}. \tag{3.4}$$

where $u_{ij}^n \approx u(\mathbf{x}_{ij},t^n)$ and $v_{ij}^n \approx v(\mathbf{x}_{ij},t^n)$ and $\partial_t f_{ij}^n = (G_{ij} - f_{ij}^n)$. The spatial derivatives $\partial_x f_{ij}^n$ and $\partial_y f_{ij}^n$ are approximated by central differences

$$\partial_x f_{ij}^n = \frac{f_{i+1j}^n - f_{i-1j}^n}{2} \\ \partial_y f_{ij}^n = \frac{f_{ij+1}^n - f_{ij-1}^n}{2}\,. \tag{3.5}$$

The discrete convolution with Gaussian function W_σ is defined as follows

$$W_\sigma * f(x,y,t) = \frac{1}{\sum_{i,j} w_{ij}} \sum_{i,j} w_{ij} f(x+i, y+j, t)\,, \tag{3.6}$$

where $-3\sigma < i < 3\sigma$ and $-3\sigma < j < 3\sigma$ and

$$w_{ij} = \frac{1}{2\pi\sigma^2} \exp^{-\frac{i^2+j^2}{2\sigma^2}}\,, \tag{3.7}$$

where σ is a standard deviation that must be chosen by users. To do so we choose a convolution matrix with the elements w_{ij} to have a size $E \times E$ where E is an odd integer and $\sigma = E/6$. The dimension E determines the neighborhood of each pixel for which the assumption about constant optical flow is considered. A proper choice of E, respectively σ, is a nontrivial requirement of the original Lucas-Kanade method. Later for some numerical experiments we discuss a proper guess of σ and an influence of different choices on results.

Once the approximations $\vec{u}_{ij}^n$ are available, we can compute f_{ij}^{n+1} by using (3.2) and proceed to next time step.

3.2. Numerical implementation of the method based on level-set motion. Formally, we can approximate (2.8) by a numerical scheme

$$\tilde{f}_{ij}^{n+1} = f_{ij}^n + S_{ij}^n |\nabla f_{ij}^n| \tag{3.8}$$

where $S_{ij}^n = \alpha_{ij}^n (G_{ij} - f_{ij}^n) \approx S(\mathbf{x}_{ij}, t^n)$. To do so we approximate firstly the gradient ∇f_{ij}^n in (3.8) using the Rouy-Tourin scheme [8]

$$\partial_x f_{ij}^n = \begin{cases} f_{ij}^n - f_{i-1j}^n & f_{i-1j}^n = \mathrm{ext}\{f_{i-1j}^n, f_{ij}^n, f_{i+1j}^n\} \\ f_{i+1j}^n - f_{ij}^n & f_{i+1j}^n = \mathrm{ext}\{f_{i-1j}^n, f_{ij}^n, f_{i+1j}^n\} \\ 0 & f_{ij}^n = \mathrm{ext}\{f_{i-1j}^n, f_{ij}^n, f_{i+1j}^n\} \end{cases} \tag{3.9}$$

$$\partial_y f_{ij}^n = \begin{cases} f_{ij}^n - f_{ij-1}^n & f_{ij-1}^n = \mathrm{ext}\{f_{ij-1}^n, f_{ij}^n, f_{ij+1}^n\} \\ f_{ij+1}^n - f_{ij}^n & f_{ij+1}^n = \mathrm{ext}\{f_{ij-1}^n, f_{ij}^n, f_{ij+1}^n\} \\ 0 & f_{ij}^n = \mathrm{ext}\{f_{ij-1}^n, f_{ij}^n, f_{ij+1}^n\} \end{cases} \tag{3.10}$$

where ext denotes a minimum or maximum with the choice

$$\mathrm{ext} = \begin{cases} \min & S_{ij}^n < 0 \\ \max & S_{ij}^n > 0\,. \end{cases} \tag{3.11}$$

Secondly we have to define the values $\alpha_{ij}^n \approx \alpha(\mathbf{x}_{ij}, t^n) > 0$ to compute S_{ij}^n in (3.8). We propose to choose maximal values of α_{ij}^n to speed up the computation such that the so called CFL condition [7] and some stopping criteria are fulfilled, namely

$$\alpha_{ij}^n = \min\left(\frac{1}{|\nabla f_{ij}^n| + \epsilon}, \frac{|\nabla f_{ij}^n|}{|G_{ij} - f_{ij}^n|(|\partial_x f_{ij}^n| + |\partial_y f_{ij}^n| + \epsilon)} \right)\,, \tag{3.12}$$

where $\epsilon > 0$ is a small number to avoid a division by zero. The parameter ϵ is set to 10^{-8} for all presented numerical experiments, and different choices, e.g. $= 10^{-4} \leq \epsilon \leq = 10^{-8}$, have no visible influence on the results.

Once the approximations in the right hand side of (3.8) are available then the values $\vec{u}_{ij}^n$ are computed by approximation of (2.7)

$$\vec{u}_{ij}^n = \begin{cases} -S_{ij}^n \frac{\nabla f_{ij}^n}{|\nabla f_{ij}^n|} & |\nabla f_{ij}^n| \neq 0 \\ \vec{0} & |\nabla f_{ij}^n| = 0\,, \end{cases} \tag{3.13}$$

and the scheme (3.2) can be used to compute f_{ij}^{n+1}.

4. Experimental results. We present the results obtained from images of cell nucleus of zebrafish, see Fig. 4.1, Fig. 4.5 and Fig. 4.8. The images were preprocessed using a segmentation of the cell nucleus. The data are originally three dimensional, but we consider only two dimensional images. Some images contain a large deformation so they are quite challenging for the optical flow estimation.

Once an estimation of $\vec{U}(\mathbf{x})$ is obtained, we determine the approximation of image $F(x - \vec{U}(\mathbf{x}))$ to check the approximation quality of numerical methods by comparing it with the original image $G(\mathbf{x})$. Namely the difference image $|G(\mathbf{x}) - F(\mathbf{x} - \vec{U}(\mathbf{x}))|$ is shown that should be white if there is no error in the approximation. The optical flow is presented graphically as $-\vec{U}(\mathbf{x})$, because we want to show from where does the position $\mathbf{x}$ in the image $G(\mathbf{x})$ comes from the image $F(\mathbf{x} - \vec{U}(\mathbf{x}))$. For a clarity, we present every fifth vector component of the optical flow in figures.

4.1. Cell movement. The first experiment includes the movement of cell nucleus. The input images are shown in Fig. 4.1 and the cell simply moves from the left to the right. The image size is 250×250.

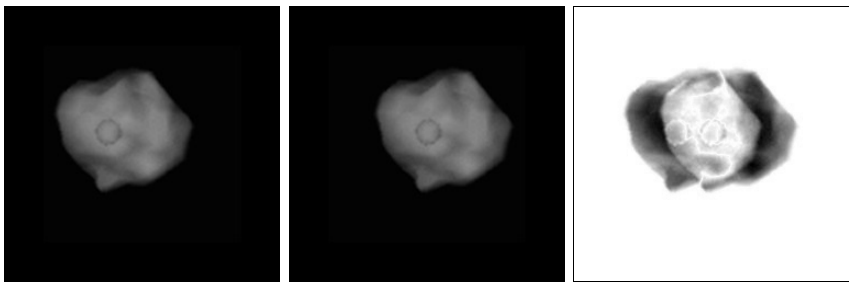


FIG. 4.1. *The input images of cell movement. From the left to the right: the first image $F(\boldsymbol{x})$, the second image $G(\boldsymbol{x})$, the difference image $|G(\boldsymbol{x}) - F(\boldsymbol{x})|$.*

Firstly, Lucas-Kanade method is used with the discrete convolution matrix of dimension 201×201 using the standard deviation $\sigma = 33.5$. In Fig. 4.2 (right) we can see a well resolved constant optical flow for this choice of σ. This size of convolution matrix was guessed from the size of a deformation caused by the optical flow visible in the difference image $|G(x) - F(x)|$.

We present computations also with a too small value of σ. In Fig. 4.3 the results of Lucas-Kanade method with standard deviation $\sigma = 16.83$ are shown when the discrete convolution matrix of the dimension 101×101 is used. From the visual inspection of optical flow in Fig. 4.3 (right) and difference image in Fig. 4.3 (middle) we can see that such convolution is not appropriate and the results are far away from expected ones.

Next we estimate the optical flow based on the level set motion. The results are presented in Fig. 4.4. The difference image in Fig. 4.4 (middle) shows that the result

is satisfactory. In Fig. 4.4 (right) we can see the obtained optical flow in normal direction that is not suitable for a visualization of movement by a constant vector field.

FIG. 4.2. *The results obtained by Lucas-Kanade method with* $\sigma = 33.5$. *From the left to the right: the image* $F(x - \vec{U}(x))$, *the difference image* $|G(\boldsymbol{x}) - F(\boldsymbol{x} - \vec{U}(\boldsymbol{x}))|$, *the optical flow* $-\vec{U}(\boldsymbol{x})$.

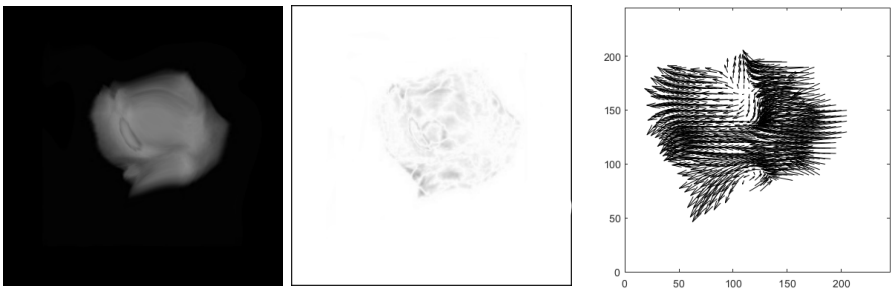


FIG. 4.3. *The results obtained by Lucas-Kanade method with* $\sigma = 16.83$. *From the left to the right: the image* $F(x - \vec{U}(x))$, *the difference image* $|G(\boldsymbol{x}) - F(\boldsymbol{x} - \vec{U}(\boldsymbol{x}))|$, *the optical flow* $-\vec{U}(\boldsymbol{x})$.

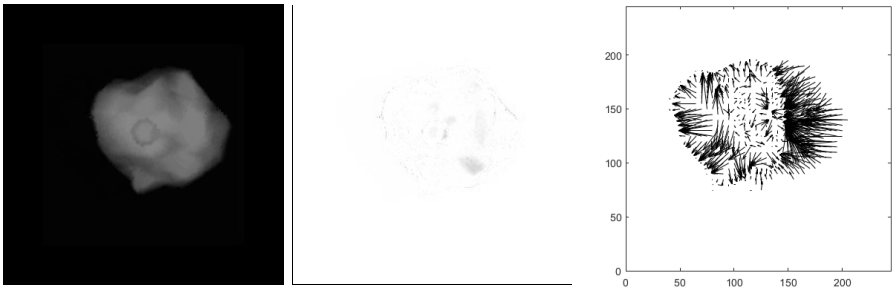


FIG. 4.4. *The results obtained by the method based on level set motion. From the left to the right: the image* $F(x - \vec{U}(x))$, *the difference image* $|G(\boldsymbol{x}) - F(\boldsymbol{x} - \vec{U}(\boldsymbol{x}))|$, *the optical flow* $-\vec{U}(\boldsymbol{x})$.

4.2. Cell deformation. In the second experiment the images in Fig. 4.5 represent a change of cell shape. The image size is 300×300.

The results are shown in Fig. 4.6 for Lucas-Kanade method and in Fig. 4.7 for method based on level set motion.

For Lucas-Kanade method the matrix of dimension 201×201 with $\sigma = 33.5$ is used as in the previous example. The optical flow in Fig. 4.6 (right) is smooth, but

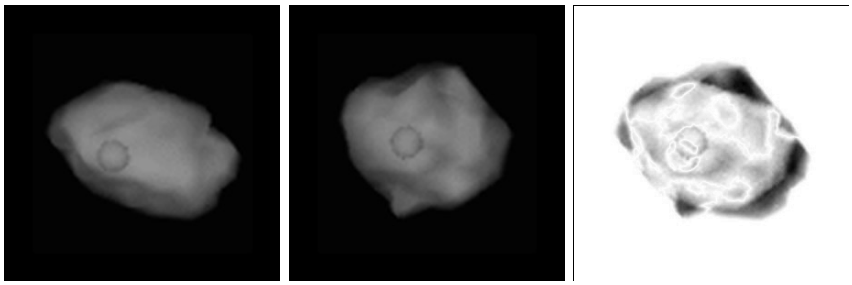


FIG. 4.5. *The input images of the deformation of cell. From the left to the right: the first image* $F(x)$*, the second image* $G(\boldsymbol{x})$*, the difference image* $|G(\boldsymbol{x}) - F(\boldsymbol{x})|$*.*

from the visual inspection of difference image in Fig. 4.6 (middle) we can see that this method can not change the shape of cell properly.

The difference image in Fig. 4.7 (middle) and the image after applying the optical flow in Fig. 4.7 (left) show us that the results are satisfactory.

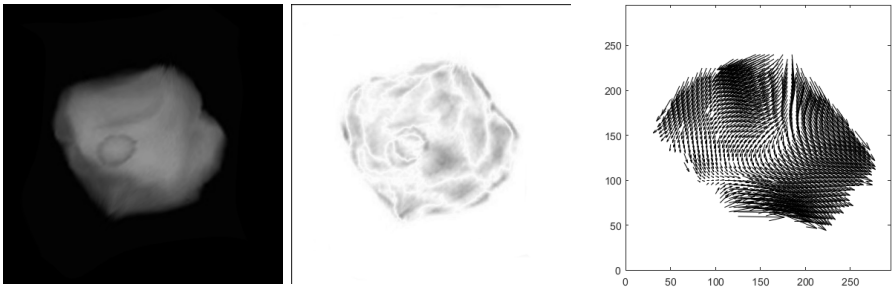


FIG. 4.6. *The results obtained by Lucas-Kanade method with* $\sigma = 33.5$*. From the left to the right: the image* $F(x - \vec{U}(\boldsymbol{x}))$*, the difference image* $|G(\boldsymbol{x}) - F(x - \vec{U}(\boldsymbol{x}))|$*, the optical flow* $-\vec{U}(\boldsymbol{x})$*.*

FIG. 4.7. *The results obtained by the method based on level-set motion. From the left to the right: the image* $F(x - \vec{U}(\boldsymbol{x}))$*, the difference image* $|G(\boldsymbol{x}) - F(x - \vec{U}(\boldsymbol{x}))|$*, the optical flow* $-\vec{U}(\boldsymbol{x})$*.*

4.3. Movement and deformation of cells. The last experiment include the motion and deformation of four cells. The input images are shown in Fig. 4.8. The size of images is 640×600.

In this case we present the results obtained by combining the Lucas-Kanade method and the method based on level-set motion.

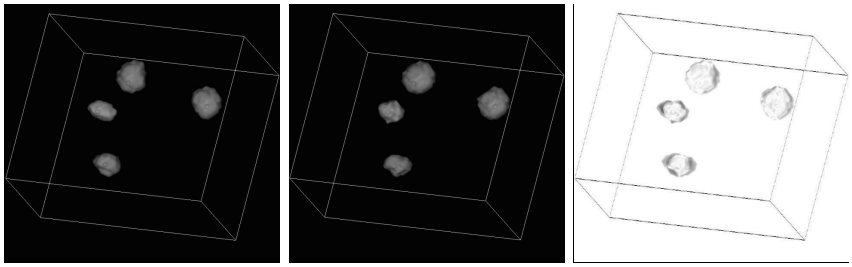


FIG. 4.8. *The input images of movement and deformation of the cells. Form left to right: the first image $F(\boldsymbol{x})$, the second image $G(\boldsymbol{x})$, the difference image $|G(\boldsymbol{x}) - F(\boldsymbol{x})|$.*

Firstly, we estimate the optical flow using the Lucas-Kanade method. The standard deviation is chosen $\sigma = 8.5$ as the deformation by the optical flow has a smaller size than in the previous examples. The resulting optical flow is shown in Fig. 4.9 (right). For a better visualisation the zooms of the optical flow in Fig. 4.9 (right) are presented for each cell in Fig. 4.10.

From the visual inspection of difference image $|G(x) - F(x - \vec{U}(x))|$ in Fig. 4.9 (middle) we can see that the results are not satisfactory as the method can move the cells but it does not change their shapes properly. This can be seen from the difference image and also from the obtained image in Fig. 4.9 (left).

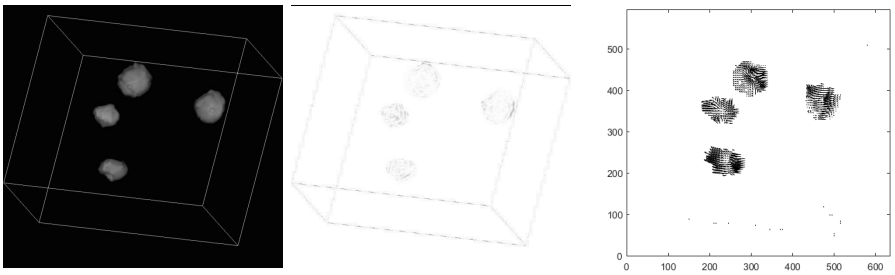


FIG. 4.9. *The results obtained by Lucas-Kanade method with $\sigma = 8.5$. From the left to the right: the image $F(x - \vec{U}(\boldsymbol{x}))$, the difference image $|G(\boldsymbol{x}) - F(\boldsymbol{x} - \vec{U}(\boldsymbol{x}))|$, the optical flow $-\vec{U}(\boldsymbol{x})$.*

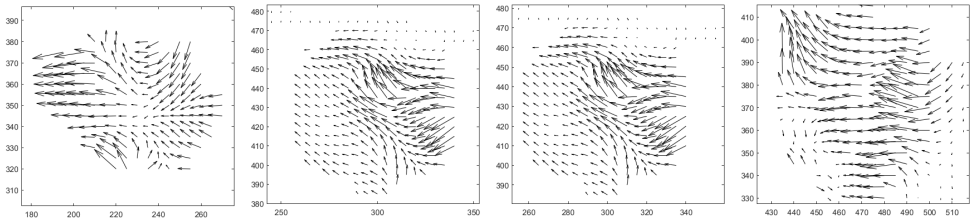


FIG. 4.10. *Zooms of the optical flow obtained by Lucas-Kanade method.*

The next step is to apply the method based on level-set motion on the obtained image from Lucas-Kanade method and to compute the total optical flow.

The results after applying two methods is shown in Fig. 4.11. Again we present the zooms of optical flow for each cells in Fig. 4.12. From the visual inspection of the difference image and image after applying optical flow in Fig. 4.11 we can see, that the results of Lucas-Kanade method were improved by the correction of method based on level set motion.

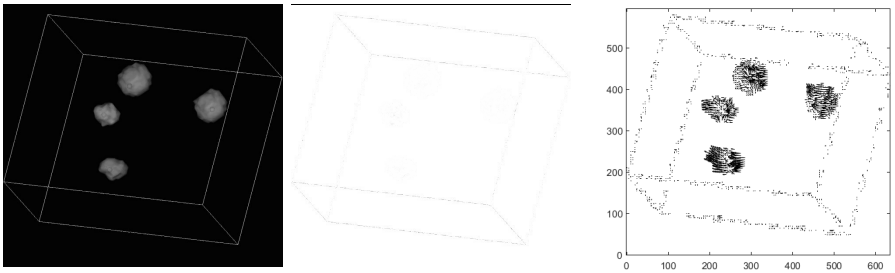


FIG. 4.11. *The results obtained by method based on level-set motion after Lucas-Kanade method. From the left to the right: the image* $F(x - \vec{U}(x))$*, the difference image* $|G(x) - F(x - \vec{U}(x))|$*, the optical flow* $-\vec{U}(\boldsymbol{x})$*.*

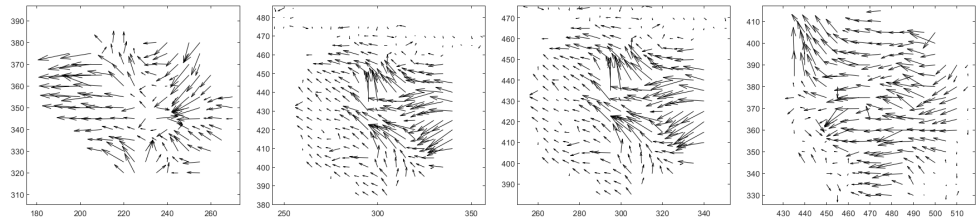


FIG. 4.12. *Zooms of the total optical flow.*

5. Conclusions. In this work we present two methods to estimate the optical flow using the backward tracking of characteristics based on two standard approaches. To study which approach is more suitable for which type of optical flow estimation we present numerical experiments and discuss the results. The Lucas-Kanade method [5] assumes that the optical flow does not vary too much in a neighborhood of each pixel when the size of such neighborhood must be set by the dimension of convolution matrix. The method gives best results when the vector field of optical flow is almost constant. The method based on the level set motion [1, 9] does not require such assumption as it estimates only the deformation by optical flow in the normal direction to isolines of image. It is appropriate when no translation is given by the optical flow and only a shape deformation can be observed between two images. In this work we present preliminary results when these two methods are combined to obtain more appropriate optical flow estimation.

REFERENCES

[1] M. Bartalmío, G. Sapiro and G. Randall, *Morphing Active Contours.*, IEEE Trans. PAMI, 22(7) (2000), pp. 733–737.

[2] A. Bruhn, J. Weickert, Ch. Schnrr, *Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods.*, Int. Journal of Comp. Vision. , 61(3) (2005), pp. 211-231

[3] B. Horn, B. Schunck, *Determining optic flow.*, Artificial Intelligence, 17(1-3) (1981), pp. 185-203.

[4] V. Kleinová, *Algoritmy extrakcie rýchlostného poľa z postupnosti obrazov.*, Diploma thesis, Faculty of Civil Engineering, Slovak University of Technology in Bratislava (2014).

[5] B. Lucas, T. Kanade, *An iterative image registration technique with an application to stereovision.*, In Int. Joint Conf. on Artificial Intel, 2 (1981), pp. 674-679.

[6] S. Osher, J.A. Sethian, *Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations* , J. Computational Physics, 79 (1988), pp. 12-49.

[7] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems.*, Cambridge University Press, 1 (2002), ISBN: 0521009243.

[8] E. Rouy, A. Tourin, *A viscosity solutions approach to shape-from-shading.*, SIAM J. Num. Anal., 29 (1992), pp. 867-884.

[9] B.C. Vemuri, J. Ye, Y. Chen, C.M. Leonard, *Image registration via level-set motion: Applications to atlas-based segmentation.*, Medical Image Analysis, 7(1) (2003), pp. 1-20.

Proceedings of EQUADIFF 2017
pp. 341–348

MATHEMATICALLY MODELLING THE DISSOLUTION OF SOLID DISPERSIONS

MARTIN MEERE*, SEAN MCGINTY †, AND GIUSEPPE PONTRELLI ‡

Abstract. A solid dispersion is a dosage form in which an active ingredient (a drug) is mixed with at least one inert solid component. The purpose of the inert component is usually to improve the bioavailability of the drug. In particular, the inert component is frequently chosen to improve the dissolution rate of a drug that is poorly soluble in water. The construction of reliable mathematical models that accurately describe the dissolution of solid dispersions would clearly assist with their rational design. However, the development of such models is challenging since a dissolving solid dispersion constitutes a non-ideal mixture, and the selection of appropriate forms for the activity coefficients that describe the interaction between the drug, the inert matrix, and the dissolution medium is delicate. In this paper, we present some preliminary ideas for modelling the dissolution of solid dispersions.

Key words. Solid Dispersion, Mathematical Model, Partial Differential Equations, Activity Coefficients

AMS subject classifications. 74N25, 82C70, 82D60

1. Introduction.

1.1. Motivation: poorly soluble drugs. Drugs that are delivered orally via a tablet should ideally be readily soluble in water. Drugs that are poorly water-soluble tend to pass through the gastrointestinal tract before they can fully dissolve, and this typically leads to poor bioavailability of the drug. Unfortunately, many drugs currently on the market or in development are poorly water-soluble, and this presents a serious challenge to the pharmaceutical industry. Many strategies have been developed to improve the solubility of drugs, such as the use of surfactants, cocrystals, lipid-based formulations, and particle size reduction. The literature on this topic is extensive, and recent reviews can be found in [1] and [2].

One particularly effective strategy to improve drug solubility is to use a *solid dispersion*. A solid dispersion typically consists of a hydrophobic drug embedded in a hydrophilic polymer matrix, where the matrix can be either in the amorphous or crystalline state. The drug is preferably in a molecularly dispersed state, but may also be present in amorphous particles or even in the crystalline form (though this is usually undesirable); see Figure 1.1.

1.2. Storage, stability and phase separation of solid dispersions. Drug loading in most dispersions greatly exceeds the equilibrium solubility in the polymer matrix for typical storage temperatures. Hence these systems are usually unstable, with phase separation eventually occurring ([7]). In such cases, the drug will eventually crystallise out or form an amorphous phase separation. However, if the dispersion is stored well below the glass transition temperature ([3]) for the polymer, and is kept dry, this can happen extremely slowly. The system is then for all

*School of Mathematics, National University of Ireland Galway, University Road, Galway, Ireland (martin.meere@nuigalway.ie).

†Division of Biomedical Engineering, University of Glasgow, Glasgow, G12 8QQ, UK (sean.mcginty@glasgow.ac.uk).

‡Istituto per le Applicazioni del Calcolo, CNR, Rome, Italy (giuseppe.pontrelli@gmail.com).

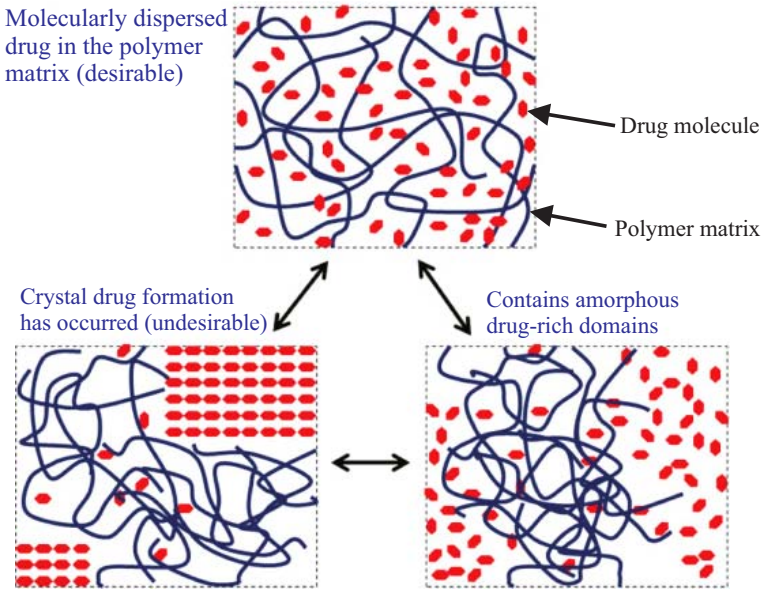


FIG. 1.1. *Adapted from [7]. In this figure, we show three possible structures for a polymer/drug dispersion. Top: Here the drug is in the molecularly dispersed state, which is usually desirable for a solid dispersion. Bottom left: Here the dispersion contains drug in the crystalline form. Bottom right: Here the dispersion contains amorphous drug-rich domains.*

practical purposes stable, and is said to be metastable. The humidity of the storage environment can be an issue because even small amounts of moisture can significantly affect the glass transition temperature. Hence polymers that have high glass transition temperatures and that are resistant to water absorption have become popular. An example of one such polymer is Hydroxypropyl Methylcellulose Acetate Succinate (HPMCAS).

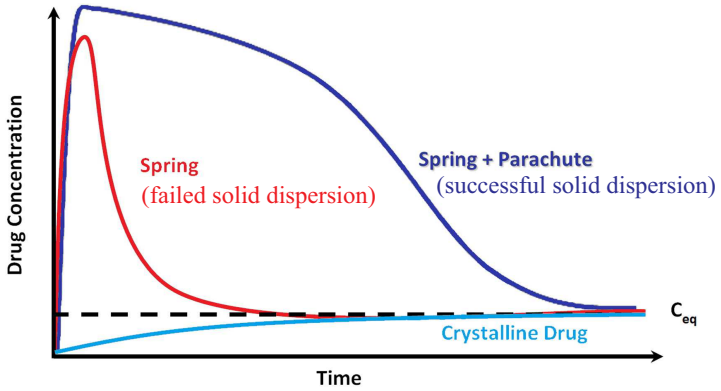


FIG. 1.2. *Adapted from [8]. The drug first dissolves along with the soluble polymer matrix to generate a supersaturated solution (spring) followed by a decline in the drug concentration in the media due to either absorption or precipitation (parachute).*

1.3. Drug release from solid dispersions. The *spring and parachute* concept is the usual strategy associated with drug release from solid dispersions. When the dispersion absorbs fluid, the dispersed drug dissolves along with the soluble polymer to create a solution with a drug concentration that is well above the drug solubility in the fluid (this is the spring). The dispersion then maintains the drug concentration at supersaturated levels for a period of hours while it is being absorbed (this is the parachute); see Figure 1.2.

Unfortunately, despite extensive research, the dissolution behaviour of solid dispersions is only partially understood. In particular, the precise mechanisms via which the polymer prolongs the supersaturation of the drug have not been fully resolved. This makes the design of successful solid dispersions a somewhat *hit and miss* affair. Clearly, the construction of reliable mathematical models that capture the key interactions between drug, polymer and solvent molecules in a dissolving solid dispersion would greatly assist with their rational design. The ultimate goal of such modelling is to identify the regions of the parameter space governing the system that lead to the desired dissolution behaviour for a given pharmaceutical product. Some previous modelling studies for solid solutions can be found in [9], [10], [11] and [12].

2. Mathematical modelling.

2.1. A multicomponent diffusion model for solid dispersions. We develop a multicomponent diffusion model for the evolution of the concentrations of the three components constituting a dissolving solid dispersion. These are the drug molecules (label 1), the polymer molecules (label 2), and the solvent molecules (label 3). For simplicity, we shall restrict our attention here to the one-dimensional form of the equations; that is, the concentrations of the species only depend on a single spatial variable x.

The chemical potential μ_i (J/mole) of species i ($i = 1, 2, 3$) gives the Gibbs free energy per mole of species i, and is given here by ([4])

$$\mu_i = \mu_{i0} + RT \ln(a_i) - \epsilon_i^2 \frac{\partial^2 N_i}{\partial x^2} \tag{2.1}$$

where μ_{i0} is the chemical potential of species i in the pure state, R (J/[K·mole]) is the gas constant, T (K) is the temperature, a_i is the activity of species i, and the term involving $\epsilon_i^2 > 0$ (m^2J/mole) penalises the formation of phase boundaries (see [5] and [6] for some discussion of this issue). Here N_i is the molar fraction of species i ($i = 1, 2, 3$), and the activities can depend on these molar fractions, so that

$$a_i = a_i(N_1, N_2, N_3).$$

The flux of species i (molar·m/s) is given by

$$J_i = c_i v_i \tag{2.2}$$

where c_i (molar), v_i (m/s) give the molar concentration and drift velocity, respectively, of species i. The drift velocity v_i gives the average velocity a particle of species i attains due to the diffusion force acting on it, and is given here by

$$v_i = M_i \mathcal{F}_i = -M_i \frac{\partial \mu_i}{\partial x} \tag{2.3}$$

where M_i (mole·s/kg), $\mathcal{F}_i$ (J/[m·mole]) give the mobility and diffusion force, respectively, for species i. Substituting (2.3) in (2.2) and using (2.1) gives

$$J_i = -M_i c_i \frac{\partial \mu_i}{\partial x} = -M_i c_i \left(\frac{RT}{a_i} \frac{\partial a_i}{\partial x} - \epsilon_i^2 \frac{\partial^3 N_i}{\partial x^3} \right)$$

and then using the fact that the activities depend on the molar fractions gives

$$J_i = -M_i c_i \left(\frac{RT}{a_i} \sum_{j=1}^{3} \frac{\partial a_i}{\partial N_j} \frac{\partial N_j}{\partial x} - \epsilon_i^2 \frac{\partial^3 N_i}{\partial x^3} \right). \tag{2.4}$$

The molar fraction is related to the molar concentration via

$$N_i = V_i c_i \tag{2.5}$$

where V_i (molar^{-1}) is the molar volume of species i. Using (2.5), we can now write (2.4) as

$$J_i = -\sum_{j=1}^{3} D_{ij} \frac{\partial c_j}{\partial x} + D_i \varepsilon_i^2 c_i \frac{\partial^3 c_i}{\partial x^3} \tag{2.6}$$

where the D_{ij} (m^2/s) are given by

$$D_{ij} = D_i \frac{V_j}{V_i} \frac{N_i}{a_i} \frac{\partial a_i}{\partial N_j} \qquad (i, j = 1, 2, 3) \tag{2.7}$$

and where

$$D_i = M_i RT \qquad \text{(Einstein relation)}$$

is the diffusion coefficient for species i. Finally, $\varepsilon_i^2 = M_i V_i \epsilon_i^2 / D_i > 0$ (m^2/molar).

Conservation of mass for species i implies that

$$\frac{\partial c_i}{\partial t} + \frac{\partial J_i}{\partial x} = 0$$

and using (2.6) now gives

$$\frac{\partial c_i}{\partial t} = \frac{\partial}{\partial x} \left(\sum_{j=1}^{3} D_{ij}(c_1, c_2, c_3) \frac{\partial c_j}{\partial x} - D_i \varepsilon_i^2 c_i \frac{\partial^3 c_i}{\partial x^3} \right) \qquad (i = 1, 2, 3) \tag{2.8}$$

where we have included the concentration dependence of the diffusion coefficients D_{ij} here to emphasise that this system is in general a coupled system of nonlinear diffusion equations. It should be noted that the equations (2.8) are not independent since $V_1 c_1 + V_2 c_2 + V_3 c_3 = 1$, and so it is sufficient to solve for two concentrationns only.

2.2. Activity coefficients. The activities a_i are usually written as

$$a_i = \gamma_i N_i$$

where the $\gamma_i = \gamma_i(N_1, N_2, N_3)$ are referred to as *activity coefficients*. Equations (2.7) now give

$$D_{ij} = D_i \frac{V_j}{V_i}\left(\delta_{ij} + \frac{N_i}{\gamma_i}\frac{\partial \gamma_i}{\partial N_j}\right) \qquad (i,j = 1,2,3) \tag{2.9}$$

where δ_{ij} is the Kronecker delta. Notice that if $\gamma_i \equiv$ constant, then the (2.9) reduce to $D_{ij} = D_i$ and the governing equations decouple to give

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial}{\partial x}\left(\frac{\partial c_i}{\partial x} - \varepsilon_i^2 c_i \frac{\partial^3 c_i}{\partial x^3}\right). \qquad (i = 1,2,3)$$

If we further have $\varepsilon_i = 0$, the governing equations reduce to a set of classical linear diffusion equations.

The details of the interactions between the species in solution are captured in the modelling by choosing appropriate forms for the activity coefficients $\gamma_i = \gamma_i(N_1, N_2, N_3)$. The construction of appropriate forms for the γ_i for various solutions is a large subject with a large literature; see, for example, the books [13] and [14].

2.3. Activity coefficients for polymer solutions.

2.3.1. Flory-Huggins Model. The Flory-Huggins model is a lattice-based model commonly used to describe the thermodynamics of polymer solutions. For a binary solution in which the subscripts 1 and 2 refer to drug and polymer molecules, respectively, this model has ([13])

$$\ln(\gamma_1) = \ln\left(\frac{\Phi_1}{N_1}\right) + 1 - \frac{\Phi_1}{N_1} + \chi_{12}\Phi_2^2$$

where

$$\Phi_1 = \frac{V_1 N_1}{V_1 N_1 + V_2 N_2}, \quad \Phi_2 = \frac{V_2 N_2}{V_1 N_1 + V_2 N_2}.$$

Here χ_{12} is the Flory-Huggins interaction parameter that quantifies the balance between polymer-polymer and polymer-solvent interactions. Note that we need only specify γ_1 here since $N_2 = 1 - N_1$.

For a ternary solution, consisting of drug molecules, polymer molecules, and solvent molecules (molar fraction N_3, molar volume V_3), we have ([13])

$$\ln(\gamma_i) = \ln\left(\frac{\Phi_i}{N_i}\right) + 1 - \frac{\Phi_i}{N_i} + 2V_i \sum_{j=1}^{3} \Phi_i b_{ij} - V_i \sum_{j,k=1}^{3} \Phi_j \Phi_k b_{jk}$$

where

$$\Phi_i = \frac{V_i N_i}{\sum_{j=1}^{3} V_j N_j} \quad \text{for} \quad i = 1,2,3,$$

and where the interaction parameters b_{ij} (molar) are such that $b_{ii} = 0$ and $b_{ij} = b_{ji}$.

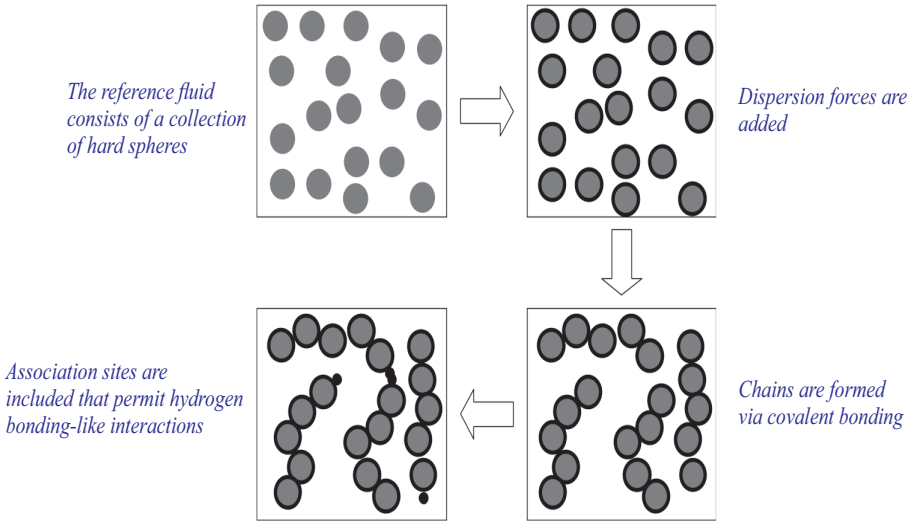


FIG. 2.1. *Adapted from [16]. In the SAFT framework, the reference fluid consists of a collection of hard spheres to which dispersion forces are added. These spheres can form chains via covalent bonding. Association sites are added to the chains that allow for the inclusion of hydrogen bonding type interactions.*

2.3.2. Statistical Associating Fluid Theory. Statistical Associating Fluid Theory (SAFT) is a sophisticated tool for developing realistic thermodynamic models for polymer solutions ([15, 16]). The theory allows for the development of tailored models for specific polymer/drug systems constituting solid dispersions. For a single component fluid system, the physical basis of the SAFT approach is illustrated in Figure 2.1. The reference fluid consists of a system of hard spheres to which weak dispersion forces are added. These spheres can form chains of a given length via covalent bonding. Finally, association sites are added to the chains to allow for hydrogen bonding-type interactions.

The Helmholtz free energy $\mathcal{A}$ (J) of the fluid in SAFT is then calculated as follows

$$\mathcal{A} = \mathcal{A}^{ideal} + \mathcal{A}^{hs} + \mathcal{A}^{disp} + \mathcal{A}^{chain} + \mathcal{A}^{assoc}$$

where

$\mathcal{A}^{ideal}$ = contribution from the ideal fluid,
$\mathcal{A}^{hs}$ = contribution from the hard sphere assumption,
$\mathcal{A}^{disp}$ = contribution from the dispersion force interactions,
$\mathcal{A}^{chain}$ = contribution from the covalent bonding,
$\mathcal{A}^{assoc}$ = contribution from the association interactions.

Expressions for each of these quantities have been calculated by the applied statistical thermodynamics community, and can be found in [16]. A full listing and explanation of these equations would occupy a number pages, and so for brevity we have omitted these details here. Once the free energies for the individual components have been calculated, the free energy for the *mixture* can then be calculated using mixing rules.

With the free energy of the mixture in hand, the activity coefficients can be calculated as follows ([12, 13, 14]). We first define the residual Helmholtz free energy

via

$$\mathcal{A}^{res} = \mathcal{A} - \mathcal{A}^{ideal} = \mathcal{A}^{hs} + \mathcal{A}^{disp} + \mathcal{A}^{chain} + \mathcal{A}^{assoc},$$

and the residual chemical potentials are then given by

$$\mu_i^{res} = \mathcal{A}^{res} + RT(\mathcal{Z} - 1) + \frac{\partial \mathcal{A}^{res}}{\partial N_i} - \sum_{j=1}^{3} N_j \frac{\partial \mathcal{A}^{res}}{\partial N_j},$$

for $i = 1, 2, 3$ as before, and where

$$\mathcal{Z} = 1 + \rho \frac{\partial(\mathcal{A}^{res}/RT)}{\partial \rho}$$

is the incompressibility factor, with ρ the density of the system. The fugacity of component i in the mixture is then given by

$$\varphi_i = \frac{1}{\mathcal{Z}} \exp\left(-\frac{\mu_i^{res}}{RT}\right).$$

Finally, the activity coefficient for component i is now given by

$$\gamma_i = \frac{\varphi_i}{\varphi_{i0}},$$

where φ_{i0} is the fugacity of the pure component i.

3. Future work. The following briefly summarises our future research plan for investigating the behaviour of solid dispersions.

- We shall begin by considering Flory-Huggins type models for polymer solutions. It is envisaged that the consideration of these simpler systems will yield mechanistic insights. Once this work has been completed, we shall use the SAFT framework to develop more realistic models for specific polymer/drug systems.
- We shall address two main problems. The first of these is the storage stability problem. For this problem, we shall develop a representative initial boundary value problem to describe the behaviour of the solid dispersion in storage. Initially, we shall consider a two component model (polymer/drug) and identify those parameter regimes that lead to stable, metastable, and unstable behaviour. It is also worth noting that because we shall be considering the full non-equilibrium problem, we should also be able to obtain information concerning the timescales over which phase separation and drug crystallization occurs. By introducing a third component for water molecules, the effect of air humidity on storage stability will also be investigated.
- The second problem we shall consider is the dissolution problem. In this case, we shall develop a representative initial boundary value problem to describe the dissolution of a solid dispersion. Particular attention will be paid to identifying the underlying mechanisms and parameter regimes that lead to the spring and parachute effect.

Acknowledgments. M. Meere thanks NUI Galway for the award of a travel grant. We thank the referee for their helpful suggestions to improve the paper.

REFERENCES

[1] H. D. Williams and N. L. Trevaskis and S. A. Charman and R. M. Shanker and W. N. Charman and C. W. Pouton and C. J. H. Porter, *Strategies to address low drug solubility in discovery and development*, Pharmacological Reviews, 65 (2013), pp. 315–499.

[2] K. T. Savjani and A. K. Gajjar and J. K. Savjani, *Drug solubility: importance and enhancement techniques*, ISRN Pharmaceutics, 2012 (2012), pp. 1–10.

[3] M. Doi, *Introduction to polymer physics*, Oxford University Press, Oxford, UK, 1996.

[4] E. B. Smith, *Basic chemical thermodynamics*, Sixth Ed., Imperial College Press, London, UK, 2014.

[5] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system. I. Interfacial free energy*, The Journal of Chemical Physics, 28 (1958), pp. 258–267.

[6] N. Provatas & K. Elder, *Phase-field methods in material science and engineering*, John Wiley & Sons, 2010.

[7] Y. Huang and W. Daib, *Fundamental aspects of solid dispersion technology for poorly soluble drugs*, Acta Pharmaceutica Sinica B, 4(1) (2014), pp. 18–25.

[8] C. Brough and R. O. Williams III, *Amorphous solid dispersions and nano-crystal technologies for poorly water-soluble drug delivery*, International Journal of Pharmaceutics, 453 (2013), pp. 157–166.

[9] D. Q. M. Craig, *The mechanisms of drug release from solid dispersions in water-soluble polymers*, International Journal of Pharmaceutics, 231 (2002), pp. 131–144.

[10] N. Ahuja and O. P. Katare and B. Singh, *Studies on dissolution enhancement and mathematical modeling of drug release of a poorly water-soluble drug using water-soluble carriers*, European Journal of Pharmaceutics and Biopharmaceutics, 65 (2007), pp. 26–38.

[11] Z. A. Langham and J. Booth and L. P. Hughes and G. K. Reynolds and S. A. C. Wren, *Mechanistic insights into the dissolution of spray-dried amorphous solid dispersions*, Journal of Pharmaceutical Sciences, 101(8) (2012), pp. 2798–2810.

[12] Y. Ji and R. Paus and A. Prudic and C. Lubbert and G. Sadowski, *A novel approach for analyzing the dissolution mechanism of solid dispersions*, Pharmaceutical Research, 32 (2015), pp. 2559–2578.

[13] G. M. Kontogeorgis and G. K. Folas, *Thermodynamic models for industrial applications*, First ed., John Wiley & Sons, Chichester, West Sussex, UK, 2010.

[14] J. M. Prausnitz and R. N. Lichtenthaler and E. G. de Azevedo, *Molecular thermodynamics of fluid-phase equilibria*, Third ed., Prentice Hall, New Jersey, 1999.

[15] I. G. Economou, *Statistical associating fluid theory: a successful model for the calculation of thermodynamic and phase equilibrium properties of complex fluid mixtures*, Industrial & Engineering Chemistry Research, 41 (2002), pp. 953–962.

[16] C. McCabe and A. Galindo, *SAFT associating fluids and fluid mixtures*, Applied Thermodynamics of Fluids, Chapter 8, Royal Society of Chemistry, pp. 215–279, 2010.

Proceedings of EQUADIFF 2017
pp. 349–358

TOWARD A MATHEMATICAL ANALYSIS FOR A MODEL OF SUSPENSION FLOWING DOWN AN INCLINED PLANE

KANAME MATSUE* AND KYOKO TOMOEDA†

Abstract. We consider the Riemann problem of the dilute approximation equations with spatiotemporally dependent volume fractions from the full model of suspension, in which the particles settle to the solid substrate and the clear liquid film flows over the sediment [Murisic et al., J. Fluid. Mech. **717**, 203–231 (2013)]. We present a method to find shock waves, rarefaction waves for the Riemann problem of this system. Our method is mainly based on [Smoller, Springer-Verlag, New York, second edition, (1994)].

Key words. hyperbolic conservation law, Riemann problem, shock wave, rarefaction, suspension, dilute approximation

AMS subject classifications. 03-06, 35L65

1. Introduction. We are concerned here with the two dimensional motion of a suspension flowing down an inclined plane under the effect of gravity. To describe the problem we choose a coordinate system (x, y), where the x-axis is along a plane with a inclination angle α $\left(0 < \alpha < \frac{\pi}{2}\right)$ and the y-axis is perpendicular to this plane. The motion of suspension is governed by the following partial differential equations

$$\begin{aligned} &\nabla p - \nabla \cdot [\mu(\phi)(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top)] = \rho(\phi)\boldsymbol{g}, \\ &\partial_t \phi + \boldsymbol{u} \cdot \nabla \phi + \nabla \cdot \boldsymbol{J} = 0, \\ &\nabla \cdot \boldsymbol{u} = 0, \qquad \text{in} \quad 0 < y < h(x,t), \quad t \geq 0. \end{aligned} \tag{1.1}$$

Here $\boldsymbol{u} = (u, v)^\top$ is the volume averaged velocity and p is the pressure of fluid and $h(x, t)$ is the total suspension thickness. ϕ is the particle volume fraction and $\boldsymbol{J} = (\boldsymbol{J}_x, \boldsymbol{J}_y)^\top$ is the particle flux and $\boldsymbol{g} = g(\sin\alpha, -\cos\alpha)^\top$ is the acceleration of gravity. $\mu(\phi)$ is the viscosity of fluid and $\rho(\phi) = \rho_p\phi + \rho_f(1-\phi)$, where ρ_f and ρ_p are the density of fluid and particles respectively. The boundary condition on the wall is the non-slip and no-penetration condition

$$\boldsymbol{u} = (0, 0)^\top, \qquad \text{at} \quad y = 0. \tag{1.2}$$

The dynamical and kinematic conditions on the free surface are

$$\begin{aligned} \left(-p\boldsymbol{I} + \mu(\phi)(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^\top)\right)\boldsymbol{n} = 0, \qquad &\text{at} \quad y = h(x,t), \\ \partial_t h + u\partial_x h - v = 0, \qquad &\text{at} \quad y = h(x,t), \end{aligned} \tag{1.3}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{n}$ is the outward unit normal vectors to the free surface. For the particle fluxes, the no-flux boundary conditions at the wall and free surface are also imposed :

$$J \cdot \boldsymbol{n} = 0, \qquad \text{at} \quad y = 0 \ \text{ and } \ y = h(x,t). \tag{1.4}$$

*Institute of Mathematics for Industry / International Institute for Carbon-Neutral Energy Research (WPI-I^{2}CNER), Kyushu University, Fukuoka 819-0395, Japan, (kmatsue@imi.kyushu-u.ac.jp)

†Institute for Fundamental Sciences, Setsunan University, Osaka 572-8508, Japan, (tomoeda@mpg.setsunan.ac.jp)

To explain the mechanisms of suspensions, some approximation equations are derived from the full model (1.1)–(1.4). Murisic et al. [4] derived the dilute approximation equation which is the system of conservation laws :

$$\partial_t h + \partial_x \left(\frac{1}{3} h^3 \right) = 0, \tag{1.5}$$

$$\partial_t n + \partial_x \left(\sqrt{\frac{2}{9C}} (nh)^{3/2} \right) = 0, \tag{1.6}$$

where $C = \frac{2(\rho_p - \rho_f) \cot \alpha}{9(\rho_f K_c)}$ is the buoyancy parameter and K_c is constant and $n = \phi h$. This dilute approximation equation focuses on the settled regime in which particles settle to the solid substrate and the clear liquid film flows over the sediment. In [4], the authors solved (1.5) exactly with the initial data $h(x, 0) = 1$ for $0 \leq x \leq 1$, $h(x, 0) = 0$ otherwise, and the exact solution for h is given by

$$h(x,t) = \begin{cases} 1 & t \leq x \leq x_\ell, \\ \sqrt{\dfrac{x}{t}} & 0 < x < \min(t, x_\ell), \\ 0 & \text{otherwise}, \end{cases}$$

for $t \geq 0$, where x_ℓ denote the liquid front position which is given by $x_\ell = 1 + \frac{t}{3}$ for $0 \leq t \leq \frac{3}{2}$, $x_\ell = \left(\frac{9t}{4}\right)^{1/3}$ for $\frac{3}{2} < t$. One of the earlier examples for solution (1.7) is given by Huppert [1] for the flow of a constant volume of viscous fluid down a constant slope. The authors in [4] also obtain the exact solution n of (1.6) with the initial data $n(x, 0) = f_0 h(x, 0)$ and some given value $f_0 \ll 1$.

Our aim in this paper is to cover the solution of the system (1.5)–(1.6) when the initial volume fraction $\phi(x, 0)$ is a variable satisfying $0 < \phi < 1$. For this system, only exact solutions obtained for the fixed initial volume fraction $\phi(x, 0) = f_0$ are treated in [4]. On the other hand, in mathematical theory, it is known that the general $m \times m$ system of the hyperbolic conservation laws

$$\partial_t U + \partial_x (F(U)) = 0$$

has a discontinuous solution such as a shock wave and a smooth solution such as a rarefaction wave, where $U = (U_1, \cdots, U_m)^\top \in \boldsymbol{R}^m$, $(x, t) \in \boldsymbol{R} \times \boldsymbol{R}_+$ and $F(U) = (F_1(U), \cdots, F_m(U))^\top$ is a vector-valued function which is C^2 in some open subset $D \subset \boldsymbol{R}^m$ (see [2], [6]). In order to cover the solution of the system (1.5)–(1.6), we consider the case where the solutions have a discontinuity, and hence we deal with the weak solution of the system which is defined by (2.2) below. Applying mathematical theories established in [2], [6] to the system (1.5)–(1.6), we give a construction method of weak solutions consisting of simple waves such as shock waves and rarefaction waves.

The organization of this paper is as follows. In Section 2, we formulate shock waves and rarefaction waves for the Riemann problem of the system (1.5)–(1.6). In Section 3, we find the admissible shock waves and rarefaction waves in settled regime by using the formula given in Section 2.

2. Preliminaries. We let

$$U = \begin{pmatrix} h \\ n \end{pmatrix}, \qquad F(U) = \begin{pmatrix} \dfrac{1}{3} h^3 \\ \sqrt{\dfrac{2}{9C}} (nh)^{3/2} \end{pmatrix},$$

so that the system (1.5) and (1.6) can be rewritten in the form

$$\partial_t U + \partial_x (F(U)) = 0. \tag{2.1}$$

It is well known that a solution to conservation laws (2.1) can become discontinuous even if the initial data is smooth. Therefore we treat the weak solution which is defined as follows :

DEFINITION 2.1 ([6]). *A bounded measurable function $U(x,t)$ is called a weak solution of the initial-value problem for* (2.1) *with bounded and measurable initial data $U(x,0)$, provided that*

$$\int_0^\infty \int_{\boldsymbol{R}} (U\psi_t + F(U)\psi_x) dxdt + \int_{\boldsymbol{R}} U(x,0)\psi(x,0)dx = 0 \tag{2.2}$$

holds for all $\psi \in C_0^1(\boldsymbol{R} \times \boldsymbol{R}_+; \boldsymbol{R}^2)$. If the weak solution $U(x,t)$ has a discontinuity along a curve $x = x(t)$, the solution U and the curve $x = x(t)$ must satisfy the Rankine-Hugoniot relations (jump conditions)

$$s(U_L - U_R) = F(U_L) - F(U_R), \tag{2.3}$$

where $U_L = U(x(t) - 0, t)$ is the limit of U approaching (x,t) from the left and $U_R = U(x(t) + 0, t)$ is the limit of U approaching (x,t) from the right, and $s = \frac{dx}{dt}$ is the propagation speed of $x(t)$.

We consider the Riemann problem for the conservation laws (2.1) with the initial data called the Riemann data

$$U(x,0) = \begin{cases} U_0 & x < 0 \\ U_2 & x > 0 \end{cases}. \tag{2.4}$$

The Jacobian matrix of F at U is

$$DF(U) = \begin{pmatrix} h^2 & 0 \\ \sqrt{\frac{1}{2C} n^3 h} & \sqrt{\frac{1}{2C} h^3 n} \end{pmatrix}$$

and district eigenvalues of $DF(U)$ are

$$\lambda_1(U) = \sqrt{\frac{1}{2C} h^3 n}, \qquad \lambda_2(U) = h^2. \tag{2.5}$$

Here we assume that h and n are real valued function of $(x,t) \in \boldsymbol{R} \times \boldsymbol{R}_+$. According to [4], set $C = 2.307$ and $n = \phi h$, where the particle volume fraction ϕ satisfies $0 \le \phi < 1$. Under these conditions, the system (2.1) is strictly hyperbolic, i.e., district eigenvalues $\lambda_j(U)$ $(j = 1, 2)$ are real-valued and $\lambda_1(U) < \lambda_2(U)$ holds for any $U \in \Omega$, where $\Omega = \{(h,n) \in \boldsymbol{R}^2 : h > 0,\ 0 \le n < h\}$. The right eigenvectors corresponding to the eigenvalues $\lambda_j(U)$ are

$$r_1(U) = \begin{pmatrix} 0 \\ t_1 \end{pmatrix}, \qquad r_2(U) = \begin{pmatrix} h^2 - \sqrt{\frac{1}{2C} h^3 n} \\ \sqrt{\frac{1}{2C} n^3 h} \end{pmatrix},$$

where $t_1 \neq 0$ is a constant. Note that $\nabla\lambda_1 \cdot r_1 = \frac{t_1}{2}\sqrt{\frac{1}{2Cn}h^3} \neq 0$ and $\nabla\lambda_2 \cdot r_2 = 2h(h^2 - \sqrt{\frac{1}{2C}h^3 n}) \neq 0$ in Ω, namely, the first and the second characteristic fields are genuinely nonlinear in Ω. In this case, the weak solution will consist of three constant states U_0, U_1, U_2; the constant states U_{j-1} and U_j $(j = 1, 2)$ are connected by either shock waves or rarefaction waves (see [2], [6]).

Fix the reference point $U_p = (h_p, n_p)$. We consider right states $U_R = U = (h, n)$ which can be connected to a left state $U_L = U_p$ followed by shock waves or rarefaction waves. If the weak solution has a jump discontinuity between the left state U_p and the right state U, then U must satisfy the Rankine-Hugoniot relation (2.3):

$$\begin{aligned} s(h - h_p) &= \frac{1}{3}\left(h^3 - h_p^3\right), \\ s(n - n_p) &= \sqrt{\frac{2}{9C}}\left((nh)^{3/2} - (n_p h_p)^{3/2}\right). \end{aligned} \tag{2.6}$$

Eliminating s from these equations, we obtain

$$(n - n_p)\left(h^2 + hh_p + h_p^2\right) = \sqrt{\frac{2}{C}}\left((nh)^{3/2} - (n_p h_p)^{3/2}\right)$$

whose graph is called the *Hugoniot locus*. In order to pick up physically relevant solutions, we further require the following k-entropy inequalities $(k = 1, 2)$

$$s < \lambda_1(U_p), \qquad \lambda_1(U) < s < \lambda_2(U), \qquad \text{(1-entropy inequality)},$$

$$\lambda_1(U_p) < s < \lambda_2(U_p), \qquad \lambda_2(U) < s, \qquad \text{(2-entropy inequality)},$$

which in this case reads

$$\sqrt{\frac{1}{2C}h^3 n} < s < \min\left\{\sqrt{\frac{1}{2C}h_p^3 n_p},\, h^2\right\}, \qquad \text{(1-entropy inequality)}, \tag{2.7}$$

$$\max\left\{\sqrt{\frac{1}{2C}h_p^3 n_p},\, h^2\right\} < s < h_p^2, \qquad \text{(2-entropy inequality)}, \tag{2.8}$$

where s is the speed of discontinuity

$$s = \left(\frac{2}{81C}\right)^{1/4}\sqrt{\frac{\left(h^2 + hh_p + h_p^2\right)\left((nh)^{3/2} - (n_p h_p)^{3/2}\right)}{n - n_p}}.$$

If U satisfies (2.6) and (2.7), then U can be connected to U_p from the right followed by a *1-shock wave*. Since the system (2.1) is strictly hyperbolic, it is clear that $\sqrt{\frac{1}{2C}h^3 n} < h^2$. Thus the 1-shock curve is given by

$$S_1(U_p) = \{(h, n) : (n - n_p)\left(h^2 + hh_p + h_p^2\right) = \sqrt{\frac{2}{C}}\left((nh)^{3/2} - (n_p h_p)^{3/2}\right), \quad h^3 n < h_p^3 n_p\}. \tag{2.9}$$

Similarly, U can be connected to U_p from the right followed by a *2-shock wave*, provided U satisfies (2.3) and (2.8). This curve is called the 2-shock curve, which is given

by

$$S_2(U_p) = \{(h,n) : (n-n_p)\left(h^2 + hh_p + h_p^2\right) = \sqrt{\frac{2}{C}}\left((nh)^{3/2} - (n_p h_p)^{3/2}\right), \quad h < h_p\}. \tag{2.10}$$

We consider candidates of right states $U_R = U = (h,n)$ which can be connected to a given left state $U_L = U_p = (h_p, n_p)$ followed by a *rarefaction wave*. Here we note that the condition for (physically relevant) rarefaction waves is that the corresponding eigenvalue (speed) λ increases from the left to the right side of the wave (see [6]), that is

$$\lambda(U_p) < \lambda(U). \tag{2.11}$$

The Riemann problem (2.1), (2.4) are invariant under the scaling $(x,t) \mapsto (\eta x, \eta t)$ for all $\eta > 0$. Therefore we seek self-similar solutions of the form $U(x,t) \equiv U(\frac{x}{t})$. If we let $\xi = \frac{x}{t}$, then we see that $U(\xi)$ satisfies the ordinary differential equation

$$(DF(U) - \xi) d_\xi U = 0,$$

where $d_\xi = \frac{d}{d\xi}$. If $d_\xi U \neq 0$, then ξ is the eigenvalue for $DF(U)$ and $d_\xi U$ is the corresponding eigenvector. Since $DF(U)$ has two real and distinct eigenvalues $\lambda_1 < \lambda_2$, there exist two families of rarefaction waves, 1-*rarefaction waves* and 2-*rarefaction waves*. For 1-rarefaction waves, the eigenvector $d_\xi U = (d_\xi h, d_\xi n)^\top$ satisfies

$$(-\lambda_1(U)\boldsymbol{I} + DF(U)) d_\xi U = \begin{pmatrix} -\sqrt{\frac{1}{2C}h^3 n} + h^2 & 0 \\ \sqrt{\frac{1}{2C}n^3 h} & 0 \end{pmatrix} \begin{pmatrix} d_\xi h \\ d_\xi n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which gives $d_\xi h = 0$. Since $d_\xi n \neq 0$, we have

$$\frac{dh}{dn} = 0.$$

We integrate this to obtain the curve passing all possible U connected to U_p followed by a 1-rarefaction wave. This curve is called the 1-rarefaction curve, which is in our case given by

$$R_1(U_p) = \{(h,n) : h = h_p,\ n > n_p\}, \tag{2.12}$$

where $n > n_p$ comes from $\lambda_1(U_p) < \lambda_1(U)$.

For 2-rarefaction waves, the eigenvector $d_\xi U$ satisfies

$$(-\lambda_2(U)\boldsymbol{I} + DF(U)) d_\xi U = \begin{pmatrix} 0 & 0 \\ \sqrt{\frac{1}{2C}n^3 h} & -h^2 + \sqrt{\frac{1}{2C}h^3 n} \end{pmatrix} \begin{pmatrix} d_\xi h \\ d_\xi n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which gives

$$\frac{dh}{dn} = \frac{h^2 - \sqrt{\frac{1}{2C}h^3 n}}{\sqrt{\frac{1}{2C}n^3 h}} = \left(\sqrt{2C}\sqrt{\frac{h}{n}} - 1\right)\frac{h}{n}.$$

We can solve this ordinary differential equation, the solution is given by

$$h = \frac{n}{(\sqrt{\frac{C}{2}} - e^A n)^2},$$

where e^A is the constant of integration. When the solution takes $U_p = (h_p, n_p)$, the constant e^A is determined as $\frac{1}{n_p}(\sqrt{\frac{C}{2}} - \sqrt{\frac{n_p}{h_p}})$ then the special solution is obtained as

$$h = \frac{n\, n_p^2}{\left(n\sqrt{\frac{n_p}{h_p}} - (n - n_p)\sqrt{\frac{C}{2}}\right)^2}.$$

The graph of this function is called the 2-rarefaction curve consisting of U which can be connected from the left state U_p by a 2-rarefaction wave. We denote by

$$R_2(U_p) = \{(h, n) : h\left(n\sqrt{\frac{n_p}{h_p}} - (n - n_p)\sqrt{\frac{C}{2}}\right)^2 = n\, n_p^2,\ h > h_p\}, \tag{2.13}$$

where the condition $h_p < h$ comes from $\lambda_2(U_p) < \lambda_2(U)$.

3. Admissble weak solutions for the settled regime. In this section we construct weak solutions of Riemann problem (2.1), (2.4) by substituting the values corresponding to the settle regime into the curves given in the previous section. We tackle the Riemann problem for situations wherein $h < h_p$ and $h > h_p$ representing a step-down and step-up function, respectively.

We begin with finding admissible wave curves connecting from the fixed left state U_0 to the right states $U = (h, n)$ when $h < h_0$. We set $U_0 = (h_0, n_0) = (1, 0.1)$ and $C = 2.307$, which are used in [4]. Then the Hugoniot locus becomes the set

$$S(U_0) : \left\{(n - 0.1)\left(h^2 + h + 1\right) = \sqrt{\frac{2}{2.307}}\left((nh)^{3/2} - (0.1)^{3/2}\right)\right\}, \tag{3.1}$$

while the 1-entropy inequality and the 2-entropy inequality are as follows, respectively :

$$\sqrt{\frac{1}{4.614}h^3 n} < s < \min\left\{\sqrt{\frac{1}{46.14}},\, h^2\right\}, \tag{3.2}$$

$$\max\left\{\sqrt{\frac{1}{46.14}},\, h^2\right\} < s < 1, \tag{3.3}$$

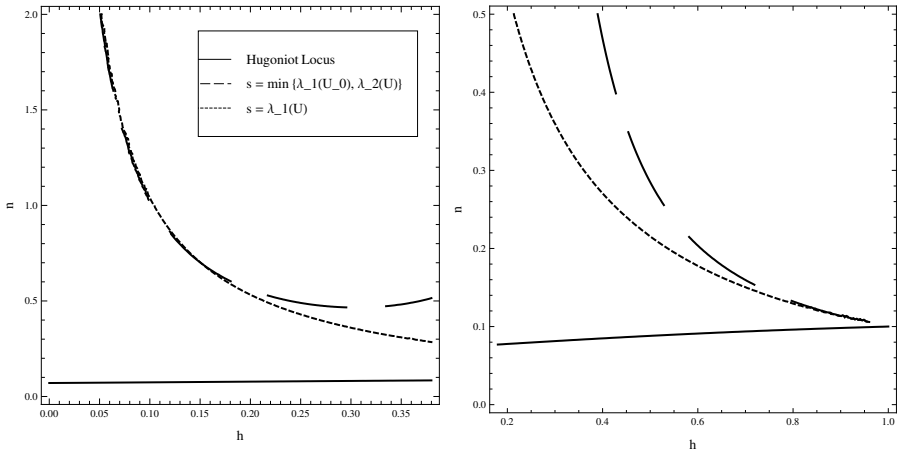
where

$$s = \left(\frac{2}{186.867}\right)^{1/4}\sqrt{\frac{(h^2 + h + 1)\left((nh)^{3/2} - (0.1)^{3/2}\right)}{n - 0.1}}. \tag{3.4}$$

We note that inequalities (3.2), (3.3) are equivalent to the following inequalities :

$$s - \sqrt{\frac{1}{4.614}h^3 n} > 0 \quad \text{and} \quad s - \min\left\{\sqrt{\frac{1}{46.14}},\, h^2\right\} < 0, \tag{3.5}$$

$$s - \max\left\{\sqrt{\frac{1}{46.14}},\, h^2\right\} > 0 \quad \text{and} \quad s - 1 < 0. \tag{3.6}$$

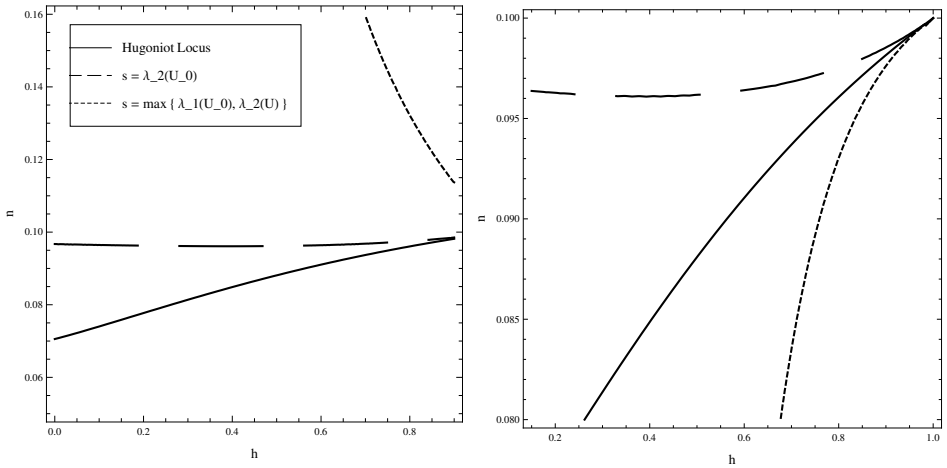


(a) $\lambda_1(U_0) \geq \lambda_2(U)$ (b) $\lambda_1(U_0) < \lambda_2(U)$

Fig. 3.1: Hugoniot locus and the 1-entropy inequality. We plot the Hugoniot locus (3.1) and implicit functions $s = \lambda_1(U)$ and $s = \min\{\lambda_1(U_0), \lambda_2(U)\}$, where $\lambda_1(U) = \sqrt{\frac{1}{4.614}h^3 n}$, $\lambda_1(U_0) = \sqrt{\frac{1}{46.14}}$, $\lambda_2(U) = h^2$. The solid, dashed and dotted curves represent the Hugoniot locus (3.1), $s = \min\{\lambda_1(U_0), \lambda_2(U)\}$ and $s = \lambda_1(U)$ respectively.

We shall examine whether there exists (h, n) satisfying (3.1) and (3.5) with phase portraits. In Figure 3.1 we plot the Hugoniot locus (3.1) and the implicit functions $s = \sqrt{\frac{1}{4.614}h^3 n}$ and $s = \min\{\sqrt{\frac{1}{46.14}}, h^2\}$, which is $s = h^2$ for the case $\sqrt{\frac{1}{46.14}} \geq h^2$ (Figure 3.1(a)) and $s = \sqrt{\frac{1}{46.14}}$ for the case $\sqrt{\frac{1}{46.14}} < h^2$ (Figure 3.1(b)). Two dashed lines in Figure 3.1 show the upper bound and lower bound for the inequality (3.5), which means that every point (h, n) within the open region between the upper graph $s = \min\{\sqrt{\frac{1}{46.14}}, h^2\}$ and the lower graph $s = \sqrt{\frac{1}{4.614}h^3 n}$ satisfies (3.5). As can be seen from the figure, (h, n) satisfying the Rankine-Hugoniot relation (3.1) does not belong to the region that the 1-entropy inequality (3.5) holds. Thus, the weak solution does not admit 1-shock waves.

Similarly, we examine whether there exists a (right) state (h, n) satisfying (3.1) and (3.6). In Figure 3.2 we plot the Hugoniot locus (3.1) and the implicit functions $s = 1$ and $s = \max\{\sqrt{\frac{1}{46.14}}, h^2\}$. When $\sqrt{\frac{1}{46.14}} \geq h^2$, every point (h, n) satisfying the Rankine-Hugoniot relation (3.1) does not belong to the region between the upper graph $s = 1$ and the lower graph $s = h^2$ (Figure 3.2(a)). On the other hand, when $\sqrt{\frac{1}{46.14}} < h^2$, the Hugoniot locus $S(U_0)$ belongs to the region between the upper graph $s = 1$ and the lower graph $s = \sqrt{\frac{1}{46.14}}$ (Figure 3.2(b)), which means that there exists (h, n) satisfying both (3.1) and (3.6). Thus, when $\sqrt{\frac{1}{46.14}} < h^2$, the 2-shock wave exists and the 2-shock curve is given by (3.1) for $h < 1$.



(a) $\lambda_1(U_0) \geq \lambda_2(U)$ (b) $\lambda_1(U_0) < \lambda_2(U)$

Fig. 3.2: Hugoniot locus and the 2-entropy inequality. In this figure we plot the Hugoniot locus (3.1) and implicit functions $s = \lambda_2(U_0)$ and $s = \max\{\lambda_1(U_0), \lambda_2(U)\}$, where $\lambda_1(U_0) = \sqrt{\frac{1}{46.14}}$, $\lambda_2(U_0) = 1$, $\lambda_2(U) = h^2$. The solid, dashed and dotted lines represent the Hugoniot locus (3.1), $s = \lambda_2(U_0)$ and $s = \max\{\lambda_1(U_0), \lambda_2(U)\}$ respectively.

As a example, we take $U_2 = (0.2, n_{2,s})$ [1], where $n_{2,s}$ is the solution of

$$1.24\,(n_{2,s} - 0.1) = \sqrt{\frac{2}{2.307}}\left((0.2\,n_{2,s})^{3/2} - (0.1)^{3/2}\right), \tag{3.7}$$

which is exactly the equation (3.1) with $U = U_2$. Then the left state $U_0 = (1, 0.1)$ and the right state U_2 is connected by a single 2-shock wave. In the range $h < 1$,(2.9) and (2.10) make no sense as 1-shock wave and 2-shock wave by the entropy inequalities, respectively.

Similarly, we find admissible wave curves in the case $h > h_0$. Fix $U_0 = (h_0, n_0) = (0.4, 0.08)$ and $C = 2.307$, and we plot the 1-rarefaction curve and 2-rarefaction curve, which are given as follows, respectively :

$$h = 0.4, \qquad n > 0.08, \tag{3.8}$$

$$h = \frac{n\,(0.08)^2}{\left(n\sqrt{0.2} - (n - 0.08)\sqrt{\frac{2.307}{2}}\right)^2}, \qquad h > 0.4, \tag{3.9}$$

which means that (3.8) makes no sense as 1-rarefaction [2], but (3.9) makes sense as 2-rarefaction by (2.11).

[1] Using Newton's method, a sample of the approximate solution for equation (3.7) is obtained as $n_{2,s} = 0.0777100325$.

[2] When $h \neq h_p$, which is typical as phenomena of fluid motion [3], 1-rarefaction waves do not exist. On the other hand, if we admit $h = h_p$, a 1-rarefaction wave connecting (h_p, n_p) and (h_p, n) with $n_p < n < h_p$ is also admitted.

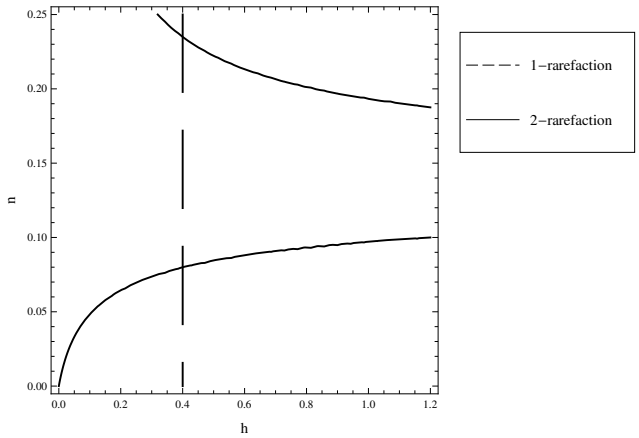


Fig. 3.3: In this figure we plot a graph of two rarefaction wave curves (3.8) and (3.9). The dashed and solid lines represent the 1-rarefaction wave curve (3.8) and the 2-rarefaction wave curve (3.9) respectively.

w_1	w_2	appear ance
1-rarefaction		△
1-rarefaction	2-rarefaction	△
1-rarefaction	1-shock wave	×
1-rarefaction	2-shock wave	×
2-rarefaction	1-rarefaction	×
2-rarefaction		◯
2-rarefaction	1-shock wave	×
2-rarefaction	2-shock wave	×

w_1	w_2	appear ance
1-shock wave	1-rarefaction	×
1-shock wave	2-rarefaction	×
1-shock wave		×
1-shock wave	2-shock wave	×
2-shock wave	1-rarefaction	×
2-shock wave	2-rarefaction	×
2-shock wave	1-shock wave	×
2-shock wave		◯

Table 3.1: Combination of solutions to appearance. w_i $(i = 1, 2)$ denote the simple wave in the i-characteristic field.

As an example, we take $U_2 = (1.0, n_{2,r})$ [3], where $n_{2,r}$ is the solution of

$$0.08\sqrt{n_{2,r}} + (n_{2,r} - 0.08)\sqrt{\frac{2.307}{2}} = n_{2,r}\sqrt{0.2}. \tag{3.10}$$

Then the left state $U_0 = (0.4, 0.08)$ and the right state U_2 is connected by a single 2-rarefaction wave.

Our argument is summarized in Table 3.1. Following the terminology "allowed sequence" of waves in [5], wave sequences consisting of shocks and rarefactions associated with *identical* characteristic fields are excluded.

4. Conclusions. In this paper we have dealt with a Riemann problem for the system of conservation laws (1.5)–(1.6) which is derived from the dilution approxima-

[3]Using Newton's method, a sample of the approximate solution for equation (3.10) is obtained as $n_{2,r} = 0.0972723141$.

tion of a suspension flow on an incline as a mathematical model in the settled regime. Murisic et al. [4] dealt only with a exact solution for the system (1.5)–(1.6), when the initial volume fraction is fixed as $\phi(x,0) \equiv f_0$ for some given $f_0 \ll 1$. On the other hand, we aim at covering the solution of this system when the initial volume fraction $\phi(x,0)$ is a variable satisfying $0 < \phi < 1$. In Sections 2 and 3, we show that the weak solution of this Riemann problem is connected by a single 2-rarefaction wave from the left state $U_0 = (h_0, n_0)$ to the right state $U_2 = (h_2, n_2)$ when $h_0 < h_2$, and connected by a single 2-shock wave when $h_0 > h_2$. To illustrate one example of these wave curves, we impose the initial conditions as follows,

$$U^r(x,0) = \begin{cases} U_0 = (0.4, 0.08) & x < 0 \\ U_2 = (1.0, n_{2,r}) & x > 0 \end{cases}, \qquad U^s(x,0) = \begin{cases} U_0 = (1.0, 0.1) & x < 0 \\ U_2 = (0.2, n_{2,s}) & x > 0 \end{cases},$$

where $n_{2,s}$ and $n_{2,r}$ is the solution of (3.7) and (3.10) respectively. We take the values of $U^r(x,0)$ and $U^s(x,0)$ to satisfy the ranges $0 \le h \le 1$ and $0 \le n \le 0.1$ of the exact solution handled in [4]. With the Riemann data $U^r(x,0)$, the weak solution consists of a single 2-rarefaction wave whose curve is shown in Figure 3.3. With the Riemann data $U^s(x,0)$, the weak solution consists of a single 2-shock wave whose curve is shown in Figure 3.2(b). The construction method given in Sections 2 and 3 may also be useful for other suspension models even if the initial volume fraction ϕ depends on x. The correspondence between rarefaction wave and shock wave obtained from (1.5)-(1.6) and experimental results in [4], as well as solutions of (general) initial value problems discussed there, will be a next issue.

5. Acknowledgements. This work is supported by 2017 IMI Joint Use Research Program CATEGORY "Short-term Visiting Researcher"in Institute of Mathematics for Industry, Kyushu University. KM was partially supported by Program for Promoting the reform of national universities (Kyushu University), Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, World Premier International Research Center Initiative (WPI), MEXT, Japan and JSPS Grant-in-Aid for Young Scientists (B) (No. JP17K14235). We finally would like to thank the reviewer for providing us with helpful comments to correct this paper and to think of successive studies.

REFERENCES

[1] H. Huppert. Flow and instability of a viscous current down a slope. Nature, **300** 427–429, (1982).
[2] P. D. Lax, Hyperbolic system of conservation laws II, Comm. Pure Appl. Math. **10** 537–566, (1957).
[3] A. Mavromoustaki, A. L. Bertozzi, Hyperbolic systems of conservation laws in gravity-driven, particle-laden thin-film flows, Journal of Engineering Mathematics **88** 29–48, (2014).
[4] N. Murisic, B. Pausader, D. Peschka, A. L. Bertozzi, Dynamics of particle settling and resuspension in viscous liquids, J. Fluid Mech. **717** 203–231, (2013).
[5] S. Schecter, D. Marchesin, B. J. Plohr, Structurally stable Riemann solutions, J. Differential Equations **126** no. 2, 303–354, (1996).
[6] J. Smoller, Shock waves and reaction-diffusion equations, **258** of Grundlehren der Mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences). Springer-Verlag, New York, second edition, (1994).

Proceedings of EQUADIFF 2017
pp. 359–368

BEHAVIOUR OF THE SUPPORT OF THE SOLUTION APPEARING IN SOME NONLINEAR DIFFUSION EQUATION WITH ABSORPTION *

KENJI TOMOEDA†

Abstract. Numerical experiments suggest interesting properties in the several fields of fluid dynamics, plasma physics and population dynamics. Among such properties, we may observe the interesting phenomena; that is, the *repeated appearance and disappearance phenomena of the region penetrated by the fluid* in the flow through a porous media with absorption. The model equation in two dimensional space is written in the form of the initial-boundary value problem for a nonlinear diffusion equation with the effect of absorption. In this paper we show some numerical examples and prove such phenomena.

Key words. nonlinear diffusion, support dynamics, finite difference scheme

AMS subject classifications. 35K65, 35B99, 65M06

1. Introduction. We are concerned with the dynamical behaviour of the region penetrated by the fluid in the filtration of the flow through an absorbing medium. The representative filtration is well known as the flow through porous media where the water evaporates. In particular, it is expected that such a seepage exhibits the *repeated appearance and disappearance phenomena of such a region*, which are caused by the interaction between the nonlinear diffusion and the penetration of the fluid from the boundary on which the flowing tide and the ebbing tide occur. To realize such phenomena we introduce the simplest model based on the following nonlinear diffusion equation with absorption in two dimensional space [7, 9] :

$$\begin{cases} v_t(t,x,y) = \Delta(v^m) - cv^p & \text{in} \quad (0,\infty)\times\Omega, \\ v(t,x,y) = \psi(t,x,y) & \text{on} \quad (0,\infty)\times\partial\Omega, \\ v(0,x,y) = v^0(x,y) & \text{on} \quad \Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded domain in $\mathbf{R}^2$ with piecewisely smooth boundary $\partial\Omega$, and satisfies the exterior sphere condition. Moreover, $v(\geq 0)$ denotes the density of the fluid, $m > 1$, $0 < p < 1$, $c > 0$, $m + p \geq 2$, and $v^0(x,y)$ and $\psi(t,x,y)$ are non-negative continuous functions. This equation is also used to describe the propagation of thermal waves in plasma physics [8].

From analytical points of view, the existence and uniqueness of a weak solution and the comparison theorem are proved by Bertsch [1].

We state some mathematical results in one dimensional case. For the initial value problem, Rosenau and Kamin [8] suggested the *support splitting phenomena* in several numerical examples, and we also constructed the initial function for which the *repeated support splitting and merging phenomena* appear [10]. Here the support means the region penetrated by the fluid; that is, the region where $v > 0$. For the

*This work was supported by Japan Society for the Promotion of Science through Grand-in-Aid for Scientific Research(C):No.16K05271.

†Graduate School of Informatics, Kyoto University, Sakyo-ward, Kyoto 606-8501, JAPAN (ktomoeda@acs.i.kyoto-u.ac.jp).

initial-boundary value problem Kersner proved the appearance of the *support splitting phenomena* [5], but he did not show that the *support merging phenomena* appear after the support splits. To investigate the occurrence of the *repeated support splitting and merging phenomena*, we construct two stationary solutions, the one is the support non-splitting solution and the other is the support splitting solution. We proved this occurrence by imposing the periodicity on the boundary value which takes the value greater than the former boundary value and less than the latter [11].

In two dimensional case, to the best of my knowledge, we are unable find any result concerned with the *repeated appearance and disappearance phenomena of the support*. By employing the argument used in the one dimensional case we try to justify the occurrence of such phenomena.

2. Stationary solutions and numerical examples. In this section we consider the profile of the stationary solution $w(x,y)(\geq 0)$ satisfying

$$\begin{cases} \Delta(w^m) - cw^p = 0 & \text{in} \quad \Omega, \\ w(t,x,y) = \varphi(x,y) & \text{on} \quad \partial\Omega, \end{cases} \tag{2.1}$$

where $m > 1$, $0 < p < 1$, $c > 0$, $m + p \geq 2$, and $\varphi(x,y)$ is a non-negative continuous function on $\partial\Omega$. Then the existence and uniqueness of the solution $w(x)$ follows (see Theorem 12.5 in [12]).

THEOREM 2.1. *The equation (2.1) has the unique solution $w(x,y)$ such that $w^m(x,y) \in C^{2,\frac{p}{m}}(\Omega) \cap C^0(\bar{\Omega})$.*

We introduce the radial solution of (2.1), which is used in the numerical computation. Put

$$\phi(x,y) = \left\{ \left(\frac{m-p}{2m} \right)^2 c(x^2+y^2) \right\}^{\frac{1}{m-p}}. \tag{2.2}$$

It is obvious that (2.2) satisfies the first equation of (2.1) with $w = \phi$ and $w(x,y) > 0$ for $(x,y) \neq (0,0)$.

To investigate the behaviour of the support of the solution (1.1) we tried numerical computation by our difference scheme, which approximates the following problem instead of (1.1):

$$\begin{cases} u_t(t,x,y) = mu\Delta u + a(u_x^2 + u_y^2) - (m-1)cu^q & \text{in} \quad (0,\infty)\times\Omega, \\ u(t,x,y) = \psi^{m-1}(t,x,y) & \text{on} \quad (0,\infty)\times\partial\Omega, \\ u(0,x,y) = (v^0)^{m-1}(x,y) & \text{on} \quad \Omega, \end{cases} \tag{2.3}$$

where $a = \frac{m}{m-1}$ and $q = \frac{m+p-2}{m-1}$, and this equation can be obtained by putting $u = v^{m-1}$ [6, 10].

We put $m = 1.5$, $p = 0.5$, $c = 6$, $\Omega = (-1.5,\ 1.5) \times (-1.5,\ 1.5)$ and the space mesh width $h = \frac{1}{32}$, and show the numerical profiles in Cases I and II.

Case I. The boundary condition $\psi(t,x,y)$ is independent of t; that is,

Example 1. $\psi(t,x,y) = \left\{\phi^{m-1}(x,y) + 0.5\right\}^{\frac{1}{m-1}}$ on $\partial\Omega$ (Left figures in Fig. 2.1),

where $v^0(x,y) = \left\{\phi^{m-1}(x,y) + 0.5 + 1.25\cos\theta(x)\cos\theta(y)\right\}^{\frac{1}{m-1}}$ on Ω;

Example 2. $\psi(t,x,y) = \left\{\phi^{m-1}(x,y) - 0.5\right\}^{\frac{1}{m-1}}$ on $\partial\Omega$ (Right figures in Fig. 2.1),

where $v^0(x,y) = \left\{[\phi^{m-1}(x,y) - 0.5]_+ + 1.5\cos^2\theta(x)\cos^2\theta(y)\right\}^{\frac{1}{m-1}}$ on Ω.

In both examples we put $\theta(\eta) = -\dfrac{\pi}{2} + \dfrac{\pi}{3}(\eta + 1.5)$ $(-1.5 \leq \eta \leq 1.5)$.

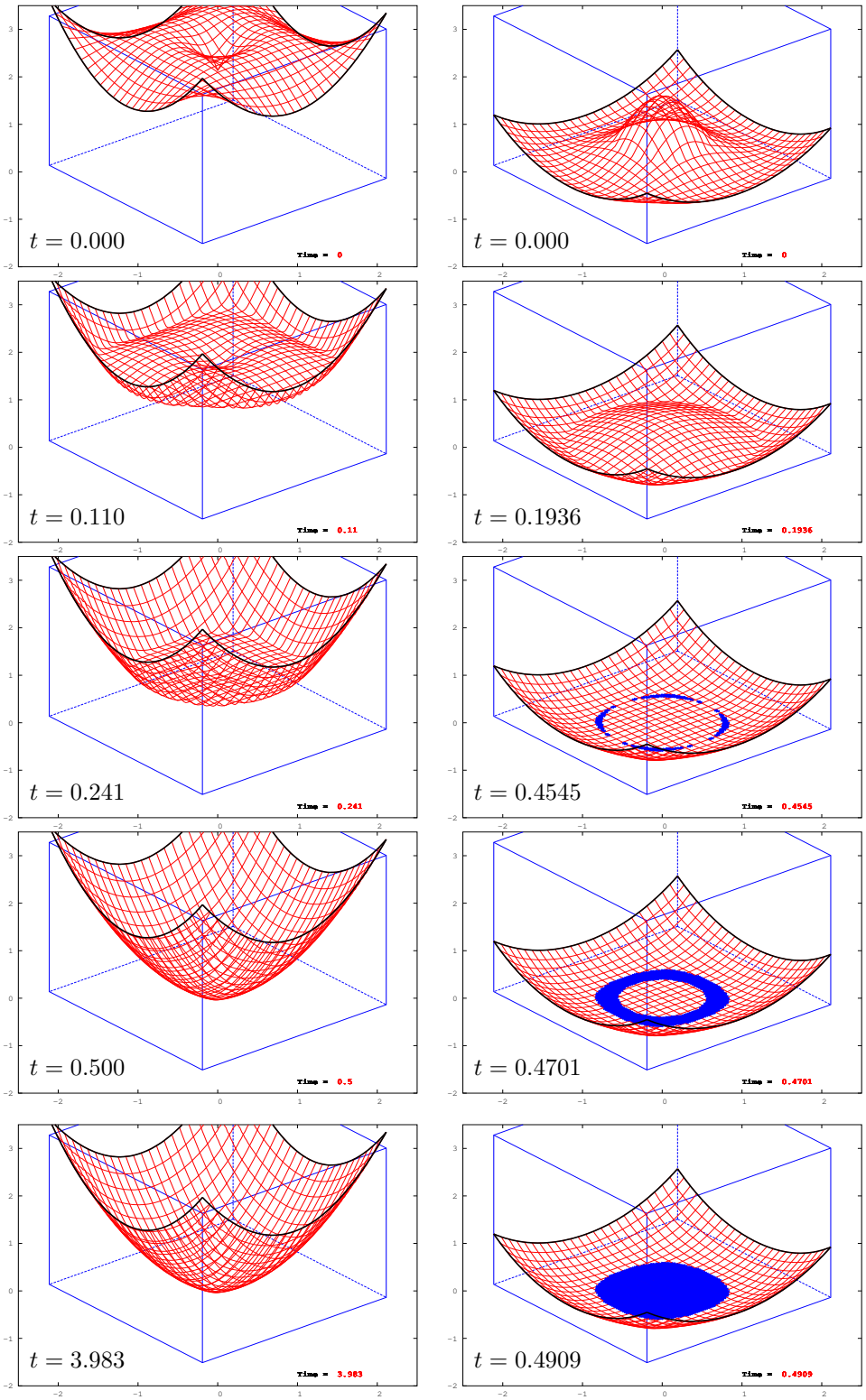


FIG. 2.1. *The non-appearance and appearance phenomena of the region where* $v = 0$

The region where $v = 0$, which is indicated in black, begins to appear in the right figure with $t = 0.4545$, but does not in the left figures. We note that such a region in the right figure with $t = 0.4909$ remains until $t = 3.972$ at which we stop computation. Thus numerical solutions converge to the stationary solutions as t increases, respectively.

Case II. We impose a period on $\psi(t, x, y)$ and put $v^0(x, y) = \phi(x, y)$ on Ω; that is,

Example 3. $\psi(t, x, y) = \left\{\phi^{m-1}(x, y) + 0.5 \sin(2\pi t)\right\}^{\frac{1}{m-1}}$ on $\partial\Omega$ (Fig. 2.2);

Example 4. $\psi(t, x, y) = \left\{\phi^{m-1}(x, y) + 0.5 \sin(8\pi t)\right\}^{\frac{1}{m-1}}$ on $\partial\Omega$ (Fig. 2.3).

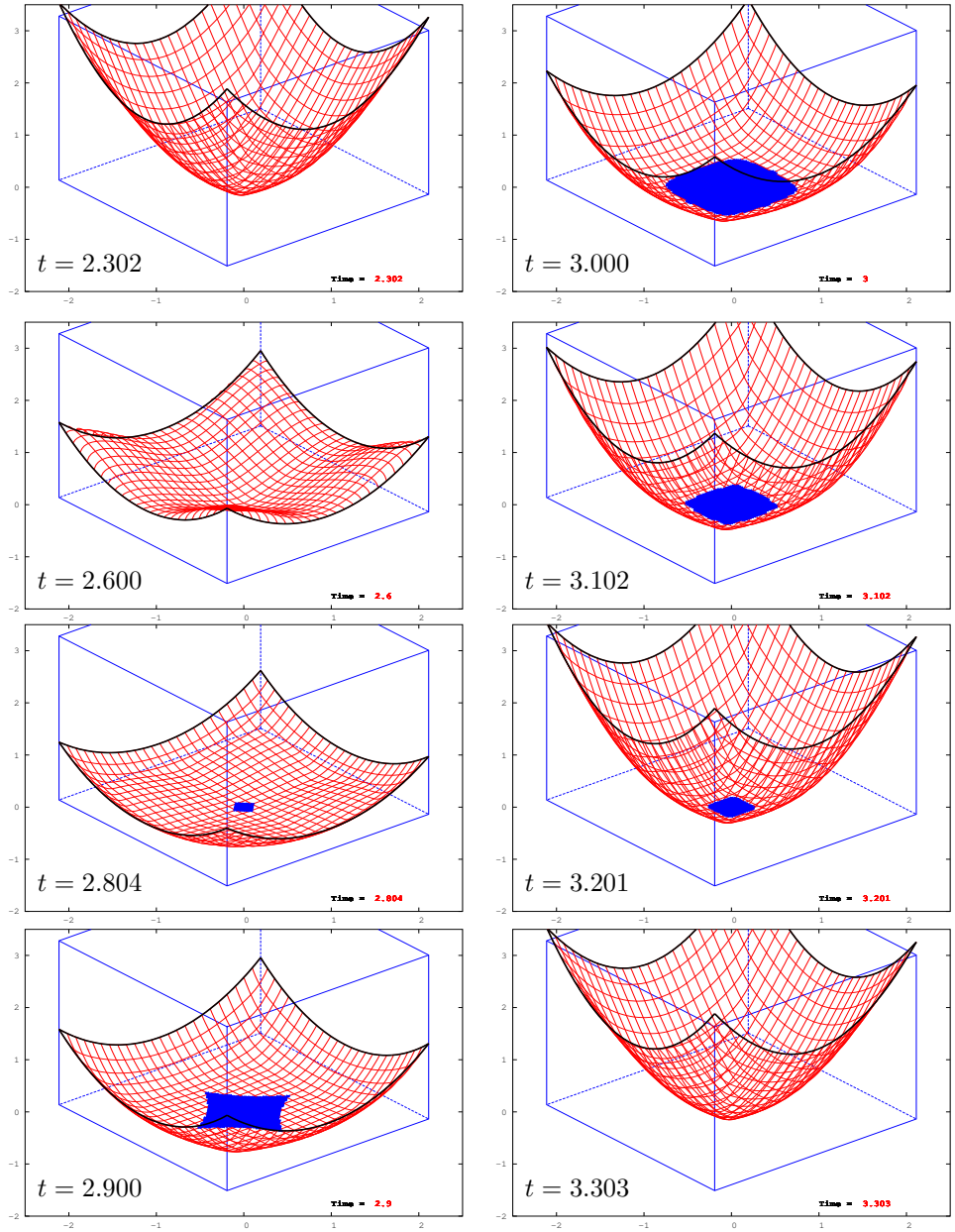


FIG. 2.2. *The repeated appearance and disappearance phenomena of the region where* $v = 0$.

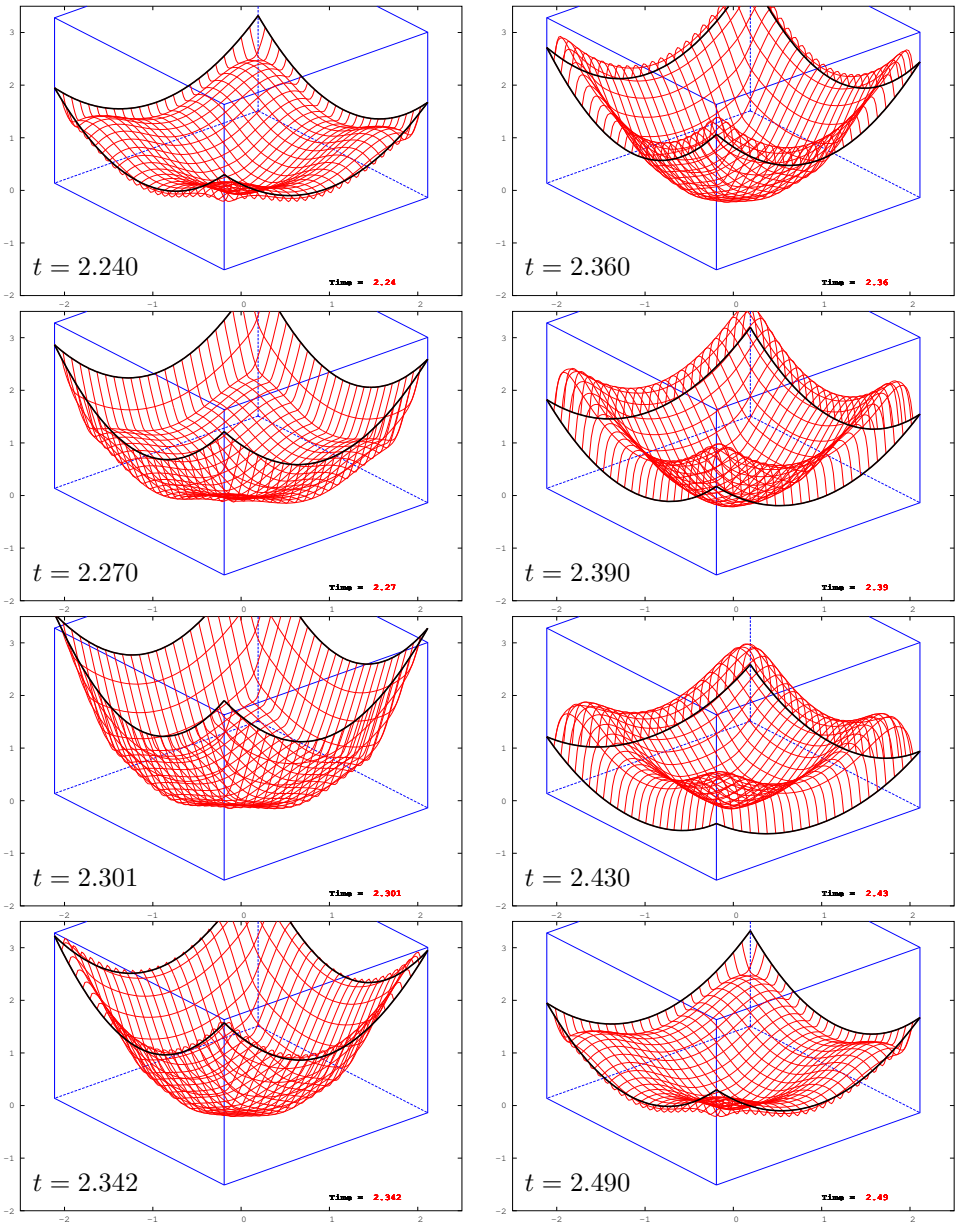


FIG. 2.3. *The appearance of the region where* $v = 0$ *is not observed.*

In Figs. 2.2 and 2.3 the initial and boundary profiles are located as equal to the same stationary solution $\phi(x, y)$. The increasing and decreasing profiles appear repeatedly as t increases in both figures. The region where $v = 0$ appears at $t = 2.804$ and disappears at $t = 3.303$ in Fig. 2.2. We may observe that the profile of the numerical solution at $t = 2.302$ approximately coincides with the one at $t = 3.303$. The numerical period $1.001 = 3.303 - 2.302$ corresponds to 1.00 of $\psi(t, x, y)$ in Example 3. Thus Fig. 2.2 shows the *repeated appearance and disappearance phenomena of the region where* $v = 0$. On the other hand, Fig. 2.3 shows the numerical period $0.250 =$

$2.490 - 2.240$, which coincides with $\frac{1}{4}$ of $\psi(t,x,y)$ in Example 4. The numerical solution is close to zero in the neighborhood of $(x,y)=(0,0)$, but not equal to zero. The region where $v=0$ never appears.

We mention our numerical method for (2.3), which is the following explicit finite difference scheme:

$$u_h^{n+1} = P_{k,h} D_{k,h} H_{k,h} u_h^n \quad (n = 0, 1, \cdots). \tag{2.4}$$

Here $u_h^n(x,y)$ is the numerical approximation to the solution $u(t_n,x,y)$. $P_{k,h}$, $D_{k,h}$ and $H_{k,h}$ approximate $u_t = mu\Delta u$, $u_t = -(m-1)cu^q$ and $u_t = a(u_x^2 + u_y^2)$, respectively, and $k \equiv k_{n+1} = t_{n+1} - t_n$ is a variable time step determined by

$$k_{n+1} = \frac{h}{2a \max(\|(u_h^n)_x\|_\infty, \|(u_h^n)_y\|_\infty)}. \tag{2.5}$$

Since $h = \frac{1}{32}$, it is observed that $k_{n+1} \approx 6.0 \times 10^{-4} \sim 5.2 \times 10^{-3}$ in Examples 1-4, which is very small. This may affect the shape of the region where $v=0$. In the right figures of Fig. 2.1 such a region looks like a square after $t = 0.4909$. Unfortunately, we are unable to analyze the appearance of such a figure. However, it is expected that these numerical solutions in Examples 1-4 qualitatively capture the *appearance and disappearance phenomena of the region where* $v=0$.

In the following section we prove the properties appearing in Figs. 2.1–2.2. At the present time it seems difficult for us to prove the occurrence of such phenomena in Fig. 2.3.

3. Stabilization.

THEOREM 3.1 (Stabilization). *Let $v(t,\cdot)$ be the solution of (1.1) with $\psi(t,x,y) = \varphi(x,y)$, where $\varphi(x,y)$ is a non-negative continuous function on $\partial\Omega$. Then $v(t,\cdot)$ converges to the unique stationary solution $w(\cdot)$ of (2.1) in $C(K)$ as $t \to \infty$, where $K \subset \Omega$ is an arbitrary fixed compact set.*

Proof. We state the proof briefly. For the solution $v(t,\cdot)$ we consider a continuous orbit $\gamma = \{v(t,\cdot) : t \geq 0\}$ in $C(K)$. By the result of DiBenedetto [2], γ is precompact in $C(K)$; that is,

$$^{\exists}\{t_n\}, ^{\exists}\hat{v}(\cdot): \quad t_n \to \infty \ \text{ and } \ v(t_n,\cdot) \to \hat{v}(\cdot) \in \omega \ \text{ in } \ C(K) \ \text{ as } \ n \to \infty,$$

where ω is the ω-limit set of γ. On the other hand, the following inequality is proved for the solutions $v_1(t,\cdot)$ and $v_2(t,\cdot)$ of (1.1) by Bertsch [1]:

$$\|v_1(t,\cdot) - v_2(t,\cdot)\|_{L^1(\Omega)} \leq e^{Mt} \|v_1(0,\cdot) - v_2(0,\cdot)\|_{L^1(\Omega)} \text{ for } t \geq 0, \tag{3.1}$$

where M is the constant number satisfying

$$(-s^p) - (-r^p) \leq M(s-r) \text{ for any } (0 \leq r \leq s). \tag{3.2}$$

In general, $M = 0$. However, taking the boundedness of the solution v of (1.1) and the stationary solution w of (2.1) into consideration, we can take $|M(v,w)| << 1$ $(M(v,w) < 0)$ depending on v and w, and obtain

$$\begin{aligned} \|v(t_n,\cdot) - w(\cdot)\|_{L^1(K)} &\leq \|v(t_n,\cdot) - w(\cdot)\|_{L^1(\Omega)} \\ &\leq e^{M(v,w)t_n} \|v(0,\cdot) - w(\cdot)\|_{L^1(\Omega)} \quad \text{for} \quad t \geq 0, \end{aligned} \tag{3.3}$$

which tends to 0 as $n \to \infty$. Thus $\hat{v}(x,y) = w(x,y)$ holds on K, and the theorem follows from the uniqueness of the stationary solution $w(x,y)$. □

Let $w_i(x,y)$ $(i=1,2)$ be two non-negative solutions of (2.1) satisfying $w_2(x,y) > w_1(x,y)$ on $\bar{\Omega}$. Assume that $w_1(x,y)$ has the non-empty region where $w_1(x,y) = 0$. Then Theorem 3.1 predicts the following result:

If the time while we keep the boundary value $\psi(t,x,y)$ of $v(t,x,y)$ greater than $w_2(x,y)$ on $\partial\Omega$ is sufficiently long, then $v(t,x,y) > 0$ on Ω. Conversely, if the time while we keep $\psi(t,x,y)$ less than $w_1(x,y)$ on $\partial\Omega$ is sufficiently long, then the region where $v(t,x,y) = 0$ appears.

Thus we may expect the *repeated appearance and disappearance phenomena of the region where $v = 0$* by imposing the period and magnitude on $\psi(t,x,y)$.

However, since Theorem 3.1 does not guarantee the appearance of the region where $v = 0$ in finite time, it is unclear in Fig. 2.1 and 2.2 whether or not such phenomena occur. So, we will prove it for the specific case in the next section.

4. Galaktionov and Vazquez's particular solution. Let $m + p = 2$ and $0 < p < 1$. Then we can construct the Galaktionov-Vazquez's particular solution which satisfies the first equation of (1.1)[3, 4]. We briefly state its construction. In the first equation of (2.3) we find $q = 0$ and have

$$u_t = mu\Delta u + a(u_x^2 + u_y^2) - (m-1)c\chi_{\{u>0\}}, \quad a = \frac{m}{m-1}. \tag{4.1}$$

We assume that this explicit solution is written in the form $u(t,x) = f(t)+g(t)h(x,y)$. Then f, g and h satisfy

$$f' + g'h = m(f+gh)g(h_{xx} + h_{yy}) + ag^2(h_x^2 + h_y^2) - (m-1)c, \tag{4.2}$$

where $'$ denotes the derivative with respect to t. Let $h(x,y) = x^2 + y^2$ in (4.2). Then we have

$$\begin{cases} f' = 4mfg - (m-1)c, \\ g' = 4(m+a)g^2. \end{cases} \tag{4.3}$$

Solving (4.3), we obtain a solution for two parameters $\varepsilon > 0$ and $\hat{\sigma} > 0$:

$$u(t,x,y) = \{E - 4(m+a)t\}^{-1} \times \left[D\{E-4(m+a)t\}^2 + G\{E-4(m+a)t\}^{\frac{1}{m}} + x^2 + y^2\right]_+, \tag{4.4}$$

$$u(0,x,y) = \varepsilon(x^2+y^2) + \hat{\sigma}, \tag{4.5}$$

where
$G \equiv G(m,c,\hat{\sigma},\varepsilon) = (\hat{\sigma} - DE)E^{\frac{m-1}{m}}$, $D \equiv D(m,c) = \dfrac{(m-1)c}{4(2m+a)}$, $E \equiv E(\varepsilon) = \varepsilon^{-1}$.

Using the same argument as used in [10], we have

LEMMA 4.1. *Let $\hat{\sigma} < DE$. Then*

$$\frac{\hat{\sigma}}{(m-1)c} < \hat{t}(m,c,\hat{\sigma},\varepsilon) < \hat{T}(m,\varepsilon) \tag{4.6}$$

holds and u satisfies

$$u(t,x,y) > 0 \text{ for } (t,x,y) \in [0,\ \hat{T}(m,\varepsilon)) \times \mathbf{R}^2 \setminus S, \tag{4.7}$$

$$u(t,x,y) = 0 \text{ for } (t,x,y) \in S, \tag{4.8}$$

$$\lim_{t \nearrow \hat{T}(m,\varepsilon)} u(t,0,0) = 0, \quad \lim_{t \nearrow \hat{T}(m,\varepsilon)} u(t,x,y) = \infty \text{ for } (x,y) \neq (0,0), \tag{4.9}$$

where

$$\hat{t}(m,c,\hat{\sigma},\varepsilon) = \frac{1}{4(m+a)} \left\{ E - \left(\frac{-G}{D} \right)^{\frac{m}{2m-1}} \right\}, \quad \hat{T}(m,\varepsilon) = \frac{E}{4(m+a)}, \tag{4.10}$$

$$S = \Big\{ (t,x,y) \mid t \in \Big[\hat{t}(m,c,\hat{\sigma},\varepsilon),\ \hat{T}(m,\varepsilon) \Big) \text{ and } x^2 + y^2 \leq \{E - 4(m+a)t\}^{\frac{1}{m}} \left[-G - D\{E - 4(m+a)t\}^{\frac{2m-1}{m}} \right] \Big\}. \tag{4.11}$$

(See Fig. 4.1 in the next page).

By the simple calculations we can show that $v(t,x,y) = u(t,x,y)^{\frac{1}{m-1}}$ satisfies the first equation of (1.1) for $(t,x) \in ((0,\ \hat{T}(m,\varepsilon)) \times \mathbf{R}^2) \backslash \partial S$. Since $1 < m < 2$, it follows that $\dfrac{1}{m-1} > 1$ and $\dfrac{m}{m-1} > 2$, which implies that $v_t(t,x,y) = \Delta(v^m)(t,x,y) = 0$ hold on $(t,x,y) \in \partial S \backslash (\hat{T}(m,\varepsilon),0,0)$. Thus $v(t,x,y)$ becomes the solution of the first equation of (1.1) on $(0,\hat{T}(m,\varepsilon)) \times \mathbf{R}^2$.

Under the specific case where $m + p = 2$ and $0 < p < 1$ we have

THEOREM 4.2. *Assume that $w(x,y)$ be the stationary solution of (2.1) with the non-empty region where $w(x,y) = 0$. Let $v(t,x,y)$ be the solution of (1.1) with $\psi(t,x,y) = w(x,y)$ on $\partial\Omega$. Then the region where $v(t,x,y) = 0$ appears in finite time.*

Proof. Without loss of generality we assume that $w(x,y) \equiv 0$ in the neighborhood of $(x,y) = (0,0)$. Let $GV(t,x,y;\ m,c,\hat{\sigma},\varepsilon)$ denote Galaktionov and Vazquez's solution written in the form of the term on the right side of (4.4). From the properties of this solution it is possible to take sufficiently large $\varepsilon(>0)$ so that there exists some constant $t_\varepsilon(>0)$ satisfying

$$GV(0,x,y;\ m,c,0,\varepsilon) \geq w^{m-1}(x,y) \quad \text{on } \bar{\Omega}, \tag{4.12}$$

$$GV(0,x,y;\ m,c,0,\varepsilon) > w^{m-1}(x,y) \quad \text{on } \partial\Omega, \tag{4.13}$$

$$GV(t,x,y;\ m,c,0,\varepsilon) > w^{m-1}(x,y) \quad \text{on } [0,\ t_\varepsilon) \times \partial\Omega. \tag{4.14}$$

Taking the positive number $\hat{\sigma} < DE$, we have $G < 0$. Then the region where $GV(t,x,y\ ;\ m,c,\hat{\sigma},\varepsilon) = 0$ appears at $t = \hat{t}(m,c,\hat{\sigma},\varepsilon)(>0)$ by Lemma 4.1. Moreover,

since $\hat{t}(m,c,\hat{\sigma},\varepsilon) \searrow 0$ as $\hat{\sigma} \searrow 0$, we take $\hat{\sigma}$ sufficiently small so that $\hat{t}(m,c,\hat{\sigma},\varepsilon) < t_\varepsilon$. We fix $\hat{\sigma}$ and ε. Then Theorem 3.1 (Stabilization) yields for sufficiently large t^*

$$GV(0,x,y;\ m,c,\hat{\sigma},\varepsilon) > u(t^*,x,y) \equiv v^{m-1}(t^*,x,y) \quad \text{on } \bar{\Omega}. \tag{4.15}$$

We have from (4.14) and (4.4)

$$\begin{aligned} GV(t,x,y;\ m,c,\hat{\sigma},\varepsilon) &> GV(t,x,y;\ m,c,0,\varepsilon) \\ &> w^{m-1}(x,y) = u(t^*+t,x,y) \\ &\equiv v^{m-1}(t^*+t,x,y) \quad \text{in } [0,\ t_\varepsilon) \times \partial\Omega. \end{aligned} \tag{4.16}$$

Applying the comparison theorem [1], which is concerned with the initial and boundary data, to (4.15) and (4.16), we obtain

$$GV(t,x,y;\ m,c,\hat{\sigma},\varepsilon) \geq u(t^*+t,x,y) \quad \text{on } [0,\ t_\varepsilon) \times \bar{\Omega}. \tag{4.17}$$

Thus the region where $v(t,\cdot) = u(t,\cdot)^{\frac{1}{m-1}} = 0$ appears at $t = t^* + \hat{t}(m,c,\hat{\sigma},\varepsilon)$, and the proof is complete. □

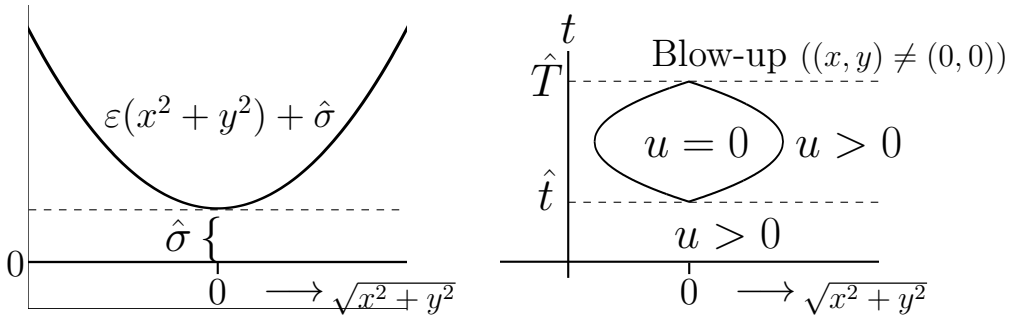


FIG. 4.1. *The initial function and the support of Galaktionov and Vazquez's solution (4.4).*

REFERENCES

[1] M. BERTSCH, *A class of degenerate diffusion equations with a singular nonlinear term*, Nonlinear Anal., 7 (1983), pp.117–127.

[2] E. DIBENEDETTO, *Continuity of weak solutions to a general porous medium equation*, Indiana Univ. Math. J., 32 (1983), pp.83–118.

[3] V. A. GALAKTIONOV AND J. L. VAZQUEZ, *Extinction for a quasilinear heat equation with absorption I. Technique of intersection comparison*, Commun. in Partial Differential Equation, 19 (1994), pp.1075–1106.

[4] V. A. GALAKTIONOV AND J. L. VAZQUEZ, *Extinction for a quasilinear heat equation with absorption II. A dynamical systems approach*, Commun. in Partial Differential Equation, 19 (1994), pp.1107–1137.

[5] R. KERSNER, *Degenerate parabolic equations with general nonlinearities*, Nonlinear Anal., 4 (1980), pp.1043–1062.

[6] T. NAKAKI AND K. TOMOEDA, *A finite difference scheme for some nonlinear diffusion equations in an absorbing medium: support splitting phenomena*, SIAM J. Numer. Anal., 40 (2002), pp.945–964.

[7] P.Y. Polubarinova-Kochina, *Theory of Ground Water Movement*, Princeton Univ. Press, 1962.
[8] P. Rosenau, S. Kamin, *Thermal waves in an absorbing and convecting medium*, Physica, 8D (1983), pp.273–283.
[9] A.E. Scheidegger, *The Physics of Flow through Porous Media*, Third edition, University of Toronto Press,1974.
[10] K. Tomoeda, *Numerically repeated support splitting and merging phenomena in a porous media equation with strong absorption*, Journal Math-for-Industry of Kyushu, 3 (2012), pp.61–68.
[11] K. Tomoeda, *Appearance of repeated support splitting and merging phenomena in a porous media equation with absorption*, Application of Mathematics in Technical and Natural Sciences (AMiTaNS'15), AIP Conference Proceedings, 1684 (2015), pp.080013-1–080013-9.
[12] D. Gilbarg and N. S. Trudinger *Elliptic Partial Differential Equations of Second Order*, Second Edition, Revised Third Printing 1998, Springer.

Proceedings of EQUADIFF 2017
pp. 369–376

AN ELEMENTARY PROOF OF ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF $U'' = VU$ *

MOTOHIRO SOBAJIMA† AND GIORGIO METAFUNE‡

Abstract. We provide an elementary proof of the asymptotic behavior of solutions of second order differential equations without successive approximation argument.

Key words. Elementary proof, second-order ordinary differential equations, asymptotic behavior.

AMS subject classifications. 34E10

1. Introduction. The asymptotic behavior of the solutions of the ordinary differential equation

$$u''(x) = V(x)u(x), \qquad x \in (0,\infty) \tag{1.1}$$

is an important tool in various fields of mathematics and mathematical physics, in particular when special functions are involved. It can be found in [3, Section 6.2] and partially in [1, Chapter 10] and in [2, Chapter IV] that if $V(x) = f(x) + g(x)$, that is,

$$u''(x) = \big(f(x) + g(x)\big)u(x), \qquad x \in (0,\infty) \tag{1.2}$$

and

$$\psi_{f,g} := |f|^{-\frac{1}{4}} \left(-\frac{d^2}{dx^2} + g\right) |f|^{-\frac{1}{4}} \text{ is absolutely integrable in } (0,\infty), \tag{1.3}$$

then two solutions of (1.2) behave like

$$u(x) \approx |f|^{-1/4} e^{\pm \int_0^x |f(s)|^{1/2}\, ds}, \quad u(x) \approx |f|^{-1/4} e^{\pm i \int_0^x |f(s)|^{1/2}\, ds}.$$

The proof is usually done treating first the cases $f = \pm 1$ and then reducing to them the general case, by the Liouville transformation. We follow the same approach but simplify the cases $f = \pm 1$ by using Gronwall's Lemma, instead of successive approximations. In order to keep the exposition at an elementary level, we avoid also Lebesgue integration and dominated convergence (which could shorten some proofs); note that we only use the notation $f \in L^1(I)$ when f is absolutely integrable in I. We consider both the behavior at infinity and near isolated singularities and apply the results to Bessel functions. We also recall that the general case

$$u''(x) + g(x)u'(x) = V(x)u(x)$$

can be reduced to the form (1.1) (with another V) by writing $u = \frac{1}{2}(\exp \int g)v$.

*This work is partially supported by Grant-in-Aid for Young Scientists Research (B) No.16K17619.

†Department of Mathematics, Faculty of Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba-ken 278-8510, Japan (msobajima1984@gmail.com).

‡Dipartimento di Matematica "Ennio De Giorgi", Università del Salento, Via Per Arnesano, 73100, Lecce, Italy (giorgio.metafune@unisalento.it).

This kind of analysis can be applied to the spectral analysis for Schrödinger operator with singular potentials (for example $S = -\Delta + V(|x|)$ with $V(r) \sim r^{-\delta}$ near the origin). Actually, the essential selfadjointness of the Schrödinger operator S can be treated by using the limit-point and limit-circle criteria (see e.g., Reed–Simon [4]) which require the behavior of two solutions to $u - u'' + \frac{N-1}{r}u + Vu = 0$. The behavior of two solutions above leads also to resolvent estimates for S. From this view-piont, the elemental consideration in the present paper helps in understanding various spectral phenomena for second-order differential operators.

2. Behavior near infinity in the simplest cases. First we consider the cases $f \equiv 1$ and $f \equiv -1$ and we prove the following results to which the general case reduces.

PROPOSITION 2.1. *If $f = 1$, $g \in L^1(0,\infty)$, then there exist two solutions u_1 and u_2 of (1.2) such that, as $x \to \infty$,*

$$e^{-x}u_1(x) \to 1, \qquad e^{-x}u_1'(x) \to 1, \tag{2.1}$$

$$e^{x}u_2(x) \to 1, \qquad e^{x}u_2'(x) \to -1. \tag{2.2}$$

PROPOSITION 2.2. *If $f = -1$, $g \in L^1(0,\infty)$, then there exist two solutions v_1 and v_2 of (1.2) such that, as $x \to \infty$,*

$$e^{-ix}u_1(x) \to 1, \qquad e^{-ix}u_1'(x) \to i, \tag{2.3}$$

$$e^{ix}u_2(x) \to 1, \qquad e^{ix}u_2'(x) \to -i. \tag{2.4}$$

By variation of parameters, every solution of (1.2) can be written as

$$u(x) = c_1e^{\zeta x} + c_2e^{-\zeta x} + \frac{1}{2\zeta}\int_a^x (e^{\zeta(x-s)} - e^{-\zeta(x-s)})g(s)u(s)\,ds, \quad x \in [a,\infty), \tag{2.5}$$

with $c_1, c_2 \in \mathbb{C}$, $\zeta = 1, i, -i$ and $a > 0$. In the following Lemma we choose $c_1 = 1, c_2 = 0$ to construct a solution which behaves like $e^{\zeta x}$ as $x \to \infty$, $\zeta = 1, i, -i$.

LEMMA 2.3. *Let $\zeta \in \{1, i, -i\}$, $a > 0$ and $g \in L^1(a,\infty)$. If $u \in C^2([a,\infty))$ satisfies*

$$u(x) = e^{\zeta x} + \frac{1}{2\zeta}\int_a^x (e^{\zeta(x-s)} - e^{-\zeta(x-s)})g(s)u(s)\,ds, \qquad x \in [a,\infty),$$

then $z(x) := e^{-\zeta x}u(x)$ satisfies

$$|z(x)| \le e^{\int_a^x |g(r)|\,dr}, \qquad x \in [a,\infty) \tag{2.6}$$

$$\|zg\|_{L^1(a,\infty)} \le e^{\|g\|_{L^1(a,\infty)}} - 1. \tag{2.7}$$

Proof. Note that

$$z(x) = 1 + \frac{1}{2\zeta}\int_a^x (1 - e^{-2\zeta(x-s)})g(s)z(s)\,ds, \quad x \in [a,\infty).$$

Since $|1 - e^{-2\zeta(x-s)}| \le 2$ for $s \le x$, we see that for $x \ge a$,

$$|z(x)| \le 1 + \left|\frac{1}{2\zeta}\int_a^x (1 - e^{-2\zeta(x-s)})g(s)z(s)\,ds\right| \le 1 + \int_a^x |g(s)|\,|z(s)|\,ds.$$

Thus Gronwall's lemma implies (2.6), in particular z is bounded on $[a,\infty)$ and then $zg \in L^1(a,\infty)$. Moreover we have

$$\|zg\|_{L^1(a,\infty)} \le \int_a^\infty |g(s)|\, e^{\int_a^s |g(r)|\,dr}\,ds = e^{\|g\|_{L^1(a,\infty)}} - 1.$$

□

Proof of Proposition 2.1. Let $a > 0$ such that $\|g\|_{L^1(a,\infty)} < \log 2$ and let u be in Lemma 2.3 with $\zeta = 1$. Then u is one solution of (1.2) with $f = 1$. Set $z(x) = e^{-x}u(x)$. Then noting that as $x \to \infty$,

$$\begin{aligned}\left|\int_a^x e^{-2(x-s)}g(s)z(s)\,ds\right| &\le \int_a^{\frac{a+x}{2}} e^{-2(x-s)}|g(s)z(s)|\,ds + \int_{\frac{a+x}{2}}^x |g(s)z(s)|\,ds\\ &\le e^{-x+a}\|gz\|_{L^1(a,\infty)} + \|gz\|_{L^1(\frac{a+x}{2},\infty)} \to 0,\end{aligned}$$

we see that z satisfies

$$\begin{aligned} z(x) &\to z_\infty := 1 + \int_a^\infty g(s)z(s)\,ds \quad \text{as } x \to \infty,\\ z'(x) &= \int_a^x e^{-2(x-s)}g(s)z(s)\,ds \to 0 \quad \text{as } x \to \infty.\end{aligned}$$

By (2.7), we deduce that $\|zg\|_{L^1(a,\infty)} < 1$. Therefore $|z_\infty - 1| \le \|zg\|_{L^1(a,\infty)} < 1$ and hence $z_\infty \neq 0$. The function $u_1(x) := z_\infty^{-1}e^x z(x)$ satisfies (2.1). Moreover, since u_1^{-2} is integrable near ∞, another solution of (1.2) is given by

$$u_2(x) = 2u_1(x)\int_x^\infty \frac{1}{u_1(s)^2}\,ds. \tag{2.8}$$

Integrating by parts we deduce that, as $x \to \infty$,

$$\begin{aligned} e^x u_2(x) &= 2z_\infty e^{2x}z(x)\int_x^\infty \frac{1}{e^{2s}[z(s)]^2}\,ds\\ &= z_\infty e^{2x}z(x)\left(-\left[\frac{1}{e^{2s}[z(s)]^2}\right]_{s=x}^{s=\infty} - 2\int_x^\infty \frac{z'(s)}{e^{2s}[z(s)]^3}\,ds\right) \to 1\end{aligned}$$

and

$$[e^x u_2(x)]' = 2z_\infty e^{2x}z'(x)\int_x^\infty \frac{1}{e^{2s}[z(s)]^2}\,ds + 2e^x u_2(x) - \frac{2z_\infty}{z(x)} \to 0.$$

□

Proof of Proposition 2.2. Let $a > 0$ such that $\|g\|_{L^1(a,\infty)} < \log 2$ and let $\tilde{u}_1$ and $\tilde{u}_2$ be as in Lemma 2.3 with $\zeta = i$ and with $\zeta = -i$, respectively. Noting that both $\tilde{u}_1$ and $\tilde{u}_2$ satisfy (1.2) with $f = -1$, and setting $z_1(x) = e^{-ix}\tilde{u}_1(x)$ and $z_2(x) = e^{ix}\tilde{u}_2(x)$, we have as $x \to \infty$

$$\begin{aligned} e^{2ix}\left(z_1(x) - 1 - \frac{1}{2i}\int_a^\infty g(s)z_1(s)\,ds\right) &\to \frac{1}{2i}\int_a^\infty e^{2is}g(s)z_1(s)\,ds,\\ e^{-2ix}\left(z_2(x) - 1 + \frac{1}{2i}\int_a^\infty g(s)z_2(s)\,ds\right) &\to -\frac{1}{2i}\int_a^\infty e^{-2is}g(s)z_2(s)\,ds\end{aligned}$$

and

$$e^{2ix}z_1'(x) \to \int_a^\infty e^{2is}g(s)z_1(s)\,ds, \qquad e^{-2ix}z_2'(x) \to \int_a^\infty e^{-2is}g(s)z_2(s)\,ds.$$

It follows that $\tilde{u}_1 \approx \xi_1 e^{ix} + \xi_2 e^{-ix}$, $\tilde{u}_1' \approx i\xi_1 e^{ix} - i\xi_2 e^{-ix}$ and $\tilde{u}_2 \approx \eta_1 e^{ix} + \eta_2 e^{-ix}$, $\tilde{u}_2' \approx i\eta_1 e^{ix} - i\eta_2 e^{-ix}$ as $x \to \infty$ where

$$\xi_1 = 1 + \frac{1}{2i}\int_a^\infty g(s)z_1(s)\,ds, \qquad \xi_2 = -\frac{1}{2i}\int_a^\infty e^{2is}g(s)z_1(s)\,ds,$$

and similarly for η_1, η_2. From (2.7) we see that $|\xi_1| > 1/2$, $|\xi_2| < 1/2$, $|\eta_1| < 1/2$ and $|\eta_2| > 1/2$ and hence $|\xi_1\eta_2 - \xi_2\eta_1| > 0$ and $\tilde{u}_1$ and $\tilde{u}_2$ are linearly independent. Therefore we can construct solutions u_1 and u_2 which satisfy (2.3) and (2.4), respectively. □

We consider now the case $f = 0$, assuming extra conditions on g.

PROPOSITION 2.4. *Assume that $xg \in L^1(0,\infty)$. Then there exist two solutions u_1 and u_2 of*

$$u''(x) = g(x)u(x) \tag{2.9}$$

such that

$$x^{-1}u_1(x) \to 1, \qquad u_1'(x) \to 1,$$
$$u_2(x) \to 1, \qquad xu_2'(x) \to 0$$

as $x \to \infty$, respectively.

Proof. Set $u(x) := xz(x)$. Then $z'' + (2/x)z' = gz$ and, assuming $z'(a) = 0$ we obtain

$$z'(x) = x^{-2}\int_a^x s^2 g(s)z(s)\,ds. \tag{2.10}$$

Then assuming $z(a) = 1$

$$\begin{aligned} |z(x) - 1| &\le \int_b^x t^{-2}\left(\int_a^t s^2|g(s)z(s)|\,ds\right)dt \\ &= \int_a^x \left(\int_s^x t^{-2}\,dt\right)s^2|g(s)z(s)|\,ds \le \int_a^x s|g(s)z(s)|\,ds. \end{aligned} \tag{2.11}$$

Gronwall's lemma yields

$$|z(x)| \le e^{\int_a^x s|g(s)|\,ds}$$

hence z is bounded and $z' \in L^1(a,\infty)$ by (2.10). As in the proof of Proposition 2.1, $z(x) \to z_\infty \neq 0$ if a is sufficiently large. Moreover, since as $x \to \infty$,

$$|xz'(x)| \le \sqrt{\frac{a}{x}}\int_a^{\sqrt{ax}} s|g(s)z(s)|\,ds + \int_{\sqrt{ax}}^x s|g(s)z(s)|\,ds \to 0,$$

$u_1(x) := z_\infty^{-1}xz(x)$ satisfies the statement. Another solution u_2 of (1.2) is given by

$$u_2(x) := u_1(x)\int_x^\infty \frac{1}{u_1(s)^2}\,ds.$$

As in the proof of Proposition 3.1 we can verify that u_2 satisfies $u_2(x) \to 1$ and $xu_2'(x) \to 0$ as $x \to \infty$. □

Observe the integrability condition for xg near ∞ is necessary. In fact, if $g(x) = cx^{-2}$ the above equation has solutions x^α if $\alpha^2 - \alpha = c$.

3. Behavior near infinity in the general case. We recall that the function $\psi_{f,g}$ is defined in (1.3) and set $v_j(x) = |f|^{1/4}u_j(x)$, $j = 1, 2$ if u_1, u_2 are solutions of (1.2). The hypothesis $|f|^{1/2}$ not summable near ∞ guarantees that the Liouville transformation Φ of Lemma 3.3 maps (a, ∞) onto $(0, \infty)$, so that the results of the previous section apply. When it is not satisfied Φ maps (a, ∞) onto a bounded interval $(0, b)$ and the behavior of the solutions of (3.5) near b is more elementary (in some cases one can use Proposition 2.4).

PROPOSITION 3.1. *Assume that $f(x) > 0$ in (a, ∞), $|f|^{1/2} \notin L^1(a, \infty)$ and $\psi_{f,g} \in L^1(a, \infty)$. Then there exist two solutions u_1 and u_2 of (1.2) such that as $x \to \infty$*

$$e^{-\int_a^x |f(r)|^{1/2}dr} v_1(x) \to 1, \qquad |f(x)|^{-1/2} e^{-\int_a^x |f(r)|^{1/2}dr} v_1'(x) \to 1, \tag{3.1}$$

$$e^{\int_a^x |f(r)|^{1/2}dr} v_2(x) \to 1, \qquad |f(x)|^{-1/2} e^{\int_a^x |f(r)|^{1/2}dr} v_2'(x) \to -1. \tag{3.2}$$

PROPOSITION 3.2. *Assume that $f(x) < 0$ in (a, ∞), $|f|^{1/2} \notin L^1(a, \infty)$ and $\psi_{f,g} \in L^1(a, \infty)$. Then there exists two solutions u_1 and u_2 of (1.2) such that asx $\to \infty$*

$$e^{-i\int_a^x |f(r)|^{1/2}dr} v_1(x) \to 1, \qquad |f(x)|^{-1/2} e^{-i\int_a^x |f(r)|^{1/2}dr} v_1'(x) \to i, \tag{3.3}$$

$$e^{i\int_a^x |f(r)|^{1/2}dr} v_2(x) \to 1, \qquad |f(x)|^{-1/2} e^{i\int_a^x |f(r)|^{1/2}dr} v_2'(x) \to -i. \tag{3.4}$$

The proof is based on the well-known Liouville transformation that we recall below.

LEMMA 3.3. *Let $a > 0$ and assume that $f \in C^2([a, \infty))$ satisfies $|f(x)| > 0$, $|f|^{1/2} \notin L^1(a, \infty)$. Define $\Phi \in C^2([a, \infty))$ by*

$$\Phi(x) := \int_a^x |f(r)|^{1/2}\, dr, \quad x \in [a, \infty).$$

Then $\Phi^{-1} : [0, \infty) \to [a, \infty)$ and if u satisfies (1.2) the function

$$w(y) := |f(\Phi^{-1}(y))|^{1/4} u(\Phi^{-1}(y)), \quad y \in [0, \infty)$$

satisfies

$$w''(y) = \left(\frac{f(\Phi^{-1}(y))}{|f(\Phi^{-1}(y))|} + \frac{\psi_{f,g}(\Phi^{-1}(y))}{|f(\Phi^{-1}(y))|^{1/2}} \right) w(y). \tag{3.5}$$

Proof. Note that $\Phi'(x) = |f(x)|^{1/2}$ and $\frac{d(\Phi^{-1})}{dy}(y) = |f(\Phi^{-1}(y))|^{-1/2}$. Setting

$w(y) = |f(\Phi^{-1}(y))|^{1/4}u(\Phi^{-1}(y))$ (and using $\xi = \Phi^{-1}(y)$ for simplicity), we have

$$\begin{aligned}
w'(y) &= \frac{d}{dx}\left[|f|^{1/4}u\right](\xi)\frac{d(\Phi^{-1})}{dy}(y)\\
&= |f(\xi)|^{-1/4}u'(\xi) + \left[|f|^{-1/2}\frac{d}{dx}|f|^{1/4}\right](\xi)u(\xi)\\
&= \left[|f|^{-1/4}u' - \frac{d}{dx}(|f|^{-1/4})u\right](\xi),\\
w''(y) &= \frac{d}{dx}\left[|f|^{-1/4}u' - \frac{d}{dx}(|f|^{-1/4})u\right](\xi)\frac{d(\Phi^{-1})}{dy}(y)\\
&= |f(\xi)|^{-3/4}u''(\xi) - \left[|f|^{-1/2}\frac{d^2}{dx^2}|f|^{-1/4}\right](\xi)u(\xi)\\
&= |f(\xi)|^{-1}(f(\xi)+g(\xi))w(y) - \left[|f|^{-3/4}\frac{d^2}{dx^2}|f|^{-1/4}\right](\xi)w(y).
\end{aligned}$$

Thus we obtain (3.5). □

Proof. [Proof of Propositions 3.1 and 3.2] It suffices to apply Propositions 2.1 and 2.2 to the respective cases $f > 0$ and $f < 0$. Set $h(y) = \psi_{f,g}(\Phi^{-1}(y))|f(\Phi^{-1}(y))|^{-1/2}$. Then

$$\int_0^b |h(y)|\,dy = \int_a^\infty |\psi_{f,g}(x)|\,dx.$$

Therefore Propositions 2.1 and 2.2 are applicable to $w'' = \pm w + hw$, respectively. Finally, using Lemma 3.3 and taking $u(x) = |f(x)|^{-1/4}w(\Phi(x))$, we obtain the respective assertions in Propositions 3.1 and 3.2. □

4. Behavior near interior singularities. If f and g have local singularities at x_0, then the behavior of solutions near x_0 is also considerable. For simplicity, we take $x_0 = 0$. The following propositions are meaningful when $|f|^{1/2}$ is not integrable near 0, in particular when $|f|^{1/2} = cx^{-1}$. We recall that $v_j(x) = |f(x)|^{1/4}u_j(x)$, $j = 1, 2$.

PROPOSITION 4.1. *Assume that $f(x) > 0$ in $(0,\infty)$ and $\psi_{f,g} \in L^1(0,\infty)$. Then there exist two solutions u_1 and u_2 of (1.2) such that as $x \downarrow 0$*

$$\begin{aligned}
e^{-\int_x^1 |f(r)|^{1/2}dr}v_1(x) &\to 1, \qquad |f(x)|^{-1/2}e^{-\int_x^1 |f(r)|^{1/2}dr}v_1'(x) \to -1,\\
e^{\int_x^1 |f(r)|^{1/2}dr}v_2(x) &\to 1, \qquad |f(x)|^{-1/2}e^{\int_x^1 |f(r)|^{1/2}dr}v_2'(x) \to 1.
\end{aligned}$$

PROPOSITION 4.2. *Assume that $f(x) < 0$ in $(0,\infty)$ and $\psi_{f,g} \in L^1(0,\infty)$. Then there exist two solutions u_1 and u_2 of (1.2) such that as $x \downarrow 0$*

$$\begin{aligned}
e^{-\int_x^1 |f(r)|^{1/2}dr}v_1(x) &\to 1, \qquad |f(x)|^{-1/2}e^{-\int_x^1 |f(r)|^{1/2}dr}v_1'(x) \to -i,\\
e^{\int_x^1 |f(r)|^{1/2}dr}v_2(x) &\to 1, \qquad |f(x)|^{-1/2}e^{\int_x^1 |f(r)|^{1/2}dr}v_2'(x) \to i.
\end{aligned}$$

Proof of Propositions 4.1 and 4.2. Setting $w(s) := su(s^{-1})$ we see that

$$\begin{aligned}
w''(s) &= s^{-3}u''(s^{-1})\\
&= s^{-3}(f(s^{-1}) + g(s^{-1}))u(s^{-1}) = s^{-4}(f(s^{-1}) + g(s^{-1}))w(s).
\end{aligned}$$

Let $\tilde{f}(s) := s^{-4} f(s^{-1})$ and $\tilde{g}(s) := s^{-4} g(s^{-1})$. Noting that

$$\begin{aligned}\psi_{\tilde{f},\tilde{g}}(s) &= s|f(s^{-1})|^{-1/4}\left(-\frac{d^2}{ds^2} + s^{-4}g(s^{-1})\right)\left(s|f(s^{-1})|^{-1/4}\right)\\ &= s^{-2}|f(s^{-1})|^{-1/4}\left(-\frac{d^2}{dx^2}|f|^{-1/4} + g|f|^{-1/4}\right)(s^{-1})\\ &= s^{-2}\psi_{f,g}(s^{-1}),\end{aligned}$$

we have $\psi_{\tilde{f},\tilde{g}} \in L^1((0,\infty))$, and hence Propositions 3.1 and 3.2 can be applied. Since

$$\int_1^s |\tilde{f}(r)|^{1/2} dr = \int_{1/s}^1 |f(t)|^{1/2} dt,$$

we obtain the respective assertions in Propositions 4.1 and 4.2. □

5. Examples from special functions. Some examples illustrate the application of the results of the previous sections.

EXAMPLE 1 (Modified Bessel functions). *We consider the modified Bessel equation of order* ν

$$u'' + \frac{u'}{r} - \left(1 + \frac{\nu^2}{r^2}\right) u = 0, \tag{5.1}$$

All solutions of (5.1) *can be written through the modified Bessel functions* I_ν *and* K_ν. *Both* I_ν *and* K_ν *are positive,* I_ν *is monotone increasing and* K_ν *is monotone decreasing (see e.g., [3, Theorem 7.8.1]). Proposition 2.1 and Proposition 4.1 give the precise behavior of* I_ν *and* K_ν *near* ∞ *and near* 0, *respectively. In fact,* (5.1) *can be written as*

$$(\sqrt{r}u)'' = \left(1 + \frac{4\nu^2 - 1}{4r^2}\right)(\sqrt{r}u). \tag{5.2}$$

Since $1/r^2$ *is integrable near* ∞, *choosing* $f = 1$ *and* $g = \frac{4\nu^2-1}{4r^2}$, *we see from Proposition 2.1 that*

$$\sqrt{r}e^{-r} I_\nu(r) \to c_1 \neq 0 \quad \textit{and} \quad \sqrt{r}e^{r} K_\nu(r) \to c_2 \neq 0 \qquad \textit{as } r \to \infty.$$

Moreover, if $\nu \neq 0$, *then choosing* $f(r) = \frac{\nu^2}{r^2}$ *and* $g(r) = 1 - \frac{1}{4r^2}$, *that is,* $\psi_{f,g}(r) = r/\nu$, *from Proposition 4.1 we have*

$$r^{-\nu} I_\nu(r) \to c_3 \neq 0 \quad \text{and} \quad r^{\nu} K_\nu(r) \to c_4 \neq 0 \qquad \textit{as } r \downarrow 0.$$

If $\nu = 0$, *then putting* $w(s) = u(e^{-s})$ *we obtain*

$$w''(s) = e^{-2s} w(s), \qquad s \in \mathbb{R}.$$

Therefore using Proposition 2.4 with $\tilde{g}(s) = e^{-2s}$ *and taking* $u(x) = w(-\log x)$, *we have*

$$I_0(r) \to c_5 \neq 0 \quad \textit{and} \quad |\log r|^{-1} K_0(r) \to c_6 \neq 0 \qquad \textit{as } r \downarrow 0.$$

EXAMPLE 2 (Fundamental solution of $\lambda - \Delta$). *For $n \geq 3$, $\lambda \geq 0$ the fundamental solution v_λ of $\lambda - \Delta$ can be computed by integrating the heat kernel:*

$$v_\lambda(r) = \int_0^\infty \frac{1}{(4\pi t)^{n/2}} e^{-\lambda t - \frac{r^2}{4t}}\, dt,$$

where $r = |x|$. Clearly $v_\lambda(r) \leq v_0(r) = cr^{2-n}$, $v_\lambda(r) \to 0$ as $r \to \infty$. The function $v = v_\lambda$ satisfies

$$v'' + \frac{n-1}{r} v' = \lambda v$$

or, setting $v = r^{(1-n)/2} w$,

$$w'' = \left(\lambda + \frac{n^2 - 1}{4r^2}\right) w.$$

Proceeding as in the example above we see that $r^{2-n} v(r) \to c_1 \neq 0$ as $r \to 0$ and $r^{(n-1)/2} e^{\sqrt{\lambda} r} v(r) \to c_2 \neq 0$ as $r \to \infty$.

EXAMPLE 3 (Bessel functions). *Next we consider the Bessel equation of order ν*

$$u'' + \frac{u'}{r} + \left(1 - \frac{\nu^2}{r^2}\right) u = 0, \tag{5.3}$$

or equivalently,

$$(\sqrt{r} u)'' = \left(-1 + \frac{4\nu^2 - 1}{4r^2}\right) (\sqrt{r} u).$$

All solutions of (5.3) can be written through the Bessel functions J_ν and Y_ν. As in Example 1, from Propositions 4.1 (for $\nu > 0$) and 2.4 (for $\nu = 0$) we obtain the behavior of J_ν and Y_ν near 0

$$r^{-\nu} J_\nu(r) \to c_1 \neq 0, \quad \text{and} \quad r^\nu Y_\nu(r) \to c_2 \neq 0 \qquad \text{as } r \downarrow 0$$

and if $\nu = 0$,

$$|\log r| J_0(r) \to c_3 \neq 0, \quad \text{and} \quad Y_0(r) \to c_4 \neq 0 \qquad \text{as } r \downarrow 0.$$

In view of Proposition 2.2 the behavior of J_ν and Y_μ near ∞ is given by

$$|\sqrt{r} J_\nu(r) - c_5 \cos(r + \theta_1)| \to 0, \quad \text{and} \quad |\sqrt{r} Y_\nu(r) - c_6 \cos(r + \theta_2)| \to 0,$$

as $r \to \infty$, where $c_5 \neq 0$, $c_6 \neq 0$ and $\theta_1, \theta_2 \in [0, \pi)$ satisfy $\theta_1 \neq \theta_2$.

REFERENCES

[1] R. Beals, R. Wong, "Special functions," A graduate text, Cambridge Studies in Advanced Mathematics **126**, Cambridge University Press, Cambridge, 2010.
[2] A. Erdélyi, "Asymptotic expansions," Dover Publications, Inc., New York, 1956.
[3] F.W.J. Olver, "Asymptotics and special functions," Computer Science and Applied Mathematics, Academic Press, New York-London, 1974.
[4] M. Reed, B. Simon, "Methods of modern mathematical physics. II. Fourier analysis, self-adjointness," Academic Press, New York-London, 1975.

Proceedings of EQUADIFF 2017
pp. 377–386

NONLINEAR TENSOR DIFFUSION IN IMAGE PROCESSING*

OĽGA STAŠOVÁ , KAROL MIKULA , ANGELA HANDLOVIČOVÁ†, AND NADINE PEYRIÉRAS‡

Abstract. This paper presents and summarize our results concerning the nonlinear tensor diffusion which enhances image structure coherence. The core of the paper comes from [3, 2, 4, 5]. First we briefly describe the diffusion model and provide its basic properties. Further we build a semi-implicit finite volume scheme for the above mentioned model with the help of a co-volume mesh. This strategy is well-known as diamond-cell method owing to the choice of co-volume as a diamond-shaped polygon, see [1]. We present here 2D as well as 3D case of a numerical scheme, see [3, 4]. Then the convergence and error estimate analysis for 2D scheme is presented, see [3, 2]. Last part is devoted to results of computational experiments. They confirm the usefulness this diffusion type not just for an image improvement but also as a pre-processed algorithm. Numerical techniques which require a good coherence of image structures (like edge detection and segmentation) achieve much better results when we use images pre-processed by such a filtration. Let us note that this diffusion technique was successfully applied within the framework of EU projects. It was used to pre-process images for the structure segmentation in zebrafish embryogenesis, see [5].

Key words. image processing, nonlinear tensor diffusion, coherence enhancing diffusion, numerical solution, semi-implicit scheme, diamond-cell finite volume method, convergence, error estimate, structure segmentation.

AMS subject classifications. 35K55, 65M12, 35B45, 68U10, 65M08.

1. Introduction. Coherence enhancing diffusion (CED), see [11], is a technique which enables to achieve an improvement of image structure connectivity. It is also helpful as a pre-processed algorithm for numerical methods in which a precise image structure coherence is desirable (e.g. edge detection, segmentation). Applying these procedures on images filtered by CED yields an enhancement of their results. The filtration process is driven by the diffusion tensor in such a way that the diffusion is strong in preferred directions, e.g. along edges (in 2D images) or along 2D edge surfaces (in 3D images) which causes a recovery of defects in image structures. Interrupted places will be completed. On the contrary, the smoothing is low in the perpendicular direction and therefore the edges are not significantly blurred.

2. Mathematical model. Let Q_T is a spatio-temporal domain, where a time interval is given by $I = [0, T]$ and Ω (subset of R2 or R3) is an image domain with the boundary $\partial\Omega$. We consider the coherence enhancing diffusion model on this domain. It has the following form, see [11, 3, 7, 4],

$$\frac{\partial u}{\partial t} - \nabla \cdot (D\nabla u) = 0 \qquad \text{in } Q_T \equiv I \times \Omega, \tag{2.1}$$

$$u(x, 0) = u_0(x) \quad \text{in } \Omega, \tag{2.2}$$

$$D\nabla u \cdot \mathbf{n} = 0 \qquad \text{on } I \times \partial\Omega, \tag{2.3}$$

*This work was supported by Grants APVV-15-0522 and VEGA 1/0608/15.

†Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 810 05 Bratislava, Slovakia (olga.stasova@stuba.sk, karol.mikula@stuba.sk, angela.handlovicova@stuba.sk).

‡Institut de Neurobiologie Alfred Fessard, CNRS-NED, Avenue de la Terrasse, 911 98 Gif-sur-Yvette, France nadine.peyrieras@inaf.cnrs-gif.fr).

where $u(x,t)$ denotes an unknown function and represents a grey level image intensity, $u_0 \in L^2(\Omega)$ and $\mathbf{n}$ denotes the outer normal unit vector to the $\partial\Omega$. The matrix D represents the so-called diffusion tensor. Its design differs in dependence on a dimension order.

2.1. 2D diffusion tensor. The construction of the 2D diffusion tensor is based on the eigenvalues and eigenvectors of the (regularized) structure tensor $J_\rho(\nabla u_{\tilde{t}}) = G_\rho * (\nabla u_{\tilde{t}} \nabla u_{\tilde{t}}{}^T) = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, where $u_{\tilde{t}}(x,t) = (G_{\tilde{t}} * u(\cdot,t))(x)$. $G_{\tilde{t}}$ and G_ρ are Gaussian convolution kernels, see [11, 3]. The matrix J_ρ is symmetric and positive semidefinite and its eigenvectors are parallel and orthogonal to $\nabla u_{\tilde{t}}$, respectively. Its eigenvalues are given by $\mu_{1,2} = \frac{1}{2}\left(a + c \pm \sqrt{(a-c)^2 + 4b^2}\right)$, $\mu_1 \geq \mu_2$. The corresponding orthogonal set of eigenvectors $(\mathbf{v}, \mathbf{w})$ to eigenvalues (μ_1, μ_2) is given as follows

$$\begin{aligned} &\mathbf{v} = (v_1, v_2), \qquad \mathbf{w} = (w_1, w_2), \qquad \mathbf{w} \perp \mathbf{v}, \qquad w_1 = -v_2, \qquad w_2 = v_1, \\ &v_1 = 2b, \qquad\qquad v_2 = c - a + \sqrt{(a-c)^2 + 4b^2}. \end{aligned} \tag{2.4}$$

The orientation of the eigenvector $\mathbf{w}$ corresponding to the smaller eigenvalue μ_2 is called the coherence orientation. This orientation has the lowest fluctuations in image intensity. The diffusion tensor D is built to steer a filtering process such that the smoothing is strong along the coherence direction $\mathbf{w}$ and increases with the coherence $(\mu_1 - \mu_2)^2$. To achieve it, we require D to possess the same eigenvectors $\mathbf{v}$ and $\mathbf{w}$ as the structure tensor $J_\rho(\nabla u_{\tilde{t}})$ and we choose the eigenvalues of D as follows

$$\begin{aligned} &\kappa_1 = \alpha, \quad \alpha \in (0,1),\ \alpha \ll 1, \\ &\kappa_2 = \begin{cases} \alpha, & \text{if } \mu_1 = \mu_2, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{(\mu_1 - \mu_2)^2}\right),\ C > 0 & \text{else.} \end{cases} \end{aligned} \tag{2.5}$$

Hence we get the diffusion tensor in the form

$$D = ABA^{-1}, \quad \text{where} \quad A = \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{pmatrix}, \tag{2.6}$$

which depends nonlinearly on partial derivatives of solution u, possesses smoothness, symmetry and uniform positive definiteness properties.

2.2. 3D diffusion tensor. The construction of the 3D diffusion tensor is based on a smoothed intensity gradient given by $\nabla u_{\tilde{t}} = (u_{x_1}, u_{x_2}, u_{x_3})^T$, where $u_{\tilde{t}}(x,t) = (G_{\tilde{t}} * u(\cdot,t))(x)$, $(\tilde{t} > 0)$ and $G_{\tilde{t}}$ is a Gaussian kernel, see [4, 7]. Provided that $\mu = ||\nabla u_{\tilde{t}}||^2 > 0$ we choose a triplet of vectors $(\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ by $\mathbf{v_1} \parallel \nabla u_{\tilde{t}}$, $\mathbf{v_2} \perp \nabla u_{\tilde{t}}$, $\mathbf{v_3} \perp \nabla u_{\tilde{t}}$, $\mathbf{v_2} \perp \mathbf{v_3}$. The direction of vector $\mathbf{v_1}$ corresponds to the direction of the largest intensity change. The other two vectors give a tangential plane to a level set of image intensity which may represent a 2D surface edge in a 3D image, provided that μ is large. It is called coherence plane $\mathcal{P}$ and represents an eigenspace corresponding to the eigenvalue 0 of the outer product $\nabla u_{\tilde{t}} \otimes \nabla u_{\tilde{t}}$. In order to enhance the coherence, the diffusion tensor D must steer the filtering process such that the diffusion is strong and increasing with the level of μ along the coherence plane $\mathcal{P}$ and is small in the perpendicular direction. We achieve it by choosing the eigenvalues of the diffusion tensor, which determine the diffusivities in the directions $\mathbf{v_1}$, $\mathbf{v_2}$ and $\mathbf{v_3}$ as

$$\kappa_1 = \alpha, \quad \alpha \in (0,1),\ \alpha \ll 1, \tag{2.7}$$

$$\kappa_2 = \kappa_3 = \begin{cases} \alpha, & \text{if } \mu = 0, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{\mu}\right), & C > 0 \quad \text{otherwise.} \end{cases}$$

Then we apply another convolution with a smoothing kernel G_ρ and get the diffusion matrix D in the form

$$(2.8)\quad D = G_\rho * D_0, \text{ where } D_0 = \begin{cases} B, & \text{if } \mu = 0, \\ PBP^{-1} & \text{otherwise,} \end{cases} \quad B = \begin{pmatrix} \kappa_1 & 0 & 0 \\ 0 & \kappa_2 & 0 \\ 0 & 0 & \kappa_2 \end{pmatrix}$$

and P denotes a transition matrix from the basis $(\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ to $(\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3})$. If $\mu > 0$, the matrix D_0 has the following form

$$\frac{1}{\mu}\begin{pmatrix} u_{x_1}^2\kappa_1 + (u_{x_2}^2 + u_{x_3}^2)\kappa_2 & u_{x_1}u_{x_2}(\kappa_1 - \kappa_2) & u_{x_1}u_{x_3}(\kappa_1 - \kappa_2) \\ u_{x_1}u_{x_2}(\kappa_1 - \kappa_2) & u_{x_2}^2\kappa_1 + (u_{x_1}^2 + u_{x_3}^2)\kappa_2 & u_{x_2}u_{x_3}(\kappa_1 - \kappa_2) \\ u_{x_1}u_{x_3}(\kappa_1 - \kappa_2) & u_{x_2}u_{x_3}(\kappa_1 - \kappa_2) & u_{x_3}^2\kappa_1 + (u_{x_1}^2 + u_{x_2}^2)\kappa_2 \end{pmatrix}$$

in the standard basis $(\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3})$. Such choice of the matrix D_0 was given in [4], it is independent on a concrete choice of $\mathbf{v_2}$ and $\mathbf{v_3}$ and can be directly and fast evaluated using the diamond-cell finite volume technique (see also next section). The $3D$ diffusion tensor satisfies the smoothness, symmetry and positive definiteness properties, see [4], as does the $2D$ diffusion tensor.

3. Diamond-cell finite volume scheme. We design the numerical scheme for CED using the finite volume method, see [6], since this discretization technique uses the piecewise constant representation of approximate solutions similarly to the structure of digital images. The restrictions of the classical five-point method for the tensor models, see [8], lead to choice of the nine-point diamond-cell method in $2D$, see [1, 3]. Similarly, we switch to 27-point scheme instead of simpler 7-point scheme in $3D$ space, see [4].

Let the image be represented by $n_1 \times n_2$ pixels (finite volumes) in $2D$ or by $n_1 \times n_2 \times n_3$ voxels in $3D$ such that it looks like a mesh with n_1 rows and n_2 columns in $2D$ or a mesh with n_1 rows, n_2 columns and n_3 layers in $3D$. Let $\Omega = (0, n_1 h) \times (0, n_2 h)$ in $2D$ or $\Omega = (0, n_1 h) \times (0, n_2 h) \times (0, n_3 h)$ in $3D$ with a pixel(voxel) size h. We consider the filtering process in a time interval $I = [0, T]$. Let the time discretization be given by $0 = t_0 \le t_1 \le \cdots \le t_{N_{max}} = T$ with $t_n = t_{n-1} + k$, where k is the length of the discrete time step. $\mathcal{T}_h$ is an admissible finite volume mesh, see [6] and further quantities and notations are given as follows: $m(W)$ is the measure of the finite volume W with boundary ∂W, $\sigma_{WE} = W \cap E = W|E$ is an edge(face) of the finite volume W, where $E \in \mathcal{T}_h$ is an adjacent finite volume to W such that the measure $m(W \cap E) \neq 0$. At several places we will replace σ_{WE} by σ to simplify notation. $m(\sigma)$ is the measure of edge (face) σ. $\mathcal{E}_W$ represents the set of edges(faces) such that $\partial W = \bigcup_{\sigma \in \mathcal{E}_W} \sigma$ and $\mathcal{E} = \bigcup_{W \in \mathcal{T}_h} \mathcal{E}_W$. The set of boundary edges(faces) is denoted by $\mathcal{E}_{ext}$, that is $\mathcal{E}_{ext} = \{\sigma \in \mathcal{E}, \sigma \subset \partial\Omega\}$ and denote $\mathcal{E}_{int} = \mathcal{E} \setminus \mathcal{E}_{ext}$. Υ is the set of pairs of adjacent finite volumes, defined by $\Upsilon = \{(W, E) \in \mathcal{T}_h^2,\ W \neq E,\ m(\sigma_{WE}) \neq 0\}$ and $\mathbf{n}_{W,\sigma}$ is the normal unit vector to σ outward to W. Let u_W^n represents a numerical solution on finite volume W, $W \in \mathcal{T}_h$ at time t_n, $n = 1, ..., N_{max}$.

Our discrete solution is given by $u_{h,k}(x,t) = \sum_{n=0}^{N_{max}} \sum_{W \in \mathcal{T}_h} u_W^n \chi_{\{x \in W\}} \chi_{\{t_{n-1} < t \le t_n\}}$, where the function $\chi_{\{A\}}$ is defined as

$$\chi_{\{A\}} = \begin{cases} 1, & \text{if A is true,} \\ 0, & \text{elsewhere.} \end{cases}$$

The finite volume approximation at the n-th time step is given by

$$u^n_{h,k}(x) = \sum_{W \in \mathcal{T}_h} u^n_W \chi\{x \in W\}$$

and initial values as $u^0_W = \frac{1}{m(W)} \int\limits_W u_0(x)dx$, $W \in \mathcal{T}_h$.

We start the scheme derivation integrating the equation (2.1) over the finite volume W, then provide a semi-implicit discretization and use the divergence theorem to have

$$\frac{u^n_W - u^{n-1}_W}{k} m(W) - \sum_{\sigma \in \mathcal{E}_W \cap \mathcal{E}_{int}} \int_\sigma (D^{n-1} \nabla u^n) \cdot \mathbf{n}_{W,\sigma} ds = 0. \tag{3.1}$$

We can define an auxiliary unknown $\phi^n_\sigma(u^n_{h,k})$ representing an approximation of the exact averaged flux $\frac{1}{m(\sigma)} \int_\sigma (D^{n-1} \nabla u^n) \cdot \mathbf{n}_{W,\sigma} ds$ for any W and $\sigma \in \mathcal{E}_W$ in order to rewrite (3.1) in the form

$$\frac{u^n_W - u^{n-1}_W}{k} - \frac{1}{m(W)} \sum_{\sigma \in \mathcal{E}_W \cap \mathcal{E}_{int}} \phi^n_\sigma(u^n_{h,k}) m(\sigma) = 0. \tag{3.2}$$

Approximation of the flux $\phi^n_\sigma(u^n_{h,k})$ is built with the help of a co-volume mesh, see e.g. [1, 3]. The $2D$ co-volume χ_σ associated to σ is constructed around each edge by joining endpoints of this edge and midpoints of finite volumes which are common to this edge, see Fig. 3.1. We create a co-volume χ_σ associated with σ around each

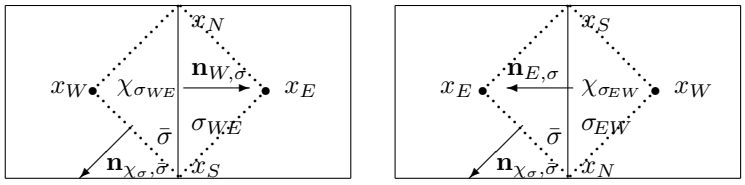


FIG. 3.1. *The co-volumes χ_σ associated to edges $\sigma = \sigma_{WE}$ (left) and $\sigma = \sigma_{EW}$ (right).*

finite volume face by joining four vertices of this face and midpoints of the finite volumes which are common to this face, see Fig. 3.2. Using this technique we obtain the scheme in the form, see [3, 4],

$$\frac{u^n_W - u^{n-1}_W}{k} - \frac{1}{m(W)} \sum_{\sigma \in \mathcal{E}_W \cap \mathcal{E}_{int}} \phi^n_\sigma(u^n_{h,k}) m(\sigma) = 0 \tag{3.3}$$

$$\text{with } 2D \ \phi^n_\sigma(u^n_{h,k}) = \bar{D}^\sigma_{11} \frac{u^n_E - u^n_W}{h} + \bar{D}^\sigma_{12} \frac{u^n_N - u^n_S}{h}, \tag{3.4}$$

$$\text{with } 3D \ \phi^n_\sigma(u^n_{h,k}) = \bar{D}^\sigma_{11} \frac{u^n_E - u^n_W}{h} + \bar{D}^\sigma_{12} \frac{u^n_{TN} + u^n_{BN} - u^n_{TS} - u^n_{BS}}{2h} + \bar{D}^\sigma_{13} \frac{u^n_{TN} + u^n_{TS} - u^n_{BN} - u^n_{BS}}{2h}. \tag{3.5}$$

where $\bar{D}^\sigma_{11}$ and $\bar{D}^\sigma_{12}$ in $2D$ $\phi^n_\sigma(u^n_{h,k})$ are elements of the matrix $D_\sigma = D^{n-1}_\sigma$ written in the basis $(\mathbf{n}_{W,\sigma}, \mathbf{t}_{W,\sigma})$, see [1], where $\mathbf{t}_{W,\sigma}$ is a unit vector parallel to σ such that $(x_N - x_S) \cdot \mathbf{t}_{W,\sigma} > 0$. The values at x_E and x_W are taken as u_E and u_W, and the

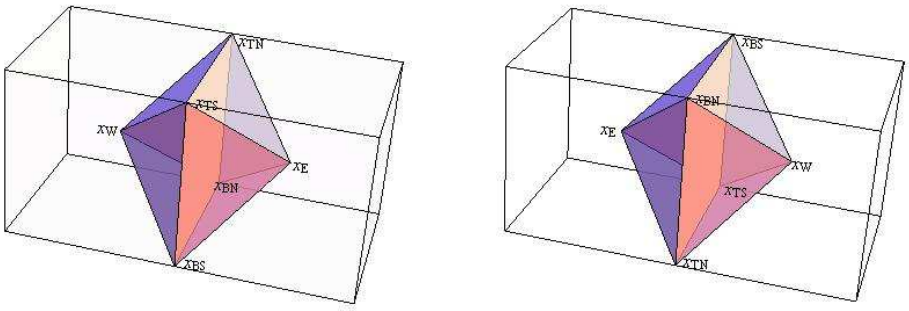


FIG. 3.2. *The co-volumes associated with the face $\sigma = \sigma_{WE}$ (left) and $\sigma = \sigma_{EW}$ (right).*

values u_S and u_N at the vertices x_N and x_S are computed as the arithmetic mean of u_W, where W are finite volumes which are common to this vertex. Further, $\bar{D}^\sigma_{11}$, $\bar{D}^\sigma_{12}$ and $\bar{D}^\sigma_{13}$ in $3D$ $\phi^n_\sigma(u^n_{h,k})$ are elements of the matrix $D_\sigma = D^{n-1}_\sigma$ written in the basis $(\mathbf{n}_{W,\sigma}, \mathbf{t1}_{W,\sigma}, \mathbf{t2}_{W,\sigma})$, where $\mathbf{t1}_{K,\sigma}$ is a unit vector parallel to $x_{TN} - x_{TS}$ such that $(x_{TN} - x_{TS}) \cdot \mathbf{t1}_{K,\sigma} > 0$ and $\mathbf{t2}_{K,\sigma}$ is a unit vector parallel to $x_{TN} - x_{BN}$ such that $(x_{TN} - x_{BN}) \cdot \mathbf{t2}_{K,\sigma} > 0$. Due to the computation of the values u_{TN}, u_{TS}, u_{BN} and u_{BS} in (3.5) as the arithmetic mean of neighbouring voxel values, we get the 27 point finite volume scheme.

4. Convergence analysis for $2D$ discrete scheme. We proved the convergence of the numerical solution of the scheme (3.3)-(3.4) to the weak solution of the problem (2.1)-(2.3) in [3]. Our convergence analysis follows the convergence proof from [8], see [3]. However, our scheme is 9-point scheme compared with the 5-point scheme from [8]. Due to this fact, we must take into account also values of corner's neighbouring volumes. They appear in the scheme in the form of a derivative in the tangential direction, since u_N and u_S are computed as arithmetical mean of their 4 adjacent volumes. In order to overcome the difficulties arising in the occurrence of u_N and u_S we bound the derivative in tangential direction by using the derivative in normal direction with the help of the following lemma.

LEMMA 4.1. (Bounding of the derivative in tangential direction) *The derivative in tangential direction is bounded by the derivative in normal direction (see Fig. 3.1) as follows*

$$\sum_{\sigma\in\mathcal{E}_{int}} \left(\frac{\bar{D}^\sigma_{12}}{\bar{D}^\sigma_{11}}\right)^2 \left(\frac{u^n_N - u^n_S}{h}\right)^2 \bar{D}^\sigma_{11} \le \gamma \sum_{\sigma\in\mathcal{E}_{int}} \left(\frac{u^n_E - u^n_W}{h}\right)^2 \bar{D}^\sigma_{11}\,, \tag{4.1}$$

$$\text{where} \quad 0 \le \gamma < 1, \quad \gamma = \max_{\sigma\in\mathcal{E}} \gamma_\sigma, \quad \gamma_\sigma = \sum_{\delta\in P_\sigma\cap\mathcal{E}_{int}} \frac{1}{4} \left(\frac{\bar{D}^\delta_{12}}{\bar{D}^\delta_{11}}\right)^2 \frac{\bar{D}^\delta_{11}}{\bar{D}^\delta_{11}}\,,$$

where edges δ and set P_σ are given in the following definition.

DEFINITION 4.2. *Let P_σ be the set of all edges δ perpendicular to σ, which have common vertex with σ and fulfill the following conditions:*
$(x_{E_\delta} - x_{W_\delta}) \cdot \mathbf{t}_{W,\sigma} > 0$ if $(x_{N_\sigma} - x_{S_\sigma}) \cdot \mathbf{t}_{W,\sigma} > 0$ and
$(x_{E_\delta} - x_{W_\delta}) \cdot \mathbf{t}_{W,\sigma} < 0$ if $(x_{N_\sigma} - x_{S_\sigma}) \cdot \mathbf{t}_{W,\sigma} < 0$,
which means that $x_{E_\delta} - x_{W_\delta}$ has the same orientation as the tangent $\mathbf{t}_{W,\sigma}$. Let us note that $x_{W_\sigma} = x^1_{W_\delta} = x^3_{E_\delta}$, for $\sigma = \sigma_{WE}$, $x_{E_\sigma} = x^2_{W_\delta} = x^4_{E_\delta}$, for $\sigma = \sigma_{WE}$, $x_{W_\sigma} = x^2_{E_\delta} = x^4_{W_\delta}$, for $\sigma = \sigma_{EW}$ and $x_{E_\sigma} = x^1_{E_\delta} = x^3_{W_\delta}$, for $\sigma = \sigma_{EW}$.

Our convergence proof is based on Kolmogorov's compactness theorem. We proved the following lemmata: Uniform boundedness, Time translate estimate, Space translate estimate and stronger Space translate estimate in [3]. Using these lemmata we know that the sequence of discrete solution $u_{h,k}$ is relatively compact in L^2, which implies that there exists a subsequence of $u_{h,k}$ which is bounded. The main theorem of the convergence analysis is given below.

THEOREM 4.3. (Convergence of the scheme) *The sequence $u_{h,k}$ converges strongly in $L^2(Q_T)$ to the unique weak solution u of (2.1)-(2.3) as $h, k \to 0$.*

The crucial ideas used in our convergence proof are the convergence of the discrete weak form to the continuous weak form, which follows from the Lipschitz continuity of diffusion tensor elements and the fact that the limit u of $u_{h,k}$ is in space $L^2(0,T;H^1(\Omega))$, which follows from stronger Space translate estimate. The detailed convergence proof can be found in [3].

5. Error estimate analysis. This section concerns with an estimate of the difference between the weak solution of the model (2.1)-(2.3) and the numerical solution satisfying the scheme (3.3)-(3.4) in dependence on spatial and time discretization step, see [2]. Subtracting the discrete form from the continuous form and rearranging it we get a relation. It can be split in several terms and each of them can be bounded. Using these estimations we can state the following theorem.

THEOREM 5.1. (Error estimate) *Let the weak solution fulfil the following regularity properties: $\nabla u \in L_\infty(Q_T)$, $u_{tt} \in L_2(Q_T)$, $u \in L_2(I, W^{2,2}(\Omega))$, $\nabla u_t \in L_2(I, L_\infty(\Omega))$. Let $e_W^n = u(x_W, t_n) - u_W^n$ and $e_{h,k}^n(x,t) = \sum_{W \in \mathcal{T}_h} e_W^n \chi_{\{x \in W\}} \chi_{\{t_{n-1} < t \le t_n\}}$. Then, there exist a constant C, such that for sufficiently small h*

$$\int_\Omega |e_{h,k}^m|^2 dx + \sum_{n=1}^{m} \int_\Omega |e_{h,k}^n - e_{h,k}^{n-1}|^2 dx + \sum_{n=1}^{m} \int_{t_{n-1}}^{t_n} \sum_{\sigma \in \mathcal{E}_{int}} \left(e_E^n - e_W^n\right)^2 dt \le C(h^2 + k)$$

for every $m = 1, ..., N_{\max}$.

One can observe that the error of the piecewise constant approximation given by our scheme in $L_\infty(I, L_2)$ is of order h. The core of error estimate proof consists of a bounding of the derivative in tangential direction by means of the derivative in normal direction, a time translate estimate for approximate solution and the Lipschitz continuity of the diffusion tensor elements with respect to the smoothed partial derivatives of the solution. The detailed error estimate proof is given in [2].

Let us note that the $3D$ convergence / error estimate analysis is still an outstanding problem since we are not yet able to extend the inequality from Lemma 4.1 to its $3D$ version.

6. Computational experiments. The goal of this section is to demonstrate benefits of our numerical technique. We performed our experiments on a $2D$ fingerprint image (type of flow-like structures image) and $3D$ image sequences coming from the two-photon laser scanning microscopy. They represent early stages of zebrafish embryogenesis.

First experiment represents the behaviour of the CED (coherence enhancing diffusion). This technique yields a coherence improvement of image flow-like structures. After several filtration steps round interrupted places become gradually elongated in the coherence direction and they will be eventually corrected, see Fig. 6.1 and Fig. 6.2. We used the following filtration parameters: a space step $0,01$, a time step $0,0001$,

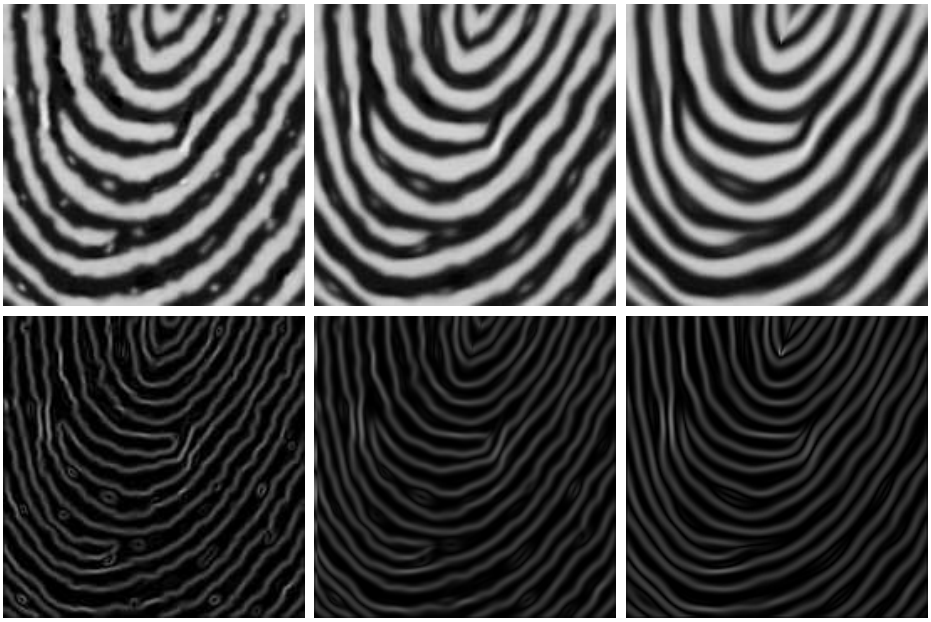


FIG. 6.1. *A fingerprint image. Top: the original image(left), the filtered image after* 5 *time steps(middle) and the filtered image after* 20 *time steps(right). Bottom: the Sobel edge detections of these images.*

$\tilde{t} = 0,000025$ and $\rho = 0,002$. Fig. 6.1 shows a fingerprint image. The original image is deteriorated by numerous redundant apertures while most of them are lost in the filtered image. Fig. 6.2 depicts damaged cell membranes. Some boundaries are almost lost in the original image, but we are able to clearly recognize them in the filtered image.

6.1. Pre-processing technique. Further, we concern our method as a pre-processing technique. We show its contribution to the subsequent image algorithms. If we pre-process images for techniques which depend on the connectivity of coherent image structures by the CED, we achieve significantly better results. We can adduce an edge detection as an example. If we compare the edge detections of an original and filtered image, see Fig. 6.1 and Fig. 6.2 (bottom), we can observe that the edge detection of the original image depicts many superfluous image structures caused by noise which are omitted in the edge detection of the filtered image. Moreover, several boundaries which are lost in the first edge detection are reconstructed in the second edge detection, see Fig. 6.2 (bottom).

The structure segmentation, see [5], is also the post-processing algorithm following CED. We use the segmentation based on the subjective surface method, see [10] and its finite volume implementation from [9]. The segmentation model has the following form

$$\partial_t u = \sqrt{\varepsilon^2 + |\nabla u|^2}\nabla . \left(g(|\nabla G_\sigma * I^0|) \frac{\nabla u}{\sqrt{\varepsilon^2 + |\nabla u|^2}} \right) \quad \text{in } Q_T \equiv I \times \Omega, \tag{6.1}$$

$$u(x,0) = u_0(x) \quad \text{in } \Omega, \tag{6.2}$$

$$u = 0 \qquad \text{on } I \times \partial\Omega, \tag{6.3}$$

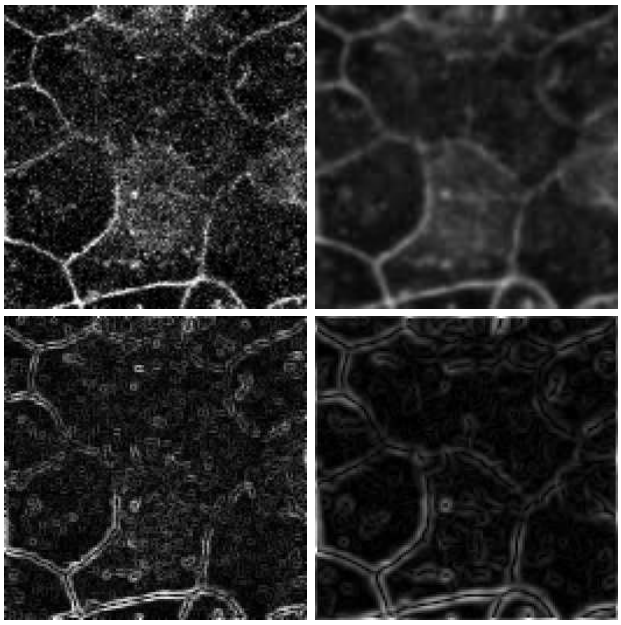


FIG. 6.2. *Cell membranes. Top: the original image(left) and the filtered image after* 5 *time steps(right). Bottom: the Sobel edge detections of these images.*

where I^0 is the image intensity and ε is the regularization parameter. The solution u denotes the evolving segmentation function. The function $g = g(|\nabla G_\sigma * I^0|)$ has the role of the edge detector. We start the segmentation imposing the initial segmentation function in an approximate center of segmented object. This function is evolved by equation (6.1) to a final steady state which gives the boundaries of the segmented object. The question is which isoline of the final steady state most precisely represents the object shape. The chosen isoline is most naturally taken as the average of maximal and minimal value of the final segmentation function.

The goal of this experiment is to segment an eye retina of a zebrafish embryo. Let us note that the structure segmentation is much more complicated than image segmentation since evolving segmentation function is restrained to achieve correct segmentation steady state by fine image objects representing inner cell structures. In order to overcome these constraints we pre-processed images for the segmentation by the CED. Even though they look too blurred they are very suitable for the structure segmentation, see Fig. 6.3. The segmentation result for the original image consists of amount of various isolines and chosen medium isoline bounds only a part of segmented object. On the contrary, the final segmentation function for the image filtered by the CED is represented by a variety of almost identical isolines and each of them precisely illustrates shape of segmented object.

In order to compare our method with other filtration techniques, we pre-processed images for segmentation by the GMCF (geodesic mean curvature flow), MCF (mean curvature flow) and PM (Perona-Malik) smoothing. Fig. 6.4 depicts their segmentation results which are much worse than the final steady state achieved by the coherence enhancing technique. It is caused by the fact that this diffusion not only smooths noise and image objects but emphasizes image structure boundaries as well.

Last experiment is devoted to results of the $3D$ CED as well as the $3D$ segmen-

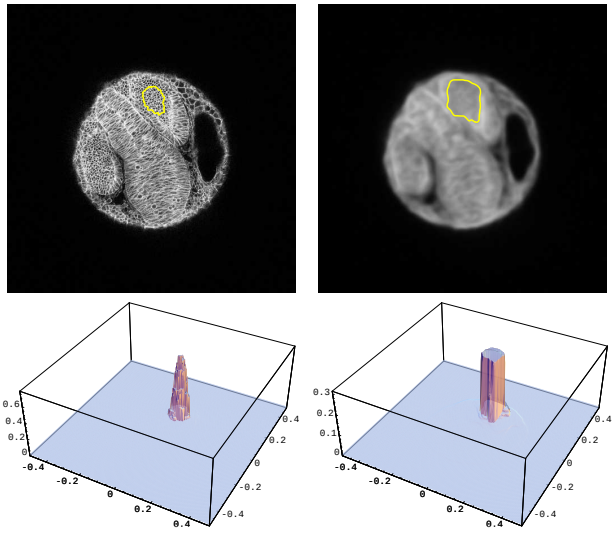


FIG. 6.3. *The eye retina segmentation using the* $2D$ *slice of* $3D$ *original image (left) and the* $2D$ *slice of* $3D$ *image filtered by* 20 *time steps of the* $3D$ *nonlinear tensor diffusion (right). Top: the averaged isoline of the final state of segmentation function is superimposed to the original and filtered slice, respectively. Bottom: the graph of the final state of segmentation function is plotted after* 2000 *segmentation time steps using the original slice and after* 200 *time steps using the filtered slice.*

tation algorithms, see Fig. 6.5. These techniques were performed on the $3D$ image detail representing two cell nuclei. One can observe that the original image is much more deteriorated by a noise than the image filtered by the CED and the noise of the filtered image is less distinct. The contours of the filtered nuclei are smoother than the nucleus contours from the original image since the diffusion tensor of the CED steers the smoothing process in such a way that the diffusion is strong along the coherence plane and very low in the perpendicular direction to this plane. Owing to the above mentioned facts we achieved more precise segmentation results for the nuclei filtered by CED, cf. Fig. 6.5(left) and (right).

The experiments mentioned before confirm the utility of this filtration as a preprocessing technique for algorithms which depend on the connectivity of coherent image structures.

REFERENCES

[1] Y. COUDIERE, J. P. VILA AND P. VILLEDIEU, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33, (1999), pp. 493–516.

[2] O. DRBLÍKOVÁ, A. HANDLOVIČOVÁ AND K. MIKULA, *Error Estimates of the Finite Volume Scheme for the Nonlinear Tensor Anisotropic Diffusion*, Applied Numerical Mathematics, 59(10) (2009), pp. 2548–2570.

[3] O. DRBLÍKOVÁ AND K. MIKULA, *Convergence Analysis of Finite Volume Scheme for Nonlinear Tensor Anisotropic Diffusion in Image Processing*, SIAM J. Numer. Anal., 46(1) (2007), pp. 37–60.

[4] O. DRBLÍKOVÁ AND K. MIKULA, *Semi-implicit Diamond-cell Finite Volume Scheme for* $3D$ *Nonlinear Tensor Diffusion in Coherence Enhancing Image Filtering*, Finite Volumes for Complex Applications, Proceedings of the 5th International Symposium on Finite Volumes for Complex Applications (FVCA5). Published in Great Britain and the United States in

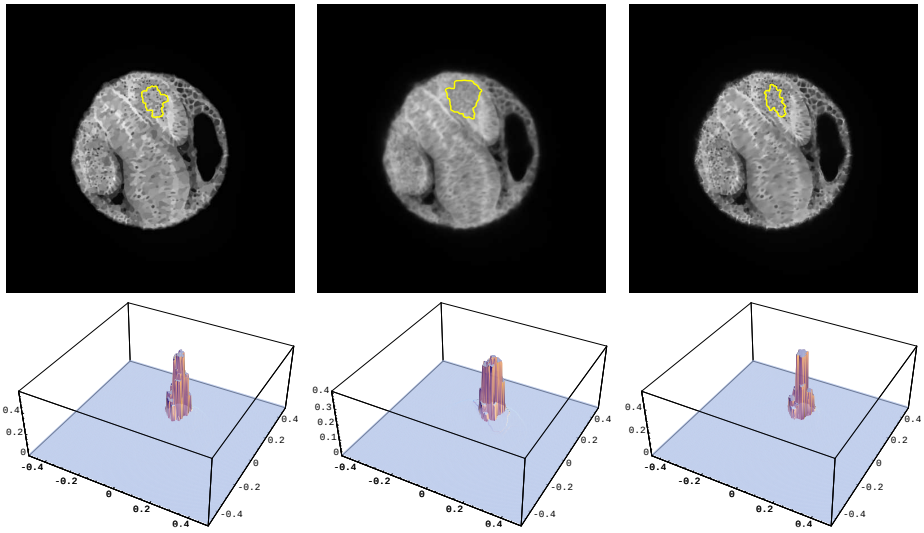


FIG. 6.4. *The eye retina segmentation using the* 2*D slice of the* 3*D image filtered by* 100 *steps of the* 3*D GMCF filtering (left),* 25 *steps of the* 3*D MCF filtering (middle) and* 20 *steps of the* 3*D PM filtering (right). Top: the averaged isoline of the final state of segmentation function is superimposed to the filtered slice. Bottom: the graph of the final state of segmentation function is plotted after* 3000 *segmentation steps using the GMCF filtering, after* 500 *segmentation steps using the MCF filtering and after* 5000 *segmentation steps using the PM filtering.*

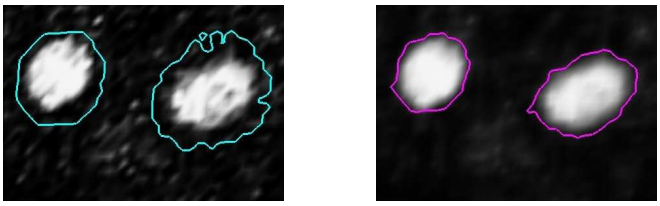


FIG. 6.5. 3*D nucleus segmentation using the* 3*D original image (left) and using the* 3*D image filtered by* 10 *time steps (right).*

2008 by ISTE Ltd and John Wiley & Sons, Inc., ISBN 978-1-84821-035-6, pp. 343–350.

[5] O. DRBLÍKOVÁ, K. MIKULA AND N. PEYRIÉRAS, *The Nonlinear Tensor Diffusion in Segmentation of Meaningful Biological Structures from Image Sequences of Zebrafish Embryogenesis*, Scale Space and Variational Methods in Computer Vision, Proceedings. Springer Berlin Heidelberg (2009), pp. 63–74.

[6] R. EYMARD, T. GALLOUËT AND R. HERBIN, *Finite Volume Methods*, in: Handbook for Numerical Analysis, Vol. 7 (Ph. Ciarlet, J. L. Lions, eds.), Elsevier (2000).

[7] E. MEIJERING, W. NIESSEN, J. WEICKERT, M. VIERGEVER, *Diffusion-Enhanced Visualization and Quantification of Vascular Anomalies in Three-Dimensional Rotational Angiography*, Results of an In-Vitro Evaluation. Medical Image Analysis, 6(3) (2002), pp. 217–235.

[8] K. MIKULA AND N. RAMAROSY, *Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing*, Numer. Math. 89 (3), (2001) pp. 561–590.

[9] K. MIKULA, A. SARTI, F. SGALLARI, *Co-volume level set method in subjective surface based medical image segmentation*, in: Handbook of Medical Image Analysis: Segmentation and Registration Models (J. Suri et al., Eds.), Springer, New York, (2005) pp. 583–626.

[10] A. SARTI, R. MALLADI, J.A. SETHIAN, *Subjective Surfaces: A Method for Completing Missing Boundaries*, Proceedings of the National Academy of Sciences of the United States of America, 12 (97), (2000) pp. 6258–6263.

[11] J. WEICKERT, *Coherence-enhancing diffusion filtering*, Int. J. Comput. Vision, Vol. 31, (1999) pp. 111–127.

Proceedings of EQUADIFF 2017
pp. 387–396

NEW EFFICIENT NUMERICAL METHOD FOR 3D POINT CLOUD SURFACE RECONSTRUCTION BY USING LEVEL SET METHODS.

BALÁZS KÓSA*, JANA HALIČKOVÁ–BREHOVSKÁ†, AND KAROL MIKULA‡

Abstract. In this article, we present a mathematical model and numerical method for surface reconstruction from 3D point cloud data, using the level-set method. The presented method solves surface reconstruction by the computation of the distance function to the shape, represented by the point cloud, using the so called Fast Sweeping Method, and the solution of advection equation with curvature term, which creates the evolution of an initial condition to the final state. A crucial point for efficiency is a construction of initial condition by a simple tagging algorithm which allows us also to highly speed up the numerical scheme when solving PDEs. For the numerical discretization of the model we suggested an unconditionally stable method, in which the semi-implicit co-volume scheme is used in curvature part and implicit upwind scheme in advective part. The method was tested on representative examples and applied to real data representing the historical and cultural objects scanned by 3D laser scanners.

Key words. point cloud, level set methods, reconstruction

AMS subject classifications. 65M06, 65Y20, 53A05, 65D17

1. Introduction. The aim of our work is to create a reliable and efficient numerical method which can easily create computerized 3D models from point cloud data that resembles the original object as much as possible. This type of data can be obtained by 3D scanning or by photogrammetric methods. The data created in this manner contains three coordinates for every scanned point. For further processing and the creation of an exact digital model of the scanned object this information is not enough. The point cloud lacks the information of the connectivity between the points, thus making the reconstruction of the surface a difficult task. Papers as [1, 2] have shown us that for solving this problem the level-set method can be applied. We follow basic ideas from these papers, but we take a different approach in the solution of the partial differential equation presented here.

In the following parts of our paper after the Mathematical Formulation of the applied level set equation in the section Algorithm for point cloud surface reconstruction we will present our method and its numerical discretization and solution. After the theoretical deduction of the method and the description of a short algorithm for computing the initial condition in Computation acceleration we suggest a way to accelerate the computational time, making the algorithm really efficient. In the last section Numerical results we present created 3D models which we obtained so far. We achieved this by implementing our method in the language C with the use of the programming environment of Visual Studio. The example pictures of the results used in this article are direct outputs from our application processed in the freely available open-source visualization software Paraview. With the help of this software we

*Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovakia (kosa@math.sk).

† Monument Board of the Slovak Republic Cesta na Červený most 6, 81406 Bratislava, Slovakia (jana.halickova@gmail.com)

‡Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Slovak University of Technology, Radlinskeho 11, 810 05 Bratislava, Slovakia (mikula@math.sk).

can easily compare the initial point cloud data and our results, to confirm that our assumptions regarding this new numerical method are right.

2. Algorithm for point cloud surface reconstruction. The level set method, which we are using is based on the solution of the advection equation with curvature term

$$u_t - \nabla d \cdot \nabla u - \delta \left|\nabla u\right| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|}\right) = 0 \qquad (x,t) \in \Omega \times [0,T] \tag{2.1}$$

where $u(x,t)$ is an unknown function, $v = -\nabla d$ is the advective velocity defined by the gradient of distance function d to the point cloud, parameter $\delta > 0$ determines influence of the curvature to the result, Ω is the computational domain and $[0,T]$ is a time interval. This equation is coupled with homogeneous Neumann boundary conditions and an initial condition which we will discuss later.

To obtain numerical solution of the model created from point cloud data, denoted by $\Omega_0 \subset \Omega$ and determined by equation (2.1), following steps have to be executed. First we have to computate the distance function to the point cloud. For computation we use the Fast sweeping method, as introduced in [3].The initialization of distance function in the Fast sweeping method is done in such way, that we prescribe exact distance to the nearest point from the cloud in the grid points next to the points in the cloud. After that we have to find a subvolume containing Ω_0, which will be used to set the initial function u^0 for the generation of the final solution of the equation. This subvolume is defined on discrete grid in subsection 2.2. The final solution (created 3D model) will be represented by an isosurface of the computated function $u\,(x,T)$ with value 0.5.

2.1. Numerical scheme for solving advection equation with curvature term. The numerical scheme is obtained by discretization of equation (2.1). We will do this analogically to the discretization used in [4].

2.1.1. Time discretization. For time discretization, we have to choose a uniform discrete time step, denoted by τ. We can replace the time derivative in (2.1) with a backward difference. Then we can formulate our semi-implicit time discretization in the following way:

Let τ be a fixed number and u^0 a function representing the initial surface of our mathematical model. Then at every discrete time $t_n = n\tau,\ n = 1, ..., N$ we search for the function u^n as the solution to equation

$$\frac{u^n - u^{n-1}}{\tau} - \nabla d \cdot \nabla u^n - \delta \left|\nabla u^{n-1}\right| \nabla \cdot \left(\frac{\nabla u^n}{|\nabla u^{n-1}|}\right) = 0 \tag{2.2}$$

2.1.2. Spatial discretization. Our discretized model consists of a 3D grid, which is built of voxels with cubic shape and an edge size h. We will interpret spatial discretization of the level set function u as numerical values $u_{i,j,k}$ at the voxel centres. In order to easily computate the gradient of the level-set equation $\left|\nabla u^{n-1}\right|$ in every time step of (2.2) we induct a 3D tetrahedral grid into the voxel structure and take a piecewise linear approximation of $u\,(x)$ on such a grid. This way we obtain a constant value of the gradient for each tetrahedron, by which we can construct in a simple and clear way the fully discrete system of equations.

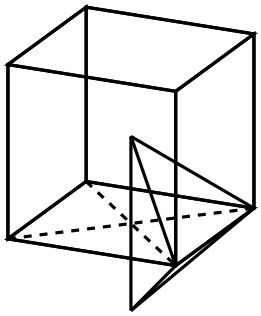


Fig. 2.1: Our initial voxel grid cell with a tetrahedral grid cell

The 3D tetrahedral finite element grid is created by the following approach. Every voxel is divided into six pyramid shaped elements with base surface given by the voxel's walls and vertex by the voxel centre. Each one of these pyramids is joined with neighbouring pyramids with whom they have a common base surface. These newly formed octahedrons are then split into four tetrahedrons as seen in Figure 2.1. In our new grid $\mathcal{T}_h$ the level-set function will be updated only at the centres of the voxels. They will represent so called degree of freedom (DF) nodes.

For the tetrahedral grid we construct a co-volume mesh, which will consist of cells p associated only with DF nodes of $\mathcal{T}_h$. We denote all neighbouring cells q of p by C_p. The cells q are all connected to the cell p by a common edge of four tetrahedrons, which is denoted by σ_{pq} with length h_{pq}. Each cell p is bounded by a plane for every $q \in C_p$ which is perpendicular to σ_{pq} and is denoted by e_{pq}. The set of tetrahedrons which have σ_{pq} as an edge are denoted by ε_{pq}. For every $T \in \varepsilon_{pq}$, c_{pq}^T is the area of the intersection of e_{pq} and T. N_p will be a set of tetrahedrons that have DF node associated with cell p as a vertex. On this grid u_h will be a piecewise linear function. Then we can use the notation $u_p = u_h(x_p)$, where x_p denotes the center coordinates of cell p.

Now that we have all notations which are needed we can begin the derivation of the spatial discretization of (2.2). We will do this by using a following modified form of the equation:

$$\frac{u^n - u^{n-1}}{\tau} + v \cdot \nabla u^n = \delta \left|\nabla u^{n-1}\right| \nabla \cdot \left(\frac{\nabla u^n}{|\nabla u^{n-1}|}\right) \tag{2.3}$$

where $v = -\nabla d$.

As the first step we will integrate (2.3) over every cell p.

$$\int_p \frac{u^n - u^{n-1}}{\tau} dx + \int_p v \cdot \nabla u^n dx = \int_p \delta \left|\nabla u^{n-1}\right| \nabla \cdot \left(\frac{\nabla u^n}{|\nabla u^{n-1}|}\right) dx \tag{2.4}$$

For the first term on the left-hand side of (2.4) we get the approximation

$$\int_p \frac{u^n - u^{n-1}}{\tau} dx = m(p) \frac{u_p^n - u_p^{n-1}}{\tau} \tag{2.5}$$

where $m(p)$ is a measure in $\mathbb{R}^d$ of the cell p.

For the second term on the left-hand side of (2.4) we are using the implicit upwind approach and get

$$\int_p v \cdot \nabla u^n dx = \sum_{q \in C_p} min\left(v_{pq}, 0\right)\left(u_q^n - u_p^n\right) \tag{2.6}$$

where $v_{pq} = h_{pq}^2 v \cdot n$.

Now what remains is the discretization of the right-hand side of (2.4). We use the divergence theorem to get

$$\int_p \delta \left|\nabla u^{n-1}\right| \nabla \cdot \left(\frac{\nabla u^n}{\left|\nabla u^{n-1}\right|}\right) dx = \delta \left|\nabla u_p^{n-1}\right| \sum_{q \in C_p} \int_{e_{pq}} \frac{1}{\left|\nabla u^{n-1}\right|} \frac{\partial u^n}{\partial n} d\sigma \tag{2.7}$$

The integral part $\int_{e_{pq}} \frac{1}{\left|\nabla u^{n-1}\right|} \frac{\partial u^n}{\partial n} d\sigma$ and $\left|\nabla u_p^{n-1}\right|$ from (2.7) will be approximated numerically using piecewise linear reconstruction of u^{n-1} on the tetrahedral grid $\mathcal{T}_h$, thus we get

$$\delta \left|\nabla u_p^{n-1}\right| \sum_{q \in C_p} \left(\sum_{T \in \varepsilon_{pq}} c_{pq}^T \frac{1}{\left|\nabla u_T^{n-1}\right|} \right) \frac{u_q^n - u_p^n}{h_{pq}}$$

$$M_p^{n-1} = \left|\nabla u_p^{n-1}\right| = \sum_{T \in N_p} \frac{m\left(T \cap p\right)}{m\left(p\right)} \left|\nabla u_T^{n-1}\right|$$

and the final form of equation (2.3) after reorganization will be

$$\begin{aligned} u_p^{n-1} = u_p^n + \frac{\tau}{m\left(p\right)} \Bigg(& \sum_{q \in C_p} min\left(v_{pq}, 0\right)\left(u_q^n - u_p^n\right) \\ & -\delta\, M_p^{n-1} \sum_{q \in C_p} \left(\sum_{T \in \varepsilon_{pq}} c_{pq}^T \frac{1}{\left|\nabla u_T^{n-1}\right|} \right) \frac{u_q^n - u_p^n}{h_{pq}} \Bigg) \end{aligned} \tag{2.8}$$

From this form, we are able to derive the system of linear equations which we will solve at every time step. For the linear equations, we will define regularized gradients by

$$\left|\nabla u_T\right|_\varepsilon = \sqrt{\varepsilon^2 + \left|\nabla u_T\right|^2} \tag{2.9}$$

After we arrange all parts of equation (2.8) we get the following coefficients

$$a_{pq}^{n-1} = \frac{\tau}{m\left(p\right)} \left(min\left(v_{pq}, 0\right) - \delta\, M_p^{n-1} \frac{1}{h_{pq}} \sum_{T \in \varepsilon_{pq}} c_{pq}^T \frac{1}{\left|\nabla u_T^{n-1}\right|_\varepsilon} \right) \tag{2.10}$$

thus, we can formulate our semi-implicit co-volume scheme:

Let $u_p^0,\ p = 1, ..., M$ be given discrete initial values of the level-set function. Then, for $n = 1, ..., N$ we look for $u_p^n,\ p = 1, ..., M$, satisfying

$$u_p^n + \frac{\tau}{m\left(p\right)} \sum_{q \in N_p} a_{pq}^{n-1}\left(u_q^n - u_p^n\right) = u_p^{n-1} \tag{2.11}$$

With addition of homogeneous Neumann boundary conditions to our fully discrete scheme we obtain a system of linear equations. Since a_{pq}^{n-1} are non-negative we can prove the following statement.

Theorem. *There exists unique solution* $(u_1^n, ..., u_M^n)$ *of* (2.11) *for any* $\tau > 0$, $\varepsilon > 0$, *and for every* $n = 1, ..., N$. *The system matrix is a strictly diagonally dominant M-matrix. For any* $\tau > 0$, $\varepsilon > 0$, *the following* L_∞ *stability holds:*

$$\min_p \ u_p^0 \le \min_p \ u_p^n \le \max_p \ u_p^n \le \max_p \ u_p^0, \ \ 1 \le n \le N. \tag{2.12}$$

The number of time steps N is determined by the difference of the solution in current and previous time steps in discrete L^2 norm. The computation is stopped if this difference is less than the prescribed tolerance, which we usually set to 10^{-6}. Then the stopping time $T = N\tau$.

If we denote the DF nodes with indexes (i, j, k) and rearrange (2.11) to obtain the coefficients for every node we can define for a DF node the equation

$$\begin{aligned} &c_{i,j,k}u_{i,j,k}^n + b_{i,j,k}u_{i,j,k-1}^n + t_{i,j,k}u_{i,j,k+1}^n + n_{i,j,k}u_{i+1,j,k}^n \\ &+s_{i,j,k}u_{i-1,j,k}^n + e_{i,j,k}u_{i,j+1,k}^n + w_{i,j,k}u_{i,j-1,k}^n = u_{i,j,k}^{n-1} \end{aligned} \tag{2.13}$$

When we collect the equations for all DF nodes and take into account Neumann boundary conditions we get the linear system which we have to solve. For the solution of this system we choose the SOR (Successive Over Relaxation) iterative method. We start the iterations by setting $u_{i,j,k}^n = u_{i,j,k}^{n-1}$, then in every iteration $l = 1, ...$ we use the following two step procedure:

$$\begin{aligned} Y = (&u_{i,j,k}^{n(0)} - b_{i,j,k}u_{i,j,k-1}^{n(l)} - t_{i,j,k}u_{i,j,k+1}^{n(l-1)} - n_{i,j,k}u_{i+1,j,k}^{n(l-1)} \\ &- s_{i,j,k}u_{i-1,j,k}^{n(l)} - e_{i,j,k}u_{i,j+1,k}^{n(l-1)} - w_{i,j,k}u_{i,j-1,k}^{n(l)})/c_{i,j,k} \\ u_{i,j,k}^{n(l)} &= u_{i,j,k}^{n(l-1)} + \omega\left(Y - u_{i,j,k}^{n(l-1)}\right) \end{aligned} \tag{2.14}$$

We define squared L_2 norm of residuum at current iteration by

$$\begin{aligned} R_l = \sum_{i,j,k} (&c_{i,j,k}u_{i,j,k}^{n(l)} + b_{i,j,k}u_{i,j,k-1}^{n(l)} + t_{i,j,k}u_{i,j,k+1}^{n(l)} + n_{i,j,k}u_{i+1,j,k}^{n(l)} \\ &+ s_{i,j,k}u_{i-1,j,k}^{n(l)} + e_{i,j,k}u_{i,j+1,k}^{n(l)} + w_{i,j,k}u_{i,j-1,k}^{n(l)} - u_{i,j,k}^{n(0)})^2 \end{aligned}$$

The iterative process is stopped if $R^l < TOL$.

2.2. Computation of the initial condition. As mentioned, this method needs an initial condition, represented by the initial function $u^0(x)$, which will be deformed to get the solution, that is the final form of the created 3D model. Theoretically any initial surface that contains the point cloud data set could be used, but an optimal initial guess is crucial for the efficiency of the method. We can find this optimal surface by identifying all points for which the value of the distance function is greater or equal to a parameter β. For simplicity let us call these points, exterior points. To find all these points we will use the following algorithm:

- Mark all points on the borders of the grid as exterior and add them to set E.
- For every point in the set E check all neighbouring points in the grid.

- If the neighbouring point is not an exterior point and its distance from the point cloud is greater or equal to β add it to the set E and mark as exterior.
- Continue until you get to the last point of E.

When we found all the exterior points we set $u^0(x)$ to be equal 0 in every exterior point and 1 in every other point. With this approach, we can find an initial surface close to the final shape as seen on the Figure 2.2.

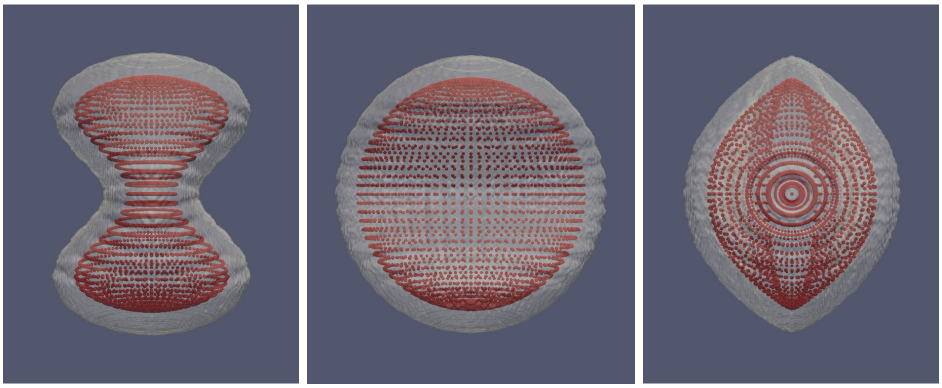


Fig. 2.2: Example for the initial condition used in our method. Object is shown from different angles.

3. Computation acceleration. The part of our algorithm which consumes the most time during computation is the solution of the linear system of equations (2.11) coupled with the computation of its coefficients. To reduce this time, we came up with the following idea. First, we construct a band around the area between the initial surface and the point cloud data. To find the surface which we want to reconstruct it is sufficient to update the values on grid cells contained in such a band, thus we can computate coefficients and evaluate the SOR method (2.14) only in this new subset of all grid cells. On Figure 3.1 we can see an example of this subset. For easier visualization, we show this on a slice with the plane $x = 0$. Here the red line marks the point cloud data, the purple line the initial surface and the white lines the borders of the created band. In the background of the picture we show the values of the distance function.

To find this area we adopted the algorithm mentioned in the previous section, which was used to find the initial surface, to this task. To obtain an outer border for the band which contains the initial surface we chose a new parameter $\gamma = 2\beta$. With this additional parameter and the introduction of a new set denoted F the algorithm for finding the band is given as follows.

- Tag all points on the borders of the grid and add them to the set E.
- For every point in the set E check all neighbouring points in the grid.
- If the neighbouring point is not tagged execute the following steps.
  - If the neighbouring point's distance from the point cloud is smaller or equal to γ add it to the set F.
  - If the neighbouring point's distance from the point cloud is greater or equal to β add it to the set E as well and tag it.
- Continue until you get to the last point of E. When we finish with set E we start a new cycle for set F.

- For every point in the set F check all neighbouring points in the grid.
- If the neighbouring point is not tagged and its distance from the point cloud is smaller or equal to γ add it to the set F.
- Continue until you get to the last point of F.

While we look for points with distance smaller or equal to γ we will cross the border with distance 0, represented by the point cloud, thus set F will contain also grid points from the inner region of the object. From the set F we can create an array consisting of values 0, for points not in the band, and 1, for points in the band. This will serve as a mask for the SOR method, thus in the computation loops we can determine if it is necessary to computate the new value or if we can skip to the next grid point.

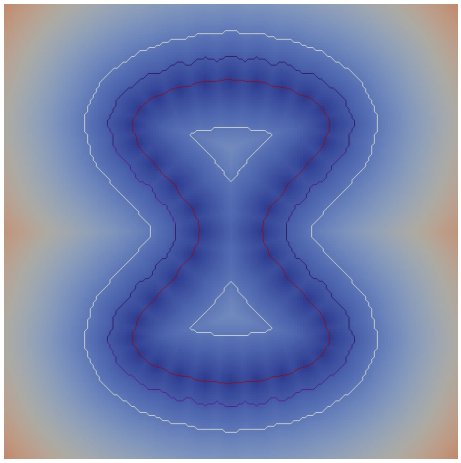


Fig. 3.1: The slice of our new computational area on the plane $x = 0$.

We measured how much time we managed to save with this new approach on real-life data sets representing a bracelet and a sealer. The tests were executed on a personal notebook with a dual core processor and 4 GB of memory. Our results are listed in the tables 3.1 and 3.2. We tested the algorithm on grids containing 40^3, 80^3 and 160^3 grid cells. All tests were performed with the same parameter β and stopping criteria for the iterations.

In the second column of the tables we recorded the number of points contained by the band. This number depends on the size and form of the original object represented by the data set. In columns three and four we see the measured times for the original and optimized implementation. In the tests, we achieved not only reduced times but also better convergence, so fewer time steps were needed. This led to computations which were 20 to 60 times faster.

Visually we cannot detect any difference between the created 3D models computed by the two methods, original and optimized. We measured the mean value of squared differences between all grid values and listed the obtained values in the third column. We can see that these values are in the tolerable range.

Number of grid cells	Points in band	CPU time (s) Original	CPU time (s) Optimized	Mean squared difference
40^3	4 636	4.269	0.261	8.90795e-7
80^3	37 640	34.247	1.677	2.27554e-8
160^3	304 456	895.68	13.385	1.92055e-8

Table 3.1: CPU times comparison for the bracelet data set

Number of grid cells	Points in band	CPU time (s) Original	CPU time (s) Optimized	Mean squared difference
40^3	6 075	13.914	0.537	1.75849e-6
80^3	48 710	88.673	3.470	4.38982e-8
160^3	392 185	2 051.402	72.846	9.36878e-9

Table 3.2: CPU times comparison for the sealer data set

4. Numerical results. In this section, we present the reconstruction of the point cloud surfaces on a representative testing example and real data. These examples are a good display of the quality of our method.

Figure 4.1 illustrates the test example. This object was used for the verification of the correct behaviour of our method during the implementation phase. The point cloud data was generated with corresponding parametric equations of the object. The representative example was created on a grid containing 80^3 cells. We can see that for this test with such a sparse grid we already got good results.

On Figure 4.2 and 4.3 we can see real-life data. These items where archaeological finds and the point cloud scans were provided by the Monuments Board of the Slovak republic to which we express our great thanks. On Figure 4.2 we can see a bracelet. The created 3D model was computed on a grid with 160^3 cells. On Figure 4.3 we can see a sealer, with a very interesting surface structure. The created 3D model was computed on a grid with 320^3 cells.

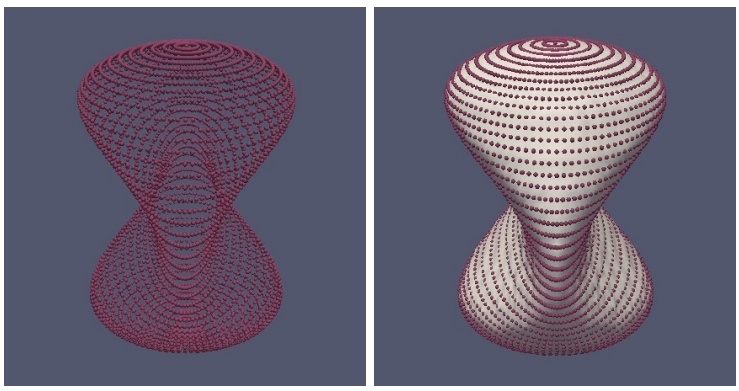


Fig. 4.1: On the left, we see the point cloud data, on the right the point cloud with the created 3D model.

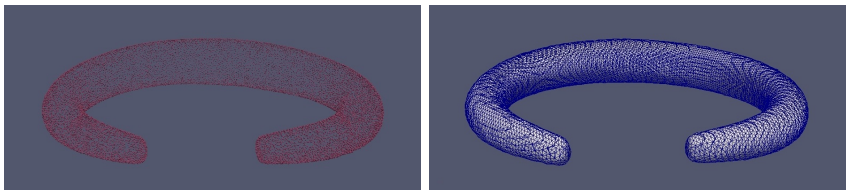


Fig. 4.2: Archaeological finds: bracelet. On the left, we see the point cloud data, on the right the final result with triangulated surface.

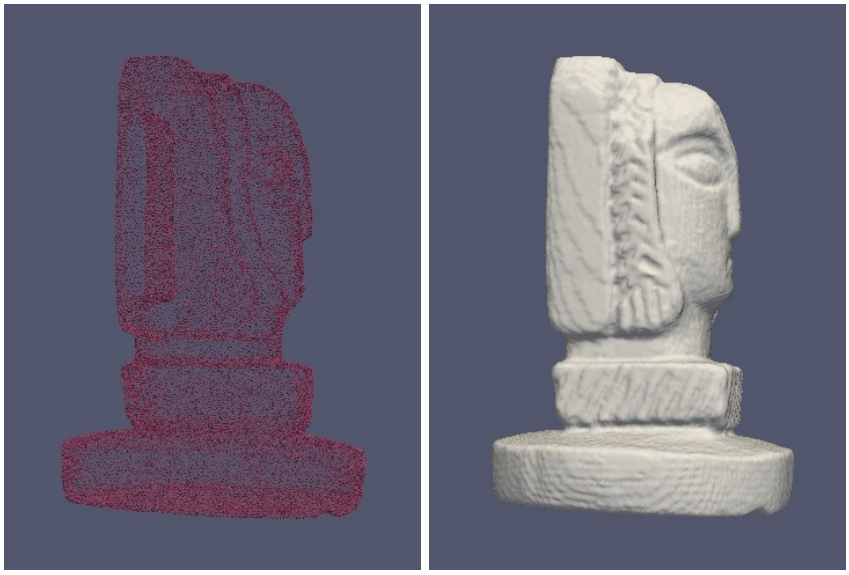


Fig. 4.3: Archaeological finds: sealer. On the left, we see the point cloud data, on the right the final result.

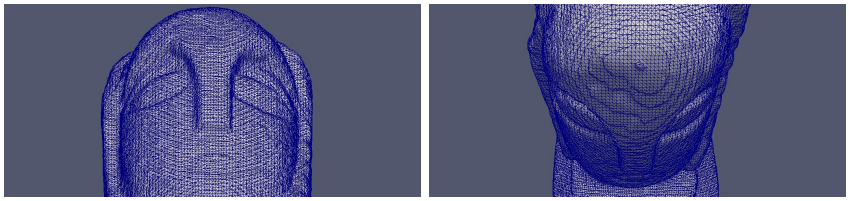


Fig. 4.4: Details of the sealer with triangulated surface.

We also tested our method on data sets with noise. In the point cloud data of the sealer we added artificial noise by changing the coordinates of 100 random points. Thanks to the curvature part of equation (2.1) this kind of noise has no effect on our created 3D model. We can observe that fact in Figure 4.5.

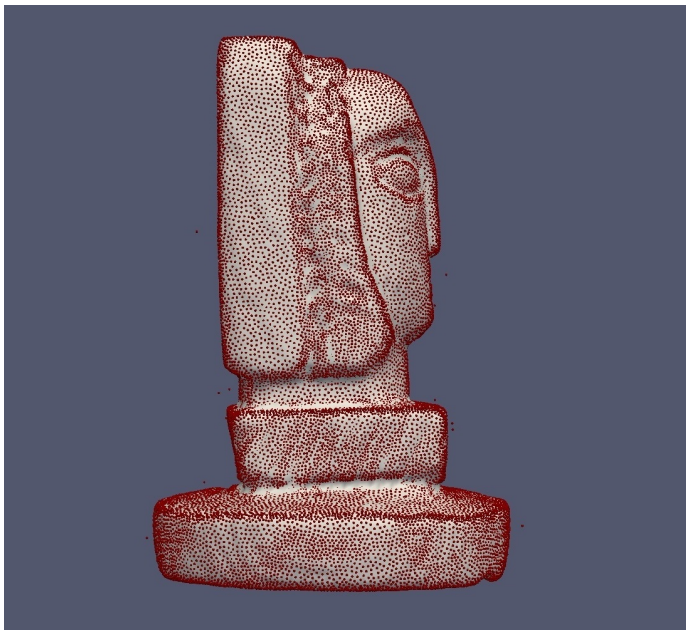


Fig. 4.5: Sealer point cloud data with noise, visualized with the final result.

5. Conclusions. In this work we presented our approach for surface reconstruction from point cloud data utilizing the level set method. We formulated the mathematical model, derived the time and spatial discretization and provided the reader with an exact description of the numerical solution. By implementing the method we obtained several interesting results for numerical tests and real-life data which we presented as examples in the last section. Our results show that for smoother objects a sparse grid already shows good result, but for an object with more detail we need more grid points. With adjusting the SOR method to our needs we achieved significant reduction of the required computational time, thus making our method more suitable for real-life application.

Acknowledgments. This work was supported by the grants APVV-15-0522 and VEGA 1/0608/15.

REFERENCES

[1] J. Haličková, K. Mikula, Level set method for surface reconstruction and its application in surveying, Journal of Surveying Engineering 143 (3). doi:10.1061/(ASCE)SU.1943-5428.0000159.

[2] H. Zhao, S. Osher, B. Merriman, M. Kang, Implicit and nonparametric shape reconstruction from unorganized data using a variational level set method, Computer Vision and Image Understanding 80 (2000) 295–319. doi:10.1006/cviu.2000.0875.

[3] H. K. Zhao, A fast sweeping method for eikonal equations, Mathematics of Computation 74 (2004) 603–627. doi:10.1090/S0025-5718-04-01678-3.

[4] S. Corsaro, K. Mikula, A. Sarti, F. Sgallari, Semi-implicit covolume method in 3d image segmentation, SIAM Journal on Scientific Computing 28 (6) (2006) 2248–2265. doi:10.1137/060651203.

Proceedings of EQUADIFF 2017
pp. 397–406

CONVERSE PROBLEM FOR THE TWO-COMPONENT RADIAL GROSS-PITAEVSKII SYSTEM WITH A LARGE COUPLING PARAMETER

JEAN-BAPTISTE CASTERAS* AND CHRISTOS SOURDIS †

Abstract. We consider strongly coupled competitive elliptic systems that arise in the study of two-component Bose-Einstein condensates. As the coupling parameter tends to infinity, solutions that remain uniformly bounded are known to converge to a segregated limiting profile, with the difference of its components satisfying a limit scalar PDE. In the case of radial symmetry, under natural non-degeneracy assumptions on a solution of the limit problem, we establish by a perturbation argument its persistence as a solution to the elliptic system.

Key words. Singular perturbation, competitive elliptic system, segregation

AMS subject classifications. 35J57

1. Introduction.

We consider coupled elliptic systems of the form

$$\Delta u_i = f_i(u_i) + g u_i \sum_{j\neq i} a_{ij} u_j^2, \text{ in } \Omega; \ u_i = 0 \text{ on } \partial\Omega, \tag{1.1}$$

$i = 1, \cdots, m$, where f_i are smooth functions with

$$f_i(0) = 0, \tag{1.2}$$

g is a real parameter, a_{ij} are nonnegative constants such that $a_{ii} > 0$, $a_{ij} = a_{ji}$, $i, j = 1, \cdots, m$, and Ω is a bounded smooth N-dimensional domain. Systems of this form arise in the study of multi-component Bose-Einstein condensates. In this context, the reaction terms are typically

$$f_i(u) = g_i u^3 - \mu_i u, \ g_i, \mu_i \in (-\infty, +\infty). \tag{1.3}$$

The coupling parameter g measures the interaction between the different components in the mixture: if $g < 0$ they attract each other, whereas if $g > 0$ they repel each other. On the other hand, the coefficients g_i in (1.3) measure the interaction between atoms in the same i-th component: if $g_i < 0$ there is attraction, whereas if $g_i > 0$ there is repulsion.

The function u_i represents the density corresponding to the i-th component in the mixture, and thus is naturally assumed to be positive. Nevertheless, the mathematical interest to (1.1) also extends to sign-changing solutions. In passing, we note that (1.1) has variational structure as it comes from a Gross-Pitaevskii energy.

In the following, we will consider the case of strong repulsion (or competition), that is $g \gg 1$. Moreover, we will focus on the case of two components, but first let us recall some of the main known results for the case of m components.

*Département de Mathématique, Université libre de Bruxelles, Campus de la Plaine CP 213, Bd. du Triomphe, 1050 Bruxelles, Belgium, supported by the Belgian Fonds de la Recherche Scientifique FNRS (jeanbaptiste.casteras@gmail.com).

†Department of Mathematics, University of Ioannina, Ioannina, 45110, Greece (sourdis@uoc.gr).

1.1. Known results. In the seminal paper [13] (see also [8] for the corresponding parabolic problem), it was shown that if a family of solutions $\mathbf{u}_g = (u_1^g, \cdots, u_m^g)$ of (1.1) remains bounded in $L^\infty(\Omega)$ as $g \to +\infty$, then it also remains bounded in $C^\alpha(\bar{\Omega})$ for any $\alpha \in (0,1)$. We also refer to [25] for a related result in planar domains. Hence, thanks to a well known compact imbedding, possibly up to a subsequence $g_n \to +\infty$, such a family converges in $C^\alpha(\bar{\Omega})$ for any $\alpha < 1$ to some limiting configuration $\mathbf{u}_\infty = (u_1^\infty, \cdots, u_m^\infty)$. In fact, it was shown in [13] that the limiting profile has Lipschitz regularity up to the boundary of Ω. Furthermore, the limiting components are segregated, that is their supports are disjoint. In its respective support, the limiting component u_i^∞ satisfies the following elliptic problem

$$\Delta u_i^\infty = f_i(u_i^\infty). \tag{1.4}$$

In the language of singular perturbations, the above limit problem is called the outer limit problem.

More recently, it was shown in [18] that such families $\mathbf{u}_g$ remain bounded, uniformly in g, even in the Lipschitz norm, at least away from the boundary of the domain and for positive solutions

The regularity properties of the sharp interface

$$\Gamma = \left\{ x \in \bar{\Omega} \ : \ u_1^\infty(x) = \cdots = u_m^\infty(x) = 0 \right\}$$

were subsequently studied in [22]. It was shown there that Γ has properties analogous to the nodal set of eigenfunctions of the Laplacian: there exists $\Sigma \subset \Gamma$ with $\mathcal{H}_{dim}(\Sigma) \leq N-2$ such that $\Gamma \setminus \Sigma$ is a finite union of smooth manifolds (we refer to [23] for a detailed description of Σ). The set Σ is referred to as the singular part of the interface Γ, whereas $\Gamma \setminus \Sigma$ as the regular part. On each side of a smooth manifold M that composes the regular part of the interface there is only one nontrivial limiting component. Moreover, across M the corresponding limiting components, say $u_\infty = u_i^\infty$ and $v_\infty = u_j^\infty$ (it holds $i \neq j$, see [9]), satisfy the following reflection law:

$$|\nabla u_\infty| = |\nabla v_\infty| \ \text{ on } M. \tag{1.5}$$

We note that the above normal derivatives are nonzero by (1.2), (1.4) and Hopf's boundary point lemma.

More refined estimates for the convergence as $g \to +\infty$ have recently been obtained in [20] and [24]. In particular, it was shown in the former reference that near a point p of M, the two corresponding components $u_g = u_i^g$, $v_g = u_j^g$ $(i \neq j)$ that survive as $g_n \to +\infty$ should behave, to main order, in the following self-similar fashion:

$$u_g(x) \sim g^{-\frac{1}{4}} U\left(g^{\frac{1}{4}} \mathrm{dist}(x,M)\right), \quad v_g(x) \sim g^{-\frac{1}{4}} V\left(g^{\frac{1}{4}} \mathrm{dist}(x,M)\right), \tag{1.6}$$

where $\mathrm{dist}(\cdot, M)$ stands for the signed distance to M, while the one-dimensional profiles $U(t), V(t)$ depend only on the point p and satisfy

$$\begin{cases} U'' = UV^2 \\ V'' = VU^2 \end{cases} \tag{1.7}$$

in the entire real line. It was shown in [4, 5] that the above problem has just a 2-parameter family of positive solutions given by

$$\mu U(\mu t + \tau), \ \mu V(\mu t + \tau),$$

with scaling parameter $\mu > 0$ and translation $\tau \in (-\infty, +\infty)$, for some fixed solution pair (U, V) which satisfies the mirror reflection symmetry

$$U(-t) \equiv V(t), \tag{1.8}$$

and enjoys the following asymptotic behaviour at respective infinities:

$$U(t) \to 0 \text{ as } t \to -\infty; \quad U'(t) \to |\nabla u_\infty(p)| > 0 \text{ as } t \to +\infty.$$

Notice that the convergence in the previous limits is super-exponentially fast. In fact, it was observed in [1] that there is an asymptotic phase $k = k(p) > 0$ in the asymptotic behaviour of U at $+\infty$. Combining all the previous information, we deduce that, for $t > 0$ large enough,

$$U(t) = |\nabla u_\infty(p)|t + k + O(e^{-c_1 t^2}) \text{ and } V(t) = O(e^{-c_2 t^2}), \tag{1.9}$$

for some positive constants c_1 and c_2. The above relations can be differentiated and, via (1.8), provide the corresponding asymptotic behaviour as $t \to -\infty$.

One also expects that the behaviour of solutions for large g near Σ should be governed by an equivariant entire solution with polynomial growth of the PDE version of system (1.7), see [5, 19], which is usually called the inner (or blow-up) limit problem.

1.2. The problem with two-components. ¿From now on, we will consider the special case of problem (1.1) with $m = 2$, which (after a rescaling) we can write as

$$\begin{cases} -\Delta u + f(u) + guv^2 = 0 & \\ & \text{in } \Omega; \\ -\Delta v + h(v) + gvu^2 = 0 & \\ & \\ u = v = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.10}$$

for some smooth functions f and h such that $f(0) = h(0) = 0$ and Ω still a bounded, smooth N-dimensional domain.

We note that the reflection law (1.5) implies that the difference

$$w = u_\infty - v_\infty$$

is smooth across the regular part of the interface. In fact, it was shown in [9] that this difference is a classical solution of the following limit problem

$$\Delta w = f(w^+) - h(-w^-) \text{ in } \Omega; \ w = 0 \text{ on } \partial\Omega, \tag{1.11}$$

where one writes

$$w = w^+ + w^- \text{ with } w^+ \geq 0 \text{ and } w^- \leq 0.$$

It is worthwhile mentioning that in the special case where $f \equiv h$ is odd, the above limit problem reduces to

$$\Delta w = f(w) \text{ in } \Omega; \ w = 0 \text{ on } \partial\Omega. \tag{1.12}$$

1.3. The converse problem. So far we have discussed how one can reach the limit problem (1.11) (and also (1.7)) starting from an appropriate family of solutions to (1.10) for large g. It is also of interest whether one can go in the opposite direction, that is under which conditions do solutions of the limit problem (1.11) generate corresponding solutions of (1.10) for large values of g.

In [10], Dancer considered (1.10) for nonlinearities as in (1.3) with $g_1, g_2 > 0$ (with the obvious correspondence with (1.1)). It was shown by variational methods that, under appropriate restrictions on μ_1, μ_2, a certain type of nodal least energy solutions of (1.11) generate corresponding solutions with positive components to (1.10) for large g. On the other hand, the authors of [26] considered the case where $g_1 = g_2 < 0$ (say -1) and $\mu_1 = \mu_2 > 0$ (say 1) in a ball in two or three dimensions. In this case, it is well known that, for any integer $m \geq 1$, the (reduced) limit problem (1.12) admits a radial nodal solution w_m with exactly an m number of sign changes. Using variational methods, they were able to show that each w_m produces a corresponding radial solution of (1.10) with positive components that shadow respectively $(w_m)^+$ and $-(w_m)^-$ as $g \to +\infty$.

At this point let us make a small detour and discuss briefly the analogous elliptic system modeling two competing populations that arises in spatial ecology. In that context, the coupling terms in both equations of (1.10) are guv, while the nonlinearities f, h are usually of logistic type. Remarkably, uniformly bounded families of solutions to both systems share essentially the same regularity properties (with respect to large g), see [6]. In particular, they have the same (outer) limit problem (1.11). For the population problem, it was shown in [7] by means of a topological degree theoretic argument that non-degenerate (in the sense that the linearized operator does not have a kernel) nodal solutions w of (1.11) give corresponding solutions (u_g, v_g) with positive components for the system with large g. The key idea for proving this is to consider the difference $u - v$ and note that this leads to a system with only one singularly perturbed equation (a standard slow-fast system in the language of dynamical systems). Interestingly enough, this result was established without making use of the analogous blow-up limit problem to (1.7). In light of the aforementioned common features of the two systems, it is natural to expect that an analogous converse result should also hold for the condensate problem (1.10), see [11].

2. Main result. We show that an analogous converse result holds for the condensate problem (1.10), provided that we restrict to the radial setting and we impose some extra but milder non-degeneracy assumptions on the solution of the limit problem (1.11).

THEOREM 2.1. *Let Ω be an N-dimensional ball or annulus, $N \geq 1$, and let $f, h \in C^4[0, \infty)$ be such that $f(0) = h(0) = 0$. Suppose that w is a radial nodal solution of the limit problem (1.11) with one sign change, which is non-degenerate in the radial class in the sense that the associated linearization does not have a nontrivial radially symmetric element in its kernel. Moreover, assume that $-w^-$ and w^+ are also non-degenerate in the radial class as solutions of (1.11) in their respective supports. Then, if g is sufficiently large, there exists a radial solution (u_g, v_g) of (1.10) with positive components such that*

$$\|v_g + w^-\|_{L^\infty(\Omega)} \leq C g^{-\frac{1}{4}}, \quad \|u_g - w^+\|_{L^\infty(\Omega)} \leq C g^{-\frac{1}{4}},$$

where the constant $C > 0$ is independent of g.

If r_0 denotes the radius of the sphere where w vanishes, and $(r - r_0)w(r) > 0$ for

$r \neq r_0$, it holds

$$\begin{cases} u_g(r) = g^{-\frac{1}{4}} U\left(g^{\frac{1}{4}}(r-r_0)\right) + O\left(g^{-\frac{1}{2}} + (r-r_0)^2\right) \\ v_g(r) = g^{-\frac{1}{4}} V\left(g^{\frac{1}{4}}(r-r_0)\right) + O\left(g^{-\frac{1}{2}} + (r-r_0)^2\right) \end{cases}$$

for $|r-r_0| \leq (\ln g) g^{-\frac{1}{4}}$, as $g \to +\infty$, where the pair (U,V) is the unique solution of (1.7) satisfying (1.8) and (1.9) with $u_\infty = w^+$ and $|p| = r_0$.

As we will describe in more detail in the sequel, our proof relies on a perturbative method. We first combine the outer and inner problems, (1.11) and (1.7) respectively, to construct a sufficiently good approximate solution to (1.10) for large g that is valid in the whole domain. Then, we can capture a genuine solution nearby by a fixed point argument owing to appropriate invertibility properties of the associated linearized operator between carefully chosen weighted spaces.

We point out that the separate non-degeneracy assumptions on $-w^-$ and w^+ were not present in the previously mentioned result of [7] for the population system. As will become apparent shortly, the underline reason for imposing them is the presence of the positive asymptotic phase k in the asymptotic behaviour of the blow-up profile (recall (1.9)). We point out that there was no such phase present in the analogous blow-up limits for the population problem. Loosely speaking, the outer and inner approximate solutions, given to main order by $\pm w^\pm$ and (1.6) with $M = \{|x| = r_0\}$, respectively, do not have the phase $k > 0$ in common (in the intermediate zone where they must match). Therefore, we need to move the outer solutions towards the inner one by a regular perturbation to compensate for the gap caused by $k > 0$ (in principle, the inner solution should control the outer ones). To be able to do so, by means of the implicit function theorem, we need to impose these non-degeneracy assumptions on $-w^-$ and w^+. We remark that the non-degeneracy assumptions for $\pm w^\pm$ are much easier to verify in practice (see for instance [16]) in comparison to that for w which is a sign-changing solution (see [21]); see also Section 4 below.

We believe that an analogous result still holds when w changes sign an arbitrary number of times, provided one imposes further analogous non-degeneracy assumptions to take into account the interaction created by adjacent zeros of $w(r)$ for $1 \ll g < \infty$.

3. Sketch of the proof. In this section, we describe briefly the main steps in the proof of Theorem 2.1. For simplicity, we will do this in a one-dimensional setting where $\Omega = (a,b)$ and $r_0 = 0$. The general radial case can be treated in a completely analogous manner.

We write v_0 instead of $-w^-$, u_0 instead of w^+, and set

$$\psi_0 = -v_0'(0) = u_0'(0) > 0.$$

3.1. Construction of the approximate solution (u_{ap}, v_{ap}). Firstly, around the origin we consider a two-parameter family of first order inner approximate solutions of the form

$$u_{in}(x) = \mu g^{-\frac{1}{4}} U(t),\ v_{in}(x) = \mu g^{-\frac{1}{4}} V(t),\ \text{where } t = \mu g^{\frac{1}{4}}(x-\xi), \tag{3.1}$$

with $\mu > 0$ and $\xi \in (-\infty, \infty)$. The remainder left by this approximation in (1.10) is of order $|x| + g^{-\frac{1}{4}}$, therefore we will use it for $|x| \leq |\ln g| g^{-\frac{1}{4}}$ (keep in mind also the super-exponential rate of convergence in (1.9)).

In $(a,0)$ and $(0,b)$ we consider one-parameter family of outer approximate solutions of the form $\left(0, v_{\tilde{\delta}}\right)$ and $(u_\delta, 0)$, respectively, through the following boundary value problems:

$$\text{(3.2)} \qquad \begin{cases} v''_{\tilde{\delta}} = h(v_{\tilde{\delta}}), & x \in (a,0), \\ \\ v_{\tilde{\delta}}(a) = 0, \quad v_{\tilde{\delta}}(0) = \tilde{\delta}, \end{cases} \qquad \begin{cases} u''_\delta = f(u_\delta), & x \in (0,b), \\ \\ u_\delta(0) = \delta, \quad u_\delta(b) = 0, \end{cases}$$

for $0 \leq \tilde{\delta}, \delta \ll 1$. We point out that such $v_{\tilde{\delta}}, u_\delta$ exist and depend smoothly on $\tilde{\delta}, \delta \geq 0$ thanks to the implicit function theorem and the assumption that v_0 and u_0 are non-degenerate solutions of the above problems for $\tilde{\delta} = 0$ and $\delta = 0$, respectively. In fact, the following asymptotic expansion holds:

$$u_\delta = u_0 + \delta u_1 + \delta^2 u_2 + \delta^3 u_3 + O(\delta^4)$$

where the u_i for $i \geq 1$ are given as solutions of linear inhomogeneous problems (which are solvable thanks to the aforementioned non-degeneracy of u_0). In particular, we have

$$-u_1'' + f'(u_0)u_1 = 0, \ x \in (0,b); \ u_1(0) = 1, \ u_1(b) = 0.$$

Naturally, an analogous expansion holds also for $v_{\tilde{\delta}}$. The outer approximate solution, made up by $(0, v_{\tilde{\delta}})$ and $(u_\delta, 0)$, will be used for $|x| \geq |\ln g| g^{-\frac{1}{4}}$. In fact, it solves (1.10) exactly except from $x = 0$. As a first order outer approximate solution (u_{out}, v_{out}) we take the pairs

$$\text{(3.3)} \qquad \left(0, v_0 + \tilde{\delta}_1 v_1\right) \text{ and } (u_0 + \delta_1 u_1, 0)$$

in $\left(a, -|\ln g| g^{-\frac{1}{4}}\right)$ and $\left(|\ln g| g^{-\frac{1}{4}}, b\right)$, respectively, with $\tilde{\delta}_1, \delta_1$ free parameters.

The main effort is placed in adjusting conveniently the four free parameters $\mu, \xi, \delta_1, \tilde{\delta}_1$ so that the above first order inner and outer approximate solutions match in an appropriate intermediate zone, which we can take as $|\ln g| g^{-\frac{1}{4}} \leq |x| \leq 2|\ln g| g^{-\frac{1}{4}}$. On the one hand, from (3.1), by virtue of (1.9) with asymptotic slope $\psi_0 > 0$ and asymptotic phase $k > 0$, the first component of the first order inner approximate solution behaves essentially as a linear function of $t = \mu g^{\frac{1}{4}}(x - \xi) \gg 1$. On the other hand, we see from (3.3) that the corresponding component of the outer approximate solution has, to main order, a linear behaviour in x near $x = 0^+$. By comparing these (say equating the powers x^0 and x^1), we get two equations to be satisfied. We point out that powers of x^2 are not present in neither the first order outer or inner approximation. We stress that an analogous property propagates to higher order powers of x when matching higher order inner and outer approximate solutions, merely by equating the powers x^0 and x^1 at each step. Doing the same on the other side for the second components, gives two more equations. The resulting system of four equations and! four unknowns, after setting $\mu = 1 + \mu_1$, reads as follows:

$$\begin{cases} \delta_1 & = g^{-\frac{1}{4}} k - \xi \psi_0, \\ \delta_1 u_1'(0) & = 2\psi_0 \mu_1, \\ \tilde{\delta}_1 & = g^{-\frac{1}{4}} k + \xi \psi_0, \\ \tilde{\delta}_1 v_1'(0) & = -2\psi_0 \mu_1. \end{cases}$$

The above system has the following unique solution, provided that $v_1'(0) \neq u_1'(0)$:

$$\mu_1 = -\frac{g^{-\frac{1}{4}} k u_1' v_1'}{\psi_0(u_1' - v_1')},\ \xi = \frac{g^{-\frac{1}{4}} k(u_1' + v_1')}{\psi_0(u_1' - v_1')},\ \delta_1 = -\frac{2g^{-\frac{1}{4}} k v_1'}{(u_1' - v_1')},\ \tilde{\delta}_1 = \frac{2g^{-\frac{1}{4}} k u_1'}{(u_1' - v_1')},$$

where here u_1', v_1' are evaluated at zero. Observe that thanks to the non-degeneracy assumption on w, we always have $v_1'(0) \neq u_1'(0)$ (otherwise, the union of v_1 and u_1 would be an element of the kernel of the linearization of (1.11)).

To improve the remainder left by (3.1) in (1.10), we consider a more refined inner approximate solution of the form

$$u_{in}(x) = \mu g^{-\frac{1}{4}} U(t) + \varphi(t),\ v_{in}(x) = \mu g^{-\frac{1}{4}} V(t) + \tilde{\varphi}(t), \tag{3.4}$$

for fluctuations $\varphi, \tilde{\varphi}$ of higher order. We point out that we will not adjust further μ and ξ, analogous parameters will appear shortly. In order to choose corrections $\varphi, \tilde{\varphi}$ for a second order inner approximate solution, we have to try (3.4) in (1.10), and then take into account the matching with the corresponding second order outer approximate solution. The latter is comprised of

$$\left(0, v_0 + (\tilde{\delta}_1 + \tilde{\delta}_2)v_1 + (\tilde{\delta}_1 + \tilde{\delta}_2)^2 v_2\right),\ \left(u_0 + (\delta_1 + \delta_2)u_1 + (\delta_1 + \delta_2)^2 u_2, 0\right) \tag{3.5}$$

in $\left(a, -|\ln g| g^{-\frac{1}{4}}\right)$ and $\left(|\ln g| g^{-\frac{1}{4}}, b\right)$, respectively, with $\delta_1, \tilde{\delta}_1$ as above and $\delta_2, \tilde{\delta}_2$ are higher order corrections to be chosen.

At first sight it seems that, to main order, the inner corrections should satisfy the following inhomogeneous linear problem in $(-\infty, +\infty)$:

$$\begin{cases} -\varphi'' + V^2\varphi + 2UV\tilde{\varphi} = -\mu^{-1} g^{-3/4} f'(0) U, \\ -\tilde{\varphi}'' + U^2\tilde{\varphi} + 2UV\varphi = -\mu^{-1} g^{-3/4} h'(0) V. \end{cases} \tag{3.6}$$

We note that the linear operator in the left side is precisely the linearization of the blow-up problem (1.7) about (U, V). It is important to note that this operator includes in its kernel the pairs (U', V') and $(tU' + U, tV' + V)$ due to the translation and scaling invariance of (1.7). In fact, it was shown in [4] that the only bounded elements in the kernel are constant multiples of (U', V'). By setting

$$(\varphi, \tilde{\varphi}) = \mu^{-1} g^{-\frac{3}{4}} \left((Z, \tilde{Z}) + (\varphi_1, \tilde{\varphi}_1)\right),$$

where $Z, \tilde{Z}$ are fixed, smooth functions such that

$$\begin{cases} Z(t) = 0,\ t \leq -1, \quad Z(t) = f'(0)\left(k\frac{t^2}{2} + \psi_0 \frac{t^3}{6}\right),\ t \geq 1, \\ \tilde{Z}(t) = h'(0)\left(k\frac{t^2}{2} - \psi_0 \frac{t^3}{6}\right),\ t \leq -1, \quad \tilde{Z}(t) = 0,\ t \geq 1, \end{cases}$$

we can transform (3.6) to an equivalent problem for $(\varphi_1, \tilde{\varphi}_1)$ with the same linear operator on the left side but with righthand side that decays super-exponential fast as $t \to \pm\infty$ and is independent of g. By the linear theory developed in [1], the resulting problem has a solution such that, for any $M > 1$, it holds

$$\begin{aligned} \varphi_1(t) = a_+ t + b + O(e^{-Mt}),\ \tilde{\varphi}_1(t) = O(e^{-Mt}) \text{ as } t \to +\infty, \\ \varphi_1(t) = O(e^{Mt}),\ \tilde{\varphi}_1(t) = a_- t + b + O(e^{Mt}) \text{ as } t \to -\infty, \end{aligned}$$

for some constants $a_\pm, b$. Therefore, we seek corrections $(\varphi, \tilde{\varphi})$ in (3.4) in the form

$$(\varphi, \tilde{\varphi}) = \mu^{-1} g^{-\frac{3}{4}} \left((Z, \tilde{Z}) + (\varphi_1, \tilde{\varphi}_1) + A(U', V') + B(tU' + U, tV' + V) \right), \tag{3.7}$$

with A, B free parameters to be determined through the matching with the outer approximation in (3.5). As before, by looking at the powers x^0, x^1, the matching amounts to solving a 4×4 linear system for $A, B, \delta_2, \tilde{\delta}_2$ which is again possible thanks to the non-degeneracy condition on w. More precisely, we find that $A = O(g^{\frac{1}{4}}), B = O(1), \delta_2 = O(g^{-\frac{1}{2}}), \tilde{\delta}_2 = O(g^{-\frac{1}{2}})$. However, it turns out that $A = O(g^{\frac{1}{4}})$ causes the second order inner approximate solution to leave a remainder of the same order as the first order one. This suggests that there should be a quasi-second order inner approximate solution given by

$$(\psi, \tilde{\psi}) = \mu^{-1} g^{-\frac{3}{4}} \left(A_1(U', V') + B_1(tU' + U, tV' + V) \right)$$

as the main correction in (3.4) for some appropriate $A_1 = O(g^{\frac{1}{4}})$ and $B_1 = O(1)$. It turns out that a successful way to go about this issue is to determine at the same time (through the previous matching considerations) the above quasi-second order inner solution, the quasi-second order outer (3.5), the genuine second order inner solution that is given by (3.7), writing $A = A_2$, $B = B_2$, with $(\varphi_1, \tilde{\varphi}_1)$ satisfying the inhomogeneous problem (3.6) with the addition of some super-exponential decaying terms of the same order in the righthand side involving A_1, B_1, and the genuine second order outer solution

$$\left(0, \ \sum_{i=0}^{3} (\tilde{\delta}_1 + \tilde{\delta}_2 + \tilde{\delta}_3)^i v_i \right), \ \left(\sum_{i=0}^{3} (\delta_1 + \delta_2 + \delta_3)^i u_i, \ 0 \right) \tag{3.8}$$

where $\delta_3, \tilde{\delta}_3$ are higher order corrections. We are led to two 4×4 linear systems for $(A_1, B_1, \delta_2, \tilde{\delta}_2)$ and the corresponding $(A_2, B_2, \delta_3, \tilde{\delta}_3)$, that are again solvable, with the flexibility of rearranging conveniently their right hand sides so that we get solutions of the desired order in g.

Finally, we can smoothly patch the (genuine) second order outer and inner approximate solutions using cutoff functions in the intermediate zone, and get a smooth global approximate solution (u_{ap}, v_{ap}) that leaves a remainder in (1.10) of order $|\ln g|^4 g^{-\frac{1}{2}}$.

3.2. The fixed point argument. We can perturb the approximate solution to a genuine one by applying the contraction mapping theorem, based on the following a-priori estimates for the associated linearized operator, expanding on ideas from [1].

PROPOSITION 3.1. *Suppose that*

$$\mathcal{L} \begin{pmatrix} \phi \\ \psi \end{pmatrix} = \begin{pmatrix} F \\ H \end{pmatrix}, \ x \in (a, b); \quad \phi(a) = \phi(b) = 0, \ \psi(a) = \psi(b) = 0,$$

where $F, H \in C[a, b]$ and

$$\mathcal{L} \begin{pmatrix} \phi \\ \\ \psi \end{pmatrix} \equiv \begin{pmatrix} -\phi'' + f'(u_{ap})\phi + g v_{ap}^2 \phi + 2g u_{ap} v_{ap} \psi \\ \\ -\psi'' + h'(v_{ap})\psi + g u_{ap}^2 \psi + 2g u_{ap} v_{ap} \phi \end{pmatrix}.$$

Then, given $\gamma \in (0, 1), \rho > 0$, there exist $C, g_0 > 0$, independent of (F, H) and (ϕ, ψ), such that

$$\|(\phi, \psi)\|_1 \leq C g^{-\frac{1}{4}} \|(F, H)\|_2,$$

$$\|(\phi,\psi)\|_1 \le Cg^{-\frac{1}{4}}\|(F,H)\|_0 + Cg^{\rho-\frac{1}{2}}\|(F,H)\|_2,$$

where

$$\|(\Phi,\Psi)\|_i = \|w_i(x)\Phi\|_{L^\infty(a,b)} + \|w_i(-x)\Psi\|_{L^\infty(a,b)}, \quad i=0,1,2,$$

with

$$w_0(x) = \begin{cases} 1+|g^{\frac{1}{4}}x|^{1+\gamma}, & x\in[0,b), \\ 1, & x\in(a,0). \end{cases} \qquad w_1(x) = \begin{cases} 1, & x\in[0,b), \\ e^{g^{\frac{1}{4}}|x|}, & x\in(a,0), \end{cases}$$

$$w_2(x) = \begin{cases} 1+|g^{\frac{1}{4}}x|^{1+\gamma}, & x\in[0,b), \\ e^{g^{\frac{1}{4}}|x|}, & x\in(a,0), \end{cases}$$

provided that $g \ge g_0$.

4. Applications of the main result. Let us now give briefly some applications of Theorem 2.1. As it was already pointed out earlier, in the case $f \equiv h$ and f is odd the limit problem becomes (1.12). It is known that when $f(u) = \lambda u - u^{2p+1}$, $\lambda \ge 0$ and p is such that

$$1 < 2p+1 < \frac{N+2}{N-2} \text{ if } N \ge 3,\ p>0 \text{ if } N=2, \tag{4.1}$$

then a radial solution w to (1.12) is unique and non-degenerate in the radial class provided that

- w is positive, $\lambda \ne 0$ and Ω is an annulus or the exterior of a ball, see [12];
- w is positive, $\lambda = 0$ and Ω is a ball or an annulus, see [14];
- w is positive, $\lambda \ne 0$ and Ω is a ball, see [2];
- w is a nodal solution with two nodal regions, $\lambda = 0$, see [15].

We also refer to [17] for more general results concerning the function f. We point out that such solutions can be shown to exist by variational methods.

Thanks to these previous results, we see that our result applies in the case $f(u) = -u^{2p+1}$ with p as in (4.1), and Ω a ball or an annulus. In a related topic, let us point out that when Ω is the whole N-dimensional space, $N \ge 3$, and $f(u) = u - |u|^{p-1}u$ with $1 < p < \frac{N+2}{N-2}$ sufficiently close to $\frac{N+2}{N-2}$, Ao, Wei and Yao [3] constructed radial solutions with $k \ge 1$ nodes to (1.12) that tend to zero as $r \to \infty$. Moreover, they established that their solutions are unique and non-degenerate. Our theorem, with only minor modifications in the proof, can produce a corresponding solution to (1.10) for large g, starting from such a one-node solution.

Acknowledgments. The second author would like to thank Peter Szmolyan and Kristian Uldall Kristiansen for inviting him in their mini-symposium "Singular perturbations and singularities: theory and applications" in Equadiff 2017 and for some interesting discussions.

REFERENCES

[1] A. Aftalion, and C. Sourdis, *Interface layer of a two-component Bose-Einstein condensate*, Commun. Contemp. Math., 19 (2017), 1650052.

[2] A. Aftalion, and F. Pacella, *Uniqueness and nondegeneracy for some nonlinear elliptic problems in a ball*, J. Differential Equations, 195 (2003), pp. 380-397.
[3] W. Ao, J. , Wei, and W. Yao, *Uniqueness and nondegeneracy of sign-changing radial solutions to an almost critical elliptic problem*, Advances in Differential Equations, 21 (2016), pp. 1049–1084.
[4] H. Berestycki, T-C. Lin, J. Wei, and C. Zhao, *On phase-separation models: asymptotics and qualitative properties*, Arch. Ration. Mech. Anal., 208 (2013), pp. 163–200.
[5] H. Berestycki, S. Terracini, K. Wang, and J. Wei, *On entire solutions of an elliptic system modeling phase separations*, Adv. Math., 243 (2013), pp. 102–126.
[6] M. Conti, S. Terracini, and G. Verzini, *Asymptotic estimates for the spatial segregation of competitive systems*, Adv. Math., 195 (2005), pp. 524-560.
[7] E. N. Dancer, and Y. Du, *Competing species equations with diffusion, large interactions, and jumping nonlinearities*, J. Differential Equations, 114 (1994), pp. 434–475.
[8] E. N. Dancer, K. Wang, and Z. Zhang, *Uniform Hölder estimate for singularly perturbed parabolic systems of Bose–Einstein condensates and competing species*, J. Differential Equations, 251 (2011), pp. 2737–2769.
[9] E. N. Dancer, K. Wang, and Z. Zhang, *The limit equation for the Gross-Pitaevskii equations and S. Terracini's conjecture*, J. Functional Analysis, 262 (2012), pp. 1087–1131.
[10] E. N. Dancer, *On the converse problem for the Gross-Pitaevskii equations with a large parameter*, Discr. Cont. Dyn. Syst., 34 (2014), pp. 2481–2493.
[11] E. N. Dancer, *Slides*, https://math.umons.ac.be/anum/pde2015/documents/Dancer.pdf
[12] P. Felmer, S. Martinez and K. Tanaka, *Uniqueness of radially symmetric positive solutions for $-\Delta u + u = u^p$ in an annulus*, J. Differential Equations, 245 (2008), pp. 1198–1209.
[13] B. Noris, H. Tavares, S. Terracini, and G. Verzini, *Uniform Hölder bounds for nonlinear Schrödinger systems with strong competition*, Comm. Pure Appl. Math., 63 (2010), pp. 267–302.
[14] F. Pacella, *Uniqueness of positive solutions of semilinear elliptic equations and related eigenvalue problems*, Milan Journal of Mathematics, 73 (2005), pp. 221–236.
[15] E. Moreira dos Santos, and F. Pacella, *Morse index of radial nodal solutions of Hénon type equations in dimension two*, Communications in Contemporary Mathematics, 19 (2017), 1650042.
[16] N. Shioji, and K. Watanabe, *A generalized Pohožaev identity and uniqueness of positive radial solutions of $\Delta u + g(r)u + h(r)u^p = 0$*, J. Differential Equations, 255 (2013), pp. 4448–4475.
[17] N. Shioji, and K. Watanabe, *Uniqueness and nondegeneracy of positive radial solutions of $div(\rho\nabla u) + \rho(-gu + hu^p) = 0$*, Calc. Var. Partial Differential Equations, 55 (2016), 42pp.
[18] N. Soave, and A. Zilio, *Uniform bounds for strongly competing systems: The optimal Lipschitz case*, Arch. Ration. Mech. Anal., 218 (2015), pp. 647–697.
[19] N. Soave, and A. Zilio, *Multidimensional entire solutions for an elliptic system modelling phase separation*, Annalysis and PDE, 9 (2016), pp. 1019-1041.
[20] N. Soave, and A. Zilio, *On phase separation in systems of coupled elliptic equations: Asymptotic analysis and geometric aspects*, Ann. Inst. H. Poincaré Anal. Non Linéaire , 34 (2017), pp. 625–654.
[21] S. Tanaka, *Uniqueness of sign-changing radial solutions for $\Delta u - u + |u|^{p-1}u = 0$ in some ball and annulus*, J. Math. Anal. Appl., 439 (2016), pp. 154–170.
[22] H. Tavares, and S. Terracini, *Regularity of the nodal set of segregated critical configurations under a weak reflection law*, Calc. Var., 45 (2012), pp. 273–317.
[23] S. Zhang, and Z. Liu, *Singularities of the nodal set of segregated configurations*, Calc. Var., 54 (2015), pp. 2017–2037.
[24] K. Wang, *Uniform Lipschitz regularity of flat segregated interfaces in a singularly perturbed problem*, Calc. Var., (2017) 56:135.
[25] J. Wei, and T. Weth, *Asymptotic behaviour of solutions of planar elliptic systems with strong competition*, Nonlinearity, 21 (2008), pp. 305–317.
[26] J. Wei, and T. Weth, *Radial solutions and phase separation in a system of two coupled Schrödinger equations*, Arch. Ration. Mech. Anal., 190 (2008), pp. 83-106.

Published by Slovak University of Technology

SPEKTRUM STU Publishing

ISBN 978-80-227-4757-8