TOPICS IN APPLIED MATHEMATICS AND MATHEMATICAL PHYSICS

PROBLEME ACTUALE ÎN MATEMATICA APLICATĂ ŞI FIZICA MATEMATICĂ

Topics in Applied Mathematics and Mathematical Physics

Edited by

Cecil Pompiliu Grünfeld

Stelian Ion

Gabriela Marinoschi

EDITURA ACADEMIEI ROMÂNE
Bucureşti, 2008

Address: EDITURA ACADEMIEI ROMÂNE

Calea 13 Septembrie nr. 13, sector 5,

050711, Bucharest, Romania,

Tel.: 4021-318 81 46, 4021-318 81 06,

Fax: 4021-318 24 44,

E-mail: edacad@ear.ro,

Internet: http://www.ear.ro

Descrierea CIP a Bibliotecii Naţionale a României

Topics in applied mathematics and mathematical physics / ed.: Cecil Pompiliu Grünfeld, Stelian Ion, Gabriela Marinoschi. - Bucureşti: Editura Academiei Române, 2008

ISBN 978-973-27-1719-6

I. Grünfeld, Cecil Pompiliu (ed.)

II. Ion, Stelian (ed.)

III. Marinoschi, Gabriela (ed.)

51:53

This book was sponsored by MEdCT under: CEEX05-D11-06 and CEEX-D11-25 research contracts.

Editor: Dan-Florin DUMITRESCU, Monica STANCIU

Technical editing: Stelian ION

Computer editing: Stelian ION

Cover design: Nicoleta NEGRUŢ

Bun de tipar: 20.11.2008. Format 16/70×100.

C.Z. pentru biblioteci mari: $\begin{cases} 51(076) = 2 \\ 53(076) = 2 \end{cases}$

C.Z. pentru biblioteci mici: 51

Table of Contents

Topics in Applied Mathematics & Mathematical Physics

Preface

The increased interest in obtaining more effective mathematical tools for both fundamental and applied sciences has led in the past years to a strong interplay between various scientific domains, in particular between applied mathematics and mathematical physics.

The present monograph contains a collection of review papers on the state of the art and new results obtained in the research activity on several topics of applied mathematics and mathematical physics. The topics are of equal interest for several research groups involved in the scientific activities of Romania. The main reason is the common mathematical concepts, analytical and numerical techniques, which have imposed themselves as particulary useful in handling various problems related to the above topics.

The proposed surveys are written by experts who attained full scientific recognition by significant contributions to mathematics, applied mathematics and mathematical physics.

At the same time, this book is the result of their joint effort in common research activities along several fruitful years, involving in this respect, "Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, "Simion Stoilow" Institute of Mathematics of the Romanian Academy, "Horia Hulubei" National Institute of Physics, and Institute of Space Sciences, all from Bucharest.

The paper "Quasi-free quantum statistical models for tunnelling junction" by N. Angelescu and M. Bundaru deals with the description of the stationary states occurring when a nanoscopic quantic system is connected to thermal reservoirs having different temperatures and activities.

"An introduction to monotonicity methods for nonlinear kinetic equations" by Cecil Grünfeld is a survey upon the recent progress on the application of monotonicity methods (with respect to the order) to investigate the existence of solutions of various Boltzmann-like nonlinear kinetic equations. To motivate the topic, we first provide several examples of Boltzmann models for complex systems, with similar monotonicity properties, which present interest in applications. These are Smoluchowski's coagulation equation, Povzner-like models with dissipative collisions and reactive collisions, respectively, a Boltzmann model for several chemical species (with reactions), and a von Neumann-Boltzmann quantum model. The common properties of the

above models can be abstracted into a very general setting. One obtains a class of nonlinear evolution equations, formulated into an abstract Lebesgue space, for which one can state general criteria for the existence, uniqueness and positivity of global (in time) solutions. The proofs extend techniques that were initially developed in the more particular context of the space-homogeneous version of the classical Boltzmann equation. Finally we show how the abstract results can be applied to our examples of Boltzmann-like models.

The paper "Estimating the number of negative eigenvalues of a relativistic Hamiltonian with regular magnetic field" by Viorel Iftimie, Marius Măntoiu and Radu Purice is concerned with the proof of the analog of the Cwickel-Lieb-Rosenblum estimation for the number of negative eigenvalues of a relativistic Hamiltonian with magnetic field $B \in C^{\infty}_{pol}(\mathbb{R}^d)$ and an electric potential $V \in L^1_{loc}(\mathbb{R}^d)$. A direct consequence is a Lieb-Thirring inequality for the sum of powers of the absolute values of the negative eigenvalues.

The lecture "Approximate inertial manifolds for nonlinear parabolic problems and approximate solutions based upon these" by Anca Veronica Ion presents the notion of approximate inertial manifold of a semi-dynamical system generated by a nonlinear evolution PDE (more precisely, a semilinear parabolic equation), as it appeared in the literature of the last twenty years. The localization of the attractors in the space of phases was a first interesting application field of the a.i.m.s. Besides, a.i.m.s found very interesting applications in the construction of some approximate solutions (and consequently in the numerical integration) of the nonlinear evolution problems. These are contained in the so-called nonlinear Galerkin and postprocessed Galerkin methods.

The chapter "Diffusion processes. Physical models and numerical approximation" by Stelian Ion deals with the numerical approximation of a class of nonlinear diffusion processes that includes the unsaturated water flow through porous media and the fast diffusion. The approximation method consists in the discretization of space derivative operators using the finite volume scheme and keeping the continuum time differentiation. Consequently, the solution of the partial differential equations is approximated by the solution of a system of ordinary differential equations. A scheme to approximate the diffusion and convective term such that one can obtain a quasi-monotone ODE system is defined. Further, it is proved that there exists a discrete comparison principle, the solutions of the discrete model are bounded and the upper and lower bounds are independent of the mesh size of triangulation. To perform the time numerical integration a class of implicit backward

differentiation formulae with adaptive time step is used. Since the implicit schemes require a nonlinear solver a method that mixes Broyden method and an inexact Newton method is constructed. The performances of the new method are illustrated by some numerical results concerning the fast diffusion equation and water infiltration through a layered soil.

The paper "On a convergent numerical method for nonlinear Boltzmann-type models" by Dorin Marinescu deals with the extensions of approximation techniques of Nambu, Babovsky and Illner for the solutions of the classical Boltzmann equation to a nonlinear generalized Boltzmann-type system of equations solving nontrivial transport flows in dilute gas mixtures. First, one proves the global existence and uniqueness of solutions. Then a weak time-discretized version of equations for positive measures is provided. To obtain an algorithm, with small numerical effort (of order $N \log N$) stochastic methods are introduced. Finally a numerical approximation scheme, converging almost surely, in some sense, to the solutions of exact equations is provided.

The first part of the paper "Mathematical models of diffusion in nonhomogeneous porous media" by Gabriela Marinoschi introduces diffusive models of water flow in saturated-unsaturated media, characterized by a space variation of the porosity. Then the analysis focuses on a model with mixed boundary conditions involving a flux on a part of the boundary and a nonhomogeneous Dirichlet condition corresponding to a singular situation (i.e., the blowing up diffusion coefficient) on the other part of the domain boundary. From the mathematical point of view, the problem resides in the study of a degenerate nonlinear variational inequality which can be reduced to a multivalued inclusion by an appropriate change of the unknown function. Finally, existence, uniqueness and other properties of the solution are established.

The editors

Bucharest, July 2008

Topics in Applied Mathematics & Mathematical Physics

Quasi-free Quantum Statistical Models for Tunnelling Junction

N. Angelescu[1], *M. Bundaru*[2] *and R. Bundaru*[3]

Contents

[1]National Institute of Physics and Nuclear Engineering "H. Hulubei", P.O. Box MG-6, Bucharest, Romania, e-mail: nangel@theory.nipne.ro
[2]National Institute of Physics and Nuclear Engineering "H. Hulubei", P.O. Box MG-6, Bucharest, Romania, e-mail: bundaru@theory.nipne.ro
[3]Institute for Space Sciences, P.O. Box MG-23, Bucharest, Romania, e-mail: bundaru@venus.nipne.ro

1. Introduction

1.1. General frame

During the last decade considerable progress has been achieved in the statistical description of non-equilibrium thermodynamic processes. While previous work concentrated and provided a reasonable understanding of situations near thermal equilibrium, such as stability of equilibrium states (approach to equilibrium) or linear response, a consistent mathematical framework initiated by Ruelle [16], is now available for accounting for the installation, at large time, of a non-equilibrium stationary state (NESS) even when the initial state of the system is far from equilibrium (see [3] for a recent review). The typical physical situation which fits in this framework is that of several reservoirs, R_i; $i = 1, ..., r$, coupled to a finite quantum system, S (sample). One has to give account for the flow of energy and particles through the sample in the large time asymptotic regime.

The isolated sample S is a usual quantum system with Hilbert space $\mathcal{H}_S$, algebra of observables $\mathcal{A}_S$ equal to the algebra of all bounded operators on $\mathcal{H}_S$, and unitary dynamics generated by the Hamiltonian H_S. The Heisenberg picture of the evolution is the automorphism group on $\mathcal{A}_S$ defined as $\alpha_S^t(A) = \exp(\mathrm{i}tH_S)A\exp(-\mathrm{i}tH_S)$. We suppose that the sample is at time $t = 0$ in an arbitrary invariant state ω_S^0, i.e. the expectation of an observable $A \in \mathcal{A}_S$ is given by a density matrix: $\omega_S^0(A) = \mathrm{tr}(\rho_S A)$ and $[\rho_S, H_S] = 0$.

The description of the reservoirs R_i is somewhat more elaborated. A reservoir is an *infinite* quantum system, which, before the coupling to the sample is switched on, is in a certain equilibrium state. Its description in the initial state fits therefore in the well-established algebraic formalism of equilibrium quantum statistical mechanics [4]. One starts with reservoirs finitely extended in some regions Λ_i of space: the pure states are the unit vectors in a Hilbert space $\mathcal{H}_{i,\Lambda_i}$, the algebra of observables $\mathcal{A}_{i,\Lambda_i}$ consists of all bounded operators on $\mathcal{H}_{i,\Lambda_i}$ and the (Heisenberg) dynamics on $\mathcal{A}_{i,\Lambda_i}$ is generated by a self-adjoint Hamiltonian H_{i,Λ_i}, $\alpha_{i,\Lambda_i}^t(A) = \exp(\mathrm{i}tH_{i,\Lambda_i})A\exp(-\mathrm{i}tH_{i,\Lambda_i})$; at given inverse temperature β_i, the finite reservoir i has one equilibrium state $\omega_{i,\beta_i,\Lambda_i}(A) = \mathrm{tr}(A\rho_{i,\beta_i,\Lambda_i})$ given by the Gibbs ansatz for the density matrix $\rho_{i,\beta_i,\Lambda_i} = (1/Z_{i,\Lambda_i}(\beta_i))\exp(-\beta_i H_{i,\Lambda_i})$, where the statistical sum $Z_{i,\Lambda_i}(\beta_i)$ is a normalizing factor. The infinite reservoir is conceived as an idealization behaving like very large reservoirs, i.e., as a limit of the above structure: The algebra of observables $\mathcal{A}_i$ is the smallest C^*-algebra containing $\mathcal{A}_{i,\Lambda_i}$ for all finite regions Λ_i, the (strongly continuous) dynamics $\alpha_i^t(\cdot)$ on it is

the strong limit (provided it exists) of the automorphism groups $\alpha^t_{i,\Lambda_i}(\cdot)$, and the equilibrium state is a limit point ω_{i,β_i} of $\omega_{i,\beta_i,\Lambda_i}$ as Λ_i increases to the infinite region L_i occupied by the reservoir R_i. The infinite reservoirs in this sense can be represented as genuine quantum systems using the so-called Gelfand-Neumark-Segal (GNS) construction. The latter consists essentially in the following: a state ω on a C^*-algebra $\mathcal{A}$ defines a sesquilinear form on it by $\langle A, B\rangle = \omega(A^*B)$; after division by the ideal $\mathcal{I}$ of all $I \in \mathcal{A}$ such that $\omega(I^*I) = 0$, $\mathcal{A}/\mathcal{I}$ becomes a pre-Hilbert space, whose completion $\mathcal{H}_\omega$ is the representation space. The representation $\pi_\omega(X)$ of an element $X \in \mathcal{A}$ is the bounded operator which sends the vector $\hat{A}$ into the vector $\widehat{XA}$; thereby, $\hat{1} =: \Omega_\omega$ is a cyclic vector for this representation, and $\omega(A) = (\Omega_\omega, \pi_\omega(A)\Omega_\omega)$. If, moreover, the state ω is invariant under the automorphism group α_t (i.e. $\omega \circ \alpha_t = \omega$), then $\pi_\omega(\alpha_t(X)) = U_\omega(-t)\pi_\omega(X)U_\omega(t)$, where $U_\omega(t) = \exp(-\mathrm{i}tH_\omega)$ is a unitary group on $\mathcal{H}_\omega$. The generator H_ω of this group, named *thermal Hamiltonian*, has Ω_ω as an eigenvector with eigenvalue 0.

To simplify the notation, we no longer mention the reference states $\omega_i^0 = \omega_{i,\beta_i}$ of the reservoirs, and simply denote $\{\mathcal{H}_i, \pi_i(\cdot), \Omega_i, H_i\}$ the GNS description for the reservoir R_i corresponding to the equilibrium state ω_i^0, i.e., respectively, the Hilbert space, the representation of the observable algebra $\mathcal{A}_i$, the cyclic vector and the thermal Hamiltonian generating the unitary implementation of the dynamical automorphism group: $\pi_i(\alpha_i^t(A)) = \exp(\mathrm{i}tH_i)A\exp(-\mathrm{i}tH_i)$. Likewise, we denote $\{\mathcal{H}_S, \pi_S(\cdot), \Omega_S, H_S\}$ the GNS representation of the sample associated to the state ω_S^0 invariant for the group α_S^t.

The composite system $S + \sum R_i$ is in turn an infinite quantum system, which is to be constructed as above from a certain reference state. The algebra of observables is taken as a C^*-tensor product of the algebras $\mathcal{A}_i$ of the reservoirs and $\mathcal{A}_S$ of the sample:

$$\mathcal{A} = \mathcal{A}_S \otimes (\otimes_i \mathcal{A}_i), \tag{1.1}$$

and the reference state is taken as the product of the initial equilibrium states ω_i^0 of the reservoirs and the α_S^t-invariant state $\omega_S^0(\cdot) = (\Omega_S, \cdot\Omega_S)$ of the sample:

$$\omega^0 = \omega_S^0 \otimes ((\otimes_i \omega_i^0). \tag{1.2}$$

On the algebra $\mathcal{A}$ one has the uncoupled dynamics described by the automorphism group $\alpha^t = \alpha_S^t \otimes ((\otimes_i \alpha_i^t)$, which leaves invariant the state ω^0: $\omega^0(\alpha^t(A)) = \omega^0(A),\ A \in \mathcal{A}$.

At time $t = 0$, a coupling between reservoirs and the sample is switched on, meaning that the dynamics of the system at positive times is given by another

automorphism group of $\mathcal{A}$, τ^t. The evolved reference state will therefore change in time, and be, at time $t > 0$, the state for which the expectation of an observable equals the ω^0-expectation of the observable evolved at time t according to the new dynamics:

$$\omega^t(A) = \omega^0(\tau^t(A)) = \omega^0(\alpha^{-t} \cdot \tau^t(A)), \tag{1.3}$$

where the second equality comes from the α^t-invariance of ω^0. Suppose a stationary (τ^t-invariant) state is approached at large time. This can be expressed as the existence of the limit of $\omega^t(A)$ when $t \to +\infty$ for all $A \in \mathcal{A}$. The latter is ensured by the existence of the limits

$$\lim_{t \to +\infty} \alpha^{-t} \cdot \tau^t(A) = \Omega_+(A), \tag{1.4}$$

i.e. by the existence of the Möller endomorphisms of the two groups. In this way, the existence of (and the convergence to) a stationary state can be presented as a scattering problem for two automorphism groups on a C^*-algebra. As a rule, τ^t is constructed as a local perturbation of α^t via a strongly convergent Dyson series; more precisely, if $\lim_{t \to 0} \frac{1}{t}(\alpha^t(A) - A) = \delta_0(A)$ for A in a dense subalgebra $\mathcal{D} \subset \mathcal{A}$, one supposes that there exists $V \in \mathcal{A}$, such that $\delta_V(A) := \lim_{t \to 0} \frac{1}{t}(\tau^t(A) - A) = \delta_0(A) + \mathrm{i}[V, A]$ for $A \in \mathcal{D}$.

As a consequence of the choice (1.2), the composite system can be realized in the tensor product of Hilbert spaces $\mathcal{H} = \mathcal{H}_S \otimes ((\otimes_i \mathcal{H}_i)$, which carries the product representation of $\mathcal{A}$, so that $\pi(\mathcal{A})$ is the C^*-tensor product of operator algebras $\pi_S(\mathcal{A}_S) \otimes ((\otimes_i \pi_i(\mathcal{A}_i))$. Thereby, the independent (uncoupled) dynamics of the reservoirs and of the sample is implemented in $\mathcal{H}$ by the unitary group $U_0(t) = \exp(-\mathrm{i}tH_0) = \exp(-\mathrm{i}tH_S) \otimes ((\otimes_i \exp(-\mathrm{i}tH_i))$. The cyclic vector $\Omega = \Omega_S \otimes ((\otimes_i \Omega_i)$ is an eigenvector of H_0 with eigenvalue 0. Also, the locally perturbed dynamics is implemented by the unitary group $U(t) = \exp(-\mathrm{i}tH)$, where

$$H = H_0 + \pi(V). \tag{1.5}$$

In this way, the problem can be reformulated as a perturbation problem for selfadjoint operators on a Hilbert space in a setting depending on the chosen reference state.

Of course, the construction of the perturbed dynamics and the proof that the Möller endomorphisms exist are to be done for the models under consideration of reservoirs, samples and couplings between them. It happens that the program outlined before can accommodate a few reservoir models of physical interest, such as spin models or free particle models obeying Fermi

statistics, and samples with finite-dimensional $\mathcal{H}_S$. One of the most restrictive assumptions is the existence of the infinite-volume dynamical group of automorphisms α^t and its assumed strong continuity. A way out to a more permissible framework for the reservoirs, R_i, is to construct as above the reference states ω_i^0 as limit points of finite-volume Gibbs states and further work within the GNS representation associated to it. In particular, a weakly continuous infinite-volume dynamics may appear as a limit of the local dynamics $\alpha^t_{\Lambda_i}(\cdot)$ viewed as automorphisms of the *weak* closures of the operator algebras $\pi_i(\mathcal{A}_i)$ representing $\mathcal{A}_i$, i.e. of the von Neumann algebras $\pi_i(\mathcal{A}_i)''$. This allows to define a representation-dependent dynamics and self-adjoint thermal Hamiltonian. Hence, the steps leading to a scattering problem in a Hilbert space are to be performed. In particular, this is the case of free-boson reservoirs, see Sec. 4. below.

1.2. Quasi-free models

In the paper we shall consider instances of concrete realizations, within a class of very simple models, of the paradigm outlined above. Essentially, we suppose that:

1. The reservoirs are free quantum identical particle systems, obeying Fermi-Dirac or Bose-Einstein statistics.

2. The perturbed (coupled) dynamics is quasi-free.

In more detail, point 1 means the following: Before taking the thermodynamic limit, i.e. when the reservoir is confined to a finite region Λ, the appropriate Fock space, which bears the Fock representation of the canonical (anti)commutation relations, can be used, whereby the number of particles $N_\Lambda = \mathrm{d}\Gamma(1)$ and Hamiltonian $H_\Lambda = d\Gamma(h^0_\Lambda)$. According to the grand-canonical prescription, H_Λ is to be replaced by $H_\Lambda - \mu N_\Lambda$ in the Gibbs ansatz for the equilibrium density matrix, where the multiplier μ is adjusted to ensure given particle density in the reservoir. In the thermodynamic limit, the C^*-algebra of observables should "contain" the local operators, i.e. functions of $a^\sharp(f)$ with f having support in some finite region. It is therefore natural to take it as the canonical (anti)commutation relations algebra, $\mathrm{CAR}(\mathcal{D})$, respectively $\mathrm{CCR}(\mathcal{D})$, over a certain subspace of the space of reservoir's one-particle states, $\mathcal{D} \subset \mathcal{H}^{(1)}$, containing at least the functions with compact support. The equilibrium states of the reservoir, i.e. the limit states of the

finite-volume Gibbs states, are well-known (see e.g. [4]), and turn out to be *quasi-free* states (i.e. states in which there are no correlations of order higher than 2) over these C^*-algebras. $\mathcal{D}$ may be extended such that the limit states be defined on the corresponding C^*-algebra. In the Fermi case $\mathcal{D} = \mathcal{H}^{(1)}$. In the Bose case, however, due to the phenomenon of Bose-Einstein condensation, $\mathcal{D} \neq \mathcal{H}^{(1)}$; in the paper, in order to avoid the domain problems, we suppose also that the Bosons live on the lattice $\mathbb{Z}^d$, leaving the general case for another publication.

The point 2 means that the evolution automorphism of the C^*-algebra is given by a unitary evolution e^{-ith} in $\mathcal{H}^{(1)}$ which leaves $\mathcal{D}$ invariant: $\tau^t(a^\sharp(f)) = a^\sharp(\mathrm{e}^{ith} f)$. As a consequence, not only the initial (reference) state ω^0, but also all $\omega^t,\ t > 0$ and the stationary state are quasi-free. Thereby, the problem is reduced to a scattering problem for the one-particle Hamiltonians, which can be explicitly solved.

In this respect, the quasi-free models are trivial, in particular they allow no interaction between particles and thus restrict consideration to simple tunneling junctions, but they turn out to be a good laboratory for conjectures concerning various phenomena and providing instances of interesting physical behavior. In particular, the coupled dynamics no longer conserves the energy and number of particles in the reservoirs, implying that, in the stationary state, there exist persistent currents of energy and particles, depending on the parameters fixing the initial equilibria of the reservoirs, and also on the geometry of the sample and its coupling to them. In this way various formulae of transport theory can be obtained beyond the linear response regime.

1.3. Summary

There is an extensive literature on quasi-free quantum systems. This work started as an attempt to systematize their application to the problems of return to equilibrium and of approach to NESS in a more abstract, comprehensive frame, as outlined in the previous subsection. In the meantime, we became aware of two recent papers with the same purpose in the Fermi case [2], [12], so we limited to the more modest aim of giving a (hopefully more friendly) presentation of their general result, of indicating its extension to the Bose case and of providing a few examples of calculation for certain interesting physical quantities.

Section 2 is concerned with the spectral and scattering problems for the one-particle Hamiltonians, as the same analysis applies to both Fermi and Bose

statistics. In order to have as far as possible explicit expressions, we consider, as an application, in subsections 2.3. and 2.4. the case of two reservoirs, in which the particles live on two d-dimensional lattices, and those in the sample on a chain of $N \geq 0$ sites; thereby, the coupling is a simple tunneling involving one site of each reservoir.

Section 3 is devoted to the Fermi statistics case, which is simpler in many respects, in particular the C^*-framework is sufficient, as the infinite-volume dynamics is a strongly continuous group of automorphisms of $CAR(\mathcal{H}^{(1)})$. A comprehensive study of this case has been performed in [2], the results of which are briefly presented. We make explicit their result for the particular setting in Section 2.3. and point out a few peculiarities of the NESS, such as the resonant character of the transport and the plateau effect for the carrier density.

Section 4 is concerned with Bose reservoirs. This brings in several new phenomena and complications. First, at high density and low temperature, Bose condensation may appear, implying the spontaneous gauge-symmetry breaking, i.e. existence of several extremal equilibrium states labeled by a phase. Moreover, the infinite volume dynamics cannot be a strongly continuous group of the CCR algebra; fortunately, as quasi-free states are regular, it is continuous in the GNS representation corresponding to equilibrium states. The interesting question here is the dependence of the NESS on the particular mixtures of phases constituting the initial equilibria of the reservoirs. This may be viewed as a caricature of the Josephson tunneling of Cooper pairs between two superconductors. The approach to equilibrium in the presence of a condensate has been analyzed by Merkli [8]. The problem of approach to a NESS, left open there, was considered by us in [1], the result of which is presented in the present, slightly more general, setting.

2. Scattering for the one-particle Hamiltonians

This section is devoted to the spectral analysis of the one-particle Hamiltonian $h = h^0 + v$, where h^0 is the one-particle Hamiltonian of the decoupled system, i.e. the direct sum of the one-particle Hamiltonians $h_i\,(i = 1, ..., r)$, h_S of the isolated reservoirs and sample and v describes the tunneling between them. We make the following assumptions:

Assumption 2.1 *The one-particle Hilbert space is an orthogonal sum*

$$\mathcal{H}^{(1)} = \mathcal{H}_S^{(1)} \oplus \mathcal{H}_R^{(1)};\ \ \mathcal{H}_R^{(1)} = \oplus_{i=1}^r \mathcal{H}_i^{(1)},$$

with $\dim \mathcal{H}_S^{(1)} = N < \infty$. *Let* $J : \mathcal{H}_R^{(1)} \to \mathcal{H}^{(1)}$ *and* $I : \mathcal{H}_S^{(1)} \to \mathcal{H}^{(1)}$ *be the natural injections:*

$$Jf = 0 \oplus f \qquad If = f \oplus 0,$$

Assumption 2.2 *In the matrix representation associated to this decomposition, the unperturbed Hamiltonian* h_0 *is block-diagonal:*

$$h^0 = h_S \oplus h^0_{\rm ac}; \;\; h^0_{\rm ac} = \oplus_{i=1}^r h_i,$$

and the perturbation v *has the following structure: There exist maps* $\tau_i : \mathcal{H}_i^{(1)} \to \mathcal{H}_S^{(1)}$, *such that*

$$v = I\tau J^* + J\tau^* I^*,$$

where

$$\tau : \mathcal{H}_R^{(1)} \to \mathcal{H}_S^{(1)}, \;\; \tau(\oplus_{i=1}^r f_i) = \sum_{i=1}^r \tau_i f_i.$$

Assumption 2.3 h_i, $i = 1, ..., r$, *have absolutely continuous spectra equal to the bounded intervals* $I_i \subset \mathbb{R}$. *Thereby, we suppose that* $\bigcup_{i=1}^r \mathrm{Int}(I_i)$ *is an interval* $(e_{\min}, e_{\max})$. *We denote* $R_i(z) = (h_i - z)^{-1}$, $(z \in \mathbb{C} \setminus I_i)$ *and* $R^0_{\rm ac} = (h_{\rm ac} - z)^{-1} = \oplus_{i=1}^r R_i(z)$. *Let* p_i, π_i *denote the right, respectively left, support of* τ_i *(i.e. the orthogonal projections onto the subspaces* $\tau_i(\mathcal{H}_i^{(1)}) \subset \mathcal{H}_S^{(1)}$, *respectively* $\tau_i^*(\mathcal{H}_S^{(1)}) \subset \mathcal{H}_i^{(1)}$*). For* **all** $x \in I_i$, *the limits*

$$\lim_{\epsilon \searrow 0} \pi_i R_i(x + \mathrm{i}\epsilon)|_{\pi_i(\mathcal{H}_i^{(1)})}$$

exist as operators in the corresponding subspaces and are continuous functions of x; *thereby, for all interior points* x *of* I_i, ,

$$\lim_{\epsilon \searrow 0} \pi_i \Im R_i(x + \mathrm{i}\epsilon)|_{\pi_i(\mathcal{H}_i^{(1)})} > 0 \quad (i = 1, ..., r).$$

2.1. Resolvent and spectrum of the perturbed Hamiltonian

The spectral decomposition of $h = h^0 + v$ is based on finding a convenient representation of the resolvent operator $R(z) = (h - z)^{-1}$. We shall use a variant of the Feshbach method, taking advantage of the fact that v has finite range, what allows summing the perturbation series in closed form.

We have to solve for f_S, f_i, $i = 1, ..., r$, the system of equations

$$\left\{ \begin{array}{lll} (h_i - z)f_i & + \qquad \tau_i^* f_S & = g_i \;\; (i = 1, ...r) \\ \sum_{i=1}^r \tau_i f_i & + (h_S - z) f_S & = g_S, \end{array} \right. \tag{2.1}$$

where $g = g_S \oplus (\oplus_{i=1}^r g_i) \in \mathcal{H}^{(1)}$ is arbitrary.

If $z \in \mathbb{C} \setminus [e_{\min}, e_{\max}]$, the first line in equation (2.1) provide f_i in terms of f_S:

$$f_i = R_i(z)(g_i - \tau_i^* f_S), \tag{2.2}$$

and the second line becomes

$$(h_{\text{eff}}(z) - z) f_S = Q(z) g, \tag{2.3}$$

where $h_{\text{eff}}(z) : \mathcal{H}_S^{(1)} \to \mathcal{H}_S^{(1)}$ and $Q(z) : \mathcal{H}^{(1)} \to \mathcal{H}_S^{(1)}$ are defined by:

$$\begin{array}{lcl} h_{\text{eff}}(z) & = & h_S - \sum_{i=1}^r \tau_i R_i(z) \tau_i^* \quad = \quad h_S - \tau R_{\text{ac}}^0(z) \tau^*, \\ Q(z) & = & I^* - \tau R_{\text{ac}}^0(z) J^*. \end{array} \tag{2.4}$$

Whenever $h_{\text{eff}}(z) - z$ is invertible, we denote $R_{\text{eff}}(z) = (h_{\text{eff}}(z) - z)^{-1}$, so that Eq. (2.3) has the unique solution

$$f_S = R_{\text{eff}}(z) Q(z) g, \tag{2.5}$$

With f_S given by Eq. (2.5) and f_i given in terms of it by Eq. (2.2), $f = f_S \oplus (\oplus_{i=1}^r f_i) = Q(\bar{z})^* f_S$ provides the solution to the system (2.1). Therefore, remarking that $\cup_{i=1}^r I_i \subset \sigma(h)$ (by the invariance of the essential spectrum under compact perturbations), the following characterization has been proved:

LEMMA **2.1** *The resolvent set of h is*

$$\rho(h) = \{ z \in \mathbb{C} \setminus [e_{\min}, e_{\max}]; \ \ker (h_{\text{eff}}(z) - z) = \{0\} \}.$$

For all $z \in \rho(h)$,

$$R(z) = J R_{\text{ac}}^0(z) J^* + Q(\bar{z})^* R_{\text{eff}}(z) Q(z). \tag{2.6}$$

The Kato-Rosenblum scattering theory [15] ensures the existence and completeness of the wave operators $W_\pm : \mathcal{H}_R^{(1)} \to \mathcal{H}^{(1)}$ for the unitary groups $\mathrm{e}^{-\mathrm{i}th}$, $\mathrm{e}^{-\mathrm{i}th_{\text{ac}}^0}$, i.e. the existence of the strong limits:

$$W_\pm := (\mathrm{s}) \lim_{t \to \pm\infty} \mathrm{e}^{\mathrm{i}th} J \mathrm{e}^{-\mathrm{i}th_{\text{ac}}^0}. \tag{2.7}$$

Hence,

LEMMA **2.2** *h has absolutely continuous spectrum $\sigma_{\rm ac}(h) = [e_{\min}, e_{\max}]$ and no singular continuous spectrum. The absolutely continuous part $h_{\rm ac}$ of h, i.e. h restricted $\mathcal{H}^{(1)}_{\rm ac}(h) = W_{\pm}(\mathcal{H}^{(1)}_R)$, is unitarily equivalent to $h^0_{\rm ac}$ via the intertwining relations $h_{\rm ac}W_{\pm} = W_{\pm}h^0_{\rm ac}$.*

Finally, we determine the point spectrum of h, $\sigma_{\rm p}(h)$.

Let $z \in \sigma_{\rm p}(h)$, and $f = f_S \oplus (\oplus_{i=1}^r f_i) \neq 0$ be an eigenvector of h with eigenvalue z. Then f is a solution of Eq. (2.1) for $g = 0$.

If, thereby, $\tau_i^* f_S = 0$ for all $i = 1, ..., r$, then $(h_i - z)f_i = 0$, $\forall i$, hence $f_i = 0$, because h_i have no point spectrum. If so, the second line in (2.1) shows that $z \in \sigma_{\rm p}(h_S)$ and that $f_S \in \ker \tau_i^*$ is a corresponding eigenvector. Conversely, if $f_S \in \cap_i \ker \tau_i^*$ is an eigenvector of h_S, then $f_S \oplus 0$ is an eigenvector of h with the same eigenvalue.

Suppose next that $\tau_i^* f_S \neq 0$ for at least one i. If $z \notin [e_{\min}, e_{\max}]$, Eq. (2.2), which expresses f_i in terms of f_S, and Eq. (2.3) show that $f_S \neq 0$ is an eigenvector of $h_{\rm eff}(z)$ with eigenvalue z. Conversely, if $\ker(h_{\rm eff}(z) - z) \ni f_S \neq 0$, then $z \in \sigma_{\rm p}(h)$ and $Q(\bar{z})^* f_S$ is an eigenvector of h with eigenvalue z (in particular, we have that $\Im z = 0$). Let us consider the family of self-adjoint operators $\{h_{\rm eff}(x);\ z = x \in (-\infty, e_{\min})\}$ and let $\lambda_1(x) \leq ... \leq \lambda_N(x)$ be the eigenvalues of $h_{\rm eff}(x)$ and $\psi(x)^{(1)}_S, ..., \psi(x)^{(N)}_S$ the corresponding eigenvectors. As remarked before, $x \in \sigma_{\rm p}(h)$ if, and only if, $x = \lambda_k(x)$ for some $k = 1, ..., N$. As $h_{\rm eff}(x)$ is a decreasing operator-valued function of x in the considered interval, all its eigenvalues $\lambda_k(x)$ are decreasing functions, hence, the equation $x = \lambda_k(x)$ has a simple solution $x = e_k^-$ if, and only if, $\lim_{x \nearrow e_{\min}} \lambda_k(x) < e_{\min}$. Then, every eigenvector of $h_{\rm eff}(e_k^-)$ with eigenvalue e_k^- can be completed to an eigenvector of h with this eigenvalue. Likewise, on $(e_{\max}, \infty)$ the equation $x = \lambda_k(x)$ has a solution e_k^+ if, and only if, $\lim_{x \searrow e_{\max}} \lambda_k(x) > e_{\max}$, implying $e_k^+ \in \sigma_{\rm p}(h)$.

Next, let $f_S \oplus f$ be an eigenvector of h corresponding to x in $(e_{\min}, e_{\max})$ and such that $\tau_i^* f_S \neq 0$ for some $i = 1, ..., r$. Let $z = x + {\rm i}y$, with $\Im z = y > 0$. We have, by the first line of equations (2.1), $f_k = R_k(x + {\rm i}y)(h_k - x - {\rm i}y)f_k = -R_k(x + {\rm i}y)\tau_k^* f_S - {\rm i}yR_k(x + {\rm i}y)f_k$, which, plugged into the second equation, implies, in particular, that

$$\begin{aligned}(f_S, (h_{\rm eff}(x + {\rm i}y) - x)f_S) &= {\rm i}y \sum_{k=1}^r (\tau_k^* f_S, R_k(x + {\rm i}y)f_k) \\ &= {\rm i}y \sum_{k=1}^r (\|f_k\|^2 - {\rm i}y(f_k, R_k(x + {\rm i}y)f_k)).\end{aligned}$$

Equating the imaginary parts of this equality, letting $y \searrow 0$ and using

$\|R_k(x+\mathrm{i}y)\| = 1/y$, we arrive at

$$\Im(f_S, \tau_k R_k(x+\mathrm{i}0)\tau_k^* f_S) = 0, \ \forall k,$$

which contradicts assumption 2.3.

In summary:

LEMMA **2.3** *The point spectrum of h in $\mathbb{R} \setminus \{e_{\min}, e_{\max}\}$ consists, besides the possible eigenvalues of h_S possessing eigenvectors $f_S \in \bigcap_{i=1}^{r} \ker \tau_i^*$, of the solutions $e_k^- \in (-\infty, e_{\min})$ and $e_k^+ \in (e_{\max}, \infty)$ of the equations $\lambda_k(x) = x$. The latter exist if, and only if, $\lambda_k(e_{\min} - 0) < e_{\min}$ and $\lambda_k(e_{\max} + 0) > e_{\max}$, respectively.*

The values $e_{\min}$ or $e_{\max}$ may be eigenvalues of h, either if they are eigenvalues of h_S with eigenvector $f_S \in \bigcap_{i=1}^{r} \ker \tau_i^*$, or if $\lambda_k(x) = x$ and the corresponding eigenvector $\psi(x)^{(k)}$ fulfills $\lim_{x' \to x} \|R_i(x')\tau_i^* \psi(x')^{(k)}\| < \infty$, $\forall i$. The latter condition, being dependent on the structure of h^0 and τ_i, is to be checked for each concrete model.

2.2. Wave operators and scattering matrix

In this subsection we derive the expressions of the wave operators and S-matrix using the formalism of stationary scattering theory [15], [17]. Namely, with the spectral representation of the unitary groups $\mathrm{e}^{-\mathrm{i}th} = \int \mathrm{e}^{-\mathrm{i}tx}\mathrm{d}E(x)$, $\mathrm{e}^{-\mathrm{i}th_i} = \int \mathrm{e}^{-\mathrm{i}tx}\mathrm{d}E_i(x)$, we can express the wave operators in terms of the resolvent $R(z)$ of h. We have

$$\begin{aligned} W_+ &= (\mathrm{s}) \lim_{\epsilon \searrow 0} \epsilon \int_0^\infty \mathrm{e}^{-t\epsilon} \exp(\mathrm{i}th) J \exp(-\mathrm{i}th^0)\mathrm{d}t \\ &= (\mathrm{s}) \lim_{\epsilon \searrow 0} \epsilon \int \mathrm{d}E(x') \int J \mathrm{d}E_{\mathrm{ac}}^0(x) \int_0^\infty \mathrm{d}t \mathrm{e}^{\mathrm{i}t(x'-x+\mathrm{i}\epsilon)} \\ &= (\mathrm{s}) \lim_{\epsilon \searrow 0} (\mathrm{i}\epsilon) \int R(x - \mathrm{i}\epsilon) J \mathrm{d}E_{\mathrm{ac}}^0(x). \end{aligned} \tag{2.8}$$

where we denoted $E_{\mathrm{ac}}^0(x) = \oplus_{i=1}^r E_i(x)$. Similar calculations are valid for W_-. Using Eq. (2.6) for $R(z)$, taking into account that $\mp\mathrm{i}\epsilon R_{\mathrm{ac}}^0(x \pm \mathrm{i}\epsilon)\mathrm{d}E_{\mathrm{ac}}^0(x) = \mathrm{d}E_{\mathrm{ac}}^0(x)$ and Assumption 2.2, the following representation is obtained:

$$W_\pm = J - (\mathrm{s}) \lim_{\epsilon \searrow 0} \int Q(x \pm \mathrm{i}\epsilon)^* R_{\mathrm{eff}}(x \mp \mathrm{i}\epsilon) \tau \mathrm{d}E_{\mathrm{ac}}^0(x). \tag{2.9}$$

Also,

$$W_{\pm}^{*} = J^{*} - (\mathrm{s}) \lim_{\epsilon' \searrow 0} \int \mathrm{d}E_{\mathrm{ac}}^{0}(x') \tau^{*} R_{\mathrm{eff}}(x' \pm \mathrm{i}\epsilon') Q(x' \pm \mathrm{i}\epsilon'). \tag{2.10}$$

Eqs. (2.9), (2.10) give for the S-matrix:

$$\begin{aligned} S = W_{+}^{*} W_{-} = 1 \quad & -J^{*} \textstyle\int Q(x - \mathrm{i}0)^{*} R_{\mathrm{eff}}(x + \mathrm{i}0) \tau \mathrm{d}E_{\mathrm{ac}}^{0}(x) \\ & - \textstyle\int \mathrm{d}E_{\mathrm{ac}}^{0}(x') \tau^{*} R_{\mathrm{eff}}(x' + \mathrm{i}0) Q(x' + \mathrm{i}0) J \\ & + \lim_{\epsilon' \searrow 0} \{ \lim_{\epsilon \searrow 0} \textstyle\int \mathrm{d}E_{\mathrm{ac}}^{0}(x') \tau^{*} R_{\mathrm{eff}}(x' + \mathrm{i}\epsilon') Q(x' + \mathrm{i}\epsilon') \\ & \qquad \times \textstyle\int Q(x - \mathrm{i}\epsilon)^{*} R_{\mathrm{eff}}(x + \mathrm{i}\epsilon) \tau \mathrm{d}E_{\mathrm{ac}}^{0}(x) \}. \end{aligned} \tag{2.11}$$

We calculate the last term using the resolvent equation, which implies

$$\begin{aligned} & Q(x' + \mathrm{i}\epsilon') Q(x - \mathrm{i}\epsilon)^{*} = 1 + \tau R_{\mathrm{ac}}^{0}(x' + \mathrm{i}\epsilon') R_{\mathrm{ac}}^{0}(x + \mathrm{i}\epsilon) \tau^{*} \\ & = 1 + (x' - x + \mathrm{i}(\epsilon' - \epsilon))^{-1} \tau [R_{\mathrm{ac}}^{0}(x' + \mathrm{i}\epsilon') - R_{\mathrm{ac}}^{0}(x + \mathrm{i}\epsilon)] \tau^{*} \\ & = (x' - x + \mathrm{i}(\epsilon' - \epsilon))^{-1} [(h_{\mathrm{eff}}(x + \mathrm{i}\epsilon) - x - \mathrm{i}\epsilon) - (h_{\mathrm{eff}}(x' + \mathrm{i}\epsilon') - x' - \mathrm{i}\epsilon')]. \end{aligned}$$

Each term of the latter expression, when plugged into Eq. (2.11), is sandwiched between R_{eff}, what, after making the obvious simplification, allows one of the integrals to be performed (e.g. $\int \mathrm{d}E_{\mathrm{ac}}^{0}(x')(x' - x + \mathrm{i}(\epsilon' - \epsilon))^{-1} \tau^{*} = R_{\mathrm{ac}}^{0}(x - \mathrm{i}(\epsilon' - \epsilon)) \tau^{*} = J^{*} Q(x - \mathrm{i}(\epsilon - \epsilon'))^{*}$). Therefore, after taking the iterated limit, the last term of Eq. (2.11) equals

$$\int J^{*} Q(x + \mathrm{i}0)^{*} R_{\mathrm{eff}}(x + \mathrm{i}0) \tau \mathrm{d}E_{\mathrm{ac}}^{0}(x) + \int \mathrm{d}E_{\mathrm{ac}}^{0}(x') \tau^{*} R_{\mathrm{eff}}(x' + \mathrm{i}0) Q(x' + \mathrm{i}0) J.$$

As $Q(z)J = -\tau R_{\mathrm{ac}}^{0}(z)$, one obtains finally

$$S = 1 + 2\mathrm{i} \int \Im(R_{\mathrm{ac}}^{0}(x + \mathrm{i}0)) \tau^{*} R_{\mathrm{eff}}(x + \mathrm{i}0) \tau \mathrm{d}E_{\mathrm{ac}}^{0}(x). \tag{2.12}$$

REMARK **2.1** *It is sometimes useful to represent the Hilbert space $\mathcal{H}_{\mathrm{ac}}(h^{0})$ as a direct integral over energy of Hilbert "eigenspaces" $\mathcal{K}_{x}$, i.e. there exists a unitary $U : \mathcal{H}_{\mathrm{ac}}(h^{0}) \to \int_{[e_{\min}, e_{\max}]}^{\oplus} \mathcal{K}_{y} \mathrm{d}y =: \mathcal{K}$, such that $U E_{\mathrm{ac}}^{0}(\Lambda) U^{*} = \chi_{\Lambda}(\cdot)$ (the operator of multiplication with the indicator of the measurable set Λ). It is clear that, for $\psi(\cdot) \in \int_{[e_{\min}, e_{\max}]}^{\oplus} \mathcal{K}_{y} \mathrm{d}y$, $(U R^{0}(z) U^{*} \psi)(y) = (y - z)^{-1} \psi(y)$. Also, $\tau U^{*} \psi = \int_{[e_{\min}, e_{\max}]} \tau_{y}(\psi(y)) \mathrm{d}y$, where $\tau_{y} : \mathcal{K}_{y} \to \mathcal{H}_{S}^{(1)}$. Eq. (2.12) shows that, in this representation, the S-matrix is diagonal, i.e. $USU^{*} = \int_{[e_{\min}, e_{\max}]}^{\oplus} S_{x} \mathrm{d}x$, where $S_{x} : \mathcal{K}_{x} \to \mathcal{K}_{x}$ equals*

$$S_{x} = 1 + 2\pi\mathrm{i} \tau_{x}^{*} R_{\mathrm{eff}}(x + \mathrm{i}0) \tau_{x} =: 1 + T_{x}. \tag{2.13}$$

T_x is called the on-shell T-matrix.

Calculating, for $f \in \mathcal{H}^{(1)}_{\rm ac}$, separately the component $I^*W_\pm f \in \mathcal{H}^{(1)}_S$ and $J^*W_\pm f \in \mathcal{H}^{(1)}_{\rm ac}$ of Eq. (2.9), one obtains

$$\begin{aligned} I^*W_\pm f &= -\int R_{\rm eff}(x \mp {\rm i}0)\tau_x(Uf)(x){\rm d}x, \\ \left[UJ^*W_\pm f\right](x) &= (Uf)(x)+ \\ &\quad + \int \tfrac{1}{x-x'\mp {\rm i}0}\tau_x^* R_{\rm eff}(x' \mp {\rm i}0)\tau_{x'}(Uf)(x'){\rm d}x'. \end{aligned} \tag{2.14}$$

Also, the action of $W^*_\pm$ on $f \in \mathcal{H}^{(1)}$ is given by

$$\begin{aligned} (UW^*_\pm f)(x) = {} & (UJ^*f)(x)- \\ & - \int \tfrac{1}{x-x'\pm {\rm i}0}\tau_x^* R_{\rm eff}(x \pm {\rm i}0)\tau_{x'}(UJ^*f)(x'){\rm d}x'- \\ & -\tau_x^* R_{\rm eff}(x \pm {\rm i}0)I^*f. \end{aligned} \tag{2.15}$$

2.3. An example: two half-infinite lattice reservoirs coupled by a wire

In this subsection we describe, as an illustration of the more general setting of the model, a particular geometry and dynamics: the system consisting of two particle reservoirs, R_1, R_2, connected by a one-dimensional wire, S.

The reservoirs, R_i, $i = 1, 2$, are taken as infinitely extended lattice quantum gases. The particles in the reservoirs live, respectively, on the two (left, respectively, right) half-infinite lattices,

$$L_i = \mathbb{Z}^d_i = \left\{ r = (r', r^d);\ r' \in \mathbb{Z}^{d-1},\ (-1)^i r^d = 1, 2, ... \right\}. \tag{2.16}$$

The Hilbert space of one-particle states in R_i is therefore

$$\mathcal{H}^{(1)}_i = l_2(L_i) = \left\{ f = (f_r)_{r\in L_i};\ \|f\|^2 = \sum_{r\in L_i} |f_r|^2 < \infty \right\}. \tag{2.17}$$

The kinetic energy operator of one particle in R_i is $1/2$ times the lattice Laplace operator with free boundary conditions, i.e.

$$(h_i f)_r = df_r - \frac{1}{2}\sum_{q\in L_i, |q-r|=1} f_q. \tag{2.18}$$

A complete set of generalized eigenvectors of h_i are $\psi^i(k) \in l_\infty(L_i),\ k \in \mathbb{T}^d_i$, where the index sets $\mathbb{T}^d_i = \{k = (k', k^d);\ k' \in [0, 2\pi)^{d-1}, k^d \in (0, \pi)\}$ are

identical (the subscript i has the only role to make the difference between the two reservoirs, e.g. by $\mathbb{T}_1^d \cup \mathbb{T}_2^d$ we mean the disjoint union of two copies this set), and

$$\psi^i(k)_r = 2(2\pi)^{-d/2} \exp(\mathrm{i}k'r') \sin(k^d|r^d|). \tag{2.19}$$

$\psi^i(k)$ corresponds to the generalized eigenvalue

$$\omega_i(k) = 2\sum_{\alpha=1}^{d} \sin^2(k^\alpha/2). \tag{2.20}$$

Again, though the two dispersion laws (2.20) are identical, we keep the label i to mark the reservoir they correspond to. Therefore the spectra of h_i are absolutely continuous and coincide with the intervals $I_1, I_2 \subset \mathbb{R}$ (both equal to $[0, 2d]$). In fact, we define the unitary operators $u_i : \mathcal{H}_i^{(1)} \to L_2(\mathbb{T}_i^d)$ by

$$u_i f = (\psi^i(\cdot), f); \tag{2.21}$$

then, $u_i h_i u_i^*$ is the operator of multiplication with the function $\omega_i(k)$ on $L_2(\mathbb{T}_i^d)$.

The sample S, providing our model of a nanowire, is a free quantum gas in which particles live on the finite set of sites $\{1, 2, ..., N\}$. The states with one particle are vectors $f = (f_1, ..., f_N) \in \mathcal{H}_S^{(1)} = l_2(\{1, 2, ..., N\}) \equiv \mathbb{C}^N$ and their evolution is controlled by the Hamiltonian

$$(h_S f)_i = (1 + e_{\mathrm{g}})f_i - 1/2(f_{i-1} + f_{i+1}),\ i = 1, ..., N\ \ (f_0 = f_{N+1} = 0), \tag{2.22}$$

where the parameter e_{g} plays the role of an adjustable gate potential. The eigenvalues of h_S are $\varepsilon_m = e_{\mathrm{g}} + 2\sin^2(q_m/2); m = 1, ..., N$, where $q_m = m\pi/(N+1)$, with eigenvectors $\psi^{(m)}$:

$$\psi_i^{(m)} = \sqrt{\frac{2}{N+1}} \sin(q_m i). \tag{2.23}$$

The one-particle Hilbert space for the entire system, $S + R_1 + R_2$ is

$$\mathcal{H}^{(1)} = \mathcal{H}_S^{(1)} \oplus \mathcal{H}_1^{(1)} \oplus \mathcal{H}_2^{(1)} = l_2(L), \text{ where } L = \{1, 2, ..., N\} \cup L_1 \cup L_2. \tag{2.24}$$

The evolution of the one-particle states for the uncoupled system is given by the one-particle Hamiltonian

$$h^0 = h_S \oplus h_1 \oplus h_2 \tag{2.25}$$

At $t = 0$, tunneling junctions are turned on between the reservoirs and the ends of the wire through the pairs of sites $(\alpha_1 = (0', -1), \{1\})$ and $(\alpha_2 = (0', 1), \{N\})$, $N > 0$. On $\mathcal{H}^{(1)}$, this is given by the one-particle operator v defined by the matrix

$$v_{r,s} = \begin{cases} t, & \text{if either } \{r,s\} = \{\alpha_1, 1\} \text{ or } \{\alpha_2, N\} \\ 0, & \text{otherwise,} \end{cases} \tag{2.26}$$

Thus, the evolution of the one-particle states in the coupled system is generated by the Hamiltonian:

$$h = h^0 + v. \tag{2.27}$$

PROPOSITION **2.1** *The model defined above fulfills the assumptions* 2.1–2.3. *Thereby, h has no eigenvalue embedded in* $(0, 2d)$.

Proof. Assumptions 2.1 and 2.2 are obvious, with $r = 2$ and τ_1, τ_2 having all matrix elements equal to 0, but for $(\tau_1)_{1,\alpha_1} = (\tau_2)_{N,\alpha_2} = t$. We have that

$$(\tau_1 R_1(z) \tau_1^*)_{i,j} = t^2 \delta_{i,1} \delta_{j,1} g(z), \tag{2.28}$$

where

$$\begin{aligned} g(z) = & \ 4(2\pi)^{-d} \textstyle\int_{\mathbb{T}^d} \sin^2(k^d)(\omega_1(k) - z)^{-1} \mathrm{d}k \\ = & \ 4(2\pi)^{-d} \int_0^{2d} (y - z)^{-1} \mathrm{d}y \int\limits_{\mathbb{T}^d(y)} \sin^2(k^d) \mathrm{d}\mu_y(k), \end{aligned} \tag{2.29}$$

where $\mathrm{d}\mu_y(k) = |\nabla\omega(k)|^{-1} \mathrm{d}\sigma_y(k)$ is the Gelfand-Leray measure on the level set $\mathbb{T}^d(y) = \{k \in \mathbb{T}^d;\, \omega(k) = y\}$ (where $\mathrm{d}\sigma_y(k)$ is the area measure on this surface). Using the Sokhotski formula $(x - \mathrm{i}0)^{-1} = \mathcal{P}(\frac{1}{x}) + \mathrm{i}\pi\delta(x)$ (where $\mathcal{P}$ denotes the principal part), we have

$$\lim_{y \searrow 0} \Im g(x + \mathrm{i}y) = 4(2\pi)^{-d} \int\limits_{\mathbb{T}^d(x)} \sin^2(k^d) \mathrm{d}\mu_x(k) > 0, \ \forall x \in (0, 2d). \tag{2.30}$$

Finally, the eigenfunctions (2.23) of h_S fulfill $\psi_1^{(m)} = \sqrt{\frac{2}{N+1}} \sin(q_m) \neq 0, \forall m = 1, ..., N$, implying that there are no eigenvalues embedded in $(0, 2d)$. □

For this model one may define the unitary U of Remark 2.1 as the composition the unitary $u_1 \oplus u_2 : \mathcal{H}_{\mathrm{ac}}(H^0) \to \oplus_{i=1}^2 L_2(\mathbb{T}_i^d)$ (where u_i are defined in Eq. (2.21)), with the unitary $v_1 \oplus v_2 : \oplus_{i=1}^2 L_2(\mathbb{T}_i^d) \to \int_0^{2d\oplus} \mathcal{K}_x \mathrm{d}x$, with $\mathcal{K}_x = \oplus_{i=1}^2 L_2(\mathbb{T}_i^d(x), \mathrm{d}\mu_{i,x}(k))$, where $(v_i f_i)(x)$ is the restriction of f_i to the

"energy shell" $\mathbb{T}_i^d(x)$ and $\mathrm{d}\mu_{i,x}$ is the Gelfand-Leray measure on the latter. Then, $\tau f = \int_0^{2d} \mathrm{d}x \tau_x(Uf(x))$, where $\tau_x : \mathcal{K}_x \to \mathcal{H}_S^{(1)}$ is given by:

$$\begin{aligned}(\tau_x\phi)_m &= \delta_{m,1} t \int_{\mathbb{T}_1^d(x)} \overline{\psi^1(k)_{\alpha_1}} \phi_1(k) \mathrm{d}\mu_{1,x}(k) + \\ &\quad + \delta_{m,N} t \int_{\mathbb{T}_2^d(x)} \overline{\psi^2(k)_{\alpha_2}} \phi_2(k) \mathrm{d}\mu_{2,x}(k),\end{aligned} \tag{2.31}$$

and $(U\tau^* f)(x) = \tau_x^* f$, where $\tau_x^* : \mathcal{H}_S^{(1)} \to \mathcal{K}_x$ is given by

$$(\tau_x^* f)(k) = t\psi^1(k)_{\alpha_1} f_1 \oplus t\psi^2(k)_{\alpha_2} f_N. \tag{2.32}$$

We remind that $\psi^i(k)_{\alpha_i} = 2(2\pi)^{-d/2} \sin(k^d)$, see Eq. (2.19).

Upon insertion of Eqs. (2.31), (2.32), the equations of the previous remark are made explicit. For instance, the T-matrix $T_x : \mathcal{K}_x \to \mathcal{K}_x$ appearing in Eq. (2.13) is an integral operator with matrix kernel:

$$T_x(k,k')_{i,j} = \frac{4\mathrm{i}}{(2\pi)^{d-1}} \sin(k^d) R_{\mathrm{eff}}(x + \mathrm{i}0)_{s_i,s_j} \sin(k'^d), \tag{2.33}$$

where $s_1 = 1,\ s_2 = N$.

2.4. An example of direct tunneling between reservoirs

The case when the reservoirs are directly coupled through a tunneling junction without any intermediate sample is special. Indeed, e.g. for two reservoirs, $\mathcal{H}^{(1)} = \mathcal{H}_{\mathrm{ac}}(h^0) = \mathcal{H}_1^{(1)} \oplus \mathcal{H}_2^{(1)}$.

In view of the application to Bose gases, where the surface effects may be drastic, we consider now the translation invariant case of lattice reservoirs, i.e. we suppose that particles live on $L_i = \mathbb{Z}^d$, $i = 1, 2$. The one-particle Hilbert spaces $\mathcal{H}_i^{(1)}$ and reservoir Hamiltonians h_i are defined by Eqs.(2.17), (2.18), respectively. Then, the generalized eigenfunctions of h_i are plane waves

$$\psi^i(k)_r = (2\pi)^{-d/2} \exp(\mathrm{i}kr),\ k \in \mathbb{T}^d = [0, 2\pi)^d, \tag{2.34}$$

with generalized eigenvalues $\omega(k)$, Eq. (2.20), and the unitaries u_i are simply the Fourier transform.

The tunneling is between the origins of L_i, i.e. we take $\alpha_i = 0 \in \mathbb{Z}^d$. Let $\pi_0 = \pi_1 \oplus \pi_2 : \mathcal{H}^{(1)} \to \mathbb{C}^2$ denote the restriction to the pair α_1, α_2 of coupled sites:

$$\pi_0(f_1 \oplus f_2) = (f_1)_0 \oplus (f_2)_0,$$

$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ be the unit matrix in $\mathbb{C}^2$ and $\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ be the first Pauli matrix (interchange of 1 and 2). The interaction can be represented as

$$v = t\pi_0^* \sigma_1 \pi_0 \tag{2.35}$$

One can simplify significantly the expressions of $R(z)$, $\Omega_\pm$, S by using the Fourier representation (2.21) on both spaces: $u = u_1 \oplus u_2 : \oplus_{i=1}^2 \mathcal{H}_i^{(1)} \to \oplus_{i=1}^2 L_2(\mathbb{T}_i^d)$. The resolvent equation $(h-z)f = g$ reduces in $\pi_0 \mathcal{H}^{(1)}$ to the equation $(\sigma_0 + t\pi_0 R^0(z)\pi_0^* \sigma_1)(\pi_0 f) = \pi_0 R^0(z) g$, which amounts to inverting a 2×2 matrix. Thereby,

$$\pi_0 R^0(z) \pi_0^* = \tilde{g}(z)\sigma_0, \tag{2.36}$$

with $\tilde{g}(z)$ given by

$$\tilde{g}(z) = (2\pi)^{-d} \int_{\mathbb{T}^d} \frac{dk}{\omega(k) - z}. \tag{2.37}$$

It should be remarked that $\Im \tilde{g}(x + \mathrm{i}0) > 0$ for all $x \in (0, 2d)$ (and is, as a matter of fact, π times the density of states of the lattice Laplaceian (2.18)) and, for $d \geq 3$, goes to 0 at the spectrum ends $x = 0, 2d$.

We obtain finally:

LEMMA **2.4** *In the direct-coupling model described above*

1. *The resolvent of $h = h^0 + v$ has the representation:*

$$R(z) = R_0(z) - tR_0(z)\pi_0^*(\sigma_1 + t\tilde{g}(z)\sigma_0)^{-1}\pi_0 R_0(z), \qquad (z \in \mathbb{C} \setminus [0, 2d],\ t^2 \tilde{g}(z)^2 \neq 1). \tag{2.38}$$

2. $\sigma_{\mathrm{ac}}(h) = [0, 2d]$.

3. *If $\lim_{x \nearrow 0} \tilde{g}(x) > 1/t$, the equation $t^2 \tilde{g}(z)^2 = 1$ has two real solutions $e_0 < 0$ and $2d - e_0$, which are simple eigenvalues of h; otherwise, $\sigma_{\mathrm{p}}(h) = \emptyset$.*

Using this representation in Eq. (2.8) (in this case, $J = 1$), one finds that the wave operators have the form $W_\pm = 1 - K_\pm$, where $uK_\pm u^*$ are integral operators in $L_2(\mathbb{T}^d) \oplus L_2(\mathbb{T}^d)$ with 2×2-matrix kernels

$$K_\pm(k, k') = \frac{t(2\pi)^{-d}}{\omega(k) - \omega(k') \pm \mathrm{i}0} (\sigma_1 + t\tilde{g}(\omega(k') \mp \mathrm{i}0)\sigma_0)^{-1}. \tag{2.39}$$

The S-matrix acquires the form $S = 1 + T$ with uTu^* having the generalized kernel

$$t(k, k') = \frac{\mathrm{i}\delta(\omega(k) - \omega(k'))}{(2\pi)^{d-1}} (\sigma_1 + t\tilde{g}(\omega(k') + \mathrm{i}0)\sigma_0)^{-1}. \tag{2.40}$$

3. Quasi-free Fermion models

3.1. The algebra of observables, the C^*-dynamics and the reference state

We consider the physical situation described in the Introduction, with r reservoirs of free Fermi gases at equilibrium, coupled via a tunneling junction with a sample consisting of free Fermi particles with a finite-dimensional one-particle state space. The dynamics is supposed quasi-free, specified by the one-particle Hamiltonian $h = h^0 + v$, fulfilling the assumptions of Sec. 2. This subsection is devoted to a precise definition of the quantum system under consideration. We use the notation of subsections 2.1., 2.2..

We start with defining the *C^*-dynamical system*:

Let $\mathcal{F}(\mathcal{H}^{(1)})$ be the antisymmetric Fock space over the one-particle space of Assumption 2.1, and denote $a^*(f)/a(f)$ the usual creation/annihilation operators of one particle in the state $f \in \mathcal{H}^{(1)}$; $a^*(f)$ is linear and $a(f)$ is antilinear with respect with $f \in \mathcal{H}^{(1)}$. The following anticommutation relations hold: for $f, g \in \mathcal{H}^{(1)}$,

$$\{a(f), a(g)\} = \{a^*(f), a^*(g)\} = 0, \ \{a(f), a^*(g)\} = (f, g). \tag{3.1}$$

It follows that $\|a(f)\| = \|a^*(f)\| = \|f\|$. The norm-closed operator algebra generated by them, denoted $CAR(\mathcal{H}^{(1)})$ (called the the algebra of canonical anticommutation relations), is taken as *the algebra of local observables* of the system. As an instance, we shall consider elements in $CAR(\mathcal{H}^{(1)})$ which are the second quantization of one-particle operators: for a trace-class operator a acting in $\mathcal{H}^{(1)}$ with canonical form $a = \sum s_k(f_k, \cdot)g_k$ (where s_k are the singular values of a), $d\Gamma(a) = \sum s_k a^*(g_k)a(f_k) \in CAR(\mathcal{H}^{(1)})$.

The one-particle Hamiltonians h^0 and h define two (strongly continuous) groups of automorphisms of $CAR(\mathcal{H}^{(1)})$ (corresponding to the uncoupled and coupled dynamics, respectively) by

$$\alpha^t(a^\sharp(f)) = a^\sharp(\mathrm{e}^{\mathrm{i}h^0 t}f), \quad \tau^t(a^\sharp(f)) = a^\sharp(\mathrm{e}^{\mathrm{i}ht}f). \tag{3.2}$$

Also, let ϕ^θ denote the gauge automorphism group of $CAR(\mathcal{H}^{(1)})$, i.e.

$$\phi^\theta(a^\sharp(f)) = a^\sharp(\mathrm{e}^{\mathrm{i}\theta}f). \tag{3.3}$$

Corresponding to the decomposition $\mathcal{H}^{(1)} = \mathcal{H}_S^{(1)} \oplus (\oplus_{i=1}^r \mathcal{H}_i^{(1)})$, one can define gauge automorphisms ϕ_i $(i = 1, ..., r)$, ϕ_S of the kinematical algebras $CAR(\mathcal{H}_i^{(1)})$ $(i = 1, ..., r)$, $CAR(\mathcal{H}_S^{(1)})$ of the reservoirs and of the sample.

The states of the system are positive linear functionals $\omega : CAR(\mathcal{H}^{(1)}) \to \mathbb{C}$ of norm $\|\omega\| = \omega(1) = 1$. A state ω is gauge invariant (i.e. $\omega \circ \phi^\theta = \omega$) if, and only if, $\omega(\prod_{i=1}^{n} a^*(g_i) \prod_{i=1}^{m} a(f_i)) = 0$, $\forall n \neq m$. For any state ω, the formula

$$\omega(a^*(g)a(f)) = (g, \rho_\omega f) \tag{3.4}$$

defines a self-adjoint operator $0 \leq \rho_\omega \leq 1$ on $\mathcal{H}^{(1)}$, called its density operator.

Given ρ self-adjoint with $0 \leq \rho \leq 1$, there exists a unique *quasi-free*, gauge-invariant state ω_ρ with density operator ρ. The higher order expectations are expressed in this state ω_ρ by

$$\omega_\rho(a^*(g_m)...a^*(g_1)a(f_1)...a(f_n)) = \delta_{m,n} \det\{(f_i, \rho g_j)\}. \tag{3.5}$$

If the initial state ω^0 of our system is quasi-free and α^t-invariant, what happens if its density operator ρ^0 commutes with h^0, its evolution ω^t under the perturbed dynamics τ^t is likewise a quasi-free state with density operator:

$$\rho^t = [\mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}]^* \rho^0 \mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}; \tag{3.6}$$

indeed, using the α^0-invariance of ω^0,

$$\begin{aligned}\omega^t(a^*(g)a(f)) := \omega^0(\tau^t(a^*(g)a(f))) = \omega^0(\alpha^{-t} \circ \tau^t(a^*(g)a(f))) = \\ = \omega^0(a^*(\mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}g)a(\mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}f)) = (\mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}g, \rho^0\mathrm{e}^{-\mathrm{i}th^0}\mathrm{e}^{\mathrm{i}th}f).\end{aligned}$$

The initial state is taken as a product state $\omega^0 = \omega_S \otimes (\otimes_{i=1}^{r} \omega_i)$, where ω_i are the equilibrium states of two lattice free Fermi gases with one-particle state spaces $\mathcal{H}_i^{(1)}$ and one-particle Hamiltonians h_i and ω_S is an invariant state of the isolated sample.

It is well-known [4] that, at given values of the temperature $\beta^{-1} \geq 0$ and chemical potential $\mu \in \mathbb{R}$, a free Fermi gas has a unique equilibrium state: it is the gauge-invariant quasi-free state with density operator $f_{\beta,\mu}(h)$, where h is the one-particle Hamiltonian, and $f_{\beta,\mu}$ is the Fermi-Dirac function:

$$f_{\beta,\mu}(x) = \frac{1}{1 + \mathrm{e}^{\beta(x-\mu)}} \tag{3.7}$$

This defines in particular the initial states of the reservoirs ω_i.

3.2. Convergence to the NESS and currents

We present here the main results of [2] within the framework defined by Assumptions 2.1–2.3. As with our assumptions no regularization is necessary,

the proof can be made considerably more transparent, so we shall sketch the argument for reader's convenience.

As all states involved are quasi-free and gauge-invariant, it is sufficient, in view of Eq. (3.5), to establish the convergence of the state on elements of the form $a(g)a^*(f)$. This means to calculate the limit density operator as a *weak* limit of the density operators ρ^t.

As shown in Sec. 2, $\mathcal{H}^{(1)} = \mathcal{H}_{\rm ac}(h) \oplus \mathcal{H}_{\rm p}(h)$, with $\mathcal{H}_{\rm p}(h)$ finite-dimensional. Let $P_{\rm ac}, P_{\rm p}$ denote the corresponding orthogonal projections. We calculate the density operator:

$$\rho_+ = ({\rm w}) \lim_{T\to+\infty} (1/T) \int_0^T \rho^t {\rm d}t. \tag{3.8}$$

For $f \in \mathcal{H}_{\rm ac}(h)$, we have, in view of Eq. (3.6),

$$\lim_{t\to+\infty} \rho^t f = W_- \rho^0 W_-^* f$$

because $\lim_{t\to+\infty} {\rm e}^{-{\rm i}th^0} {\rm e}^{{\rm i}th} f = W_-^* f$ exists. On the other hand, if $f \in \mathcal{H}_{\rm p}(h)$, it is a finite combination of eigenvectors, so, we can suppose that f is an eigenvector of h with eigenvalue e,

$$({\rm w}) \lim_{t\to+\infty} P_{\rm ac} {\rm e}^{-{\rm i}th} \rho^0 {\rm e}^{{\rm i}th} f = ({\rm w}) \lim_{t\to+\infty} P_{\rm ac} {\rm e}^{-{\rm i}t(h-e)} (\rho^0 f) = 0$$

by the Riemann-Lebesgue lemma, while, for any eigenvector g of h with eigenvalue e',

$$\lim_{T\to+\infty} (1/T) \int_0^T (g, \rho^t f) {\rm d}t = \lim_{T\to+\infty} (1/T) \int_0^T {\rm e}^{{\rm i}t(e-e')} (g, \rho^0 f) {\rm d}t = \delta_{e,e'} (g, \rho^0 f).$$

In summary,

PROPOSITION **3.1** *The following limit exists for $A \in CAR(\mathcal{H}^{(1)})$*

$$\lim_{T\to+\infty} (1/T) \int_0^T \omega^t(A) {\rm d}t = \omega_+(A) \tag{3.9}$$

and is the quasi-free gauge invariant state of density operator

$$\rho_+ = W_- \rho^0 W_-^* + \sum_{e\in\sigma_{\rm p}(h)} P_e \rho^0 P_e, \tag{3.10}$$

where P_e is the projection onto the eigenspace of h corresponding to the eigenvalue e. Thereby, the restriction of ω_+ to $CAR(\mathcal{H}_{\rm ac}(h))$ is the quasi-free state of density $W_- \rho^0 W_-^*$, and we have

$$\lim_{t\to+\infty} \omega^t(A) = \omega_+(A), \ A \in CAR(\mathcal{H}_{\rm ac}(h)). \tag{3.11}$$

Clearly, the state ω_+ is τ^t-invariant, in particular, for any trace-class operator a on $\mathcal{H}^{(1)}$, $\frac{\mathrm{d}}{\mathrm{d}t}\omega_+(\tau^t(d\Gamma(a))) = 0$, implying that $\mathrm{tr}(\rho_+[h,a]) = 0$. However, if a is not a trace-class operator (but $\rho_+[h,a]$ is trace-class), it may happen that $\mathrm{tr}(\rho_+[h,a]) \neq 0$. This is the case for the extensive conserved charges of the isolated reservoirs, and it expresses the existence of the steady currents in the NESS ω_+ constructed above.

Each of the reservoirs R_i has two conserved quantities, the energy and the particle number, which correspond formally to $d\Gamma(h_0 P_i)$ and $d\Gamma(P_i)$, where P_i is the projection of $\mathcal{H}^{(1)}$ onto $\mathcal{H}^{(1)}_i$. This is expressed by the invariance of their equilibrium states ω_i under the dynamical and gauge automorphism groups, α_i^t and ϕ_i^θ, of the isolated reservoirs. The energy and particle currents from the reservoirs R_i is calculated as the ω_+-expectation of the corresponding fluxes $I_{i,\mathrm{en}} = d\Gamma(-\mathrm{i}[h, h^0 P_i]) = d\Gamma(-\mathrm{i}[v, h^0 P_i])$ and $I_{i,\mathrm{part}} = d\Gamma(-\mathrm{i}[h, P_i]) = d\Gamma(-\mathrm{i}[v, P_i])$, respectively. Remark that, because v is a finite range operator, the commutators are trace-class in $\mathcal{H}^{(1)}$, so the proposition 3.1 applies. As $P_e h = h P_e = e P_e$, the sum over the point spectrum in Eq. (3.10) does not contribute to any of the two currents $J = \omega_+(I)$. Hence,

PROPOSITION **3.2** *The energy and particle currents from the reservoirs R_i are calculated according to the formulas*

$$\begin{aligned} J_{i,\mathrm{en}} &= -\mathrm{tr}(\rho^+ \mathrm{i}[v, h_0 P_i]) = -\mathrm{tr}(W_-\rho^0 W_-^* \mathrm{i}[v, h_0 P_i]), \\ J_{i,\mathrm{part}} &= -\mathrm{tr}(\rho^+ \mathrm{i}[v, P_i]) = -\mathrm{tr}(W_-\rho^0 W_-^* \mathrm{i}[v, P_i]). \end{aligned} \tag{3.12}$$

We shall next bring formulas (3.12) to a form, known as Landauer-Büttiker formulas, which make clear that the currents depend in fact only on the on-shell T-matrix T_x. We start with a statement [2] relative to a larger class of conserved reservoir observables.

PROPOSITION **3.3** *Let a be a bounded self-adjoint operator in $\mathcal{H}^{(1)}_{\mathrm{ac}}(h^0)$ commuting with h^0, so that, in the representation of Remark 2.1, $UaU^* = \int^\oplus a(x)\mathrm{d}x$, with $a(x)$ bounded self-adjoint operators in $\mathcal{K}_x$. We denote $\hat{a} = JaJ^*$ its counterpart in $\mathcal{H}^{(1)}$. Let*

$$J(a) := \omega_+(d\Gamma(-\mathrm{i}[h, \hat{a}])) = -\mathrm{tr}_{\mathcal{H}_{\mathrm{ac}}(h)}(W_-\rho^0 W_-^* \mathrm{i}[h, \hat{a}]) \tag{3.13}$$

be the "current" associated to a. Then,

$$J(a) = -\int \mathrm{tr}_{\mathcal{K}_x}\{\rho^0(x)[a(x)T_x + T_x^* a(x) + T_x^* a(x) T_x]\}\frac{\mathrm{d}x}{2\pi}. \tag{3.14}$$

Proof. The equality in Eq. (3.13), meaning that the sum over the point spectrum of h in Eq. (3.10) vanishes, is shown in the same way as for Eq. (3.12).

As, by Assumption 2.2, $v = J\tau^* I^* + I\tau J^*$, the commutator in the r.h.s. of (3.13) equals $[h, \hat{a}] = [v, \hat{a}] = I\tau a J^* - Ja\tau^* I^*$, which has finite-range. Using the permutation invariance of the trace,

$$\mathrm{tr}_{\mathcal{H}_{\mathrm{ac}}(h)}(W_-\rho^0 W_-^*[v, \hat{a}]) = \mathrm{tr}_{\mathcal{K}}(U\rho^0 W_-^*[v, \hat{a}]W_- U^*).$$

We show that the operator under trace is an integral operator on $\mathcal{K}$, i.e. of the form $K\psi(x) = \int \mathrm{d}y k(x, y)\psi(y)$, where $k(x, y) : \mathcal{K}_y \to \mathcal{K}_x$ are continuous, trace-class-operator valued functions. Therefore, the trace can be calculated as $\int \mathrm{d}x \mathrm{tr}_{\mathcal{K}_x} k(x, x)$.

To this aim, we factorize the two terms of the commutator as

$$\begin{aligned} UW_-^*[v, \hat{a}]W_- U^* = \; & (UW_-^* I\tau U^*)(UaU^*)(UJ^* W_- U^*) \\ & -(UW_-^* JU^*)(UaU^*)(U\tau^* I^* W_- U^*). \end{aligned}$$

Remembering the representation of τ, τ^* in Remark 2.1 and the expressions (2.14), (2.15) of W_-, W_-^*, the generalized kernels of the operators in brackets are

$$(UW_-^* JU^*)(x, y) = \delta(x - y) + (y - x + \mathrm{i}0)^{-1}\tau_x^* R_{\mathrm{eff}}(x - \mathrm{i}0)\tau_y;$$

$$(UJ^* W_- U^*)(x, y) = \delta(x - y) + (x - y - \mathrm{i}0)^{-1}\tau_x^* R_{\mathrm{eff}}(y + \mathrm{i}0)\tau_y;$$

$$(UW_-^* I\tau U^*)(x, y) = -\tau_x^* R_{\mathrm{eff}}(x - \mathrm{i}0)\tau_y;$$

$$(U\tau^* I^* W_- U^*)(x, y) = -\tau_x^* R_{\mathrm{eff}}(y + \mathrm{i}0)\tau_y.$$

The kernel $k(x, y)$ is obtained as the composition of the kernels of the factors. The continuity with respect with x, y is a consequence of Assumption 2.2. The diagonal $k(x, x)$ equals

$$\begin{aligned} & -\tau_x^* R_{\mathrm{eff}}(x - \mathrm{i}0)\tau_x a(x) + a(x)\tau_x^* R_{\mathrm{eff}}(x - \mathrm{i}0)\tau_x - \\ & - \int \mathrm{d}x' \tau_x^* R_{\mathrm{eff}}(x - \mathrm{i}0)\tau_x' a(x')\tau_{x'}^* R_{\mathrm{eff}}(x + \mathrm{i}0)\tau_x \times \\ & \qquad \times [(x' - x - \mathrm{i}0)^{-1} - (x' - x + \mathrm{i}0)^{-1}] \\ & = \tfrac{1}{2\pi \mathrm{i}}[T_x^* a(x) + a(x)T_x + T_x^* a(x)T_x], \end{aligned}$$

where we used the Sokhotski formula $(x - \mathrm{i}0)^{-1} = \mathcal{P}\left(\frac{1}{x}\right) + \mathrm{i}\pi\delta(x)$ and the definition (2.13) of the T-matrix. Insertion of this calculation in Eq. (3.13) gives Eq. (3.14). □

We take now into account the decomposition $\mathcal{H}^{(1)}_{\rm ac}(h^0) = \bigoplus_i \mathcal{H}^{(1)}_i$. For an energy $x \in [e_{\min}, e_{\max}]$, we have $\mathcal{K}_x = \bigoplus_i \mathcal{K}_{x,i}$; thereby, if $x \notin I_i$, $\mathcal{K}_{x,i} = \{0\}$. Accordingly, the operators under $\mathrm{tr}_{\mathcal{K}_x}$ in Eq. (3.14) have matrix representations. The density $\rho^0(x)$ is the diagonal matrix with $\rho^0(x)_{i,i} = f_{\beta_i,\mu_i}(x) \cdot 1$. Also, $(T_x)_{i,j} = 2\pi \mathrm{i}(\tau_i^*)_x R_{\rm eff}(x + \mathrm{i}0)(\tau_j)_x$, which vanishes for $x \notin I_i \cap I_j$. What concerns $a(x)$, as we are interested in observables associated with the isolated reservoirs, we suppose that its matrix has block-diagonal form: $a(x)_{i,j} = \delta_{i,j} a_i(x)$. In this case,

$$\begin{aligned}&\mathrm{tr}_{\mathcal{K}_x}\{\rho^0(x)[a(x)T_x + T_x^* a(x) + T_x^* a(x) T_x]\} = \\ &\sum_{i=1}^r f_{\beta_i,\mu_i}(x) \mathrm{tr}_{\mathcal{K}_{x,i}}\{a_i(x)(T_x)_{i,i} + (T_x^*)_{i,i} a_i(x) + \sum_{j=1}^r (T_x^*)_{i,j} a_j(x) (T_x)_{j,i}\}.\end{aligned} \tag{3.15}$$

This can be further simplified using the unitarity of the S-matrix:

$$(T_x)_{i,i} + (T_x^*)_{i,i} + \sum_{j=1}^r (T_x)_{i,j}(T_x^*)_{j,i} = 0$$

and the permutation invariance of the trace, whence

$$\begin{aligned}&\sum_{i=1}^r f_{\beta_i,\mu_i}(x) \mathrm{tr}_{\mathcal{K}_{x,i}}\{a_i(x)(T_x)_{i,i} + (T_x^*)_{i,i} a_i(x)\} \\ &= -\sum_{i=1}^r f_{\beta_i,\mu_i}(x) \mathrm{tr}_{\mathcal{K}_{x,i}}\{a_i(x) \sum_{j=1}^r (T_x)_{i,j}(T_x^*)_{j,i} \\ &= -\sum_{j=1}^r f_{\beta_j,\mu_j}(x) \mathrm{tr}_{\mathcal{K}_{x,i}}\{\sum_{j=1}^r (T_x^*)_{i,j} a_j(x)(T_x)_{j,i}.\end{aligned}$$

Hence,

COROLLARY **3.1** *For a self-adjoint operator a in $\mathcal{H}^{(1)}_{\rm ac}(h^0)$ such that $a(x)_{i,j} = \delta_{i,j} a_i(x), \forall x$,*

$$J(a) = \sum_{i,j=1}^r \int [f_{\beta_i,\mu_i}(x) - f_{\beta_j,\mu_j}(x)] \mathrm{tr}_{\mathcal{K}_{x,i}}\{a_i(x)(T_x)_{i,j}(T_x^*)_{j,i}\} \mathrm{d}x. \tag{3.16}$$

Thereby, $(T_x)_{i,j} \neq 0$ only for $x \in I_i \cap I_j$.

In particular, defining the transmission probability between reservoirs R_i and R_j as $t_{i,j}(x) := \mathrm{tr}_{\mathcal{K}_{x,i}}\{(T_x)_{i,j}(T_x^*)_{j,i}\}$,

$$\begin{aligned}J_{i,\rm en} &= \sum_{j=1}^r \int [f_{\beta_i,\mu_i}(x) - f_{\beta_j,\mu_j}(x)] x t_{i,j}(x), \\ J_{i,\rm part} &= \sum_{j=1}^r \int [f_{\beta_i,\mu_i}(x) - f_{\beta_j,\mu_j}(x)] t_{i,j}(x).\end{aligned} \tag{3.17}$$

3.3. Consequences for the model of Sec. 2.3

We specialize here to the case of two reservoirs ($r = 2$) of free lattice Fermi gases described in Sec. 2.3. and draw a few conclusions about its behavior as a function of the dimension of the lattices d_i and of the wire length N.

- *The currents*, Eq. (3.17), *are a sum of two currents*, each obtained when one of the two reservoirs is put in turn in the Fock state (corresponding to the density matrix $f_{+\infty,-\infty}(h_i) = 0$. One may consider therefore only the particle current

$$J_{1,\mathrm{part}}(\beta,\mu) = \int f_{\beta,\mu}(x) t_{1,2}(x). \tag{3.18}$$

- *The transmission probability*

$$t_{1,2}(x) = \int_{\mathbb{T}^d(x)} \mathrm{d}\mu_x(k) \int_{\mathbb{T}^d(x)} \mathrm{d}\mu_x(k) |T(k,k')_{1,2}|^2$$

has a resonant structure. In view of Eq. (2.33), one has to study the energy dependence of the matrix element $R_{\mathrm{eff}}(x + \mathrm{i}0)_{1,N}$. By analytic perturbation theory, as h_S has simple eigenvalues ε_m, the eigenvalues $\lambda_m(x)$, $m = 1, ..., N$ of $h_{\mathrm{eff}}(x + \mathrm{i}0)$ are simple for sufficiently small tunneling constant t. Let $\psi^{(m)}(x)$ be the corresponding eigenvectors; then $\bar{\psi}^{(m)}(x)$ is the dual basis (i.e. $(\bar{\psi}^{(m)}(x), \psi^{(m')}(x)) = \delta_{m,m'}$. Hence,

$$R_{\mathrm{eff}}(x + \mathrm{i}0)_{1,N} \sim \sum_{m=1}^{N} (\lambda_m(x) - x)^{-1} \psi_1^{(m)}(x) \psi_N^{(m)}(x).$$

To lowest order in t, $\lambda_m(x) \sim \varepsilon_m - \frac{2}{N+1} t^2 g(x + \mathrm{i}0) \sin^2 q_m$, where we used Eq. (2.28) and the explicit form (2.23) of the eigenvectors $\psi^{(m)}$ at $t = 0$, which puts into evidence "resonances" at $x = \varepsilon_m - \frac{2}{N+1} t^2 \Re g(x + \mathrm{i}0) \sin^2 q_m$ of "width" $\frac{2}{N+1} t^2 \Im g(x + \mathrm{i}0) \sin^2 q_m$.

- *The density profile*

$$n(r) = \omega_+(a^*(\delta_r) a(\delta_r)) = \sum (P_e \delta_r, \rho^0 P_e \delta_r) + (W_-^* \delta_r, \rho^0 W_-^* \delta_r) \tag{3.19}$$

is a sum over reservoirs of density profiles corresponding to the other reservoir put in its Fock state (due to the block structure of $\rho^0 = \sum_i^{\oplus} \rho_i$). We calculate the second term of (3.19) with $\rho_2 = 0$. We need

therefore $P_1 W_-^* \delta_r$, where P_1 is the projection onto $\mathcal{H}_1^{(1)}$. In view of Eq. (2.15), we have

$$(UP_1 W_-^* \delta_r)(x) = -t\psi^1(k)_{\alpha_1} R_{\text{eff}}(x - \mathrm{i}0)_{1,r},$$

if $r \in \{1, ..., N\}$,

$$(UP_1 W_-^* \delta_r)(x) = \psi^1(k)_r + t^2 \psi^1(k)_{\alpha_1} R_{\text{eff}}(x - \mathrm{i}0)_{1,1} R_1(x + \mathrm{i}0)_{\alpha_1, r}$$

if $r \in L_1$, and

$$(UP_1 W_-^* \delta_r)(x) = t^2 \psi^1(k)_{\alpha_1} R_{\text{eff}}(x - \mathrm{i}0)_{1,N} R_2(x + \mathrm{i}0)_{\alpha_2, r},$$

if $r \in L_2$.

In particular, the density profile inside R_2 (the initially void reservoir), is given by

$$t^4 \int \mathrm{d}k f_{\beta_1,\mu_1}(\omega_1(k)) |\psi^1(k)_{\alpha_1} R_{\text{eff}}(\omega_1(k) - \mathrm{i}0)_{1,N}|^2 |R_2(\omega_1(k) + \mathrm{i}0)_{\alpha_2,r}|^2.$$

It is to be remarked that, if $d_2 = 1$ (which is the model of infinite leads used in [6]), the density of transmitted particles has a nonzero limit as $r \to \infty$; this seems improper for a reservoir, which is expected to keep unchanged its "conserved charges" even after coupling it to other reservoirs. For $d_2 > 1$, the density decays like $|r|^{-1}$ irrespective of d_2 [14].

4. Quasi-free Boson models

4.1. The algebra of observables and the reference state

The kinematical C^*-algebra of the model is the canonical commutation relation algebra $CCR(\mathcal{D})$ over a suitable subspace $\mathcal{D} \subset \mathcal{H}^{(1)}$, which is left invariant by the one-particle evolution groups: $\exp(\mathrm{i}th^0)\mathcal{D} = \mathcal{D}$, $\exp(\mathrm{i}th)\mathcal{D} = \mathcal{D}$.

$CCR(\mathcal{D})$ is generated by the Weyl operators $\{\mathcal{W}(f);\ f \in \mathcal{D}\}$, satisfying

$$\mathcal{W}(f)\mathcal{W}(g) = \mathrm{e}^{-\frac{\mathrm{i}}{2}\Im(f,g)} \mathcal{W}(f+g). \tag{4.1}$$

The defining equation (4.1) implies that $\mathcal{W}(0) = 1$ and $\mathcal{W}(f)$ are unitaries ($\mathcal{W}(f)^* \mathcal{W}(f) = 1$). According to a theorem by Slawny, such a C^*-algebra

is unique up to an isomorphism; in particular, it can be shown (using the well-known Fock representation) that $\|\mathcal{W}(f) - 1\| \geq \sqrt{2}$ for $f \neq 0$, implying that the application $f \mapsto \mathcal{W}(f)$ cannot be norm-continuous [13].

To any state ω on $CCR(\mathcal{D})$ a function $E : \mathcal{D} \to \mathbb{C}$ is associated by

$$E(f) = \omega(\mathcal{W}(f)), \tag{4.2}$$

named its generating functional. E satisfies: (i) normalization: $E(0) = 1$, (ii) unitarity: $\overline{E(f)} = E(-f)$, and (iii) positivity:

$$\sum_{i,j=1}^{n} z_i E(f_i - f_j) \mathrm{e}^{-\frac{\mathrm{i}}{2}\Im(f_i, f_j)} \bar{z}_j \geq 0, \ \forall n, \forall z_i \in \mathbb{C}, f_i \in \mathcal{D}\,(i = 1, ..., n).$$

Conversely, any E with these properties defines a unique state by Eq. (4.2). Therefore, in describing the initial and evolved states of our model, it will be sufficient to specify the corresponding generating functionals.

A state ω is quasi-free if, and only if, E has the particular form

$$E(f) = \exp\left(\mathrm{i}\sqrt{2}\Re\langle l, f\rangle - \frac{1}{4}Q(f, f)\right), \tag{4.3}$$

where $l \in \mathcal{D}'$ is a linear form and $Q(\cdot, \cdot) \geq 1$ a quadratic form on $\mathcal{D} \times \mathcal{D}$. Quasi-free states ω are regular, i.e. in the associated GNS representation π_ω, for any $f \in \mathcal{D}$, the unitary group $\mathbb{R} \ni t \mapsto \pi_\omega(\mathcal{W}(tf))$ is weakly continuous. Hence, $\forall f \in \mathcal{D}$, there exist self-adjoint operators $\varphi(f)$ – "field operators", such that $\pi_\omega(\mathcal{W}(tf)) = \exp(\mathrm{i}t\varphi(f))$. The fields $\varphi(f)$ are real-linear functions of f. In terms of the fields $\varphi(f)$ one can define creation and annihilation operators by $a^*(f) = 2^{-1/2}(\varphi(f) - \mathrm{i}\varphi(\mathrm{i}f))$, $a(f) = 2^{-1/2}(\varphi(f) + \mathrm{i}\varphi(\mathrm{i}f))$. Then, denoting Ω_ω the cyclic vector of π, one has the following

PROPOSITION **4.1** *In a quasi-free state with generating functional* (4.3)*, Ω_ω is in the domain of all powers of $a^\sharp(f)$, $f \in \mathcal{D}$, and the following relations hold:*

$$\begin{gathered}(\Omega_\omega, a^*(f)\Omega_\omega) = \overline{(\Omega_\omega, a(f)\Omega_\omega)} = \langle l, f\rangle, \\ (\Omega_\omega, a^*(g)a(f)\Omega_\omega) - (\Omega_\omega, a^*(g)\Omega_\omega)(\Omega_\omega, a(f)\Omega_\omega) = Q(f, g);\end{gathered} \tag{4.4}$$

all other truncated expectations vanish.

The time evolutions α^t, τ^t, for the uncoupled, respectively, coupled reservoirs and sample are the groups of Bogoliubov automorphisms on $CCR(\mathcal{D})$ defined

by their action on $\mathcal{W}(f)$:

$$\alpha^t(\mathcal{W}(f)) = \mathcal{W}(\mathrm{e}^{\mathrm{i}h^0 t} f),$$
$$\tau^t(\mathcal{W}(f)) = \mathcal{W}(\mathrm{e}^{\mathrm{i}ht} f). \tag{4.5}$$

In view of the canonical commutation relations (4.1), Eq. (4.5) is sufficient to uniquely define the action of τ^t on all elements of $CCR(\mathcal{D})$. By the remark above, the two automorphism groups are *not* strongly continuous. However, in a quasi-free representation they are implemented by weakly continuous unitary groups. Moreover, the evolution of a quasi-free initial state under a dynamics of the form (4.5) is likewise quasi-free. This means that the evolved state at time $t > 0$ of Boson systems, which, at $t = 0$, were in a quasi-free state, is uniquely determined by the evolved one-point and two-point functions, i.e. by $\langle l_t, f\rangle = \langle l, \mathrm{e}^{\mathrm{i}ht} f\rangle$ and $Q_t(f,g) = Q(\mathrm{e}^{\mathrm{i}ht} f, \mathrm{e}^{\mathrm{i}ht} g)$. In this respect, their study parallels the study of Fermi systems in the previous section and the counterpart of proposition 3.1 holds true. There appear, however, subtleties related to the choice of the initial (reference) state; in particular, unlike in the Fermi case, the domain $\mathcal{D}$ (i.e. the kinematical algebra $CCR(\mathcal{D})$) depends on the reference state. In order to keep the exposition at a reasonable level of complexity, we shall explain them only for the model in Sec. 2.4., i.e. direct tunneling between reservoirs on $\mathbb{Z}^d$ with no intermediate sample. The consideration of the general frame (given by assumptions 2.1–2.3, supplemented with special requirements about the existence of a density of energy levels in the infinite volume limit) is left for another publication.

The equilibrium states of a free Bose gas are quasi-free; they have been studied in detail in the literature [4]. The peculiarity of the free Bose gas is that, under certain conditions, it shows a phase transition at low temperature and high density. It happens that, in the multi-phase region, the canonical and grand-canonical are inequivalent. As we are interested in particle flows between reservoirs, it is natural to use the canonical description for the reservoirs.

We remind below the expressions of the generating functionals for the canonical equilibrium states for our model of reservoir, obtained by an easy adaptation of the derivation by Cannon [4], [11] for the continuum Bose gas.

We start by describing one reservoir R, consisting of a free lattice Bose gas living on $\mathbb{Z}^d$.

Let β, ρ be fixed positive numbers and define:

$$\rho_{\mathrm{cr}}(\beta) = (2\pi)^{-d} \int\limits_{\mathbb{T}_1^d} \frac{1}{\mathrm{e}^{\beta\omega(k)} - 1} \mathrm{d}^d k \leq +\infty, \tag{4.6}$$

where $\omega(k)$ is the dispersion law Eq. (2.20). As $\omega(k) \approx \frac{1}{2}|k|^2$ around its minimum at $k = 0$, one has that $\rho_{\rm cr}(\beta)$ is finite for $d \geq 3$ and is infinite for $d = 1, 2$.

For $\rho < \rho_{\rm cr}(\beta)$, the fugacity z is defined to be the unique solution $z(\beta, \rho)$ of the equation

$$\rho = (2\pi)^{-d} \int\limits_{\mathbb{T}^d} \frac{z}{\mathrm{e}^{\beta\omega(k)} - z} \mathrm{d}^d k,$$

while, for $\rho \geq \rho_{\rm cr}(\beta)$, put $z(\beta, \rho) = 1$. The momentum distribution for $k \neq 0$ at the given β, ρ is proportional to

$$n_{\beta,\rho}(k) = \frac{z(\beta, \rho)}{\mathrm{e}^{\beta\omega(k)} - z(\beta, \rho)}, \tag{4.7}$$

while the condensate density is given by

$$\rho_0 = \max\{0, \rho - \rho_{\rm cr}(\beta)\}. \tag{4.8}$$

Then, the generating functional of the canonical equilibrium state at β, ρ is given by the formula

$$E_{\beta,\rho}(f) = \exp\left\{-\frac{\|f\|^2}{4} - \frac{1}{2}(uf, n_{\beta,\rho}\, uf)\right\} J_0(\sqrt{2(2\pi)^d \rho_0}\, |(uf)(0)|), \tag{4.9}$$

where u is the Fourier transform and J_0 is the Bessel function.

For $\rho \leq \rho_{\rm cr}(\beta)$, the canonical state defined by Eq. (4.9) is extremal, however, if $\rho_{\rm cr}(\beta) < \infty$ and $\rho > \rho_{\rm cr}(\beta)$, it has a nontrivial decomposition into extremal states indexed by a phase $\mathrm{e}^{\mathrm{i}\theta}$:

$$E_{\beta,\rho}(f) = (2\pi)^{-1} \int_0^{2\pi} E^\theta_{\beta,\rho}(f) \mathrm{d}\theta, \tag{4.10}$$

where

$$E^\theta_{\beta,\rho}(f) = \exp\left\{-\frac{\|f\|^2}{4} - \frac{(uf, n_{\beta,\rho}\, uf)}{2} - \frac{\mathrm{i}\sqrt{2\rho_0}}{(2\pi)^{d/2}} \Re(\mathrm{e}^{-\mathrm{i}\theta}(uf)(0))\right\}. \tag{4.11}$$

Thereby, the test function space $\mathcal{D}$ should be chosen such that the functionals (4.11) are well defined for $f \in \mathcal{D}$, e.g. taking $\mathcal{D} = l^1(\mathbb{Z}^d)$ would suffice. Indeed, with this choice uf is continuous on $\mathbb{T}^d$, ensuring both the integrability of $n_{\beta,\rho}|uf|^2$ and the existence of $(uf)(0)$. We shall impose, however a stronger condition ensuring that uf is Hölder-continuous, and take $\mathcal{D}$ as

the space $l^1(\mathbb{Z}^d;|x|^\epsilon)$ for some $\epsilon>0$, consisting of functions $f:\mathbb{Z}^d\to\mathbb{C}$ for which $\|f\|_{\mathcal{D}}:=\sum\limits_{x\in\mathbb{Z}^d}|x|^\epsilon|f_x|<\infty$.

Using the matrix notation associated with the direct sum $\mathcal{H}^{(1)}=\mathcal{H}_1^{(1)}\oplus\mathcal{H}_2^{(1)}$, we take $f=f_1\oplus f_2\in\mathcal{D}_1\oplus\mathcal{D}_2$ (where $\mathcal{D}_i$ are copies of $\mathcal{D}$) and the initial state ω^0 as a product of canonical equilibrium states of R_i at temperatures β_i and densities ρ_i $(i=1,2)$, respectively:

$$\omega^0(\mathcal{W}(f))=E_0(f)=E_{\beta_1,\rho_1}(f_1)E_{\beta_2,\rho_2}(f_2), \tag{4.12}$$

where $E_{\beta_i,\rho_i}(f_i)$ are arbitrary mixtures (with probability measures $\mathrm{d}\mu_{1,2}(\theta_{1,2})$) of the extremal state generating functionals (4.11). Denoting $\rho_{0,i}$ the condensate densities in R_i and

$$\tilde{n}_0=\begin{pmatrix} n_{\beta_1,\rho_1} & 0\\ 0 & n_{\beta_2,\rho_2}\end{pmatrix},\ \tilde{\rho}_0(\theta_1,\theta_2)=\begin{pmatrix}\sqrt{2\rho_{0,1}}\mathrm{e}^{-\mathrm{i}\theta_1} & \sqrt{2\rho_{0,2}}\mathrm{e}^{-\mathrm{i}\theta_2}\end{pmatrix}, \tag{4.13}$$

we have

$$E_0(f)=\int\mathrm{d}\mu_1(\theta_1)\mathrm{d}\mu_2(\theta_2)E_0^{\theta_1,\theta_2}(f), \tag{4.14}$$

where

$$E_0^{\theta_1,\theta_2}(f)=\exp\left\{-\frac{\|f\|^2}{4}-\frac{(uf,\tilde{n}_0\,uf)}{2}-\frac{\mathrm{i}}{(2\pi)^{d/2}}\Re(\tilde{\rho}_0(\theta_1,\theta_2)(uf)(0))\right\}. \tag{4.15}$$

In particular, the canonical states (4.9) are obtained for $\mathrm{d}\mu_i(\theta)=(2\pi)^{-1}\mathrm{d}\theta$.

4.2. The approach to, and properties of, the NESS

We are interested in the time evolution of an initial state ω^0 as defined by Eq. (4.14) (which is α^t-invariant) under the coupled dynamics τ^t, Eq. (4.5). We have

$$\omega^t(\mathcal{W}(f))=\omega^0(\mathcal{W}(\exp(\mathrm{i}th)f)=\omega^0(\mathcal{W}(\exp(-\mathrm{i}th^0)\exp(\mathrm{i}th)f). \tag{4.16}$$

Using the analysis done in Sec. 2.4., we obtain the following convergence result, which defines the stationary state.

PROPOSITION **4.2** *Under the condition above, the following limit exists and defines a quasi-free invariant state* ω_{stat}: $\forall f\in\mathcal{D}$,

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T\omega^t(\mathcal{W}(f))\mathrm{d}t=E_{\mathrm{stat}}(f). \tag{4.17}$$

Corresponding to the decomposition (4.14) *of the initial state,*

$$E_{\rm stat}(f) = \int {\rm d}\mu_1(\theta_1){\rm d}\mu_2(\theta_2) E^{\theta_1,\theta_2}_{\rm stat}(f), \tag{4.18}$$

where

$$E^{\theta_1,\theta_2}_{\rm stat}(f) = E^{\theta_1,\theta_2}_0(W_-^* P_{\rm ac} f) E^{\theta_1,\theta_2}_{\rm (p)}(P_{\rm p} f). \tag{4.19}$$

Thereby, the limit in mean is necessary only for the contribution of the point spectrum, i.e. for $f = P_{\rm ac} f$, *the limit* $\lim\limits_{t\to\infty} \omega^t(\mathcal{W}(f))$ *exists and equals* $\int {\rm d}\mu_1(\theta_1){\rm d}\mu_2(\theta_2) E^{\theta_1,\theta_2}_0(W_-^* P_{\rm ac} f)$.

Proof. We isolate, in the quadratic and linear forms appearing at the exponent in $E^{\theta_1,\theta_2}_0({\rm e}^{{\rm i}ht} f)$, the terms which do not depend on $P_{\rm ac} f$, i.e. $T_{\rm p}(t) := -\frac{1}{4}\|P_{\rm p} f\|^2 - \frac{1}{2}(u{\rm e}^{{\rm i}ht} P_{\rm p} f, \tilde{n}_0\, u{\rm e}^{{\rm i}ht} P_{\rm p} f) - {\rm i}(2\pi)^{-3/2}\Re(\tilde{\rho}_0(\theta_1,\theta_2)(u{\rm e}^{{\rm i}ht} P_{\rm p} f)(0))$. The t-dependence of $T_{\rm p}(t)$ comes from exponentials of the form ${\rm e}^{{\rm i}e_0 t}$, ${\rm e}^{{\rm i}(2d-e_0)t}$ and ${\rm e}^{{\rm i}2(d-e_0)t}$, where $e_0, 2d-e_0$ are the two eigenvalues of h. Therefore, ${\rm e}^{T_{\rm p}(t)}$ is an almost-periodic function, what ensures that $\lim\limits_{T\to\infty} \frac{1}{T}\int\limits_0^T {\rm e}^{T_{\rm p}(t)}{\rm d}t =: E^{\theta_1,\theta_2}_{\rm (p)}(P_{\rm p} f)$ exists. Remark that $(P_{\rm p} f)_r$ decays exponentially as $r\to\infty$, therefore, if $f\in\mathcal{D}$, $P_{\rm ac} f\in\mathcal{D}$ as well. Hence, $\int_{\mathbb{T}^d(x)} (uP_{\rm ac} f)(k){\rm d}\mu_x(k)$ is Hölder continuous of x, therefore, by the Privalov theorem [7],

$$\begin{aligned} &(uW_-^* P_{\rm ac} f)(k) = (uP_{\rm ac} f)(k) - \\ &-\tfrac{t}{(2\pi)^d}(\sigma_1 + t\tilde{g}(\omega(k)-{\rm i}0)\sigma_0)^{-1} \textstyle\int_{\mathbb{T}^d} \frac{(uP_{\rm ac} f)(k'){\rm d}k'}{\omega(k')-\omega(k)+{\rm i}0} \end{aligned} \tag{4.20}$$

is likewise Hölder continuous of $\omega(k)$ and, as such, belongs to the domain of E^{θ_1,θ_2}_0. By an analysis like that in the proof of Proposition 3.1, the remaining terms have (usual) limits as $t\to\infty$, which proves the assertion. □

In view of the explicit forms (4.15) of the functionals E^{θ_1,θ_2}_0, Proposition 4.2 provides a detailed description of the stationary state and allows the calculation of various quantities of physical interest.

We report below the analytic results for the energy and particle currents. We point out that, like in the Fermi case, the point spectrum of h gives no contribution to the currents and the contribution of the absolutely continuous spectrum may be expressed in terms of the S-matrix alone (Landauer-Büttiker-like formula). We shall not repeat here the proof of the latter, but perform the direct calculation based on Eq. (4.19). Thereby, if $d\geq 3$,

we suppose, for simplicity, that we are in the weak coupling regime, where $\sigma_{\rm p}(h) = \emptyset$.

In calculating the currents between pure phases of the reservoirs, we take advantage that the initial state, being a product of extremal equilibrium states, can be approximated by finite-volume states (possibly with weak symmetry-breaking perturbations), what allows to substantiate expressions (of the currents from a reservoir in an extremal state) similar to those in the Fermi case [1]. As a preparation, we calculate, using Eq. (4.20), $W_-^* f$ for a few local functions f appearing in these expressions:

• For $(\delta_0^1)_r = \delta_{0,r} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and δ_0^2 defined analogously for the second reservoir,

$$(uP_j W_-^* \delta_0^i)(k) = \frac{1}{(2\pi)^{d/2}} \left\{ \delta_{i,j} - t\tilde{g}(\omega(k) - {\rm i}0)[(\sigma_1 + t\tilde{g}(\omega(k) - {\rm i}0))^{-1}]_{j,i} \right\},$$

where P_j projects onto the reservoir j and we used the definition (2.37) of $\tilde{g}$;

• For $(h_0^1)_r = (d\delta_{x,0} - \frac{1}{2}\delta_{|x|,1}) \begin{pmatrix} 1 \\ 0 \end{pmatrix}$,

$$\begin{aligned} &(uP_j W_-^* h_0^1)(k) = \tfrac{1}{(2\pi)^{d/2}} \{\omega(k)\delta_{j,1} - \\ &-t[(\sigma_1 + t\tilde{g}(\omega(k) - {\rm i}0))^{-1}]_{j,1}[1 + \omega(k)\tilde{g}(\omega(k) - {\rm i}0)] \} . \end{aligned}$$

PROPOSITION **4.3** *In the direct tunneling model of Section* 2.4*, the currents flowing from* R_1 *in the stationary state* $\omega_{stat}^{\theta_1,\theta_2}$ *arising from extremal initial states are given by:*

1. *The particle current:*

$$\begin{aligned} J^1_{\rm part}(\theta_1,\theta_2) &= 2t\Im\omega_0^{\theta_1,\theta_2}(a_0^*(W_-^*(\delta_0^1))a_0(W_-^*(\delta_0^2))) \\ &= \frac{2t}{(2\pi)^d}\int (n_1(k) - n_2(k)) \frac{\Im\tilde{g}(\omega(k) - {\rm i}0))}{|1 - t^2\tilde{g}(\omega(k) - {\rm i}0))^2|^2} d^3k \\ &+ \frac{2t}{(2\pi)^d} \frac{\sqrt{\rho_{01}\rho_{02}}}{1 - \tilde{g}(0)^2} \sin(\theta_2 - \theta_1) \end{aligned}$$

2. *The energy current:*

$$\begin{aligned} J^1_{\rm en}(\theta_1,\theta_2) &= 2t\Im\omega_0^{\theta_1,\theta_2}(a_0^*(W_-^*(h_0^1))a_0(W_-^*(\delta_0^2))) \\ &= \frac{2t}{(2\pi)^d}\int (n_1(k) - n_2(k)) \frac{\omega(k)\Im\tilde{g}(\omega(k) - {\rm i}0))}{|1 - t^2\tilde{g}(\omega(k) - {\rm i}0))^2} |^2 d^3k. \end{aligned}$$

Several remarks are in order:

If both reservoirs are condensed, i.e. $\rho_{0,1}$, and $\rho_{0,2}$ are both different from zero, the particle current shows a peculiar dependence on the phase difference. This is not true for the energy current, where the second term, coming from the expectations of the creation/annihilation operators does not contribute (as expected, as the $k = 0$ states carry no energy). Also, if $\rho_{0,1}\rho_{0,2} \neq 0$ and $\beta_1 = \beta_2$, then $n_1(k) = n_2(k)$, in which case the integral terms in the currents, representing the contribution of the excited states, vanish, therefore particles are exchanged only between the $k = 0$ states, and there is no energy flow.

In order to obtain the currents in the canonical state, we have still to integrate the expressions of the currents over the phases θ_i of the two condensates. This has the effect that the particle currents between the $k = 0$ states are averaged out, and only the first term in the expression of the particle current survives. In particular, there is no current if the temperatures are equal and either $\rho_1 = \rho_2 \leq \rho_{\rm cr}(\beta)$, or both densities are overcritical (irrespective of their values).

As a matter of fact, Proposition 4.3 implies that the presence of the condensates in the reservoirs has little influence on the currents, as long as one considers non-symmetry-breaking states. We conjecture that this holds true for more general junctions.

References

[1] ANGELESCU, N., BUNDARU, M., *On the transport between condensed phases*, J. Phys. A: Math. Theor. **40** (2007), pp. 5565–5573.

[2] ASCHBACHER, W., JAKŠIĆ, V., PAUTRAT, Y. and PILLET, C.-A., *Transport properties of quasi-free Fermions*, J. Math. Phys., **48** (2007), 032101.

[3] ASCHBACHER, W., JAKŠIĆ, V., PAUTRAT, Y. and PILLET, C.-A., *Topics in Non-Equilibrium Quantum Statistical Mechanics*, Lecture Notes in Mathematics, Vol. **1882**, pp. 1–66 (Springer, New York, 2006).

[4] BRATTELI, O. and ROBINSON, D. W., *Operator Algebras and Quantum Statistical Mechanics II* (Springer, New York, 1981).

[5] CANNON, J. T., *Infinite volume limits of the canonical free Bose gas on the Weyl algebra*, Comm. Math. Phys. **29** (1973), pp. 89–104.

[6] CORNEAN, H. D., JENSEN, A. and MOLDOVEANU, V., *A rigorous proof of the Landauer-Büttiker formula*, J. Math. Phys. **46** (2005), 042106.

[7] GAKHOV, F. D., *Boundary Value Problems* (Pergamon, New York, 1966).

[8] MERKLI, M., *Stability of equilibria with a condensate*, Comm. Math. Phys. **257** (2005), pp. 621–640.

[9] MERKLI, M., FRÖHLICH, J. *Another return to "return to equilibrium"*, Comm. Math. Phys. **251** (2004), pp. 235–262.

[10] MERKLI, M., MÜCK, M. and SIGAL, I. M. *Instability of equilibrium states for coupled heat reservoirs at different temperature*, J. Funct. Anal. **243** (2007), pp. 87–120.

[11] MERKLI, M., *The ideal quantum gas*, Lecture Notes in Mathematics, Vol. **1880** (Springer, New York, 2006).

[12] NENCIU, G., *Independent electron model for open quantum systems: Landauer-Büttiker formula and strict positivity of the entropy production*, arXiv:math-ph/0610074.

[13] PETZ, D., *An Invitation to the Algebra of Canonical Commutation Relations*, Leuven Notes in Mathematical and Theoretical Physics, **2** (Leuven Univ. Press, 1990).

[14] POULIN, P., *Green's function of generalized Laplacians*, CRM Proceedings and Lecture Notes **42** (American Mathematical Society, Providence, 2007).

[15] REED, M. and SIMON, B., *Methods of Modern Mathematical Physics*, Vols. **3**, **4** (Academic Press, New York, 1983).

[16] RUELLE, D., *Natural nonequilibrium states in quantum statistical mechanics* J. Statist. Phys. **98** (2000), pp. 57–75.

[17] YAFAEV, D. R., *Mathematical Scattering Theory: General Theory*, Translations of Mathematical Monographs **105** (American Mathematical Society, Providence, 1992).

Topics in Applied Mathematics & Mathematical Physics

An Introduction to Monotonicity Methods for Nonlinear Kinetic Equations

Cecil Pompiliu Grünfeld [1]

Contents

[1]Institute of Space Sciences & "Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: grunfeld@venus.nipne.ro

This work was supported by the research contract CEEX05-D11-06/03.10.2005 between the Institute of Space Sciences-Bucharest and the Institute of Atomic Physics, Bucharest.

1. Introduction

Many nonlinear kinetic equations for complex systems appear as generalization of the classical Boltzmann equation (see, e.g. [4]). The last years have been marked by an increased interest in the mathematical properties of such models. This can be explained by various applications not only in physics, astrophysics and chemistry (e.g. studies of simple and complex/reacting fluids, granular media, coagulation-fragmentation, formation of planetary rings, galaxy collision) but also in modeling evolution processes in immunology, traffic flow, communication networks, etc.

In many situations, the above equations are phenomenological or microscopic models that describe the evolution of various *populations* (macroscopic systems) of many well individualized, *objects* (e.g. rarefied gas particles, cells networks signals etc.) interacting among themselves. The interactions are (localized) *microscopic* processes: a) any interaction has a very short duration, with respect to the time-scale of the macroscopic evolution; b) the number of partners of any interaction is very small, with respect to the total number of the components of the population. Depending on the model, an interaction may change the state, nature and/or the number of the participants in interaction. This may result in modifications of the values of the physical quantities characterizing the states of the interacting objects. However, such modifications must be consistent with certain *balance* laws (e.g. conservation /dissipation laws) imposed by the peculiarities of the microscopic processes.

The problem of the existence and uniqueness of solutions of the above models is not only of an academic interest. Indeed, good criteria for the existence of general solutions and a detailed study of the properties of the solutions can be particularly useful in obtaining effective convergent numerical schemes for the models.

The above models present some mathematical properties, similar to those of the classical Boltzmann equation, in particular similar monotonicity properties (with respect to the order). This made possible to extend nontrivially monotonicity methods, initially introduced for the classical Boltzmann equation, [2] (see also [28]) to study these models [18], [27], [9], [7]. Recently the ideas of [2] and [28]) have been reconsidered nontrivially within a more general, abstract framework, [11], [12], [13]. The present work is a survey of the recent progress in the domain, and includes five sections and an Appendix. This Introduction is the first Section. The next Section, is a brief presentation, at formal level, of some relevant examples of Boltzmann models for complex systems. In Section 3, we introduce a class of abstract evolution

problems, as a generalization of the examples considered in Section 2. Then we develop the general existence theory based on monotonicity arguments. Section 4 is devoted to applications. Finally, Section 5 contains conclusions and open problems.

2. Boltzmann-like kinetic models

In this section we present several nonlinear models with nonlinear singularities, that exhibit similar isotonicity properties. In very general terms, these equations are essentially described by nonlinear evolution equations of the form

$$\frac{\mathrm{d}f}{\mathrm{d}t} = Af + Q(t,f), \quad t > 0, \tag{2.1}$$

formulated in the positive cone of some suitable ordered function space X, usually an ordered Banach space. The unknown $f = f(t)$ characterizes the state of the macroscopic system at time t. The two terms of the r.h.s. of Eq.(2.1), Af (possibly $A = 0$) and $Q(t,f)$ describe the free motion and the contribution of the interaction processes, respectively. From a mathematical point of view, A is the generator of a evolution linear group in X, while $Q(t,\cdot)$ is a nonlinear integral operator.

In many situations, we can write $Q(t,\cdot) = Q^+(t,\cdot) - Q^-(t,\cdot)$, where $Q^+(t,\cdot)$ and $Q^-(t,\cdot)$ are *positive* and *isotone* with respect to the order of X. Moreover, $Q^+(t,\cdot)$ and $Q^-(t,\cdot)$ satisfy certain relations -macroscopic balance laws- determined by the microscopic balance properties.

In this work we are interested in solving the initial value problem (i.v.p.) for Eq.(2.1), which can take various formulations, depending on the model.

2.1. Smoluchowski's coagulation equation

Smoluchowski's coagulation equation, [21, 25] (see also, e.g., [1], for a recent review), describes the irreversible evolution of particles that may coalesce into larger clusters. The continuous version of the Smoluchowski's equation reads

$$\frac{\partial}{\partial t} f = Q_c(f) = Q_c^+(f) - Q_c^-(f) \tag{2.2}$$

for the unknown $f(t,y) \geq 0$, the density of clusters of size $y \in \mathbb{R}_+ := [0,\infty)$ at time $t \geq 0$. Here

$$Q_c^+(g)(y) = \frac{1}{2}\int_0^y q(y-y_*,y_*)g(y-y_*)g(y_*)\mathrm{d}y_*, \tag{2.3}$$

$$Q_c^-(g)(y) = g(y)\int_0^\infty q(y,y_*)g(y_*)\mathrm{d}y_*, \tag{2.4}$$

with the (coagulation) kernel $q : \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ a symmetric, measurable function.

We assume that there exist the constants $q_0, q_1 \geq 0$ and $0 \leq \alpha \leq \beta$, such that

$$q(y,y_*) \leq q_0 + q_1(y^\alpha y_*^\beta + y^\beta y_*^\alpha) \quad (y,y_* \geq 0), \tag{2.5}$$

where

$$\alpha + \beta \leq 1. \tag{2.6}$$

Condition (2.5) includes the case when either $q_0 = 0$ or $q_1 = 0$. Without loss of generality, we can assume that $q_1 > 0$ (indeed the situation when q is bounded by a constant can be considered as a particularization of (2.5) to the case where $q_1 > 0$ and $\alpha = \beta = 0$).

The following property of the Smoluchowski's model is essential for our analysis. Formally, if $g, \psi : \mathbb{R}_+ \mapsto \mathbb{R}$ are measurable, then

$$\int_0^\infty \psi(y)\left[Q_c^+(g)(y) - Q_c^-(g)(y)\right]\mathrm{d}y =$$

$$= \frac{1}{2}\int_0^\infty\int_0^\infty \widetilde{\psi}(y,y_*)q(y,y_*)g(y)g(y_*)\mathrm{d}y\mathrm{d}y_*, \tag{2.7}$$

(provided that the integrals exist), where

$$\widetilde{\psi}(y,y_*) := \psi(y+y_*) - \psi(y) - \psi(y_*). \tag{2.8}$$

Property (2.7) follows from the change of variables $(y,y_*) \to (y-y_*,y_*)$ in the first term of the l.h.s. of (2.7), and then applying Fubini's theorem.

In particular, if $\psi(y) = y$ in (2.7), then

$$\int_0^\infty Q_c(g)(y)y\mathrm{d}y = 0. \tag{2.9}$$

This gives formally the mass conservation for Eq. (2.2).

Similar considerations as before can be made for the discrete version of the Smoluchowski equation

$$\dot{c}_j = \frac{1}{2}\sum_{k=1}^{j-1} Q_{j-k,k}(c(t)) - \sum_{k=1}^{\infty} Q_{j,k}(c(t)), \quad c_j(0) = c_{j,0} \geq 0 \quad (j = 1, 2, ...), \tag{2.10}$$

where $Q_{j,k}(c) := q(k,j)c_k c_j$, is defined by the same symmetric coagulation kernel introduced before, subject to (2.5), (2.6), and the component $c_j(t) \geq 0$ of $c(t) := (c_j(t))$ is interpreted as the concentration of clusters of size j at time $t \geq 0$.

2.2. Povzner-like model with dissipative collisions

The model describes a rarefied mono-component fluid of particles of unit mass, evolving in the free space with dissipative (conservative) binary collisions, i.e., collisions resulting in the loss (conservation) of the kinetic energy of the encounters.

According to the model, [7], the post-collision velocities $\mathbf{v}'$, $\mathbf{w}'$ are related to the pre-collision velocities $\mathbf{v}$ and $\mathbf{w}$ by

$$\mathbf{v}' = \mathbf{v} - (1-\beta(\mathbf{n}))\langle \mathbf{v}-\mathbf{w}, \mathbf{n}\rangle \mathbf{n}, \quad \mathbf{w}' = \mathbf{w} + (1-\beta(\mathbf{n}))\langle \mathbf{v}-\mathbf{w}, \mathbf{n}\rangle \mathbf{n}, \tag{2.11}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean product in $\mathbb{R}^3$ and $\mathbf{n} \in \Omega$ - the unit sphere in $\mathbb{R}^3$. Here, $\beta : \Omega \mapsto [0, 1/2)$ is a given measurable function. The total momentum is conserved in collisions, $\mathbf{v}' + \mathbf{w}' = \mathbf{v} + \mathbf{w}$, but the kinetic energy is lost

$$|\mathbf{v}'|^2 + |\mathbf{w}'|^2 = |\mathbf{v}|^2 + |\mathbf{w}|^2 - 2\beta(\mathbf{n})(1-\beta(\mathbf{n}))\,|\langle \mathbf{v}-\mathbf{w}, \mathbf{n}\rangle|^2, \tag{2.12}$$

excepting the case $\beta = 0$, when the collisions become elastic.

For each fixed $\mathbf{n} \in \Omega$, the transformation $\mathbb{R}^3 \times \mathbb{R}^3 \ni (\mathbf{v},\ \mathbf{w}) \mapsto (\mathbf{v}', \mathbf{w}') \in \mathbb{R}^3 \times \mathbb{R}^3$ is invertible. The inversion formulae are

$$\hat{\mathbf{v}} = \mathbf{v} - \left(\frac{1-\beta(\mathbf{n})}{1-2\beta(\mathbf{n})}\right)\langle \mathbf{v}-\mathbf{w}, \mathbf{n}\rangle \mathbf{n}, \quad \hat{\mathbf{w}} = \mathbf{w} + \left(\frac{1-\beta(\mathbf{n})}{1-2\beta(\mathbf{n})}\right)\langle \mathbf{v}-\mathbf{w}, \mathbf{n}\rangle \mathbf{n}. \tag{2.13}$$

Formally the above model reads

$$\frac{\partial}{\partial t} f = -\mathbf{v}\cdot\nabla_{\mathbf{x}} f + Q_d^+(f) - Q_d^-(f) \tag{2.14}$$

where $f = f(t, \mathbf{x}, \mathbf{v})$ is the one-particle distribution function, depending on time $t \geq 0$, position $\mathbf{x} \in \mathbb{R}^3$, and velocity $\mathbf{v} \in \mathbb{R}^3$ of the so-called test particle,

Q_d^+ and Q_d^- are the so-called nonlinear gain and loss operators, respectively, and describe the influence of the collisions on the evolution of f. They are formally given by

$$Q_d^+(g)(\mathbf{x},\mathbf{v}) =$$

$$= \int_0^R \mathrm{d}r \int_{\Omega\times\mathbb{R}^3} \frac{|\langle \mathbf{n}, \mathbf{v}-\mathbf{w}\rangle|^\gamma}{(1-2\beta(\mathbf{n}))^{1+\gamma}} P(r,\mathbf{n}) g(\mathbf{x},\hat{\mathbf{v}}) g(\mathbf{x}+r\mathbf{n},\hat{\mathbf{w}}) \mathrm{d}\mathbf{n}\,\mathrm{d}\mathbf{w} \quad (2.15)$$

and

$$Q_d^-(g)(\mathbf{x},\mathbf{v}) = g(\mathbf{x},\mathbf{v}) \int_0^R \mathrm{d}r \int_{\Omega\times\mathbb{R}^3} |\langle \mathbf{n}, \mathbf{v}-\mathbf{w}\rangle|^\gamma P(r,\mathbf{n}) g(\mathbf{x}+r\mathbf{n},\mathbf{w}) \mathrm{d}\mathbf{n}\,\mathrm{d}\mathbf{w}, \quad (2.16)$$

respectively, where $P : \mathbb{R}_+ \times \Omega \mapsto \mathbb{R}_+$ is a given measurable function with $P(r,\mathbf{n}) = P(r,-\mathbf{n})$ assumed to satisfy

$$P(r,\mathbf{n}) \le c_0 r^2 \quad (r \ge 0,\ \mathbf{n} \in \Omega), \quad (2.17)$$

for some constants $c_0 > 0$, $0 \le \gamma \le 1$, and $R > 0$, specific to the collision processes.

The basic property of the model is the formal identity

$$\int_{\mathbb{R}^3} \psi(\mathbf{v}) \left[Q_d^+(g) - Q_d^-(g)\right] \mathrm{d}\mathbf{v} =$$

$$= \int_{\Omega\times\mathbb{R}^3\times\mathbb{R}^3} \widetilde{\psi}(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}') \frac{|\langle \mathbf{n}, \mathbf{w}-\mathbf{v}\rangle|^\gamma}{2} P(r,\mathbf{n}) g(\mathbf{x},\mathbf{v}) g(\mathbf{x}+r\mathbf{n},\mathbf{w}) \mathrm{d}\mathbf{n}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}, \quad (2.18)$$

where $\psi : \mathbb{R}^3 \mapsto \mathbb{R}$ and $g : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ are measurable functions such that (2.18) is well defined, and

$$\widetilde{\psi}(\mathbf{v},\mathbf{w},\mathbf{v}',\mathbf{w}') := \psi(\mathbf{v}') + \psi(\mathbf{w}') - \psi(\mathbf{v}) - \psi(\mathbf{w}), \quad (2.19)$$

with $\mathbf{v}'$ and $\mathbf{w}'$ given by (2.11). We deduce easily (2.18), performing the change of variable $(v,w) \to (\hat{v},\hat{w})$ in the first term of the l.h.s (2.18).

If $\beta \equiv 0$, then (2.14) yields a version of the so-called generalized Boltzmann equation with binary elastic (conservative) collisions, analyzed in [3].

2.3. Povzner-like model with chemical reactions

We recall here a Povzner-like model with chemical reactions introduced in [8] for a reacting gas mixture of N species A_i and mass m_i, $1 \le i \le N$, without interaction with photon fields. We assume binary reactions

$$A_i + A_j \to A_k + A_l, \quad 1 \le i,\, j,\, k,\, l \le N, \quad (2.20)$$

where case $i = j = k = l$ corresponds to non-reactive (elastic) processes. According to the model of [8], for each species i, the gas particles have one internal energy state, say $E_i \geq 0$, $1 \leq i \leq N$. It is assumed that the reactions are consistent with the conservation of mass, momentum and total energy, i.e., $m_i + m_j = m_k + m_l$, and $m_i\mathbf{v} + m_j\mathbf{w} = m_k\mathbf{v}' + m_l\mathbf{w}'$, as well as

$$\frac{m_i |\mathbf{v}|^2}{2} + E_i + \frac{m_j |\mathbf{w}|^2}{2} + E_j = \frac{m_k |\mathbf{v}'|^2}{2} + E_k + \frac{m_l |\mathbf{w}'|^2}{2} + E_l, \tag{2.21}$$

where $(\mathbf{v}, \mathbf{w})$ are the pre-reaction velocities of the particles (i, j) and $(\mathbf{v}', \mathbf{w}')$ are the post-reaction velocities of the particles (k, l)

The conservation relations give

$$\frac{m_k m_l |\mathbf{v}' - \mathbf{w}'|^2}{2(m_k + m_l)} = \frac{m_i m_j |\mathbf{v} - \mathbf{w}|^2}{2(m_i + m_j)} + E_i + E_j - E_k - E_l := t_{kl,ij}(\mathbf{v}, \mathbf{w}) \tag{2.22}$$

and obviously, (2.20) occurs, provided that

$$t_{kl,ij}(\mathbf{v}, \mathbf{w}) \geq 0. \tag{2.23}$$

It can be easily seen that $(\mathbf{v}', \mathbf{w}')$ can be represented in terms of the pre-reaction velocities $(\mathbf{v},\ \mathbf{w})$ and of the unit vector $\mathbf{n} = (\mathbf{v}' - \mathbf{w}') |\mathbf{v}' - \mathbf{w}'|^{-1}$ as

$$\mathbf{v}' = \frac{m_i\mathbf{v} + m_j\mathbf{w}}{m_i + m_j} + \frac{2^{1/2}(m_l)^{1/2}}{m_k^{1/2}(m_i + m_j)^{1/2}} t_{kl,ij}(\mathbf{v}, \mathbf{w})^{1/2}\mathbf{n} := \mathbf{v}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) \tag{2.24}$$

and

$$\mathbf{w}' = \frac{m_i\mathbf{v} + m_j\mathbf{w}}{m_i + m_j} - \frac{2^{1/2}(m_k)^{1/2}}{m_l^{1/2}(m_i + m_j)^{1/2}} t_{kl,ij}(\mathbf{v}, \mathbf{w})^{1/2}\mathbf{n} := \mathbf{w}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) \tag{2.25}$$

It is convenient to extend the definitions of $\mathbf{v}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n})$ and $\mathbf{w}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n})$ by setting

$$\mathbf{v}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = \mathbf{w}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = \frac{m_i\mathbf{v} + m_j\mathbf{w}}{m_i + m_j} \tag{2.26}$$

whenever $t_{kl,ij}(\mathbf{v}, \mathbf{w}) < 0$. By virtue of the above formulae one has

$$\mathbf{v}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = \mathbf{v}_{kl,ji}(\mathbf{w}, \mathbf{v}, \mathbf{n}) = \mathbf{w}_{lk,ij}(\mathbf{v}, \mathbf{w},\ -\mathbf{n}) \tag{2.27}$$

and

$$\mathbf{w}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}) = \mathbf{w}_{kl,ji}(\mathbf{w}, \mathbf{v}, \mathbf{n}) = \mathbf{v}_{lk,ij}(\mathbf{v}, \mathbf{w},\ -\mathbf{n}). \tag{2.28}$$

Each species $1 \leq i \leq N$ is described by the one-particle distribution function $f_i = f_i(t, \mathbf{x}, \mathbf{v})$ depending on time $t \geq 0$, position $\mathbf{x}$ and velocity $\mathbf{v}$.

Assuming molecular chaos and (instant) point localized reactions, the kinetic model is derived following the original argument for the classical Boltzmann equation. The obtained model reads, [8],

$$\frac{\partial}{\partial t} f_i = -\mathbf{v} \cdot \nabla_{\mathbf{x}} f_i + Q_i^+(f) - Q_i^-(f), \quad 1 \leq i \leq N, \tag{2.29}$$

where $f = (f_1, ..., f_N)$ and, formally,

$$Q_i^+(g)(\mathbf{x}, \mathbf{v}) =$$

$$= \sum_{j,k,l=1}^{N} \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{S}^2} p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) g_k(t, \mathbf{x}, \mathbf{v}_{kl,ij}) g_l(t, \mathbf{x} + \mathbf{y}, \mathbf{w}_{kl,ij}) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}, \tag{2.30}$$

$$Q_i^-(g)(\mathbf{x}, \mathbf{v}) =$$

$$= \sum_{j,k,l=1}^{N} \int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \Omega} r_{kl,ij}(\mathbf{y}, \ \mathbf{v}, \mathbf{w}, \mathbf{n}) g_i(t, \mathbf{x}, \mathbf{v}) g_j(t, \mathbf{x} + \mathbf{y}, \mathbf{w}) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}. \tag{2.31}$$

Here, $g := (g_1, ...g_N)$ with $g_i : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_+$, $\Omega := \{\mathbf{n} \in \mathbb{R}^3 : |\mathbf{n}| = 1\}$, $g_k(\cdot, \cdot, \mathbf{v}_{kl,ij}) = g_k(\cdot, \cdot, \mathbf{v}_{kl,ij}(\mathbf{v}, \mathbf{w})$, $g_l(\cdot, \cdot, \mathbf{w}_{kl,ij}) = g_l(\cdot, \cdot, \mathbf{w}_{kl,ij}(\mathbf{v}, \mathbf{w}, \mathbf{n}))$. Moreover, $p_{kl,ij}, r_{kl,ij} : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \times \Omega \to [0, \infty)$, are given measurable maps with the property that if $(\mathbf{v}, \mathbf{w}) \notin \mathcal{D}_{ij,kl} := \{(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^3 \times \mathbb{R}^3 : t_{ij,kl}(\mathbf{v}, \mathbf{w}) \geq 0\}$, then

$$p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = 0. \tag{2.32}$$

One assumes that the following properties are satisfied a.e.:

$$p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = 0 \quad (\mathbf{y} > R), \tag{2.33}$$

$$p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = p_{kl,ij}(-\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}),$$

$$r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = r_{kl,ij}(-\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}), \tag{2.34}$$

$$p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = p_{kl,ji}(\mathbf{y}, \mathbf{w}, \mathbf{v}, \mathbf{n}) = p_{lk,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, -\mathbf{n}), \tag{2.35}$$

$$r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) = r_{kl,ji}(\mathbf{y}, \mathbf{w}, \mathbf{v}, \mathbf{n}) = r_{lk,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, -\mathbf{n}). \tag{2.36}$$

Moreover,

$$\int_{\mathbb{R}^3 \times \mathbb{R}^3 \times \Omega} \varphi(\mathbf{v}, \mathbf{w}) p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) \psi(\mathbf{v}_{kl,ij}, \mathbf{w}_{kl,ij}) \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} =$$

$$= \int_{\mathbb{R}^3\times\mathbb{R}^3\times\Omega} \varphi(\mathbf{v}_{ij,kl}, \mathbf{w}_{ij,kl}) r_{ij,kl}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) \psi(\mathbf{v}, \mathbf{w}) \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} \tag{2.37}$$

for all $(\psi, \varphi) : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, provided that whichever side of (2.37) is defined.

The kernels $p_{kl,ij}, r_{kl,ij} : \mathbb{R}^3 \times \mathbb{R}^3 \times \Omega \to [0, \infty)$ carry the information of the reaction processes. For a gas composed by one species of particles with elastic collisions, the above system of equations reduces to the so-called generalized Boltzmann equation.

Our main hypothesis is as follows:

Assumption 2.1 *There exist constants $c_q > 0$ and $0 \leq q \leq 1$ such that*

$$\int_\Omega r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) \mathrm{d}\mathbf{n} \leq c_q \left[1 + |\mathbf{v}|^2 + |\mathbf{w}|^2\right]^q. \tag{2.38}$$

Observe that since $r_{kl,ij}$ and $p_{kl,ij}$ are related by (2.37), then the above hypothesis is also an implicit condition on $p_{kl,ij}$.

Under Assumption (2.38), one can show that, at least, formally,

$$\sum_{i=1}^N \int_{\mathbb{R}^3\times\mathbb{R}^3} [Q_i^+(g)(\mathbf{x}, \mathbf{v}) - Q_i^-(g)(\mathbf{x}, \mathbf{v})] h_i(\mathbf{x}, \mathbf{v}) \mathrm{d}\mathbf{v}\mathrm{d}\mathbf{x} =$$

$$= \frac{1}{4} \sum_{i,j,k,l=1}^N \int_{\mathcal{D}} [p_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) g_k(\mathbf{x}, \mathbf{v}_{kl,ij}) g_l(\mathbf{x} + \mathbf{y}, \mathbf{w}_{kl,ij})$$

$$- r_{kl,ij}(\mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{n}) g_i(\mathbf{x}, \mathbf{v}) g_j(\mathbf{x} + \mathbf{y}, \mathbf{w})]$$

$$\times [h_i(\mathbf{x}, \mathbf{v}) + h_j(\mathbf{x} + \mathbf{y}, \mathbf{w}) - h_k(\mathbf{x}, \mathbf{v}_{kl,ij}) - h_l(\mathbf{x} + \mathbf{y}, \mathbf{w}_{kl,ij})] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} \tag{2.39}$$

for all $g=(g_1, ...g_N)$ and $h=(h_1, ...h_N)$, with $g_i, h_i \geq 0$, for which the integrals are defined. Here, $\mathcal{D} := \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \times \Omega$. The last property follows by applying (2.27), (2.28), (2.32)–(2.37), as well as the invariance properties of the sums in (2.39), with respect to the change of variables $(\mathbf{x}, \mathbf{y}, \mathbf{n}) \to (\mathbf{x}', \mathbf{y}', \mathbf{n}') := (\mathbf{x} + \mathbf{y}, -\mathbf{y}, -\mathbf{n})$, and a suitable interchanges of summation indices.

At least, at formal level, property (2.39) implies the bulk conservation for mass, momentum, and total energy,

$$\sum_{i=1}^N \int_{\mathbb{R}^3\times\mathbb{R}^3} \Psi_i^{(j)}(\mathbf{x}, \mathbf{v}) f_i(t, \mathbf{x}, \mathbf{v}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} = \sum_{i=1}^N \int_{\mathbb{R}^3\times\mathbb{R}^3} \Psi_i^{(j)}(\mathbf{x}, \mathbf{v}) f_i(0, \mathbf{x}, \mathbf{v}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} \tag{2.40}$$

$(0 \leq j \leq 4)$, where $f_i(t)$ are the components of the solution f of Eq. (2.29), and

$$\Psi_i^{(0)}(\mathbf{x}, \mathbf{v}) := m_i, \quad \Psi_i^{(4)}(\mathbf{x}, \mathbf{v}) := m_i |\mathbf{v}|^2 /2 + E_i, \quad \Psi_i^{(j)}(\mathbf{x}, \mathbf{v}) := m_i v_j \tag{2.41}$$

$(j = 1, 2, 3)$, with v_j are the components of $\mathbf{v}$.

2.4. A model with inelastic collisions and chemical reactions

In this example, we consider an abstract system of a Boltzmann-like phenomenological equations, [9, 10, 14], for a multi-component reacting gas of particles with internal states and discrete values of the internal energy. Thinking a real gas mixture of particles with internal structure as a mixture of several chemical species of mass points with unique internal state, one can assume that any gas particle of the model has only one internal state. Specifically, the model refers to a gas consisting of N chemical species. A particle of species $n = 1, 2, ..., N$ is characterized by mass $m_n > 0$ and internal energy E_n. Without loss of generality, one can assume that $E_n \geq 0$, $1 \leq n \leq N$. It is assumed that the chemical reactions are induced by inelastic (possibly) multi-body, instant collisions. A reaction is identified with a couple $(\alpha, \beta) \in \mathcal{M} \times \mathcal{M}$, where $\mathcal{M} := \{\gamma = (\gamma_n)_{1 \leq n \leq N} \mid \gamma_n \in \{0, 1, \ldots, K\}\}$ is a multi-index set. Here $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathcal{M}$ and $\beta = (\beta_1, \ldots, \beta_N) \in \mathcal{M}$ designate the pre-collision and post-collision channels, respectively, with $0 \leq \alpha_n, \beta_n \leq K$ participants of species n; $1 \leq n \leq N$. Any couple of the form $(\gamma, \gamma) \in \mathcal{M} \times \mathcal{M}$ is identified with a multi-body elastic collision with γ_n collision partners of species n; $1 \leq n \leq N$. The number of particles in some channel $\gamma \in \mathcal{M}$ is $|\gamma| := \sum_{i=1}^{N} \gamma_i$. The family of chemical species participating in channel γ is denoted by $\mathcal{N}(\gamma) := \{i : \gamma_i > 0, 1 \leq i \leq N\}$.

Let M_γ, $\mathbf{V}_\gamma(\mathbf{w})$ and $W_\gamma(\mathbf{w})$ denote the total mass, velocity of the mass center and total energy, respectively, for the particles in channel γ, i.e.,

$$M_\gamma := \sum_{i=1}^{N} \gamma_i m_i, \tag{2.42}$$

$$\mathbf{V}_\gamma(\mathbf{w}) := \frac{1}{M_\gamma} \sum_{i \in \mathcal{N}(\gamma)} \sum_{j=1}^{\gamma_i} m_i \mathbf{w}_{i,j}, \tag{2.43}$$

$$W_\gamma(\mathbf{w}) := \sum_{i \in \mathcal{N}(\gamma)} \sum_{j=1}^{\gamma_i} (2^{-1} m_i \mathbf{w}_{i,j}^2 + E_i), \tag{2.44}$$

where $\mathbf{w} = ((\mathbf{w}_{k,i})_{i\in\{1,\dots,\alpha_k\}})_{k\in\mathcal{N}(\gamma)}$ represents the ensemble of velocities of the particles in channel γ. Then, the kinetic energy of the particles (with velocities $\mathbf{w}$) in channel γ, relative to the frame of the mass center, reads

$$W_{r,\gamma}(\mathbf{w}) = W_\gamma(\mathbf{w}) - \frac{M_\gamma \mathbf{V}_\gamma(\mathbf{w})^2}{2} - \sum_{i=1}^{N} \gamma_i E_i. \tag{2.45}$$

Obviously, $W_{r,\gamma}(\mathbf{w}) \geq 0$.

A gas reaction (α, β) may take place only if it is consistent with the conservation of mass, momentum and energy, i.e.,

$$M_\alpha = M_\beta, \quad \mathbf{V}_\alpha(\mathbf{w}) = \mathbf{V}_\beta(\mathbf{u}), \quad W_\alpha(\mathbf{w}) = W_\beta(\mathbf{u}). \tag{2.46}$$

We will assume here that elastic collisions are always present. Therefore, the set $\mathcal{C}_M := \{(\alpha, \beta) \in \mathcal{M} \times \mathcal{M} : M_\alpha = M_\beta\}$ is nonempty.

The Boltzmann-like system of equations for the above model is

$$\frac{\partial}{\partial t} f_i = Q_i^+(f) - Q_i^-(f). \tag{2.47}$$

Here the unknown $f_i : \mathbb{R}_+ \times \mathbb{R}^3 \mapsto \mathbb{R}_+$ is the one particle distribution functions $f_i = f_i(t, \mathbf{v})$ (t-time, $\mathbf{v}$-velocity) of the particles of species $1 \leq i \leq N$. In Eq. (2.47), $Q_i^+(f)$ and $Q_i^-(f)$, with $f := (f_1, \dots, f_N)$, are the so-called loss and gain (nonlinear) operators for the particles of species i, respectively. Formally,

$$Q_i^+(g)(\mathbf{v}) = \sum_{\alpha,\beta\in\mathcal{M}} \alpha_i \int\limits_{\mathbb{R}^{3|\alpha|-3}\times\Omega_\beta} \Big[p_{\beta,\alpha}(\mathbf{w}, \mathbf{n})(g^\beta \circ \mathbf{u}_{\beta,\alpha})(\mathbf{w}, \mathbf{n})\Big]_{\mathbf{w}_{i,\alpha_i}=\mathbf{v}} \mathrm{d}\tilde{\mathbf{w}}_i \mathrm{d}\mathbf{n}, \tag{2.48}$$

$$Q_i^-(g)(\mathbf{v}) = \sum_{\alpha,\beta\in\mathcal{M}} \alpha_i \int\limits_{\mathbb{R}^{3|\alpha|-3}\times\Omega_\beta} [r_{\beta,\alpha}(\mathbf{w}, \mathbf{n}) g^\alpha(\mathbf{w})]_{\mathbf{w}_{i,\alpha_i}=\mathbf{v}} \, \mathrm{d}\tilde{\mathbf{w}}_i \mathrm{d}\mathbf{n}, \tag{2.49}$$

where

$$g^\gamma(\mathbf{w}) := \prod_{i\in\mathcal{N}(\gamma)} \prod_{j=1}^{\gamma_i} g_i(\mathbf{w}_{i,j}), \quad \gamma \in \mathcal{M}, \tag{2.50}$$

Ω_γ is the unit sphere in $\mathbb{R}^{3|\gamma|-3}$, with $\gamma \in \mathcal{M}$, and $\mathrm{d}\tilde{\mathbf{w}}_i$ is the Euclidean element of area on $\big\{\mathbf{w} \in \mathbb{R}^{3|\alpha|} \mid \mathbf{w}_{i,\alpha_i} = \mathbf{v}\big\}$. Here, the functions $\mathbf{u}_{\beta,\alpha} \in C(\mathbb{R}^{3|\alpha|} \times \Omega_\beta; \mathbb{R}^{3|\beta|})$, and the measurable functions $r_{\beta,\alpha}$, $p_{\beta,\alpha} : \mathbb{R}^{3|\alpha|} \times \Omega_\beta \mapsto \mathbb{R}_+$ are given.

The following conditions are assumed ([9, 11, 14]):

(B_1) $r_{\beta,\alpha} = p_{\beta,\alpha} = 0$ unless: $|\alpha| \geq 2$, $|\beta| \geq 2$, $(\alpha, \beta) \in \mathcal{C}_M$, and $\mathbf{w} \in D^+_{\beta,\alpha} := \left\{ \mathbf{w}' \in \mathbb{R}^{3|\alpha|} : W_{r,\alpha}(\mathbf{w}') + \sum_{i=1}^{N} (\alpha_i - \beta_i) E_i \geq 0 \right\}$.

(B_2) For each $i \in \mathcal{N}(\alpha)$ fixed, $p_{\beta,\alpha}(\mathbf{w}, \mathbf{n})$, $r_{\beta,\alpha}(\mathbf{w}, \mathbf{n})$, and $u_{\beta,\alpha}(\mathbf{w})$ are invariant with respect to the interchange of the components $\mathbf{w}_{i,1}, ..., \mathbf{w}_{i,\alpha_i}$ of $\mathbf{w}$.

(B_3) If $(\alpha, \beta) \in \mathcal{C}_M$, $\mathbf{w} \in D^+_{\beta,\alpha}$, then

$$(V_\beta \circ \mathbf{u}_{\beta,\alpha})(\mathbf{w}, \mathbf{n}) = V_\alpha(\mathbf{w}), \quad (W_\beta \circ \mathbf{u}_{\beta,\alpha})(\mathbf{w}, \mathbf{n}) = W_\alpha(\mathbf{w}), \tag{2.51}$$

for all $\mathbf{n} \in \Omega_\beta$, and

$$\int\limits_{\mathbb{R}^{3|\alpha|} \times \Omega_\beta} p_{\beta,\alpha}(\mathbf{w}, \mathbf{n}) \varphi(\mathbf{w}, \mathbf{n}) (\psi \circ \mathbf{u}_{\beta,\alpha})(\mathbf{w}, \mathbf{n}) \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} =$$

$$= \int\limits_{\mathbb{R}^{3|\beta|} \times \Omega_\alpha} r_{\alpha,\beta}(\mathbf{w}, \mathbf{n}) (\varphi \circ \mathbf{u}_{\alpha,\beta})(\mathbf{w}, \mathbf{n}) \psi(\mathbf{w}, \mathbf{n}) \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}, \tag{2.52}$$

for all $\varphi : \mathbb{R}^{3|\alpha|} \mapsto \mathbb{R}$ and $\psi : \mathbb{R}^{3|\beta|} \mapsto \mathbb{R}$, for which the integrals are well defined.

We suppose that the reactions are reversible, i.e., if $r_{\beta,\alpha} \neq 0$ for some (α, β), then also $r_{\alpha,\beta} \neq 0$.

From (3.9), it follows that $p_{\beta,\alpha}$ and $r_{\beta,\alpha}$ are related one to another. Indeed, a more explicit relationship between $p_{\beta,\alpha}$ and $r_{\beta,\alpha}$ can be derived, as it results from a general example constructed in [9, 14]. Note also here that if one assumes a mono-component gas of particles with binary elastic collisions (i.e., $N = 1$, $K = 2$, and $p_{\beta,\alpha} = r_{\beta,\alpha} = 0$ unless $\alpha = \beta = (1, 1)$), then Eq. (2.47) reduces to the space homogeneous classical Boltzmann equation

$$\frac{\partial}{\partial t} f = Q^+(f) - Q^-(f), \tag{2.53}$$

where

$$Q^+(f)(\mathbf{v}) = \int\limits_{\mathbb{R}^3 \times \Omega} q(\mathbf{v}, \mathbf{w}, \mathbf{n}) f(\mathbf{v}') f(\mathbf{w}') \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}, \tag{2.54}$$

$$Q^-(f)(\mathbf{v}) = \int\limits_{\mathbb{R}^3 \times \Omega} q(\mathbf{v}, \mathbf{w}, \mathbf{n}) f(\mathbf{v}) f(\mathbf{w}) \mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}. \tag{2.55}$$

The notations are $f = f(t, \mathbf{v})$ – the one-particle distribution function, $\mathbf{v}' = \mathbf{v} - \langle \mathbf{v} - \mathbf{w}, \mathbf{n} \rangle \mathbf{n}$, $\mathbf{w}' = \mathbf{w} + \langle \mathbf{v} - \mathbf{w}, \mathbf{n} \rangle \mathbf{n}$, and $\mathbf{n} \in \Omega$ – the unit sphere in $\mathbb{R}^3$. Here, the Boltzmann collision law q is a positive measurable function (depending, in our case, on $\mathbf{v}$ and $\mathbf{w}$ through the variable $\mathbf{v} - \mathbf{w}$).

The last condition of the model concerns the behavior of $r_{\beta,\alpha}$ (see [9]):

Assumption 2.2 *There are some constants* $0 \leq q \leq 1$ *and* $c_q > 0$ *such that*

$$\nu_{\beta,\alpha}(\mathbf{w}) := \int_{\Omega_\beta} r_{\beta,\alpha}(\mathbf{w}, \mathbf{n}) \mathrm{d}\mathbf{n} \leq c_q (1 + W_\alpha(\mathbf{w}))^q \quad (\mathbf{w} \in \mathbb{R}^{|\alpha|}, \, a.e.), \quad (2.56)$$

for all $\alpha, \beta \in \mathcal{M}$.

Obviously, $\nu_{\beta,\alpha}(\mathbf{w}) = 0$, unless $(\alpha, \beta) \in \mathcal{C}_M$.

A consequence of (B_1), (B_2) and (2.56) is the key equality

$$\sum_{i=1}^{N} \int_{\mathbb{R}^3} \Psi_i^{(j)}(\mathbf{v}) \left[Q_i^+(g)(\mathbf{v}) - Q_i^-(g)(\mathbf{v}) \right] \mathrm{d}\mathbf{v} = 0 \quad (0 \leq j \leq 4), \quad (2.57)$$

for all $g = (g_1, ..., g_N)$ with $(1 + |\mathbf{v}|^2)^{1+q} g_i \in L^1(\mathbb{R}^3; \mathrm{d}\mathbf{v})$, $i = 1, 2, ..., N$. Here,

$$\Psi_i^{(0)}(\mathbf{v}) := m_i, \quad \Psi_i^{(4)}(\mathbf{v}) := \frac{1}{2} m_i |\mathbf{v}|^2 + E_i, \quad \Psi_i^{(j)}(\mathbf{v}) := m_i v_j \quad (1 \leq i \leq N), \quad (2.58)$$

where v_j is the j-component, $j = 1, 2, 3$, of $\mathbf{v}$. Equality (2.57) implies, at lest formally, the bulk conservation of mass, momentum and total energy.

2.5. A nonlinear von Neumann-Boltzmann equation

Besides classical models, we can also consider "quantum" kinetic models with monotonicity properties similar to classical ones.

Let $X = \mathcal{T}(\mathcal{H})$ be the space of trace class selfadjoint operators in some separable Hilbert space $\mathcal{H}$. On X, we consider the order $F \leq G$ iff $(f, Ff) \leq (f, Gf)$, $\forall f \in \mathcal{D}(F) \cap \mathcal{D}(G)$. Let $\|F\| := Tr(|F|)$ be the norm on X.

For some orthogonal base $\{e_0, e_1, ...\} \subset \mathcal{H}$, define the selfadjoint operator

$$H = \sum_{i \geq 0} \mu_i (e_i, \cdot) e_i, \quad (2.59)$$

where $\{\mu_n\}_n \subset \mathbb{R}$. Let $\{U^t\}_{t\in\mathbb{R}}$ denote the continuous group of positive isometries on X, given by $U^t(F) := \exp(-\mathrm{i}Ht)F\exp(\mathrm{i}Ht)$, $\mathrm{i} = \sqrt{-1}$. Consider a second sequence, $0 \leq \lambda_0 < \lambda_1 < \lambda_2 \leq \ ... \ \lambda_{n-1} \leq \lambda_n \ ... \ \nearrow \infty$, as $n \to \infty$. Let $\left\{V^t\right\}_{t\geq 0}$ be the C_0 semigroup on X, defined by

$$(e_i, V^t(F)e_j) := (V^t(F))_{i,j} = \exp[-(1+\lambda_i\delta_{i,j})t]F_{i,j} \tag{2.60}$$

where $F_{i,j} := (e_i, Fe_j)$, and let the infinitesimal generator of $\left\{V^t\right\}_{t\geq 0}$ be denoted by $(-\Lambda)$. Then

$$(\Lambda)_{i,j}(F) := (1+\lambda_i\delta_{i,j})F_{i,j}, \tag{2.61}$$

hence $\Lambda \geq \mathbb{I}$. Clearly, U^t leaves $\mathcal{D}(\Lambda) \cap X_+$ invariant and $U^t\Lambda = \Lambda U^t$ on $\mathcal{D}(\Lambda) \cap X_+$.

Now we can consider the following example of nonlinear von Neumann-Boltzmann equation X (see also [12]):

$$\frac{\mathrm{d}F}{\mathrm{d}t} + \mathrm{i}[H, F] = Q^+(F) - Q^-(F) \tag{2.62}$$

with $Q^{\pm} : \mathcal{D}(\Lambda) \subset X \to X$ given by

$$Q^-(F) := F_{0,0}\mathrm{Tr}(\Lambda F)(\sum_{i=0}^{2} P_i), \tag{2.63}$$

and

$$Q^+(F) := Q^-(F) + L(F), \tag{2.64}$$

where $P_i := (e_i, \cdot)e_i$ and

$$L(F) := F_{0,0}\mathrm{Tr}(\Lambda F)(\sum_{i=0}^{2} \varepsilon_i P_i). \tag{2.65}$$

Here, $\varepsilon_0 = \varepsilon\,(\lambda_1-\lambda_0)^{-1}(\lambda_2-\lambda_0)^{-1}$, $\varepsilon_1 = -\varepsilon(\lambda_1-\lambda_0)^{-1}\,(\lambda_2-\lambda_1)^{-1}$, $\varepsilon_2 = \varepsilon(\lambda_2-\lambda_0)^{-1}\,(\lambda_2-\lambda_1)^{-1}$ and $0 < \varepsilon < (\lambda_0-\lambda_1)\,(\lambda_0-\lambda_2)$. Thus $Q^{\pm}$ are positive operators, and a simple computation gives

$$\mathrm{Tr}Q^+(F) = \mathrm{Tr}Q^-(F) \tag{2.66}$$

for $0 \leq F \in \mathcal{D}(\Lambda)$, and

$$\mathrm{Tr}(\Lambda Q^+)(F) = \mathrm{Tr}(\Lambda Q^-)(F) \tag{2.67}$$

for $0 \leq F \in \mathcal{D}(\Lambda^2)$, so that both $\mathrm{Tr}F(t)$ and $\mathrm{Tr}(\Lambda F)(t)$ remain constant with time.

3. General theory

3.1. A monotonicity result for the classical Boltzmann equation

Before proceeding to a more general analysis, we start with a relevant example - the Arkeryd's monotonicity result for the Boltzmann equation ([2]).

Specifically, in [2], the main interest is to solve the Cauchy problem for the space homogeneous Boltzmann equation (2.47) in the positive cone L^1_+ of $L^1 = L^1(\mathbb{R}^3, \mathrm{d}\mathbf{v})$, namely

$$\frac{\mathrm{d}}{\mathrm{d}t} f = Q(f) \equiv Q^+(f) - Q^-(f), \ f(0) = f_0 \geq 0 \ (t \geq 0) \tag{3.1}$$

with $Q^{\pm}$ defined by (2.54) and (2.55), respectively.

The basic hypothesis is that the collision kernel q satisfies

$$q(\mathbf{v}, \mathbf{w}, \mathbf{n}) \leq C_q(1 + |\mathbf{v}|^{\lambda} + |\mathbf{w}|^{\lambda}) \quad (0 \leq \lambda \leq 2), \tag{3.2}$$

for some constant $C_q > 0$. The initial data f_0 is supposed to satisfy (at least) the condition of finite mass and energy, i.e. $\|f_0\|_2 < \infty$, where

$$\|g\|_l := \int (1 + |\mathbf{v}|^2)^{\frac{l}{2}} \, |g(\mathbf{v})| \, \mathrm{d}\mathbf{v}. \tag{3.3}$$

Unfortunately, under condition (3.2), the operators $Q^{\pm}$ are too singular to allow for applying general methods to the above problem. The idea of [2] is to approximate $Q^{\pm}$ by collision-like operators $Q^{\pm}_m$ with bounded (hence simpler) kernels $q_m(\mathbf{v}, \mathbf{w}) := \min\{q(\mathbf{v}, \mathbf{w}), m\}$, $m = 1, 2, \ldots$.

Thus one starts by solving the simple model

$$\frac{\mathrm{d}}{\mathrm{d}t} f = Q_m(f) \equiv Q^+_m(f) - Q^-_m(f), \ f(0) = f_0 \ (t \geq 0). \tag{3.4}$$

Note that, since (3.4) is a Boltzmann-type equation, then for "many" $g \in L^1$,

$$\int \varphi_i(\mathbf{v}) Q_m(g) \mathrm{d}\mathbf{v} = 0, \tag{3.5}$$

where $\varphi_0(\mathbf{v}) = 1$, $\varphi_i(\mathbf{v}) = \mathbf{v}_i$, $i = 1, 2, 3$, $\varphi_4(\mathbf{v}) = |\mathbf{v}|^2$. An immediate consequence is that for any solution $f = f(t, \mathbf{v})$ of (3.4),

$$\|f(t)\|_0 = \|f_0\|_0 \ (t \geq 0). \tag{3.6}$$

Moreover, if also $\|f(t)\|_2 < \infty$, then

$$\|f(t)\|_2 = \|f_0\|_2 \,. \tag{3.7}$$

Writing the solution of (3.4) as f_m, one could hope that if $m \to \infty$, then f_m converges somehow to a solution of the original problem (3.1). Another key point in the analysis is to use the above equalities as a priori estimates in order to replace (3.4) with other (somehow equivalent) equations, more suitable for monotone iteration with respect to the natural order of L^1.

Thus, one can first prove the following result ([2]).

PROPOSITION **3.1** *There exists a unique non-negative solution $f_m(t, \mathbf{v}) \in L^1$ of (3.4) for every $0 \le f_0 \in L^1$.*

Proof. By (3.6), the positive solutions (in L^1) of (3.4) are exactly the positive solutions of the equation

$$\frac{\mathrm{d}}{\mathrm{d}t} f + C \|f_0\|_0 f = Q_m(f) + C \|f(t)\|_0 f, \;\; f(0) = f_0 \;\; (t \ge 0), \tag{3.8}$$

which satisfy equality (3.6). Here $C > 0$ is some constant. Let $v(t) := \exp(-C \|f_0\|_0 t)$. Since the operators $Q_m^{\pm}$ are locally Lipschitz in L^1, (3.8) has a unique local solution $f_m(t)$, which is also a unique local solution to the mild equation

$$f(t) = v(t) f_0 + \int_0^t v(t-s)[Q_m(f)(s) + C \|f(s)\|_0 f(s)]\mathrm{d}s. \tag{3.9}$$

Define the sequence $\{f_m^n\}_n$ by

$$f_m^1 = 0, \quad f_m^n = v(t) f_0 + \int_0^t v(t-s)[Q_m(f_m^n)(s) + C \|f_m^n(s)\|_0 f_m^n(s)]\mathrm{d}s. \tag{3.10}$$

If C is sufficiently large, then the operator $X \ni g \to Q_m(g) + C \|g\|_0 g \in X$ is positive. Then the sequence $\{f_m^n(t)\}_n$ is positive and increasing in L^1. A simple induction, making use of (3.5), gives $\|f_m^n(t)\|_0 \le \|f_0\|_0$. Then by the monotone completeness of L^1 (Levi's theorem) $\{f_m^n(t)\}_n$ is convergent, its limit $g^m(t)$ satisfies (3.9), and $\|g_m(t)\|_0 \le \|f_0\|_0$. But by virtue of the uniqueness of the aforementioned local solution $f_m(t)$ (of both (3.8) and (3.9)), clearly $g_m(t) = f_m(t) \ge 0$ for t small enough. Moreover, $g^m(t)$ extends $f_m(t)$, as the unique solution of (3.8), for all $t \ge 0$. It remains to show that

this solution satisfies (3.6). To this end, one integrates (3.8), with f_m as solution, and rearrange conveniently the resulting expression as

$$f_m + \int_0^t [Q_m^-(f_m)(s) + C \|f_0\|_0 f_m(s)]\mathrm{d}s =$$

$$= f_0 + \int_0^t [Q_m^+(f_m)(s) + C \|f_m(s)\|_0 f_m(s)]\mathrm{d}s. \tag{3.11}$$

As $f_m(t)$, $Q_m^{\pm}(f_m)(t) \geq 0$, invoking the additivity of the L^1 norm, and the property $\|f_m(t)\|_0 \leq \|f_0\|_0$, one finally obtains

$$0 \leq \|f_0\|_0 - \|f_m(t)\|_0 \leq C \|f_0\|_0 \int_0^t (\|f_0\|_0 - \|f_m(s)\|_0)\mathrm{d}s. \tag{3.12}$$

Thus by Gronwall's inequality,

$$\|f_m(t)\|_0 = \|f_0\|_0 , \quad (t \geq 0) \tag{3.13}$$

so the proof is concluded. □

An induction involving (3.10), and making use of (3.5) also shows ([2]) that if f_m is as in Prop. 3.1, and $(1 + |\mathbf{v}|^2)f_0 \in L^1$, then $(1 + |\mathbf{v}|^2)f_m \in L^1$, and

$$\|f_m(t)\|_2 = \|f_0\|_2 \ (t \geq 0). \tag{3.14}$$

Another important property is the following estimation, uniform with respect to m (see [2]): for any $t_* > 0$,

$$\|f_m(t)\|_l \ \leq K \|f_0\|_l \quad (0 \leq t \leq t_*), \quad l \geq 4, \tag{3.15}$$

for some number $0 < K = K(t_*, \|f_0\|_2 , C_q, l)$. The proof (see the slightly more general Prop. 1.3 of [2]) is inductive, and applies (3.10) and the basic inequality

$$\int_{\mathbb{R}^3} (1 + |\mathbf{v}|^2)^{\frac{l}{2}} Q_m(f_m)\mathrm{d}\mathbf{v} \leq$$

$$\leq \frac{3}{2} C_q \beta_l [\|f_m(t)\|_{l+\lambda-\theta} \|f_m(t)\|_\theta + \|f_m(t)\|_{l-\theta} \|f_m(t)\|_{\lambda+\theta} , \tag{3.16}$$

valid for some $\beta_l > 0$ and for any $0 \leq \theta \leq 2$. Inequality (3.16) follows (see, e.g., [2]) from an elementary inequality due to Povzner, [23], and will be also called Povzner inequality[2].

One can prove that f_m converges to a solution of (3.1), under a stronger condition on f_0 than in Prop. 3.1. Indeed, one has ([2])

[2]Povzner-like inequalities can be also proved for the models presented in the previous sections.

PROPOSITION **3.2** *If $\|f_0\|_l < \infty$ for some $l \geq 4$, then there exists a unique solution $f \geq 0$ of problem (3.1) such that $(1+|\mathbf{v}|^l)f(t) \in L^1$. Moreover, $\|f(t)\|_2 = \|f_0\|_2$ ($t \geq 0$), and for any $t_* > 0$, there is some number $K = K(t_*, \|f_0\|_2, l)$ such that $\|f(t)\|_l \leq K \|f_0\|_l \quad (0 \leq t \leq t_*)$.*

Proof. Consider the equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}f + hf = Q_m^a(f),\ f(0) = f_0\ (t \geq 0), \tag{3.17}$$

where $h(\mathbf{v}) := C(1+|\mathbf{v}|^2)\,\|f_0(\mathbf{v})\|_2$ and $Q_m^a(f) := Q_m + hf$.

If f_m is as in Prop. 3.1, but f_0 is as in Prop. 3.2, then f_m is also the unique positive solution of Eq. (3.17), which satisfies (3.14). Further, consider

$$\frac{\mathrm{d}}{\mathrm{d}t}f + hf = Q_m^b(f),\ f(0) = f_0\ (t \geq 0), \tag{3.18}$$

where $Q_m^b(f) := Q_m^+(f) - Q^-(f) + hf$.

Let $V(t) := \exp(-th)$. One can introduce recurrences similar to (3.10),

$$\widetilde{f}_m^{1,i} = 0, \quad \widetilde{f}_m^{n+1,i} = V(t)f_0 + \int_0^t V(t-s)Q_m^i(\widetilde{f}_m^{n,i})(s)\mathrm{d}s \quad (n \geq 1); \quad i = a, b. \tag{3.19}$$

Under condition (3.2), if $C > 0$ is sufficiently large, the operators Q_m^i are positive and isotone so that the sequences $\left\{\widetilde{f_m}^{n,i}(t)\right\}_n$ are positive and increasing ($i = a, b$). Moreover, if $0 \leq (1+|\mathbf{v}|^2)g \in L^1$, then $Q_m^a\,(g) \geq Q_m^b(g)$ and $Q_m^b\,(g) \geq Q_j^b(g)$ for all m, $0 \leq j \leq m$. Using the above properties, one finds by induction that

$$0 \leq \widetilde{f}_j^{\,n,b}(t) \leq \widetilde{f_m}^{n,b}(t) \leq \widetilde{f}_m^{n,a}(t) \leq f_m(t); \quad 0 \leq j \leq m. \tag{3.20}$$

Hence, the increasing sequences $\left\{\widetilde{f_m}^{n,i}(t)\right\}_n$ are convergent. Note that if we set $f_m^b(t) := \lim_{n\to\infty} \widetilde{f_m}^{n,b}(t)$, then $0 \leq f_j^b(t) \leq f_m^b(t) \leq f_m(t); \quad 0 \leq j \leq m$. Then $\left\{f_m^b(t)\right\}_n$ is increasing and $\left\|f_m^b(t)\right\|_2 \leq \|f_0\|_2$, hence $\left\{f_m^b(t)\right\}_n$ converges to some limit $f(t)$, as $m \to \infty$, and

$$\|f(t)\|_2 \leq \|f_0\|_2\,. \tag{3.21}$$

Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}t}f + hf = Q(f) + hf \tag{3.22}$$

and, by (3.15),

$$\|f(t)\|_l \leq K \|f_0\|_l \quad (0 \leq t \leq t_*), \quad l \geq 4. \tag{3.23}$$

Thus f is a solution of (3.1) if there is equality in (3.21). This can be proved by estimating $s_m := f_m - f_m^b(t)$. Indeed, as f_m is the solution of (3.17), (3.18), one can write

$$\frac{\mathrm{d}}{\mathrm{d}t} s_m + h s_m = Q_m^a(f_m) - Q_m^b(f_m^b). \tag{3.24}$$

A short computation, which takes advantage that s_m is non-negative, and applies (3.23), gives (under hypothesis (3.2))

$$\|s_m(t)\|_2 \leq tCK \|f_0\|_4 \sup_{0 \leq s \leq t_*} \|s_m(s)\|_2 + o(1) \tag{3.25}$$

as $m \to \infty$ (with $C > 0$ sufficiently large, and K, t_* as in (3.23)).

Then for t sufficiently small, $\|s_m(t)\|_2 \to 0$ as $m \to \infty$, hence $\|f(t)\|_2 = \lim_{m\to\infty} \|f_m^b(t)\|_2 = \lim_{m\to\infty} \|f_m(t)\|_2 = \|f_0\|_2$.

To prove the uniqueness part of the proposition, observe that if $g \geq 0$ satisfies Eq. (3.1), and if $\|g(t)\|_2 \leq \infty$, then $\|g(t)\|_2 = \|f_0\|_2$. But g also satisfies the mild form of (3.22). Then $g \geq f$, by the construction of f. □

Variants of Arkeryd's monotonicity argument were successfully applied to other models close to the classical Boltzmann equation, [18], [27], [9], [7]. Thus, developing the above line of reasoning within a more general framework has become a tempting task. But this is not trivial, and requires new ideas (as will be seen in this section). Indeed, for instance, too key issues of Arkeryd's analysis seem rather specific to the model considered in [2]: a) choice of a priori estimates; b) construction of suitable regular operator approximations of the Boltzmann collision operators.

3.2. An abstract model

We begin with some terminology and facts related to Banach lattices ([17, 24]).

The frame of our analysis is a separable AL-space X with norm $\|\cdot\|$, order $\leq$, and positive cone X_+. We recall that an (AL) space, is a Banach lattice whose norm satisfies

$$\|g + h\| = \|g\| + \|h\| \quad (g, h \in X_+). \tag{3.26}$$

As X is an AL-space, if $h : \mathbb{R} \mapsto X_+$ is Bochner integrable, then property (3.26) gives

$$\left\| \int_{\mathcal{S}} h(s)\mathrm{d}s \right\| = \int_{\mathcal{S}} \|h(s)\| \, \mathrm{d}s \tag{3.27}$$

for any measurable set $\mathcal{S}$ of $\mathbb{R}$, the integral being in the sense of Lebesgue.

Examples of AL-spaces are L^1-real and the real subspace of self-adjoint trace-class operators (with trace norm)[3].

Related to the order of X, we shall also use the standard notations $(g \geq h) \Leftrightarrow (h \leq g)$, as well as $(g < h) \Leftrightarrow (\ h > g) \Leftrightarrow (g \leq h$ and $\ g \neq h)$. AL-spaces are *monotone complete*, in the sense that any increasing (i.e., directed $\leq$) norm-bounded family converges. The norm of an AL-space is *order continuous*, i.e., any directed $\geq$ filters that converges to 0 is also norm convergent to 0 . A map $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$, with $\mathcal{D}(\Gamma) \cap X_+ \neq \emptyset$, is called *positive (strictly positive)* if $0 \leq \Gamma g$ for $0 \leq g \in \mathcal{D}(\Gamma)$ (if $0 < \Gamma g$ for $0 < g \in \mathcal{D}(\Gamma)$). Further, $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is called *isotone (strictly isotone)* if $\Gamma g \leq \Gamma h$, whenever $g \leq h$ (if $\Gamma g < \Gamma h$, whenever $g < h$), $g, h \in \mathcal{D}(\Gamma)$. Obviously, if $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is isotone, $0 \in \mathcal{D}(\Gamma)$ and $0 \leq \Gamma(0)$, then Γ is positive. We say that a subset $\mathcal{M} \subset X$ is *p-saturated* (positively saturated) if $\mathcal{M} \cap X_+ \neq \emptyset$, and from $0 \leq g \leq h \in \mathcal{M}$, it follows that $g \in \mathcal{M}$. An operator $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ will be called *o-closed* (closed with respect to the order) if for any increasing sequence $\{g_n\} \subset \mathcal{D}(\Gamma)$ such that $\{g_n\}$ is increasing and convergent (in symbols, $\nearrow$) to some g, and $\{\Gamma g_n\}$ is Cauchy, one has $g \in \mathcal{D}(\Gamma)$ and $\lim_{n\to\infty} \Gamma g_n = \Gamma g$. Clearly, any closed mapping is o-closed.

We recall (see, e.g., [16]) that if $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is a closed linear operator, then

$$\Gamma \int_{\mathbb{S}} h(s)\mathrm{d}s = \int_{\mathbb{S}} \Gamma h(s)\mathrm{d}s. \tag{3.28}$$

for any function h Bochner integrable on some measurable set $\mathbb{S} \in \mathbb{R}$, with values in $\mathcal{D}(\Gamma)$, and such that Γh is Bochner integrable.

We recall that a *positive C_0 semigroup* on X is a C_0 semigroup of positive linear operators on X. If $\left\{S^t\right\}_{t\geq 0}$ is a positive C_0 semigroup on X, then its infinitesimal generator G is densely defined and closed (as the infinitesimal generator of a C_0 semigroup). Moreover, G^k is densely defined and closed, $k = 2, 3, \ldots$. Additional useful properties are collected in the following lemma.

Let I denote the identity on X. Set $\mathcal{D}_+^\infty(G) := \cap_{k=1}^\infty \mathcal{D}(G^k) \cap X_+$.

[3]Actually, according to Kakutani's theorem, [24], every AL-space is isometrically isomorphic (as an ordered vector space) to a space of type L^1.

LEMMA **3.1** ([11])

a) *The sets* $\mathcal{D}(G^k) \cap X_+$, $k = 1, 2, ...,$ *and* $\mathcal{D}_+^\infty(G)$ *are dense in* X_+.

b) *Suppose that there is some number* $\gamma \geq 0$ *such that*

$$(G + \gamma I)g \leq 0 \quad (g \in \mathcal{D}(G) \cap X_+). \tag{3.29}$$

Then $\mathcal{D}(G^k) \cap X_+$, $k = 1, 2, ...,$ *and* $\mathcal{D}_+^\infty(G)$ *are p-saturated. Moreover, for any* $h \in X_+$,

$$0 \leq S^t h \leq \exp(-\gamma t)h \quad (t \geq 0), \tag{3.30}$$

and there is an increasing sequence $\{h_n\} \subset \mathcal{D}_+^\infty$, *such that* $h_n \nearrow h$ *as* $n \to \infty$.

Motivated by the examples of the previous section, it is of interest to consider the following abstract i.v.p., [11],

$$\frac{\mathrm{d}f}{\mathrm{d}t} = Q(t, f) = Q^+(t, f) - Q^-(t, f), \quad f(0) = f_0 \in X_+ \quad (t > 0), \tag{3.31}$$

formulated in X_+ (the particular autonomous case is not excluded).

In Eq. (3.31), Q^+ and Q^- are mappings defined from $\mathbb{R}_+ \times \mathcal{D}$ to X, for some $\mathcal{D} \subset X$ such that $\mathcal{D} \cap X_+$ is dense in X_+.

The following properties are assumed for $Q^\pm$:

a) For a.e. $t \geq 0$, the operators $Q^\pm(t, \cdot) : \mathcal{D} \mapsto X$ are positive and isotone.

b) The mappings $\mathbb{R}_+ \ni t \mapsto Q^\pm(t, g(t)) \in X_+$ are measurable for any Lebesgue measurable function $g : \mathbb{R}_+ \mapsto X$ that satisfies $g(t) \in \mathcal{D} \cap X_+$ a.e. on $\mathbb{R}_+$.

c) For a.e. $t \geq 0$, the operators $Q^\pm(t, \cdot)$ are o-closed and their common domain $\mathcal{D}$ is p-saturated.

We are interested in the existence and uniqueness of positive (i.e., in X_+) strong solutions of Eq. (3.31) under additional hypotheses which abstract further properties of the Boltzmann model.

We recall that a function $f : \mathbb{R}_+ \mapsto X$ is a strong solution of Eq. (3.31), if it is absolutely continuous on $\mathbb{R}_+$, differentiable a.e. on $\mathbb{R}_+$, satisfies Eq. (3.31) a.e. on $\mathbb{R}_+$, and verifies the initial condition. Equivalently, f is a strong solution of problem (3.31) if it is solution of the integral equation

$$f(t) = f_0 + \int_0^t Q(s, f(s))\mathrm{d}s \quad (t \geq 0), \tag{3.32}$$

where the integral is in the sense of Bochner.

We also consider the following problem related to Eq. (3.31)

$$\frac{\mathrm{d}f}{\mathrm{d}t} = Af + Q(t,f), \quad f(0) = f_0 \in X_+ \quad (t > 0), \tag{3.33}$$

with Q as in Eq. (3.31). Here A is the infinitesimal generator of a C_0 group of positive linear isometries on X, which commutes with Λ.

We are interested in the existence and uniqueness of mild solutions of Eq. (3.31) in X_+, i.e, solutions of the integral equation

$$f(t) = U^t f_0 + \int_0^t U^{t-s} Q(s, f(s)) \mathrm{d}s \quad (t \geq 0) \tag{3.34}$$

in X_+, where $\left\{U^t\right\}_{t\in\mathbb{R}}$ is the C_0 group of positive linear isometries on X, generated by A (the integral is in the sense of Bochner).

As the above model is still too general for developing an existence theory of solutions, additional hypotheses are needed. The examples of the previous section suggest to assume some sort of dissipation (conservation) property, [11]. This claims the existence of a positive, densely defined, closed linear operator $\Lambda : \mathcal{D}(\Lambda) \subset X \mapsto X$ such that, for any positive solution $f(t) \in \mathcal{D}(\Lambda^2)$ of Eq. (3.31), the quantity $\|\Lambda f(t)\|$ is dissipated (conserved), i.e., is decreasing (constant) in t, and $\left\|\Lambda^2 f(t)\right\|$ is locally bounded in t. The "law of decrease" of $\|\Lambda f(t)\|$ can be used as a "natural" a priori estimate[4]. In particular,

$$\|\Lambda f(t)\| \leq \|\Lambda f_0\| \quad (t \geq 0). \tag{3.35}$$

To be precise, we introduce the following "dissipation" property ([11]). Let $\mathcal{M}$ be a subset of $\mathcal{D} \cap X_+$ dense in X_+.

DEFINITION **3.1** ([11]) *A closed positive linear operator $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is called of type D on $\mathcal{M}$ (with respect to Eq. (3.31)) if $\mathcal{M} \subset \mathcal{D}(\Gamma)$, $Q^{\pm}(t, \mathcal{M}) \subset \mathcal{D}(\Gamma)$ a.e. on $\mathbb{R}_+$, and for any $g \in \mathcal{M}$,*

$$0 \leq \Delta(t, g; \Gamma, Q) := \left\|\Gamma Q^-(t,g)\right\| - \left\|\Gamma Q^+(t,g)\right\| \quad (t \geq 0 \quad a.e.). \tag{3.36}$$

If Γ is of type D on $\mathcal{M}$, then the following property can be easily established by making use of (3.27) and (3.28).

LEMMA **3.2** ([11]) *Let g_0, $g(t)$, $h(t) \in \mathcal{M}$, $t \geq 0$ a.e., with $Q^{\pm}(\cdot, h(\cdot))$, $\Gamma Q^{\pm}(\cdot, h(\cdot)) \in L^1_{loc}(\mathbb{R}_+; X_+)$, and*

$$g(t) \leq g_0 + \int_0^t Q(s, h(s)) \mathrm{d}s \quad (t \geq 0). \tag{3.37}$$

[4]This can take various forms in applications, depending on the form of Λ and Q, e.g., conservation energy, in the case of the model of [2].

Then

$$\|\Gamma g(t)\| + \int_0^t \Delta(s, h(s); \Gamma, Q)\mathrm{d}s \leq \|\Gamma g_0\| \quad (t \geq 0). \tag{3.38}$$

Moreover, (3.38) *holds with equality sign for any* $t \geq 0$*, provided that there is equality in* (3.37) *for all* $t \geq 0$*.*

On the other hand, in determining the behavior of $\left\|\Lambda^2 f(t)\right\|$, a major role appears to be played by the Povzner inequality (3.16). This has to be somehow included in the model.

Now we are in position to complete the setting of Eq. (3.31) with additional hypotheses, making more precise the above considerations.

Specifically, we assume that there is a linear operator $\Lambda : \mathcal{D}(\Lambda) \subset X \mapsto X$, with $\mathcal{D}(\Lambda) \subset \mathcal{D}$ and $Q^{\pm}(t, \mathcal{D}(\Lambda^k) \cap X_+) \subset \mathcal{D}(\Lambda^{k-1})$, $t \geq 0$ a.e., $k = 2, 3$, such that:

(A_0) The operator $(-\Lambda)$ is the infinitesimal generator of a C_0 semigroup of positive linear operators on X, and there is a number $\lambda_0 > 0$ such that

$$(\Lambda - \lambda_0 I)g \geq 0 \quad (g \in \mathcal{D}(\Lambda) \cap X_+). \tag{3.39}$$

(A_1) For a.e. $t \geq 0$,

$$\Delta(t, g) := \Delta(t, g; \Lambda, Q) \geq 0 \quad (g \in \mathcal{D}(\Lambda^2) \cap X_+), \tag{3.40}$$

and the map $\mathcal{D}(\Lambda^2) \cap X_+ \ni g \mapsto \Delta(t, g) \in \mathbb{R}_+$ is isotone.

(A_2) There exists a non-decreasing convex function $a : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that

$$a(\|\Lambda g\|)\Lambda g - Q^-(t, g) \geq 0, \quad (g \in \mathcal{D}(\Lambda) \cap X_+, \quad t \geq a.e.), \tag{3.41}$$

and for a.e. $t \geq 0$, the map $\mathcal{D}(\Lambda) \cap X_+ \ni g \mapsto a(\|\Lambda g\|)\Lambda g - Q^-(t, g) \in X$ is isotone.

(A_3) There exists a non-decreasing function $\rho : \mathbb{R}_+ \mapsto \mathbb{R}_+$, and there is an operator $\Lambda_1 : \mathcal{D}(\Lambda_1) \subset X \mapsto X$ of type D on $\mathcal{D}(\Lambda^2) \cap X_+$ such that

$$-\Delta(t, g; \Lambda^2, Q) \leq \rho(\|\Lambda_1 g\|) \left\|\Lambda^2 g\right\| \quad (g \in \mathcal{D}(\Lambda^3) \cap X_+, \, t \geq 0 \; a.e.). \tag{3.42}$$

Some remarks are in order.

First, observe that if $g \in \mathcal{D}(\Lambda^2) \cap X_+$, then by (3.39), (3.40) and (3.41) we have the simple inequalities

$$\|g\| \leq \lambda_0^{-1} \|\Lambda g\| \leq \lambda_0^{-2} \left\|\Lambda^2 g\right\| \tag{3.43}$$

and

$$\left\|Q^{\pm}(t,g)\right\| \leq \lambda_0^{-1} \left\|\Lambda Q^{\pm}(t,g)\right\| \leq \lambda_0^{-1} \left\|\Lambda Q^{-}(t,g)\right\| \leq$$

$$\leq a(\|\Lambda g\|)\lambda_0^{-1} \left\|\Lambda^2 g\right\| \leq a(\lambda_0^{-1} \left\|\Lambda^2 g\right\|)\lambda_0^{-1} \left\|\Lambda^2 g\right\| \quad (t \geq 0 \quad a.e.), \tag{3.44}$$

with the following obvious consequences.

REMARK **3.1** *$Q^{\pm}(t,0) = 0$ and $\Delta(t,0) = 0$ a.e. on $\mathbb{R}_+$.*

Let $\Lambda^0 := I$.

REMARK **3.2** *If $g : \mathbb{R}_+ \mapsto X_+$ is measurable, with $g(t) \in \mathcal{D}(\Lambda^2)$, $t \geq 0$, a.e., and $\left\|\Lambda^2 g\right\| \in L^{\infty}_{loc}(\mathbb{R}_+)$, then g, $\Lambda^{k+1} g$, and $\Lambda^k Q^{\pm}(\cdot, g(\cdot))$ are in $L^1_{loc}(\mathbb{R}_+; X_+)$, $k = 0, 1$.*

Lemma 3.1a) and (A_0) imply that $\mathcal{D}(\Lambda^k) \cap X_+$, $k = 1, 2, ...$, and $\mathcal{D}^{\infty}_+ := \mathcal{D}^{\infty}_+(\Lambda)$ are p-saturated and dense in X_+. Obviously, (3.39) shows that Λ is positive. Thus, by (3.40), the operator Λ is of type D on $\mathcal{D}(\Lambda^2) \cap X_+$. This has the following important consequence.

If $f(t) \in \mathcal{D}(\Lambda^2)$, $t \geq 0$, a.e., and if $Q^{\pm}(\cdot, f(\cdot))$, $\Lambda Q^{\pm}(\cdot, f(\cdot)) \in L^1(\mathbb{R}_+; X_+)$, then by (3.38), applied with equality sign,

$$\|\Lambda f(t)\| + \int_0^t \Delta(s, f(s))\mathrm{d}s = \|\Lambda f_0\| \quad (t \geq 0). \tag{3.45}$$

Thus $\|\Lambda f(t)\|$ is decreasing in time and satisfies (3.35). In particular, if $\Delta(t,g) = 0$ for all $g \in \mathcal{D}(\Lambda^2) \cap X_+$, $t \geq 0$ a.e., then $\|\Lambda f(t)\|$ is conserved for all $t \geq 0$.

Observe that inequality (3.42) is of the form

$$-\Delta(t, g; \Gamma, Q) \leq \rho_{\Gamma}(\|\Lambda_1 g\|) \|\Gamma g\| \quad (g \in \mathcal{M}_1, \, t \geq 0 \; a.e.), \tag{3.46}$$

where $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is some positive linear operator, and $\mathcal{M}_1 \subset \mathcal{D}(\Gamma) \cap \mathcal{D}(\Lambda^2) \cap X_+$ is such that $Q^{\pm}(t, \mathcal{M}_1) \subset \mathcal{D}(\Gamma)$, $t \geq 0$ a.e., while $\rho_{\Gamma} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is some non-decreasing function.

Formula (3.45) generalizes a priori estimates introduced in e.g., [2, 7, 8, 9, 27]. Formula (3.46) can be regarded as an abstract correspondent to the Povzner inequality, [2, 23].

We finally remark that the above setting does not exclude the case $\Lambda_1 = \Lambda$ when, obviously, some of the above conditions become redundant.

3.3. General results on the existence of solutions

We are now in position to state some results ([11], [13]) on the existence of solutions to our abstract model. The proofs will be sketches in the next subsection (for more details, the reader is referred to [11] and [13]). First we consider problem (3.31).

THEOREM **3.1** *Let either of the following two sets of conditions be fulfilled:*

a) $Q^+(t,\mathcal{D}_+^\infty)\subset\mathcal{D}_+^\infty$, $t\geq 0$ *a.e.*, $\Lambda^k Q^+(\cdot,\mathcal{D}_+^\infty)\subset L^1_{loc}(\mathbb{R}_+;X_+)$, $k=1,2,\ldots$. *In problem* (3.31), $f_0\in\mathcal{D}(\Lambda^2)\cap X_+$.

b) *The operators* $Q^\pm$ *do not depend explicitly on* t. *In problem* (3.31), $f_0\in\mathcal{D}(\Lambda^3)\cap X_+$.

Then there exists a unique positive strong solution of the i.v.p. (3.31) *such that* $f(t)\in\mathcal{D}(\Lambda^2)$ *for any* $t\geq 0$, *and* $\left\|\Lambda^2 f(\cdot)\right\|$ *is locally bounded on* $\mathbb{R}_+$.

Moreover, $f,\Lambda f\in C(\mathbb{R}_+;X_+)$. *Furthermore,* f *satisfies Eq.* (3.45) *and*

$$\left\|\Lambda^2 f(t)\right\|\leq\exp(\rho(\|\Lambda_1 f_0\|)t)\,\left\|\Lambda^2 f_0\right\|\quad(t\geq 0).\tag{3.47}$$

Note here that Theorem 3.1a) is also applicable to the autonomous case, but, clearly, its conditions are different from those of Theorem 3.1b).

Theorem 3.1 has an immediate noticeable consequence, as follows:

Consider Eq. (4.22) and let $\left\{U^t\right\}_{t\in\mathbb{R}}$ be the C_0 group of positive linear isometries on X, generated by A.

If f is a solution of (3.34), then setting $F(t):=U^{-t}f(t)$ in (3.34), we get

$$F(t)=f_0+\int_0^t Q_U(s,F(s))\mathrm{d}s\quad(t\geq 0),\tag{3.48}$$

hence, by differentiation,

$$\frac{\mathrm{d}}{\mathrm{d}t}F=Q_U(t,F)=Q_U^+(t,F)-Q_U^-(t,F),\quad F(0)=f_0\quad(t\geq 0\quad a.e.),\tag{3.49}$$

where $Q_U(t,\cdot):=U^{-t}Q(t,U^t\cdot)$ and $Q_U^\pm(t,\cdot):=U^{-t}Q^\pm(t,U^t\cdot)$.

Suppose that $U^t\mathcal{D}(\Lambda)=\mathcal{D}(\Lambda)$ and $U^t\Lambda=\Lambda U^t$ on $\mathcal{D}(\Lambda)$ for every $t>0$. Also, let $U^t\mathcal{D}(\Lambda_1)=\mathcal{D}(\Lambda_1)$ and $U^t\Lambda_1=\Lambda_1U^t$ on $\mathcal{D}(\Lambda_1)$ for all $t>0$.

Now $Q_U^\pm$ and Q_U are well defined as maps from $\mathbb{R}_+\times\mathcal{D}(\Lambda)$ to X, the last equation is of the form (3.31), and we can state the following consequence ([11]) of Theorem 3.1a):

COROLLARY **3.1** *Let* $Q^+(t,\mathcal{D}_+^\infty) \subset \mathcal{D}_+^\infty$, $t \geq 0$ *a.e., and* $\Lambda^k Q^+(\cdot, U^\cdot g) \in L^1_{loc}(\mathbb{R}_+; X_+)$ *for all* $g \in \mathcal{D}_+^\infty$, $k = 1, 2, \ldots$. *Suppose that* $f_0 \in \mathcal{D}(\Lambda^2) \cap X_+$ *in* (4.22). *Then problem* (4.22) *has a unique positive mild solution* f *such that* $f(t) \in \mathcal{D}(\Lambda^2)$ *for any* $t \geq 0$ *and* $\|\Lambda^2 f(\cdot)\|$ *is locally bounded on* $\mathbb{R}_+$. *Moreover,* $f, \Lambda f \in C(\mathbb{R}_+; X_+)$. *Furthermore,* f *satisfies* (3.45) *and* (3.47).

The following result, [13], extends the existence of strong solutions of Eq. (3.31) to the case of initial datum $f_0 \in \mathcal{D}(\Lambda) \cap X_+$ (instead of $\mathcal{D}(\Lambda^2) \cap X_+$, as assumed in Theorem 3.1).

THEOREM **3.2** *Under the assumptions of Theorem* 3.1a) *on* Λ *and* $Q^\pm$, *let* $f_0 \in \mathcal{D}(\Lambda) \cap X_+$ *in Eq.* (3.31). *Then there exists a strong solution,* $f \in C([0,\infty); X_+)$, *of the i.v.p.* (3.31). *Moreover, for any* $t \geq 0$, $f(t) \in \mathcal{D}(\Lambda)$, $\|\Lambda f(t)\| \leq \|\Lambda f_0\|$, *and*

$$\|f(t)\| = \|f_0\| + \int_0^t \left\|Q^+(s, f(s))\right\| - \left\|Q^-(s, f(s))\right\| \mathrm{d}s. \tag{3.50}$$

Note here that if f is as in Theorem 3.2, we know only that $f \in \mathcal{D}(\Lambda) \cap X_+$. Then $\Delta(t, f)$ and $\Lambda^2 f$ may not be not well-defined. Therefore, we cannot obtain inequalities of the form (3.45) (except the case when $\Delta = 0$ on $\mathcal{D}(\Lambda^2) \cap X_+$,) or like (3.47), at the level of abstraction of the theorem.

Also remark that Theorem 3.2 leaves open the question on the uniqueness of the solution in the general case (under the conditions of the theorem).

However, uniqueness can be proved under additional conditions, [13].

PROPOSITION **3.3** *If* $\Delta(t, g) = 0$ *for all* $g \in \mathcal{D}(\Lambda^2) \cap X_+$, t *– a.e., then*

$$\|\Lambda f(t)\| = \|\Lambda f_0\| \quad (t \geq 0), \tag{3.51}$$

and there is a unique solution of the i.v.p. (3.31) *as in Theorem* 3.2, *which satisfies* (3.51).

A similar result like Corollary 3.1 can be formulated for Theorem 3.2.

The following proposition yields additional useful estimates, [11], for the solutions of Eq. (3.31). For simplicity, we remain in the conditions of Theorem 3.1a). However, similar results are valid when Theorem 3.1b) holds, as can be seen by inspecting the proof of the proposition.

Assume that $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ is a closed, positive linear operator. Let f be a solution of problem (3.31), provided by Theorem 3.1a).

PROPOSITION **3.4** a) *Suppose that* Γ *is of type D on* $\mathcal{D}_+^\infty$. *Then* $f(t) \in \mathcal{D}(\Gamma)$, $t \geq 0$, *and*

$$\|\Gamma f(t)\| \leq \|\Gamma f_0\| \quad (t \geq 0). \tag{3.52}$$

b) *Suppose that* Γ *and* ρ_Γ *are as in* (3.46), *with* $\mathcal{M}_1 \supseteq \mathcal{D}_+^\infty$. *Then* $f(t) \in \mathcal{D}(\Gamma)$, $t \geq 0$, *and*

$$\|\Gamma f(t)\| \leq \exp(\rho_\Gamma(\|\Lambda_1 f_0\|)t)\ \|\Gamma f_0\| \quad (t \geq 0). \tag{3.53}$$

In applications, the choice of Λ and Λ_1 may be not unique. In some cases, the role of Λ_1 and Γ may be played by suitable powers of Λ, while, in other examples, $\Lambda = \Lambda_1 = \Gamma$.

A correspondent to Prop. 3.4, applicable to Corollary 3.1, can be readily obtained. The modifications in the reformulation of the proposition are obvious and include additional hypotheses for the commutation of U^t with Γ, etc.

3.4. Proofs

Sketch of the proof of Theorem 3.1

In the following, we give an insight into the rather lengthy argument of Theorem 3.1 (see [11] for a detailed proof), and explain the role of assumptions (A_0)-(A_3).

We start by observing that if $f_0 = 0$ in (3.31), then, by Remark 3.1, clearly $f(t) \equiv 0$ is a solution to Eq. (3.31). It is the unique strong solution in $\mathcal{D}(\Lambda^2) \cap X_+$, as it follows from (3.45). Moreover, if $0 \neq f_0 \in \mathcal{D}(\Lambda^2) \cap X_+$, but $a(\|\Lambda f_0\|) = 0$, then $Q^\pm(t, f_0) = 0$, for a.e. $t \geq 0$, by (3.44), hence $f(t) \equiv f_0$ is a solution to (3.31). It is the unique solution in $\mathcal{D}(\Lambda^2) \cap X_+$, because any other solution $f^*(t) \in \mathcal{D}(\Lambda^2) \cap X_+$ must be a.e. constant. Indeed, applying (3.45), and invoking the positivity and monotonicity of a, we obtain $0 \leq a(\|\Lambda f^*(t)\|) \leq a(\|\Lambda f_0\|) = 0$. This leads (again by (3.44)) to $Q^\pm(t, f(t)) = 0$ a.e.

Therefore, one can assume below that $f_0 \neq 0$ and $a(\|\Lambda f_0\|) \neq 0$.

We first refer to the **existence** part of the theorem. Inspired from [2], one can consider the problem

$$\frac{\mathrm{d}}{\mathrm{d}t} f + a(\|\Lambda f_0\|)\Lambda f = B(t, f, f), \quad f(0) = f_0 \in X_+ \quad (t \geq 0). \tag{3.54}$$

Here a is as in (A_2), and B is formally defined by

$$B(t,g,h) := Q(t,g(t)) + a\left(\|\Lambda g(t)\| + \int_0^t \Delta(s,h(s))\mathrm{d}s\right)\Lambda g(t) \quad (t \geq 0 \quad a.e.) \tag{3.55}$$

for all $g(t) \in \mathcal{D}(\Lambda) \cap X_+$ and $h(t) \in \mathcal{D}(\Lambda^2) \cap X_+$ with $\Lambda Q^{\pm}(\cdot, h(\cdot)) \in L^1_{loc}(\mathbb{R}_+; X_+)$.

By (3.45), any strong positive solution of Eq. (3.31) is also a solution to (3.54). Conversely, any positive strong solution of problem (3.54) is a solution of Eq. (3.31), provided that it satisfies (3.45).

Recall now that, by (A_0) and Lemma 3.1b), the operator $L = -a(\|\Lambda f_0\|)\Lambda$ is the infinitesimal generator of a C_0 positive semigroup $\{V^t\}_{t\geq 0}$, and

$$0 \leq V^t h \leq \exp(-a(\|\Lambda f_0\|)\lambda_0 t)h \leq h \quad (h \in X_+). \tag{3.56}$$

Thus any solution of Eq. (3.54) is also a solution of the mild problem

$$f(t) = V^t f_0 + \int_0^t V^{t-s} B(s,f,f)\mathrm{d}s\,, \tag{3.57}$$

the integral being in the sense of Bochner.

Eq. (3.57) is useful for monotone iteration. Indeed, $\{V^t\}_{t\geq 0}$ is positive, and one can prove[5] the following properties ([11]).

LEMMA **3.3** *Let g_i, h_i, $i = 1,2$, satisfy the conditions of Remark 3.2. Suppose that $g_1(t) \leq g_2(t)$ and $h_1(t) \leq h_2(t)$ a.e. on $\mathbb{R}_+$. Then $B(\cdot, g_i, h_j) \in L^1_{loc}(\mathbb{R}_+; X_+)$, $i,j = 1,2$. In addition, for a.e. $t \geq 0$,*

$$0 \leq B(t,g_1,h_1) \leq B(t,g_2,h_2). \tag{3.58}$$

Thus, formally, by (3.57) one could consider the following iteration, hopefully, increasing:

$$f_1(t) = 0, \quad f_2(t) = V^t f_0, \tag{3.59}$$

$$f_n(t) = V^t f_0 + \int_0^t V^{t-s} B(s, f_{n-1}, f_{n-2})\mathrm{d}s \quad (n = 3,4,...). \tag{3.60}$$

Note that if $\{f_n(t)\}_n$ is sufficiently regular, by differentiation, (3.60) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f_n(t) = B(t, f_{n-1}, f_{n-2}) - a(\|\Lambda f_0\|)\Lambda f_n(t) \quad (t > 0 \quad a.e., \quad n \geq 3), \tag{3.61}$$

[5] See the Appendix.

and integrating (3.61) one has

$$f_n(t) = f_0 + \int_0^t Q(s, f_{n-1}(s))\mathrm{d}s +$$

$$+ \int_0^t a\left(\|\Lambda f_{n-1}(s)\| + \int_0^s \Delta(\tau, f_{n-2}(\tau))\mathrm{d}\tau\right)\Lambda f_{n-1}(s)\mathrm{d}s.$$

$$- \int_0^t a(\|\Lambda f_0\|)\Lambda f_n(s)\mathrm{d}s. \tag{3.62}$$

However, in general, $B(\cdot, g, h)$ does not exist for all g, $h \in X$. Hence we need give a meaning to (3.60), at least for f_0 in a sufficiently large set. Here comes the role of $\mathcal{D}_+^\infty$ (of $\mathcal{D}(\Lambda^3) \cap X_+$). Indeed, if $f_0 \in \mathcal{D}_+^\infty$ ($f_0 \in \mathcal{D}(\Lambda^3) \cap X_+$), then one can show that $f_n(t) \in \mathcal{D}_+^\infty$ ($f_0 \in \mathcal{D}(\Lambda^3) \cap X_+$), and is sufficiently regular. This is clarified in the lemma bellow, which summarizes the main results[6] of [11] on the properties of $\{f_n(t)\}_n$.

LEMMA **3.4** a) *In addition, to the conditions of Theorem* 3.1a), *let* $f_0 \in \mathcal{D}_+^\infty$. *Then* $f_n(t)$, $Q^\pm(t, f_n(t)) \in \mathcal{D}_+^\infty$ *a.e. on* $\mathbb{R}_+$. *Moreover,* $\Lambda^k Q^\pm(\cdot, f_n(\cdot)) \in L^1_{loc}(\mathbb{R}_+; X_+)$, $k = 0, 1, ...,$ $n = 1, 2, ...$.

b) *Assume the conditions of Theorem* 3.1b). *Then* $f_n(t) \in \mathcal{D}(\Lambda^3) \cap X_+$ *and* $Q^\pm(f_n(t)) \in \mathcal{D}(\Lambda^2) \cap X_+$; $t \geq 0$. *Moreover,* $\Lambda^k Q^\pm(f_n) \in L^1_{loc}(\mathbb{R}_+; X_+)$, $k = 0, 1, 2,$, $n = 1, 2, ...$.

c) *In both cases* a) *and* b), $\Lambda^k f_n \in C(\mathbb{R}_+; X_+)$, $k = 0, 1, 2$, *and* f_n *is a.e. differentiable on* $\mathbb{R}_+$ *and satisfies* (3.61) *(and* (3.62)*). Moreover, for any* $t \geq 0$, *the sequence* $\{f_n(t)\}_n$ *is increasing.*

d) *If* $f_n(t)$ *is as in* a) *or* b), *and* $n \geq 2$, *then*

$$f_n(t) \leq f_0 + \int_0^t Q(s, f_{n-1}(s))\mathrm{d}s \tag{3.63}$$

and

$$\|\Lambda f_n(t)\| + \int_0^t \Delta(s, f_{n-1}(s))\mathrm{d}s \leq \|\Lambda f_0\|. \tag{3.64}$$

e) *If* $f_n(t)$ *is as in* a) *or* b), *and* Γ *is an operator of type D on* $\mathcal{D}_+^\infty$, *(on* $\mathcal{D}(\Lambda^2) \cap X_+$*) then for any* $t \geq 0$,

$$\|\Gamma f_n(t)\| \leq \|\Gamma f_0\| \quad (n = 1, 2, ...). \tag{3.65}$$

[6] See the Appendix for a proof.

In particular,

$$\left\|\Lambda^2 f_n(t)\right\| \leq \exp(\rho(\|\Lambda_1 f_0\|)t)\left\|\Lambda^2 f_0\right\| \quad (t \geq 0, \quad n = 1, 2, ...), \tag{3.66}$$

with ρ *as in* (3.42).

f) *Suppose that* $f_n(t)$ *is as in* a) *(as in* b)*). Let* $\Gamma : \mathcal{D}(\Gamma) \subset X \mapsto X$ *be some closed, positive linear operator, satisfying* (3.46)*, with* $\mathcal{M}_1 \supseteq \mathcal{D}_+^\infty$ *(with* $\mathcal{M}_1 \supseteq \mathcal{D}(\Lambda^3) \cap X_+$*). Then for any* $t \geq 0$*,*

$$\|\Gamma f_n(t)\| \leq \exp(\rho_\Gamma(\|\Lambda_1 f_0\|)t)\|\Gamma f_0\| \quad (n = 1, 2, ...), \tag{3.67}$$

with ρ_Γ *as in* (3.46).

By the above lemma, $\{f_n(t)\}_n$ is increasing, and the key inequality (3.64) shows that $\{f_n(t)\}_n$ is norm bounded[7]. Thus $\{f_n(t)\}_n$ is convergent, because X is monotone complete. One expects the limit to satisfy (3.54) (and (3.57), too). The proof hinges on the application of Lebesgue's dominated convergence theorem to (3.62) (as the operators $Q^\pm$ are o-closed, and Λ is closed). To this end, the limit of $\{f_n(t)\}_n$ must be in $\mathcal{D}(\Lambda^2)$, which follows from (3.66). Now, to prove that the limit of $\{f_n(t)\}_n$ is a strong solution to (3.31), it remains to show that the above limit satisfies (3.45). This is done by applying Gronwall's Lemma to an inequality to be obtained from (3.62) (by using (3.66) and the convexity of a). But the above procedure provides the existence part of the Theorem 3.1a) only for $f_0 \in \mathcal{D}_+^\infty$, hence one more step is needed. Since $\mathcal{D}_+^\infty$ is dense in X_+ (cf. Lemma 3.1), any initial datum as in the assumptions of Theorem 3.1a), can be approximated by elements of $\mathcal{D}_+^\infty$. This leads to a monotone scheme approximating (3.60) and one can apply successively Lebesgue's convergence theorem. In details, one proceeds as follows.

Step A. If in addition to the conditions of Theorem 3.1 a), one assumes $f_0 \in \mathcal{D}_+^\infty$ then Lemma 3.4 applies. As Λ^k is closed, clearly, by (3.39) and the monotone completeness of X, it follows that there is some $f(t) \in \mathcal{D}(\Lambda^k)$ such that $\Lambda^k f_n(t) \nearrow \Lambda^k f(t)$ as $n \to \infty$, $t \geq 0$, $k = 0, 1, 2$. Consequently, $f(t)$ satisfies (3.47). Moreover, Remark 3.2 implies that $\Lambda^k f$, $k = 0, 1, 2$, $Q^\pm(\cdot, f(\cdot))$, and $\Lambda Q^\pm(\cdot, f(\cdot))$ are in $L^1_{loc}(\mathbb{R}_+; X_+)$. Then, applying Lebesgue's dominated convergence theorem in (3.62) and (3.64), we get

$$f(t) = f_0 + \int_0^t Q(s, f(s))\mathrm{d}s +$$

[7]Inequality (3.64) motivates the construction (3.60) as a second-order recurrence. Indeed, except for the case $\Delta \equiv 0$, an inequality of the form (3.64) could not be proved if (3.60) was redefined with $B(s, f_{n-1}, f_{n-1})$ instead of $B(s, f_{n-1}, f_{n-2})$.

$$+\int_0^t \left[a\left(\|\Lambda f(s)\| + \int_0^s \Delta(\tau, f(\tau))\mathrm{d}\tau\right) - a(\|\Lambda f_0\|)\right] \Lambda f(s)\mathrm{d}s \quad (t \geq 0) \tag{3.68}$$

(i.e., f is a strong solution of Eq.(3.54)) and, also,

$$0 \leq \psi(t) := \|\Lambda f_0\| - \|\Lambda f(t)\| - \int_0^t \Delta(s, f(s))\mathrm{d}s \quad (t \geq 0). \tag{3.69}$$

Obviously, (3.68) implies $f, \Lambda f \in C(\mathbb{R}_+; X_+)$.

Note now the usefulness of (3.68): to prove that f is a strong solution of (3.31), it is sufficient to show that $\psi \equiv 0$ (which means exactly (3.45)).

To this end, first observe that since, by (A_2), a is non-decreasing and locally Lipschitz, then inequality (3.69) implies that there is a number $0 < c = c(\|\Lambda f_0\|)$, depending only on $\|\Lambda f_0\|$, such that

$$0 \leq a(\|\Lambda f_0\|) - a\left(\|\Lambda f(t)\| + \int_0^t \Delta(\tau, f(\tau))\mathrm{d}\tau\right) < c\psi(t). \tag{3.70}$$

Further rewriting Eq. (3.68) conveniently, and applying Λ to the resulting equation, one can invoke (3.26) and (3.27) to obtain

$$\psi(t) = \int_0^t \left[a(\|\Lambda f_0\|) - a\left(\|\Lambda f(s)\| + \int_0^s \Delta(\tau, f(\tau))\mathrm{d}\tau\right)\right] \left\|\Lambda^2 f(s)\right\| \mathrm{d}s. \tag{3.71}$$

As $f(t)$ satisfies (3.47), introducing (3.70) in (3.71), we find

$$0 \leq \psi(t) \leq c\int_0^t \psi(s) \left\|\Lambda^2 f(s)\right\| \mathrm{d}s \leq c_T \int_0^t \psi(s)\mathrm{d}s \quad (0 \leq t \leq T), \tag{3.72}$$

for each $T > 0$. Here, $c_T > 0$ is a number depending only on T and f_0.

Now the Gronwall inequality implies $\psi(t) = 0$, $0 \leq t \leq T$, for any $T > 0$. This concludes the existence part of the proof of the Theorem 3.1a), in the case $f_0 \in \mathcal{D}_+^\infty$).

Step B. We use the result of the previous step to prove the existence part of Theorem 3.1 a), in the case $f_0 \in \mathcal{D}(\Lambda^2) \cap X_+$, as follows. First note that by Lemma 3.1b), there is an increasing sequence $\{f_{0,i}\} \subset \mathcal{D}_+^\infty$ such that $f_{0,i} \nearrow f_0$, as $i \to \infty$. Then, by Step A, there is a sequence of strong solutions $\{F_i\}_i$ of Eq. (3.31) with $F_i(0) = f_{0,i}$, satisfying the properties of the theorem. In particular,

$$\left\|\Lambda^2 F_i(t)\right\| \leq \exp\left[\rho(\|\Lambda_1 f_{0,i}\|)\right] \left\|\Lambda^2 f_{0,i}\right\| \quad (t \geq 0). \tag{3.73}$$

In addition,

$$F_i(t) = f_{0,i} + \int_0^t Q(s, F_i(s))\mathrm{d}s, \tag{3.74}$$

$$\Lambda F_i(t) = \Lambda f_{0,i} + \int_0^t \Lambda Q(s, F_i(s))\mathrm{d}s, \tag{3.75}$$

and

$$\|\Lambda F_i(t)\| + \int_0^t \Delta(s, F_i(s))\mathrm{d}s = \|\Lambda f_{0,i}\|. \tag{3.76}$$

Moreover, by Step A, each F_i is the limit of an increasing sequence $\{f_{n,i}(t)\}_n$ defined by (3.60) with $f_{n,i}(0) = f_{0,i}$. But the positivity of V^t and Lemma 3.3 imply that if $f_{0,i} \le f_{0,j}$, then $f_{n,i}(t) \le f_{n,j}(t)$ for all n and $t \ge 0$. Then the sequence $\{F_i\}$ is increasing.

Furthermore, since $\|\Lambda_1 f_{0,i}\| \le \|\Lambda_1 f_0\|$, $\left\|\Lambda^2 f_{0,i}\right\| \le \left\|\Lambda^2 f_0\right\|$, and since ρ is non-decreasing, it follows from inequality (3.73) that

$$\left\|\Lambda^2 F_i(t)\right\| \le \exp(\rho(\|\Lambda_1 f_0\|)t)\left\|\Lambda^2 f_0\right\| \quad (t \ge 0). \tag{3.77}$$

Now a convergence argument, as in the beginning of Step A, implies that there is an element $f \in L^1_{loc}(\mathbb{R}_+; X_+)$, with the properties stated in Remark 3.2, such that $F_i(t) \nearrow f(t)$ as $i \to \infty$, a.e. It remains to apply, say, Lebesgue's convergence theorem in (3.74)–(3.76) to conclude the existence part of Theorem 3.1a).

Existence in case b). In this case, Lemma 3.4 applies, corresponding to the fulfillment of the conditions of Theorem 3.1b). Then, the proof is as in Step A of case a).

Finally, we prove the **uniqueness** part of Theorem 3.1.

Let f be the solution of Eq. (3.31) provided by the existence part of this proof, and recall that it satisfies Eq. (3.45). If F is *another* positive solution of Eq. (3.31) with regularity properties as in Theorem 3.1, then F satisfies Eq. (3.45), too, hence

$$\|\Lambda f(t)\| + \int_0^t \Delta(s, f(s))\mathrm{d}s = \|\Lambda f_0\| = \|\Lambda F(t)\| + \int_0^t \Delta(s, F(s))\mathrm{d}s.$$

By Lebesgue's convergence theorem applied to (3.60), clearly, f also solves Eq. (3.57). On the other hand, F is a solution to (3.57). But $f \le F$, because of the form of (3.60), so that

$$\|\Lambda f(t)\| + \int_0^t \Delta(s, f(s))\mathrm{d}s < \|\Lambda F(t)\| + \int_0^t \Delta(s, F(s))\mathrm{d}s$$

on some subset of $\mathbb{R}_+$ with nonzero Lebesgue measure. □

Proof of Theorem 3.2

As in the proof of Theorem 3.1, to exclude trivial situations, we suppose the $\|f_0\| \neq 0$ or $a(\|f_0\|) \neq 0$. By Lemma 3.1, there is a sequence $\{f_{n,0}\}_n \subset \mathcal{D}_+^\infty$ such that $f_{n,0} \nearrow f_0$ as $n \to \infty$. Then by Theorem 3.1a) the i.v.p. (3.31) with initial condition $f_{n,0}$ has a unique positive solutions $F_n \in \mathcal{D}(\Lambda^2) \cap X_+$ such that (3.31) provided by Theorem 3.1 with initial datum $f_{n,0}$ forms an increasing sequence such that $F_n, \Lambda F_n \in C(\mathbb{R}_+; X_+)$,

$$F_n(t) = f_{n,0} + \int_0^t Q^+(s, F_n(s))\mathrm{d}s - \int_0^t Q^-(s, F_n(s))\mathrm{d}s \quad (t \geq 0). \tag{3.78}$$

and

$$\|\Lambda F_n(t)\| + \int_0^t \Delta(s, F_n(s)\mathrm{d}s = \|\Lambda f_{n,0}\| \quad (t \geq 0). \tag{3.79}$$

But $\Delta(s, F_n(s) \geq 0$ so that

$$\|\Lambda F_n(t)\| \leq \|\Lambda f_{n,0}\| \leq \|\Lambda f_0\| \quad (t \geq 0). \tag{3.80}$$

Note now that F_n, $f_{n,0}$, $Q^\pm(t, F_n(t))$ are positive. Then (3.26) and (3.27) imply

$$\|F_n(t)\| = \|f_{n,0}\| + \int_0^t \left\|Q^+(s, F_n(s))\right\| \mathrm{d}s - \int_0^t \left\|Q^-(s, F_n(s))\right\| \mathrm{d}s \quad (t \geq 0), \tag{3.81}$$

To prove the theorem, we need show that $\{F_n(t)\}_n$ and $\{Q^\pm(t, F_n(t))\}_n$ are convergent, and, then we need to interchange the limits conveniently in (3.78) and (3.81).

To this end, first observe that since $\{f_{n,0}\}_n$ is positive and increasing, and each F_n is the limit of a sequence of the form (3.60), we obtain by a simple induction (which uses the positivity and isotonicity of B in (3.60)) that $\{F_n(t)\}_n$ is increasing. Thus, by (A_0), the positive sequence $\{\Lambda F_n(t)\}_n$ is also increasing. Then (A_0) and (3.80) give $\|F_n(t)\| \leq {\lambda_0}^{-1} \|\Lambda F_n(t)\| \leq {\lambda_0}^{-1} \|\Lambda f_{n,0}\| \leq {\lambda_0}^{-1} \|\Lambda f_0\|$. Hence, for each $t \geq 0$, both $\{F_n(t)\}_n$ and $\{\Lambda F_n(t)\}_n$ are convergent, because X is monotone complete. Moreover, as Λ is closed, the limit $f(t)$ of $\{F_n(t)\}_n$ satisfies $f(t) \in \mathcal{D}(\Lambda) \cap X_+$, and we have $\Lambda F_n(t) \nearrow \Lambda f(t)$ as $n \to \infty$. Then, also $\{Q^\pm(t, F_n(t))\}_n$ are increasing, and $Q^\pm(t, F_n(t)) \leq Q^\pm(t, f(t))$ a.e. In particular, $\|Q^\pm(t, F_n(t))\| \leq \|Q^\pm(t, f(t))\|$ a.e. Consequently, $Q^\pm(t, F_n(t)) \nearrow Q^\pm(t, f(t))$ as $n \to \infty$, t -a.e., because X is monotone complete and $Q^\pm(t, \cdot)$ are o-closed t-a.e.

Now, applying (A_2) and (3.80) we get

$$\left\|Q^-(t, f(t))\right\| = \lim_{n\to\infty} \left\|Q^-(t, F_n(t))\right\| \leq a(\|\Lambda f_0\|)\, \|\Lambda f_0\| \tag{3.82}$$

a.e., hence $Q^-(\cdot, f) \in L^1_{loc}(\mathbb{R}_+; X_+)$.

Thus we can take the limit $n \to \infty$ in (3.78) and (3.81), and we can apply, say, Lebesgue's theorem to the second term of (3.78) and (3.81), respectively. We obtain

$$f(t) = f_0 + \lim_{n\to\infty} \int_0^t Q^+(s, F_n(s))\mathrm{d}s - \int_0^t Q^-(s, f(s))\mathrm{d}s, \tag{3.83}$$

and, by (3.26),

$$\|f(t)\| = \|f_0\| + \lim_{n\to\infty} \int_0^t \left\|Q^+(s, F_n(s))\right\| \mathrm{d}s - \int_0^t \left\|Q^-(s, f(s))\right\| \mathrm{d}s. \tag{3.84}$$

Since $\|f(t)\| < \infty$ for $t \geq 0$, and $Q^-(\cdot, f) \in L^1_{loc}(\mathbb{R}_+; X_+)$, by (3.84), for each $t \geq 0$,

$$\lim_{n\to\infty} \int_0^t \left\|Q^+(s, F_n(s))\right\| \mathrm{d}s < \infty. \tag{3.85}$$

Hence, applying, e.g., the monotone convergence theorem, it follows that $Q^+(\cdot, f)$ is Bochner integrable and we can finally pass to the limit under the integral sign in (3.83), (3.84), (3.80), and in (3.79), to conclude the proof of theorem. □

Proof of Proposition 3.3

Equality (3.51) follows observing that $\Delta(s, F_n(s)) \equiv 0$ in (3.79), and taking the ∞ limit. As in the uniqueness part of the proof of Theorem 3.1, the solution f of (3.31) provided by Theorem 3.2 also solves the mild problem (3.57) (but here, $\Delta(t, f) = 0$ in the expression (3.55) of B, by virtue of (3.51)). Now the uniqueness follows by an argument similar to the one used in the uniqueness part of the proof of Theorem 3.1, taking now advantage of the property $\Delta(s, F_n(s)) \equiv 0$ (hence of (3.51)). □

Proof of Proposition 3.4

a) Let f_0, $\{f_{0,i}\}, \{f_{n,i}(t)\}_n$, and $\{F_i(t)\}_i$ be as in Step B of the proof of Theorem 3.1a). Then for each i, the sequence $\{\Gamma f_{n,i}(t)\}_n$ is positive and increasing. Moreover, it is norm-bounded because

$$\|\Gamma f_{n,i}(t)\| \leq \|\Gamma f_0\| \quad (t \geq 0), \tag{3.86}$$

as a consequence of (3.65) and of the property $\Gamma f_{0,i} \leq \Gamma f_0$.

As X is monotone complete, it follows that $\{\Gamma f_{n,i}(t)\}_n$ is convergent for all i.

Recall that Γ is closed, and $f_{n,i}(t) \nearrow F_i(t)$ as $n \to \infty$, for all i. Consequently, $F_i(t) \in \mathcal{D}(\Gamma)$ and $\Gamma f_{n,i}(t) \nearrow \Gamma F_i(t)$ as $n \to \infty$, $i = 1, 2, ...$ In addition, $\|\Gamma F_i\| \leq \|\Gamma f_0\|$, $t \geq 0$,$i = 1, 2, ...$ Then, reasoning as before, we conclude that $f(t) \in \mathcal{D}(\Gamma)$, $\Gamma F_i(t) \nearrow \Gamma f(t)$ as $i \to \infty$, and that $\|\Gamma f\|$ satisfies (3.52).

b) The proof of (3.53) follows as in a), with the only remark that instead of (3.86), we make use of the inequalities

$$\|\Gamma f_{n,i}(t)\| \leq \exp(\rho_\Gamma(\|\Lambda_1 f_{0,i}\|)t)\ \|\Gamma f_{0,i}\| \leq \exp(\rho_\Gamma(\|\Lambda_1 f_0\|)t)\,\|\Gamma f_0\| \quad (t \geq 0), \tag{3.87}$$

which are immediate by (3.67), because ρ_Γ is non-decreasing. □

4. Applications

4.1. Smoluchowski's coagulation equation

For $k \geq 0$, let $L^1_k := L^1_k(\mathbb{R}_+; \mathrm{d}y)$ be the space of real measurable functions $g : \mathbb{R}_+ \mapsto \mathbb{R}$ such that

$$\|g\|_{L^1_k} := \int_{\mathbb{R}_+} (1+y)^k\, |g(y)|\, \mathrm{d}y < \infty. \tag{4.1}$$

Denote $L^1_{k,+} = \{g \in L^1_k : g \geq 0\}$. Consider problem (2.2) in the space $X = L^1(\mathbb{R}_+; \mathrm{d}y)$ (equipped with the usual norm $\|\cdot\| = \|\cdot\|_{L^1}$, and with the natural order $\leq$).

Consider L^1_k as a subset of X. Let $i = 0, 1$ and define the positive linear operators $\Lambda_{c,i} : \mathcal{D}(\Lambda_{c,i}) \subset X \mapsto X$ by $\mathcal{D}(\Lambda_{c,i}) = L^1_{\gamma_i}$, $(\Lambda_{c,i} g)(y) := \lambda_i(y) g(y)$, with $\lambda_i(y) := (1+y)^{\gamma_i}$, $y \geq 0$ a.e., where $\gamma_0 = \beta$ and $\gamma_1 = \alpha + \beta$.

Note that (2.3) and (2.4) define Q^+_c and Q^-_c as positive and isotone nonlinear operators in X, respectively, with the common domain $\mathcal{D}_c := L^1_\beta$.

Then the i.v.p. for (2.2) can be formulated in X as

$$\frac{\mathrm{d}}{\mathrm{d}t} f = Q_c(f) = Q^+_c(f) - Q^-_c(f) \quad f(0) = f_0, \quad t > 0. \tag{4.2}$$

In this case, one can apply Theorem 3.1a). The only point is to check that $\Lambda_{c,i}$ $(i = 0, 1)$ and $Q^\pm_c$ verify inequalities of the form (3.40) and (3.42). Indeed, if $g \in L^1_{2\beta,+}$, then starting from (2.7), we find

$$0 \leq \left\|\Lambda_{c,i} Q^-_c(g)\right\| - \left\|\Lambda_{c,i} Q^+_c(g)\right\| =$$

$$= \frac{1}{2}\int_{\mathbb{R}^2_+} [(1+y)^{\gamma_i} + (1+y_*)^{\gamma_i} - (1+y+y_*)^{\gamma_i}] q(y,y_*)g(y)g(y_*)\mathrm{d}y\mathrm{d}y_*, \tag{4.3}$$

because $0 \le \gamma_i \le 1$, and

$$\frac{(1+y)^\gamma + (1+y_*)^\gamma}{(1+y+y_*)^\gamma} \ge \inf_{x\ge 0} \frac{1+x^\gamma}{(1+x)^\gamma} = 1 \quad (0 \le \gamma \le 1, \quad y, y' \ge 0). \tag{4.4}$$

Inequality (4.3) shows that $g \mapsto \Delta_c(g) := \|\Lambda_{c,0} Q_c^-(g)\| - \|\Lambda_{c,0} Q_c^+(g)\|$ defines a positive isotone map $\Delta_c : \mathcal{D}(\Delta_c) \mapsto \mathbb{R}$ with domain $\mathcal{D}(\Delta_c) = L^1_{2\beta,+}$.

Starting again from (2.7), we find that if $g \in L^1_{3\beta,+}$, then

$$\left\|\Lambda^2_{c,0} Q_c^+(g)\right\| - \left\|\Lambda^2_{c,0} Q_c^-(g)\right\| =$$

$$= \frac{1}{2}\int_{\mathbb{R}^2_+} \left[(1+y+y_*)^{2\beta} - (1+y)^{2\beta} - (1+y_*)^{2\beta}\right] q(y,y_*)g(y)g(y_*)\mathrm{d}y\mathrm{d}y_*. \tag{4.5}$$

If $0 \le \beta \le 1/2$, applying again (4.4) in (4.5), we get

$$\left\|\Lambda^2_{c,0} Q_c^+(g)\right\| - \left\|\Lambda^2_{c,0} Q_c^-(g)\right\| \le 0, \tag{4.6}$$

which is of the form (3.42) with $\rho \equiv 0$.

If $1/2 < \beta \le 1$, then to estimate (4.5), we apply the following form ([11]) of Povzner's algebraic inequality, which can be easily proved[8]:

$$(1+y+y_*)^{2\beta} - (1+y)^{2\beta} - (1+y_*)^{2\beta} \le 2(1+y)^\beta (1+y_*)^\beta \quad (y, y_* \ge 0). \tag{4.7}$$

Thus, applying (4.7) in (4.5), we find that there is a number $c > 0$ such that

$$\left\|\Lambda^2_{c,0} Q_c^+(g)\right\| - \left\|\Lambda^2_{c,0} Q_c^-(g)\right\| \le c \left\|\Lambda_{c,1} g\right\| \left\|\Lambda^2_{c,0} g\right\|. \tag{4.8}$$

Clearly, inequality (4.8) is of the form (3.42) with $\rho(x) = cx$.

Let $a_c(x) := a_0 x$, for some constant $a_0 > 0$. If a_0 is sufficiently large, then the map $L^1_{\beta,+} \ni g \mapsto a_0 \left\|\Lambda_{c,0} g\right\| \Lambda_{c,0} g - Q_c^-(g) \in X$ has the properties required in (A_2).

It appears that $Q_c^\pm$, $\Lambda_{c,0}$, $\Lambda_{c,1}$ and a_c verify the conditions of Theorem 3.1a) for $Q^\pm$, Λ, Λ_1 and a, respectively, provided that a_0 is sufficiently large. Consequently, one can apply Theorem 3.1a) to the i.v.p. (4.2). We obtain

[8] Indeed, (4.7) is equivalent to $\zeta(x) = 2x^\beta + 1 + x^{2\beta} - (1+x)^{2\beta} \ge 0$ for all $x > 0$. However, as $\zeta(x^{-1}) = x^{-2\beta}\zeta(x)$, to prove that $\zeta(x) \ge 0$ for $x > 0$, we need only show that $\zeta(x) \ge 0$ on $(0, 1]$, which is immediate, because $1/2 < \beta \le 1$.

THEOREM **4.1** *Let $f_0 \in L^1_{2\beta,+}$ in problem* (4.2). *Then Eq.* (4.2) *has a unique strong solution f such that $f(t) \in L^1_{2\beta,+}$, $t \geq 0$, and $\|f(t)\|_{L^1_{2\beta}}$ is locally bounded on $\mathbb{R}_+$. In addition $f, (1+y)^\beta f \in C(\mathbb{R}_+; L^1(\mathbb{R}_+, \mathrm{d}y))$,*

$$\|f(t)\|_{L^1_\beta} + \int_0^t \Delta_c(f(s))\mathrm{d}s = \|f_0\|_{L^1_\beta} \quad (t \geq 0), \tag{4.9}$$

and there is a constant $c > 0$ such that

$$\|f(t)\|_{L^1_{2\beta}} \leq \exp(c\,\|f_0\|_{L^1_{\alpha+\beta}}\, t)\,\|f_0\|_{L^1_{2\beta}} \quad (t \geq 0). \tag{4.10}$$

Note here that if $0 \leq 2\beta < 1$, then Theorem 4.1 allows for the existence of solutions with infinite initial mass (see also [22]) i.e., $f_0 \in L^1_{2\beta,+}$, but $f_0 \notin L^1_1$. The theorem does not imply directly the mass conservation, except for the case $q_1 > 0$, $\beta = 1$ and $\alpha = 0$. However, if $f_0 \in L^1_{2\beta,+} \cap L^1_1$, then the solution $f(t)$ has finite mass: indeed, if $\Gamma : L^1_1 \subset L^1 \mapsto L^1$ is defined by $(\Gamma g)(y) = yg(y)$ a.e. on $\mathbb{R}_+$, then clearly, Γ is of type D on $\cap_{k=1}^\infty L^1_{k\beta,+}$, hence Prop. 3.4a) applies, so that $f \in L^1_{2\beta,+} \cap L^1_1$, and $\|\Gamma f(t)\| \leq \|\Gamma f_0\|$.

Theorem 4.1 remains valid in the case of the discrete Smoluchowski equation (2.10), with obvious change in formulation[9].

4.2. Povzner-like model with dissipative collisions

Let $X = L^1(\mathbb{R}^3 \times \mathbb{R}^3; \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}) = L^1$, equipped with the norm $\|\cdot\| := \|\cdot\|_{L^1}$ and the natural order $\leq$. Denote by $L^1_k := L^1_k(\mathbb{R}^3 \times \mathbb{R}^3; \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v})$, $k \in \mathbb{R}$, the space of measurable functions on $g : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ satisfying

$$\|g\|_{L^1_k} := \int_{\mathbb{R}_+} (1 + |\mathbf{v}|^2)^{\frac{k}{2}}\, |g(\mathbf{x}, \mathbf{v})|\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} < \infty. \tag{4.11}$$

As before, $L^1_{k,+}$ denotes the positive cone in L^1_k. It can be seen that (2.15) and (2.16) define $Q^\pm_d$ as positive and isotone operators on the common domain $\mathcal{D} := L^1_\gamma$. This follows easily if we perform the change of variable $(0, R] \times \Omega \ni (r, \mathbf{n}) \mapsto \mathbf{y} := r\mathbf{n} \in \{\mathbf{z} \in \mathbb{R}^3 : |\mathbf{z}| \leq R\}$ in (2.15) and (2.16), and then take into account (2.17).

Now, formulated in X, the i.v.p. (2.14) reads

$$\frac{\mathrm{d}}{\mathrm{d}t} f = Af + Q^+_d(f) - Q^-_d(f), \quad f(0) = f_0 \geq 0, \tag{4.12}$$

[9]Note that L^1_r, defined before, must be now replaced by $l^1_r(\mathbb{R}) = \{c = (c_j) : c_j \in \mathbb{R},\, j = 1, 2, ...,\, \|c\|_r := \sum_{j=1}^\infty j^r\, |c_j| < \infty\}$, $r \geq 0$.

where $f = f(t, \mathbf{x}, \mathbf{v})$ is the one-particle distribution function, A is the infinitesimal generator of the C_0 group $(U^t f)(\mathbf{x}, \mathbf{v}) := f(\mathbf{x} - t\mathbf{v}, \mathbf{v})$, a.e.

Let the positive linear operator $\Lambda_d : L_2^1 \mapsto X$ be defined by $(\Lambda_d g)(\mathbf{x}, \mathbf{v}) := \lambda(\mathbf{v}) g(\mathbf{x}, \mathbf{v})$ a.e. on $\mathbb{R}^3 \times \mathbb{R}^3$, with $\lambda(\mathbf{v}) := (1 + |\mathbf{v}|^2)$. Define $a_d(x) := c_0 x$ for some constant $c_0 > 0$. If c_0 is sufficiently large, then a_d, Λ_d and $Q_d^{\pm}$ verify the conditions of Corollary 3.1 for a, $\Lambda = \Lambda_1$ and $Q^{\pm}$, respectively.

Indeed, the operators $Q_d^{\pm}$ are p-saturated. Moreover, they are o-closed, by the monotone convergence theorem. It is immediate that the domain conditions imposed in Corollary 3.1 are satisfied. Further, applying (2.12) in (2.18), we obtain an inequality of the form (3.40), i.e., if $g \in L_{4,+}^1$, then

$$0 \leq \Delta_d(g) := \left\|\Lambda_d Q_d^-(g)\right\| - \left\|\Lambda_d Q_d^+(g)\right\| =$$

$$= \int_0^R \mathrm{d}r \int_{\Omega \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3} \pi(r, \mathbf{n}, \mathbf{v}, \mathbf{w}, \mathbf{x}) g(\mathbf{x}, \mathbf{v}) g(\mathbf{x} + r\mathbf{n}, \mathbf{w}) \mathrm{d}\mathbf{n}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{x}, \tag{4.13}$$

where $\pi(r, \mathbf{n}, \mathbf{v}, \mathbf{w}, \mathbf{x}) := \beta(\mathbf{n})(1 - \beta(\mathbf{n}))\, |\langle \mathbf{n}, \mathbf{v} - \mathbf{w} \rangle|^{2+\gamma} P(r, \mathbf{n})$. Remark here that the map $L_{4,+}^1 \ni g \mapsto \Delta_d(g) \in \mathbb{R}$ is positive and isotone. Moreover, for c_0 sufficiently large, the map $L_{2,+}^1 \ni g \mapsto c_0 \|\Lambda_d g\| \Lambda_d g - Q_d^-(g) \in X$ is also positive and isotone. Further, to obtain an inequality of the form (3.42), note that (2.12) gives $\lambda(\mathbf{v}')^2 + \lambda(\mathbf{w}')^2 \leq (2 + |\mathbf{v}'|^2 + |\mathbf{w}'|^2)^2 \leq (2 + |\mathbf{v}|^2 + |\mathbf{w}|^2)^2 = \lambda(\mathbf{v})^2 + \lambda(\mathbf{w})^2 + 2\lambda(\mathbf{v})\lambda(\mathbf{w})$, which can be applied in (2.18) to conclude easily that there are two constants c_1, $c > 0$ such that

$$\left\|\Lambda_d^2 Q_c^+(g)\right\| - \left\|\Lambda_d^2 Q_d^-(g)\right\| \leq$$

$$\leq c_1 \int_0^R \mathrm{d}r \int_{\Omega \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3} r^2 \lambda(\mathbf{v}) \lambda(\mathbf{w})^{1+\frac{\gamma}{2}} g(\mathbf{x}, \mathbf{v}) g(\mathbf{x} + r\mathbf{n}, \mathbf{w}) \mathrm{d}\mathbf{n}\mathrm{d}\mathbf{v}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{x} \leq$$

$$\leq c \|\Lambda_d g\| \left\|\Lambda_d^2 g\right\|, \tag{4.14}$$

for all $g \in L_{6,+}^1$. Finally, it is obvious that the group U^t (generated by A) commutes with the semigroup V^t generated by Λ_d, and $\Lambda^k Q^+(U^{\cdot} g) \in L_{loc}^1(\mathbb{R}_+; X_+)$ for all $g \in \cap_{n=1}^{\infty} L_{n,+}^1$, $k = 1, 2, ...$

Therefore, by Corollary 3.1, we have the following result ([11]):

THEOREM **4.2** *Let $f_0 \in L_{4,+}^1$ in problem (4.12). Then Eq. (4.12) has a unique positive mild solution f such that $f(t) \in L_{4,+}^1$, $t \geq 0$, and $\|f(t)\|_{L_4^1}$ is locally bounded on $\mathbb{R}_+$. In addition, f, $(1 + |\mathbf{v}|^2) f \in C(\mathbb{R}_+; L^1)$,*

$$\|f(t)\|_{L_2^1} + \int_0^t \Delta_d(f(s)) \mathrm{d}s = \|f_0\|_{L_2^1} \quad (t \geq 0), \tag{4.15}$$

and there is a constant $c > 0$ such that

$$\|f(t)\|_{L^1_4} \leq \exp(c\,\|f_0\|_{L^1_2}\,t)\,\|f_0\|_{L^1_4} \quad (t \geq 0). \tag{4.16}$$

The argument of Theorem 4.2 can be repeated with obvious modifications to provide a similar result for the space-homogeneous version of Eq. (2.14), which coincides with the force-free, three dimensional space-homogeneous Boltzmann model for granular flows, [5, 6].

4.3. Povzner-like model with chemical reactions

Let $X := L^1(\mathbb{R}^3 \times \mathbb{R}^3; d\mathbf{x}\mathrm{d}\mathbf{v})^N$ be equipped with the order $\leq$ induced by the order of the components (i.e., the natural order of L^1). The norm on X is defined as

$$\|g\| := \sum_{i=1}^{N} \int\limits_{\mathbb{R}^3\times\mathbb{R}^3} |g_i(\mathbf{x},\mathbf{v})|\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} = \sum_{i=1}^{N} \|g_i\|_{L^1}\,. \tag{4.17}$$

Denote by $L^1_k := L^1_k(\mathbb{R}^3\times\mathbb{R}^3; \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v})$, $k \in \mathbb{R}$, the space of measurable functions $g : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ satisfying

$$\|g\|_{L^1_k} := \int\limits_{\mathbb{R}^3\times\mathbb{R}^3} (1 + |\mathbf{v}|^2)^{\frac{k}{2}}\,|g(\mathbf{x},\mathbf{v})|\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} \tag{4.18}$$

and let $L^1_{k,+}$ be the positive cone in L^1_k.

It is natural to formulate the i.v.p. (2.29) in the space X.

Under the conditions of the model, (2.30) and (2.31) define Q^+_i and Q^-_i, $1 \leq i \leq N$, as operators from the common domain $(L^1_2)^N \subset X$ to $L^1(\mathbb{R}^3; \mathrm{d}\mathbf{v})$. Defining the operators $Q^\pm_B : (L^1_2)^N \subset X \mapsto X$ by $Q^\pm_B = (Q^\pm_1, ..., Q^\pm_N)$, we can write the i.v.p. for Eq. (2.29) in X as

$$\frac{\mathrm{d}}{\mathrm{d}t}f + A = Q^+_B(t,f) - Q^-_B(t,f), \quad 0 \leq f(0) = f_0 \in X \quad (t > 0), \tag{4.19}$$

where A is the infinitesimal generator of the C_0 group of isometries $\{U^t\}_{t\in\mathbb{R}}$ on X, given by $(U^t f)(\mathbf{x},\mathbf{v}) := f((\mathbf{x} - t\mathbf{v}, \mathbf{v})$.

Define the positive closed linear operator $\Lambda_B : (L^1_2)^N \mapsto X$ by $(\Lambda_B g)_i(\mathbf{v}) = \lambda_i(\mathbf{v})g(\mathbf{v})$ a.e. on $\mathbb{R}^3 \times \mathbb{R}^3$, where $\lambda_i(\mathbf{v}) := m_i + m_i\,|\mathbf{v}|^2/2 + E_i$, $1 \leq i \leq N$. One can state the following result ([12]):

THEOREM **4.3** *Suppose that in problem* (4.19), $f_{0,i} \in L^1_{4,+}$, $1 \leq i \leq N$. *Then Eq.* (4.19) *has a unique mild solution* $f(t) = (f_1, ..., f_N)$ *such that* $f_i(t) \in L^1_{4,+}$, $t \geq 0$, *and* $\|f_i(t)\|_{L^1_4}$ *is locally bounded on* $\mathbb{R}_+$, $1 \leq i \leq N$. *In addition,* f_i, $(1 + |\mathbf{v}|^2)f_i \in C(\mathbb{R}_+; L^1)$, $1 \leq i \leq N$,

$$\|\Lambda_B f(t)\| = \|\Lambda_B f_0\| \quad (t \geq 0), \tag{4.20}$$

and there is a constant $\rho_0 > 0$ *such that*

$$\left\|\Lambda_B^2 f(t)\right\| \leq \exp(\rho_0 \|\Lambda_B f_0\| t) \left\|\Lambda_B^2 f_0\right\| \quad (t \geq 0). \tag{4.21}$$

The above result follows by applying Theorem 3.1 in the case $\Lambda = \Lambda_1 = \Lambda_B$. Indeed, the domain conditions of Theorem 3.1, as well as properties (A_0), (A_1) can be immediately checked (with $\Delta = 0$, owing to (2.38). Next, let $a_0 > 0$ be some constant, and define $a(x) := a_0 x$. Owing to (2.38), for a_0 sufficiently large, the map $L^1_{2,+} \ni g \to a_0 \|\Lambda_B g\| \Lambda_B g - Q^-(g) \in X$ satisfies (A_2). Finally, note that, as a consequence of (2.39) (and of (2.37)), there exists a number $\rho_0 > 0$ such that

$$\sum_{i=1}^N \int_{\mathbb{R}^3} (\Psi_i^{(0)} + \Psi_i^{(4)})^2 \left[Q_i^+(g) - Q_i^-(g)\right] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v} \leq$$

$$\leq \rho_0 \left\|(1 + |\mathbf{v}|^4)g\right\|_{L^1} \left\|(1 + |\mathbf{v}|^2)g\right\|_{L^1}, \tag{4.22}$$

for, say, all $g \in (L^1_{6+})^N$.

Then inequality (3.13) gives exactly (A_3) with $\rho(x) := \rho_0 x$.

4.4. Boltzmann model with inelastic collisions and reactions

Let $X := (L^1(\mathbb{R}^3; \mathrm{d}\mathbf{v}))^N$ be equipped with the order $\leq$ induced by the order of the components (i.e., the natural order of L^1). The norm on X is defined as

$$\|g\| := \sum_{i=1}^N \int_{\mathbb{R}^3} |g_i(\mathbf{v})| \, \mathrm{d}\mathbf{v} = \sum_{i=1}^N \|g_i\|_{L^1}. \tag{4.23}$$

Denote by $L^1_k := L^1_k(\mathbb{R}^3; \mathrm{d}\mathbf{v})$, $k \in \mathbb{R}$, the space of measurable functions $g : \mathbb{R}^3 \times \mathbb{R}^3 \mapsto \mathbb{R}$ satisfying

$$\|g\|_{L^1_k} := \int_{\mathbb{R}_+} (1 + |\mathbf{v}|^2)^{\frac{k}{2}} |g(\mathbf{v})| \, \mathrm{d}\mathbf{v} < \infty \tag{4.24}$$

and let $L^1_{k,+}$ be the positive cone in L^1_k.

It is natural to formulate the i.v.p. for Eq. (2.47) in the space X. Under the above conditions, (2.48) and (2.49) define Q_i^+ and Q_i^-, $1 \le i \le N$, respectively, as operators from the common domain $\mathcal{D} = (L^1_2)^N \subset X$ to $L^1(\mathbb{R}^3; d\mathbf{v})$. Defining $Q_B^\pm : \mathcal{D} \subset X \mapsto X$ by $Q_B^\pm = (Q_1^\pm, ..., Q_N^\pm)$, we can write the i.v.p. for Eq. (2.47) in X

$$\frac{\mathrm{d}}{\mathrm{d}t} f = Q_B^+(f) - Q_B^-(f), \quad f(0) = f_0 = (f_{0,1}, ..., f_{0,N}) \in X_+. \tag{4.25}$$

We shall prove the existence of solutions to problem (4.25), by applying Theorem 3.1a) (in the case $\Lambda = \Lambda_1$). To this end, let the positive closed linear operator $\Lambda_B : (L^1_2)^N \mapsto X$ be defined on components by $(\Lambda_B g)_i(\mathbf{v}) = \lambda_i(\mathbf{v}) g(\mathbf{v})$ a.e. on $\mathbb{R}^3 \times \mathbb{R}^3$, where $\lambda_i(\mathbf{v}) := m_i + m_i |\mathbf{v}|^2 /2 + E_i$, $1 \le i \le N$. Denote $l_\gamma(\mathbf{w}) := \sum_{i \in \mathcal{N}(\gamma)} \sum_{j=1}^{\gamma_i} \lambda_i(\mathbf{w}_{i,j})$; $\gamma \in \mathcal{M}$. Then clearly, $l_\gamma(\mathbf{w}) = M_\gamma + W_\gamma(\mathbf{w})$, hence

$$0 \le W_\gamma(\mathbf{w}) < l_\gamma(\mathbf{w}). \tag{4.26}$$

In addition, defining $\lambda^\gamma(\mathbf{w}) := \prod_{i \in \mathcal{N}(\gamma)} \prod_{j=1}^{\gamma_i} \lambda_i(\mathbf{w}_{i,j})$, $\gamma \in \mathcal{M}$, we have

$$l_\gamma(\mathbf{w}) \le |\gamma|\, E^{1-|\gamma|} \lambda^\gamma(\mathbf{w}), \tag{4.27}$$

where $E := \min\{m_i + E_i : 1 \le i \le N\}$. It is useful to remark that, since $W_\gamma(\mathbf{w}) \ge E\,|\gamma| > 0$, and $0 \le q \le 1$, then by (2.56), (4.26) and (4.27),

$$\nu_{\beta,\alpha}(\mathbf{w}) \le C \lambda^\alpha(\mathbf{w}) \quad (\mathbf{w} \in \mathbb{R}^{|\alpha|}, \, a.e.), \tag{4.28}$$

for all $\alpha, \beta \in \mathcal{M}$. Here $C = C(E, K) > 0$ is a number depending on E and K (recall that K is the maximum number of partners in a reaction channel).

To apply Theorem 3.1a) to (4.25), first remark that $Q_B^\pm$ and Λ_B verify the domain conditions imposed to $Q^\pm$ and Λ by the theorem. Moreover, Λ_B has the properties required for Λ in (A_0). Further, observe that formula (2.57) provides a correspondent to (3.40), specifically,

$$\Delta_B(g) := \left\| \Lambda_B Q_B^-(g) \right\| - \left\| \Lambda_B Q_B^+(g) \right\| = 0 \quad (g \in (L^1_{4,+})^N). \tag{4.29}$$

To obtain a correspondent to (3.42), let $s_\gamma(\mathbf{w}) := \sum_{i \in \mathcal{N}(\gamma)} \sum_{j=1}^{\gamma_i} \lambda_i(\, \mathbf{w}_{i,j})^2$. Next, using the definition of Q_B^+ and property (B_2), and applying the obvious inequality $s_\alpha(\mathbf{w}) \le l_\alpha(\mathbf{w})^2$, we find that if $g \in (L^1_{6,+})^N$, then

$$\left\| \Lambda_B^2 Q_B^+(g) \right\| = \sum_{\alpha,\beta \in \mathcal{M}} \int_{\mathbb{R}^{3|\alpha|} \times \Omega_\beta} s_\alpha(\mathbf{w}) p_{\beta,\alpha}(\mathbf{w}, \mathbf{n}) (g^\beta \circ u_{\beta,\alpha})(\mathbf{w}, \mathbf{n}) \mathrm{d}\mathbf{w} \mathrm{d}\mathbf{n} \le$$

$$\leq \sum_{\alpha,\beta\in\mathcal{M}} \int\limits_{\mathbb{R}^{3|\alpha|}\times\Omega_\beta} l_\alpha(\mathbf{w})^2 p_{\beta,\alpha}(\mathbf{w},\mathbf{n})(g^\beta\circ u_{\beta,\alpha})(\mathbf{w},\mathbf{n})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}. \tag{4.30}$$

We apply property (3.9) in the last integral. Then interchanging α and β, we get

$$\left\|\Lambda_B^2 Q_B^+(g)\right\| \leq \sum_{\alpha,\beta\in\ \mathcal{M}} \int\limits_{\mathbb{R}^{3|\alpha|}\times\Omega_\beta} (l_\beta\circ u_{\beta,\alpha})^2(\mathbf{w},\mathbf{n}) r_{\beta,\alpha}(\mathbf{w},\mathbf{n}) g^\alpha(\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}. \tag{4.31}$$

Since $l_\beta(\mathbf{w}) = M_\beta + W_\beta(\mathbf{w})$, property (B_3) implies that $(l_\beta\circ u_{\beta,\alpha})(\mathbf{w},\mathbf{n}) = l_\alpha(\mathbf{w})$ for all $(\alpha,\beta)\in\mathcal{C}_M$, $\mathbf{w}\in D^+_{\beta,\alpha}$. This and (B_1) enable us to deduce from (4.31) that

$$\left\|\Lambda_B^2 Q_B^+(g)\right\| \leq \sum_{\alpha,\beta\in\mathcal{M}_{\mathbb{R}^{3|\alpha|}\times\Omega_\beta}} \int l_\alpha(\mathbf{w})^2 r_{\beta,\alpha}(\mathbf{w},\mathbf{n}) g^\alpha(\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}. \tag{4.32}$$

Now, using the definitions of $l_\alpha(\mathbf{w})$ and Q_B^-, and then, taking advantage of (2.56) and (4.26), we obtain from (4.32)

$$\left\|\Lambda_B^2 Q_B^+(g)\right\| \leq$$

$$\leq \sum_{\alpha,\beta\in\mathcal{M}_{\mathbb{R}^{3|\alpha|}\times\Omega_\beta}} \int s_\alpha(\mathbf{w}) r_{\beta,\alpha}(\mathbf{w},\mathbf{n}) g^\alpha(\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} + \rho_B(\|\Lambda_B g\|)\left\|\Lambda_B^2 g\right\| =$$

$$= \left\|\Lambda_B^2 Q_B^-(g)\right\| + \rho_B(\|(\Lambda_B g\|)\left\|\Lambda_B^2 g\right\|, \tag{4.33}$$

where ρ_B is a positive non-decreasing (polynomial) function.

Therefore, the last inequality is the required correspondent to (3.42) (in the case $\Lambda = \Lambda_1$).

Further, let $a_0 > 0$ be some constant, and define $a(x) := a_0\sum_{p=1}^{NK} x^p$, $x \geq 0$. Therefore, $a(\|\Lambda_B g\|) = a_0\sum_{p=1}^{NK}\|\Lambda_B g\|^p$. But each term $\|\Lambda_B g\|^p$ in the r.h.s of the last equality can be expressed by (4.23), and the resulting expression can be expanded by the multinomial formula. Then, after some elementary algebra we get the following useful expression

$$a(\|\Lambda_B g\|) = a_0 \sum_{\gamma\in\mathcal{M},\,|\gamma|\geq 1} c_{\gamma,i} \int\limits_{\mathbb{R}^{3|\gamma|}} \lambda^\gamma(\mathbf{w}) g^\gamma(\mathbf{w})\mathrm{d}\mathbf{w}, \tag{4.34}$$

where $c_{\gamma,i} > 0$ are strictly positive, constant coefficients, $\gamma\in\mathcal{M}$, $|\gamma|\geq 1$, $1\leq i\leq N$.

We show that if a_0 is large enough, then $(L^1_{2,+})^N \ni g \mapsto a(\|\Lambda_B f\|)\Lambda_B g - Q_B^-(g) \in X$ is positive and isotone. To this end, first note that one can write

$$Q_i^-(g)(\mathbf{v}) = R_i(g)(\mathbf{v})\, g_i(\mathbf{v}), \quad (g \in (L^1_{2,+})^N,\, \mathbf{v} \in \mathbb{R}^3\, a.e.,\, 1 \le i \le N), \tag{4.35}$$

where

$$R_i(g)(\mathbf{v}) := \sum_{\alpha,\beta\in\mathcal{M}} \alpha_i \int\limits_{\mathbb{R}^{3|\alpha|-3}} \left[\nu_{\beta,\alpha}(\mathbf{w}) \prod_{\substack{s\in\mathcal{N}(\alpha)\\(s,j)\neq(i,\alpha_i)}} \prod_{j=1}^{\alpha_s} g_s(\mathbf{w}_{s,j}) \right]_{\mathbf{w}_{i,\alpha_i}=\mathbf{v}} \mathrm{d}\tilde{\mathbf{w}}_i, \tag{4.36}$$

with $\nu_{\beta,\alpha}$ as in (2.56). Hence,

$$a(\|\Lambda_B g\|)(\Lambda_B g)_i(\mathbf{v}) - Q_i^-(g)(\mathbf{v}) = [a(\|\Lambda_B g\|)\lambda_i(\mathbf{v}) - R_i(g)(\mathbf{v})]\, g_i(\mathbf{v}). \tag{4.37}$$

It is convenient to set

$$R_i^A(g)(\mathbf{v}) := C \sum_{\alpha,\beta\in\mathcal{M}} \alpha_i \int\limits_{\mathbb{R}^{3|\alpha|-3}} \left[\lambda^\alpha(\mathbf{w}) \prod_{\substack{s\in\mathcal{N}(\alpha)\\(s,j)\neq(i,\alpha_i)}} \prod_{j=1}^{\alpha_s} g_s(\mathbf{w}_{s,j}) \right]_{\mathbf{w}_{i,\alpha_i}=\mathbf{v}} \mathrm{d}\tilde{\mathbf{w}}_i, \tag{4.38}$$

with C as in (4.28). Summing on β in (4.38), using the explicit form of $\lambda^\alpha(\mathbf{w})$, and invoking property (B_1), we are easily led to

$$R_i^A(g)(\mathbf{v}) = C\lambda_i(\mathbf{v}) \sum_{\gamma\in\mathcal{M},\,|\gamma|\ge 1} q_{\gamma,i} \int\limits_{\mathbb{R}^{3|\gamma|}} \lambda^\gamma(\mathbf{w}) g^\gamma(\mathbf{w})\, \mathrm{d}\mathbf{w}, \tag{4.39}$$

where $q_{\gamma,i} \ge 0$ are constant coefficients, $\gamma \in \mathcal{M}$, $|\gamma| \ge 1$, $1 \le i \le N$.

We introduce (4.34) and (4.38) in (4.37). Consequently, for $\mathbf{v} \in \mathbb{R}^3$ a.e.,

$$a(\|\Lambda_B g\|)(\Lambda_B g)_i(\mathbf{v}) - Q_i^-(g)(\mathbf{v}) = [R_i^A(g)(\mathbf{v}) - R_i(g)(\mathbf{v})]g_i(\mathbf{v}) + T_i(g)(\mathbf{v}), \tag{4.40}$$

where

$$T_i(g)(\mathbf{v}) := \lambda_i(\mathbf{v})g_i(\mathbf{v}) \sum_{\gamma\in\mathcal{M},\,|\gamma|\ge 1} (a_0 c_{\gamma,i} - C q_{\gamma,i}) \int\limits_{\mathbb{R}^{3|\gamma|}} \lambda^\gamma(\mathbf{w}) g^\gamma(\mathbf{w}) \mathrm{d}\mathbf{w}. \tag{4.41}$$

Now we compare (4.36) and (4.38), by taking advantage of (4.28). It follows that the map $(L^1_{2,+})^N \ni g \mapsto [R_i^A(g) - R_i(g)]g_i \in L^1$ is positive and

isotone, $1 \leq i \leq N$. Moreover, because of the form of $T_i(g)$, if $a_0 > 0$ is sufficiently large, then the mapping $(L^1_{2,+})^N \ni g \mapsto T_i(g)(\mathbf{v}) \in L^1$ is positive and isotone for all i. In this case, by virtue of (4.40), the map $(L^1_{2,+})^N \ni g \mapsto a(\|\Lambda_B g\|)\Lambda_B g - Q^-_B(g) \in X$ is also positive and isotone.

In conclusion, the conditions of Theorem 3.1a) are fulfilled (in the case $\Lambda = \Lambda_1$), so that we are in position to state the following result ([11]):

THEOREM **4.4** *Suppose that in problem* (4.25), $f_{0,i} \in L^1_{4,+}$, $1 \leq i \leq N$. *Then Eq.* (4.25) *has a unique strong solution* $f(t) = (f_1, ..., f_N)$ *such that* $f_i(t) \in L^1_{4,+}$, $t \geq 0$, *and* $\|f_i(t)\|_{L^1_4}$ *is locally bounded on* $\mathbb{R}_+$, $1 \leq i \leq N$. *In addition,* f_i, $(1 + |\mathbf{v}|^2)f_i \in C(\mathbb{R}_+; L^1)$, $1 \leq i \leq N$,

$$\|\Lambda_B f(t)\| = \|\Lambda_B f_0\| \quad (t \geq 0), \tag{4.42}$$

and there is a non-decreasing function $\rho_B : \mathbb{R}_+ \mapsto \mathbb{R}_+$ *such that*

$$\left\|\Lambda^2_B f(t)\right\| \leq \exp(\rho_B(\|f_0\|)t) \left\|\Lambda^2_B f_0\right\| \quad (t \geq 0). \tag{4.43}$$

Theorem 4.4 does not state the conservation of mass, momentum and energy, but the conservation (in arbitrary units) of the quantity mass+(total) energy. However, the properties of $f(t)$, cf. Theorem 4.4, allow for checking immediately the separate conservation for each of the above quantities.

Theorem 4.4 reduces to the main monotonicity result of [2] when Eq. (4.25) is particularized to the case of the classical Boltzmann equation. Moreover, in that case, using suitable additional Povzner-like estimations, we can re-obtain the general moment estimations of [2], as application of Prop. 3.4b).

Finally, remark that similar analyses as for Theorems 4.2 and 4.4 can be developed for the main model considered, e.g., in [27].

4.5. Nonlinear von Neumann-Boltzmann equation

As Λ is unbounded (by construction), the existence of solutions to problem (2.62) seems not immediate from general considerations.

However, one can show that the conditions of Theorem 3.1 are fulfilled with $a(x) = x$.

First recall that $\mathrm{Tr}[\Lambda^k(Q^+ - Q^-)](F) = 0$, for all $0 \leq F \in \mathcal{D}(\Lambda^k) \cap X_+$, $k = 0, 1$. Then observe that, since $\Lambda \geq \mathbb{I}$, it follows easily that $\mathrm{Tr}[\Lambda^2(Q^+ - Q^-)](F) \leq \varepsilon \mathrm{Tr}(\Lambda F)\mathrm{Tr}F \leq \varepsilon \mathrm{Tr}(\Lambda F)\mathrm{Tr}(\Lambda^2 F)$ for all $0 \leq F \in \mathcal{D}(\Lambda^3) \cap X_+$.

So we can now formulate our existence result ([12]):

THEOREM **4.5** *Suppose that in problem* (2.62), $0 \leq F_0 \in \mathcal{D}(\Lambda^2)$. *Then Eq.* (2.62) *has a unique mild solution* $0 \leq F(t) \in \mathcal{D}(\Lambda^2)$, *and* $\mathrm{Tr}F(t)$ *is locally bounded. Moreover,* $F, \Lambda F \in C(\mathbb{R}_+; X)$, $\mathrm{Tr}F(t) = \mathrm{Tr}F_0$, $\mathrm{Tr}(\Lambda F)(t) = \mathrm{Tr}(\Lambda F_0)$ *and* $\mathrm{Tr}(\Lambda^2 F)(t) \leq \exp(t\varepsilon \mathrm{Tr}(\Lambda F_0))\mathrm{Tr}(\Lambda^2 F_0) \quad (t \geq 0)$.

5. Concluding remarks

The results of the previous section of applications can be easily completed taking advantage of Theorem 3.2. As an example, the previous Theorem 4.1 can be completed as follows

PROPOSITION **5.1** *Let* $f_0 \in L^1_{\beta,+}$ *in problem* (4.2). *Then Eq.* (4.2) *has a strong solution* $f(t) \in L^1_{\beta,+}$, $t \geq 0$.

As mentioned before, the uniqueness is no longer ensured in the latter case. Theorem 3.2 extends the main existence result of [11]. The other general existence results formulated in [11] can be similarly completed, with obvious modifications. This allows to reconsider the applications of [11], accordingly, in an obvious manner.

Prop. 3.3 provides uniqueness of the solutions in the special case when Δ vanishes on a rather large set. This can be applied, for instance, to the space-homogeneous Boltzmann equation with hard potentials, to obtain a similar existence result as in, e.g., [20]. However, in a more general case, the uniqueness problem, under the conditions of Theorem 3.2, remains open. Here we can however remark that the regularity conditions required in the theorem might be necessary to ensure the uniqueness of the *strong* solutions. Indeed, examples of non-unique (but) less regular solutions of the Boltzmann equation have been recently discovered, [26], [19].

In this chapter, we presented various examples of existence results for generalized Boltzmann models obtained by monotonicity methods. The above methods are potentially applicable to investigate other evolution problems.

On the other hand, the results presented in this review describe only partially the properties of the models considered. They must be completed by a thorough study of other properties of the models, e.g. the existence of stationary or/and equilibrium solutions, Lyapunov functionals, H-theorems (see e.g. [7]), asymptotic properties, construction of effective numerical methods.

6. Appendix

1) *Sketch of the Proof of Lemma* 3.3

Property $B(\cdot, g_i, h_j) \in L^1_{loc}(\mathbb{R}_+; X_+)$, $i, j = 1, 2$, follows from (A_1), (A_2) and Remark 3.2.

To prove (3.58), let

$$y_i(t) := \int_0^t \Delta(s, h_i(s))\mathrm{d}s \quad (i = 1, 2). \tag{6.1}$$

Clearly, $0 \leq y_1(t) \leq y_2(t)$, because of the isotonicity of $\Delta(t, \cdot)$ (cf. (A_1)). Further, define $F(x, y) := a(x + y) - a(x)$, with a as in (A_2). The properties of a (cf.(A_2)) imply

$$F(x^*, y) - F(x, y) = \int_0^y \left[a'(x^* + \xi) - a'(x + \xi)\right] \mathrm{d}\xi \geq 0 \tag{6.2}$$

for all $0 \leq x \leq x^*$ and $y \geq 0$. Then one can show easily (invoking (A_2), the isotonicity of $Q^+(t, \cdot)$ and the obvious inequality $\Lambda g_1(t) \leq \Lambda g_2(t)$) that

$$\begin{aligned} 0 \leq B(t, g_1, h_1) = B(t, g_1, 0) + F\left(\|\Lambda g_1(t)\|, y_1(t)\right) \Lambda g_1(t) \leq \\ \leq B(t, g_2, 0) + F\left(\|\Lambda g_1(t)\|, y_1(t)\right) \Lambda g_2(t) \end{aligned} \tag{6.3}$$

and

$$0 \leq F\left(\|\Lambda g_1(t)\|, y_1(t)\right) \leq F\left(\|\Lambda g_2(t)\|, y_1(t)\right) \leq F\left(\|\Lambda g_2(t)\|, y_2(t)\right). \tag{6.4}$$

Inequalities (6.3) and (6.4) can be now easily combined to obtain (3.58). □

2) *Sketch of the Proof of Lemma* 3.4

a) Since $\mathcal{D}^\infty_+$ is p-saturated and $\Lambda^k Q^\pm(t, \cdot)$ are positive and isotone, the key point is to show that for each $T > 0$ and $n = 1, 2, ...$, there is $g_{n,T} \in \mathcal{D}^\infty_+$ such that

$$0 \leq f_n(t) \leq g_{n,T} \quad (0 \leq t \leq T \quad a.e.). \tag{6.5}$$

Then (3.41) gives $Q^-(t, g_{n,T}) \in \mathcal{D}^\infty_+$ a.e. on $\mathbb{R}_+$, hence $\Lambda^k Q^-(\cdot, g_{n,T}) \in L^1_{loc}(\mathbb{R}_+; X_+)$ for all $k = 0, 1, 2, ...$ The same properties hold for $Q^+(t, g_{n,T})$ and $\Lambda^k Q^+(\cdot, g_{n,T})$, respectively (by virtue of the assumptions of Theorem 3.1a) and by (3.44)).

Inequality (6.5) can be proved by induction.

Indeed, note that (6.5) is trivially verified for $n = 1$ by $g_{1,T} := 0$, and for $n = 2$ by $g_{2,T} := f_0$. Further, at the induction step, assuming that (6.5) is

fulfilled for $n = 1, 2, ..q-1$ (with $q \geq 3$) applying, in essence, the properties of Δ, a, and (3.28), one first obtains

$$\Lambda^k \int_0^t B(s, g_{n-1,T}, g_{n-2,T})\mathrm{d}s = \int_0^t \Lambda^k B(s, g_{n-1,T}, g_{n-2,T})\mathrm{d}s \quad (0 \leq t \leq T), \tag{6.6}$$

for all $k = 1, 2, ...$ and $n = 1, 2, ..., q-1$. Then observe that $f_{q-1}(t) \leq g_{q-1,T}$ and $f_{q-2}(t) \leq g_{q-2,T}$ satisfy the conditions of Lemma 3.3 for $g_1 \leq g_2$ and $h_1 \leq h_2$, respectively. Thus, applying conveniently (3.56) and (3.58) in (3.60), and invoking (6.6), we get

$$0 \leq f_q(t) \leq f_0 + \int_0^T B(s, g_{q-1,T}, g_{q-2,T})\mathrm{d}s := g_{q,T} \in \mathcal{D}_+^\infty \quad (0 \leq t \leq T). \tag{6.7}$$

b) As before, it is sufficient to show by induction that property (6.5) is verified by $g_{n,T} \in \mathcal{D}(\Lambda^3) \cap X_+$.

First note that if $g_{1,T} = 0$ and $g_{2,T} = f_0$, then (6.5) is trivially verified for $n = 1, 2$, respectively.

The induction step is simpler than in a), because now one can make use of the fact that V^t is C_0. Then, $\int_0^t V^s h\mathrm{d}s \in \mathcal{D}(\Lambda)$ for all $h \in X$, $t \geq 0$, which, in our case, implies (for any $0 \leq t \leq T$)

$$\int_0^t V^{t-s} B(T, g_{q-1,T}, g_{q-2,T})\mathrm{d}s = \int_0^t V^s B(T, g_{q-1,T}, g_{q-2,T})\mathrm{d}s \in \mathcal{D}(\Lambda^3) \cap X_+. \tag{6.8}$$

Since, in our case, $B(t, g_{q-1,T}, g_{q-2,T}) \leq B(T, g_{q-1,T}, g_{q-2,T})$, we conclude the induction step, using property (6.8) with the key inequality

$$0 \leq f_q(t) \leq f_0 + \int_0^t V^{t-s} B(T, g_{q-1,T}, g_{q-2,T})\mathrm{d}s \quad (0 \leq t \leq T), \tag{6.9}$$

which follows, in essence, by Lemma 3.3, and by applying (3.56) and (3.58) in (3.60).

c) The statement follows from simple regularity considerations and some direct computation.

d) Obviously, $0 = f_1(t) \leq f_2(t) \leq f_3(t)$ a.e.. Then a straightforward induction, applying (3.58), shows that $\{f_n(t)\}$ is a.e. increasing.

For the rest of the proof, note that (3.63) implies (3.64). Inequality (3.63) can be proved by induction. Indeed, since $0 = f_1 \leq f_2(t) \leq f_0$, and $\Delta(t, 0) = 0$ a.e. (cf. Remark 3.1), formula (3.63) is trivially verified for $n = 2$. Let $q \geq 3$

and suppose inequality (3.63) to be valid for $n = 2, 3, ..., q-1$. If $n = q$ in (3.62), then the positivity of a and $0 \leq \Lambda f_{q-1}(t) \leq \Lambda f_q(t)$ give

$$f_q(t) \leq f_0 + \int_0^t Q(s, f_{q-1}(s))\mathrm{d}s +$$

$$+ \int_0^t \left[a\left(\|\Lambda f_{q-1}(s)\| + \int_0^s \Delta(\tau, f_{q-2}(\tau))d\tau \right) - a\left(\|\Lambda f_0\|\right) \right] \Lambda f_q(s)\mathrm{d}s. \tag{6.10}$$

According to the induction hypothesis, (3.63) holds true for $n = q-1$. Hence (3.64) is also valid for $n = q-1$, as concluded before. Then $a(\|\Lambda f_{q-1}(s)\| + \int_0^s \Delta(\tau, f_{q-2}(\tau))d\tau)) \leq a\left(\|\Lambda f_0\|\right)$, because a is non-decreasing. As $\Lambda f_q(s)$ is positive, clearly the integral term containing $\Lambda f_q(s)$, in the r.h.s. of (6.10) is negative. Then (3.63) becomes true for $n = q$.

e) Note that $Q^{\pm}(t, f_n(t)) \in \mathcal{D}(\Gamma)$, for a.e. $t \geq 0$. Also, $\Gamma Q^{\pm}(\cdot, f_n(\cdot)) \in L^1_{loc}(\mathbb{R}_+; X_+)$. Indeed, let $T > 0$ and $g_{n,T} \geq f_n(t)$ be as in a). If Γ is of type D on $\mathcal{D}^{\infty}_+$ (on $\mathcal{D}(\Lambda^2) \cap X_+$), then (3.36) and (3.41) give $\|\Gamma Q^{\pm}(t, f_n(t))\| \leq \|\Gamma Q^{\pm}(t, g_{n,T})\| \leq \|\Gamma Q^{-}(t, g_{n,T})\| \leq a(\|g_{n,T}\|)\,\|\Gamma \Lambda g_{n,T}\|$ for a.e. $0 \leq t \leq T$. On the other hand, if Γ satisfies (3.46), then (3.41) implies

$$\left\|\Gamma Q^{+}(t, f_n(t))\right\| \leq \left\|\Gamma Q^{-}(t, f_n(t))\right\| + \rho_{\Gamma}(\|\Lambda_1 g_{n,T}\|)\,\|\Gamma g_{n,T}\| \leq$$

$$\leq a(\|g_{n,T}\|)\,\|\Gamma \Lambda g_{n,T}\| + \rho_{\Gamma}(\|\Lambda_1 g_{n,T}\|)\,\|\Gamma g_{n,T}\| \quad (0 \leq t \leq T \quad a.e.).$$

But (3.63) is of the form (3.37), and the above considerations show that Lemma 3.2 applies (with Γ instead of Λ). Hence,

$$\|\Gamma f_n(t)\| + \int_0^t \Delta(s, f_{n-1}(s); \Gamma, Q)\mathrm{d}s \leq \|\Gamma f_0\| \quad (t \geq 0, \quad n \geq 2). \tag{6.11}$$

Now the proof can be immediately concluded: if $n = 1$, then formula (3.65) is trivially satisfied; if $n \geq 2$, then (3.65) is directly implied by (6.11).

To obtain (3.66) observe that Λ^2 satisfies the conditions for Γ in e).

f) First apply inequality (3.46) in (6.11). It follows that

$$\|\Gamma f_n(t)\| \leq \|\Gamma f_0\| + \int_0^t \rho_{\Gamma}(\|\Lambda_1 f_{n-1}(s)\|)\,\|\,\Gamma f_{n-1}(s)\|\,\mathrm{d}s \quad (t \geq 0, \quad n \geq 2). \tag{6.12}$$

But Λ_1 satisfies the conditions of e) in the present lemma, hence $\|\Lambda_1 f_n(t)\| \leq \|\Lambda_1 f_0\|$, $t \geq 0$, $n = 1, 2, ...$. Introducing the last inequality in (4.16), we obtain

$$\|\Gamma f_n(t)\| \leq \|\Gamma f_0\| + \rho_{\Gamma}(\|\Lambda_1 f_0\|) \int_0^t \|\Gamma f_{n-1}(s)\|\,\mathrm{d}s \quad (t \geq 0, \quad n \geq 2). \tag{6.13}$$

Finally, since (3.67) is obviously satisfied for $n = 1, 2$, a straightforward (Gronwall type) induction in (6.13) concludes the proof. □

References

[1] ALDOUS, D. J., *Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists*, Bernoulli, **5** (1999), pp. 3–48.

[2] ARKERYD, L., *On the Boltzmann equation I & II*, Arch. Ration. Mech. Anal., **45** (1972), pp. 1–34.

[3] BELLOMO, N., POLEWCZAK, J., *The generalized Boltzmann equation solution and exponential trend to equilibrium*, Transport Theory Statist. Phys., **26** (1997), pp. 661–677.

[4] BELLOMO, N., PULVIRENTI M., Eds., *Modeling in Applied Sciences: A Kinetic Theory Approach*, Series:, Model. Simul. Sci. Eng. Technol., Birkhäuser, Boston, 2000.

[5] BENEDETTO, D., CAGLIOTI, E., PULVIRENTI, M., *Collective Behaviour of One-Dimensional Granular Media*, in Modeling in Applied Sciences: A Kinetic Theory Approach, Series: Model. Simul. Sci. Eng. Technol., pp. 81–110, Bellomo, N., Pulvirenti, M., Eds., Birkhäuser, Boston, 2000.

[6] BOBYLEV, A. V., CERGINANI, C. *Self-Similar asymptotics for the Boltzmann equation with inelastic and elastic interactions*, J. Satist. Phys., **110** (2003), pp. 333–375.

[7] DE ANGELIS, E., GRÜNFELD, C. P., *The Cauchy problem for the generalized Boltzmann equation with dissipative collisions*, Appl. Math. Lett., **14** (2001), pp. 941–947.

[8] DE ANGELIS, E., GRÜNFELD, C. P., *Modeling and analytic problems for a generalized Boltzmann equation for a multicomponent reacting gas*, Nonlinear Anal. Real World Appl., **4** (2003), pp. 189–202.

[9] GRÜNFELD, C. P., *Nonlinear Kinetic Models with Chemical Reactions*, in Modeling in Applied Sciences: A Kinetic Theory Approach, Series: Model. Simul. Sci. Eng. Technol., pp. 173–224, Bellomo, N., Pulvirenti, M., Eds., Birkhäuser, Boston, 2000.

[10] Grünfeld, C. P., *On a class of kinetic equations for reacting gas mixtures with multiple collisions*, C. R. Acad. Sci. Paris Sér. I Math., **316** (1993), pp. 953–958.

[11] Grünfeld, C. P., *A Nonlinear Evolution Equation in an Ordered Space, Arising from Kinetic Theory*, Commun. in Contemp. Math., **9** (2007), pp. 217–251.

[12] Grünfeld, C. P., *On an Evolution Equation in an Ordered Space*, Anal. Univ. Buc., series Mathematics, **LV** (2006), pp. 79–90.

[13] Grünfeld, C. P., *On a Class of Nonlinear Evolution Equations in an Abstract Lebesgue Space*, to appear in Proceedings of the Fifth International Conference on Dynamic Systems and Applications, May 30-June 2, 2007, Atlanta, USA, Dynamic Publishers, Inc.

[14] Grünfeld, C. P., Georgescu, E., *On a class of kinetic equations for reacting gas mixtures*, Mat. Fiz. Anal. Geom., **2** (1995), pp. 408–435.

[15] Grünfeld, C. P., Marinescu, D., *On the numerical simulation of a class of reactive Boltzmann type equations*, Transport Theory Statist. Phys., **26** (1997), pp. 287–318.

[16] Hille, E., Phillips, R. S., *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, 1974.

[17] Kantorovich, L. V., Akilov, G. P., *Functional Analysis*, Pergamon Press, Oxford – Elmsford – N.Y., 1982.

[18] Lachowicz, M. Pulvirenti, M., *A stochastic system of particles modelling the Euler equations*, Arch. Ration. Mech. Anal., **109** (1990), pp. 81–93.

[19] Lu, X., Wennberg, B., *Solutions with increasing energy for the spatially homogeneous Boltzmann equation*, Nonlinear Anal. Real World Appl, **3** (2002), pp. 243–258.

[20] Mischler, S., Wennberg, B., *On the spatially homogeneous Boltzmann equation*, Ann. de l'I.H.P., section C, **16** (1999), pp. 467–501.

[21] Müller, H. *Zur allgemeinen Theorie der raschen Koagulation*, Kolloid-Beih., **27** (1928), pp. 223–250.

[22] Norris, J. R., *Smoluchowski's coagulation equation: Uniqueness, nonuniqueness and a hydrodynamic limit for the stochastic coalescent*, Ann. Appl. Probab., **9** (1999), pp. 78–109.

[23] POVZNER, A. YA., *The Boltzmann equation in the kinetic theory of gases*, Mat. Sb. (N. S.), **58**(100) (1962), pp. 65–86.

[24] SCHAEFER, H. H., *Banach Lattices and Positive Operators*, Springer-Verlag, N. Y. – Heidelberg, 1974.

[25] von SMOLUCHOWSKI, M., *Drei Vorträge über Diffusion, Brownsche Molekular-bewegung und Koagulation von Kolloidteilchen*, Phys. Z., **17** (1916), pp. 557–571, 585–599.

[26] WENNBERG, B., *An example of nonuniqueness for solutions to the homogeneous Boltzmann equation*, J. Statist. Phys., **95** (1999), pp. 1–2.

[27] WIESEN, B., *On a phenomenological generalized Boltzmann equation*, J. Math. Phys. **33** (1992) pp. 1786–1798.

[28] WILD, E., *On Boltzmann's equation in the kinetic theory of gases*, Proc. Camb. Philos. Soc., **47** (1951), pp. 602–609.

Topics in Applied Mathematics & Mathematical Physics

Estimating the number of negative eigenvalues of a relativistic Hamiltonian with regular magnetic field

V. Iftimie[*] [1] [2], *M. Măntoiu*[†] [1] [3] *and R. Purice*[*] [1] [4]

Contents

[1] Institute of Mathematics "Simion Stoilow" of the Romanian Academy,
[2] The Faculty of Mathematics and Informatics of the Bucharest University, e-mail: Viorel.Iftimie@imar.ro
[3] Departamento de Matematicas, Facultad de Ciencias, Universidad de Chile, Santiago de Chile, e-mail: Marius.Mantoiu@imar.ro
[4] Laboratoire Europeen Associe CNRS *Math-Mode*, e-mail: Radu.Purice@imar.ro
[*] Partial support from the Contract no. 2-CEx06-11-18/2006
[†] Partial support from Contract no 2-CEx06-11-34/25.07.2006 and from Proyecto Fondecyt no. 1085162.

1. Introduction

For the Schrödinger operator $-\Delta + V$ on $L^2(\mathbb{R}^d)$ $(d \geq 3)$, one has the well-known CLR (Cwikel-Lieb-Rosenblum) estimation for $N(V)$, *the number of negative eigenvalues*:

$$N(V) \ \leq \ c(d) \int_{\mathbb{R}^d} \mathrm{d}x \ |V_-(x)|^{d/2} \, . \tag{1.1}$$

V is the multiplication operator with the function $V \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ and $V_- := (|V|-V)/2 \in L^{d/2}(\mathbb{R}^d)$; the constant $c(d) > 0$ only depends on the dimension $d \geq 3$ (see [47], Th. XII.12).

There exist at least four different proofs of this inequality. Rosenblum [35] uses "piece-wise polynomial approximation in Sobolev spaces". Lieb [25] relies on the Feynman-Kac formula. Cwickel [4] uses ideas from interpolation theory. Finally, Li and Yau [31] make a heat kernel analysis.

The inequality (1.1) has been extended in [1] and [48] to the case of operators with magnetic fields $(-i\nabla - A)^2 + V$, where the components of the vector potential $A = (A_1, \ldots, A_d)$ belong to $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. The basic ingredient of the proof is the Feynman-Kac-Ito formula. Melgaard and Rosenblum [41] generalizes this result (by a different method) to a class of differential operators of second order with variable coefficients. The idea for treating the relativistic Hamiltonian (without a magnetic field), by replacing Brownian motion with a Lévy process, appears in [5] and we follow it in our work giving all the technical details. Some similar results but for a different Hamiltonian and with different techniques have been obtained recently in [8].

Our aim in this paper is to obtain an estimation of the type (1.1) for an operator that is a good candidate for a relativistic Hamiltonian with magnetic field (for scalar particles); it is gauge covariant and obtained through a quantization procedure from the classical candidate. We shall make use of a "magnetic pseudodifferential calculus" that has been introduced and developed in some previous papers [34], [35], [27], [28], [36], [38], [24].

Let us denote by $C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$ the family of functions $f \in C^\infty(\mathbb{R}^d)$ for which all the derivatives $\partial^\alpha f$, $\alpha \in \mathbb{N}^d$ have polynomial growth.

Let B be a magnetic field (a 2-form) with components $B_{jk} \in C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$. It is known that it can be expressed as the differential $B = dA$ of a vector potential (a 1-form) $A = (A_1, \ldots, A_d)$ with $A_j \in C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$, $j = 1, \ldots, d$; an

example is the transversal gauge:

$$A_j(x) = -\sum_{k=1}^{n} \int_0^1 \mathrm{d}s\; B_{jk}(sx) s x_k.$$

We denote by

$$\Gamma^A(x,y) := \int_0^1 \mathrm{d}s\, A((1-s)x + sy) = \int_{[x,y]} A, \quad x, y \in \mathbb{R}^d. \tag{1.2}$$

the circulation of A along the segment $[x,y]$, $x, y \in \mathbb{R}^d$. If a is a symbol on $\mathbb{R}^d$, one defines by an oscillatory integral the linear continuous operator $\mathfrak{Op}^A(a) : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}^*(\mathbb{R}^d)$ by

$$\left[\mathfrak{Op}^A(a)\right](x) := (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathrm{d}y\, \mathrm{d}\xi\, \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi} \mathrm{e}^{-\mathrm{i}\int_{[x,y]} A} a\left(\frac{x+y}{2}, \xi\right) u(y), \tag{1.3}$$

The correspondence $a \mapsto \mathfrak{Op}^A(a)$ is meant to be a quantization and could be regarded as a functional calculus $\mathfrak{Op}^A(a) = a(Q, \Pi^A)$ for the family of non-commuting operators $(Q_1, \ldots, Q_d; \Pi_1^A, \ldots, \Pi_d^A)$, where Q is the position operator, $\Pi^A := D - A(Q)$ is the magnetic momentum, with $D := -\mathrm{i}\nabla$.

If a belongs to the Schwartz space $\mathcal{S}(\mathbb{R}^{2d})$, then $\mathfrak{Op}^A(a)$ acts continuously in the spaces $\mathcal{S}(\mathbb{R}^d)$ and $\mathcal{S}^*(\mathbb{R}^d)$, respectively. It enjoys the important physical property of being gauge covariant: if $\varphi \in C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$ is a real function, A and $A' := A + d\varphi$ define the same magnetic field and one prove easily that $\mathfrak{Op}^{A'}(a) = \mathrm{e}^{\mathrm{i}\varphi} \mathfrak{Op}^A(a) \mathrm{e}^{-\mathrm{i}\varphi}$. The property is not shared by the quantization $a \mapsto \mathfrak{Op}_A(a) := \mathfrak{Op}(a \circ \nu_A)$, where $\mathfrak{Op}$ is the usual Weyl quantization and $\nu_A : \mathbb{R}^d \to \mathbb{R}^d$, $\nu_A(x,\xi) := (x, \xi - A(a))$ is an implementation of "the minimal coupling".

We mention that in the references quoted above, a symbolic calculus is developed for the magnetic pseudodifferential operators (1.3). In particular, a symbol composition $(a,b) \mapsto a \sharp^B b$ is defined and studied, verifying $\mathfrak{Op}^A(a)\mathfrak{Op}^A(b) = \mathfrak{Op}^A(a \sharp^B b)$. It depends only on the magnetic field B, no choice of a gauge being needed. The formalism has a C^*-algebraic interpretation in terms of twisted crossed products, cf. [35], [37], [39] and it has been used in [40] for the spectral theory of quantum Hamiltonians with anisotropic potentials and magnetic fields.

We shall denote by H_A the unbounded operator in $L^2(\mathbb{R}^d)$ defined on $C_0^\infty(\mathbb{R}^d)$ by $H_A u := \mathfrak{Op}^A(h)u$, with $h(x,\xi) \equiv h(\xi) := \langle \xi \rangle - 1 = (1 + |\xi|^2)^{1/2} - 1$. One

can express it as

$$(H_A u)(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathrm{d}y\, \mathrm{d}\xi\, \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi} h\left(\xi - \Gamma^A(x,y)\right) u(y). \tag{1.4}$$

H_A is a symmetric operator and, as seen below, essentially self-adjoint on $C_0^\infty(\mathbb{R}^d)$. Also denoting its closure by H_A, we will have $H_A \geq 0$.

Ichinose and Tamura [19], [20], using the quantization $a \mapsto (Op)_A(a)$, study another relativistic Hamiltonian with magnetic field defined by

$$\left(H'_A u\right)(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathrm{d}y\, \mathrm{d}\xi\, \mathrm{e}^{\mathrm{i}(x-y)\cdot\xi} h\left(\xi - A\left(\frac{x+y}{2}\right)\right) u(y), \tag{1.5}$$

for which they prove many interesting properties. Unfortunately, H'_A is not gauge covariant (cf. [24]). Many of the properties of H'_A also hold for H_A (by replacing $A\left(\frac{x+y}{2}\right)$ with $\Gamma^A(x,y)$ in the statements and proofs) and this will be used in the sequel.

Aside the magnetic field $B = dA$, we shall also consider an electric potential $V \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$, real function expressed as $V = V_+ - V_-$, $V_\pm \geq 0$, such that $V_- \in L^{d+k}(\mathbb{R}^d) \cap L^{d/2+k}(\mathbb{R}^d)$ for some $k \geq 0$. We are interested in the operator $H(A,V) := H_A + V$; it will be shown that it is well-defined in form sense as a self-adjoint operator in $L^2(\mathbb{R}^d)$, with essential spectrum included into the positive real axis. Taking advantage of gauge covariance, we denote by $N(B,V)$ the number of strictly negative eigenvalues of $H(A,V)$ (multiplicity counted); it only depends on the potential V and the magnetic field B.

The main result of the article is

THEOREM **1.1** *Let $B = dA$ be a magnetic field with $B_{jk} \in C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$, $A_j \in C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$ and let $V = V_+ - V_- \in L^1_{\mathrm{loc}(\mathbb{R}^d)}$ be a real function with $V_\pm \geq 0$ and $V_- \in L^d(\mathbb{R}^d) \cap L^{d/2}(\mathbb{R}^d)$. Then there exists a constant C_d, only depending on the dimension $d \geq 3$, such that*

$$N(B,V) \leq C_d \left(\int_{\mathbb{R}^d} \mathrm{d}x\, V_-(x)^d + \int_{\mathbb{R}^d} \mathrm{d}x\, V_-(x)^{d/2} \right). \tag{1.6}$$

A standard consequence is the next Lieb-Thirring-type estimation:

COROLLARY **1.1** *We assume that the components of B belong to $C^\infty_{\mathrm{pol}}(\mathbb{R}^d)$ and that $V = V_+ - V_- \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$ is a real function with $V_\pm \geq 0$ and $V_- \in L^{d+k}(\mathbb{R}^d) \cap L^{d/2+k}(\mathbb{R}^d)$, $k > 0$. We denote by $\lambda_1 \leq \lambda_2 \leq \ldots$ the*

strictly negative eigenvalues of $H(A,V)$ (with multiplicity). For any $d \geq 2$ there exists a constant $C_d(k)$ such that

$$\sum_j |\lambda_j|^k \leq C_d(k) \left(\int_{\mathbb{R}^d} \mathrm{d}x \, V_-(x)^{d+k} + \int_{\mathbb{R}^d} \mathrm{d}x \, V_-(x)^{d/2+k} \right). \tag{1.7}$$

Sections 2, 3, 4 will contain essentially known facts (usually presented without proofs), needed for checking Theorem 1.1. So, in Section 2 we introduce the Feller semigroup ([20], [17], [26]) associated to the operator $H_0 := \langle D \rangle - 1$. In the third section we define properly the operator $H(A,V)$ and study its basic properties. In Section 4 we recall some probabilistic results, as the Markov process associated to the semigroup defined by H_0 ([25], [6], [26]) and the Feynman-Kac-Itô formula adapted to a Lévy process ([20]).

In Section 5 we prove Theorem 1.1 for $B = 0$, using some of Lieb's ideas for the non-relativistic case (see [48]) in the setting proposed in [5]. The last section contains the proof of Theorem 1.1 with magnetic field as well as Corollary 1.1. The main ingredient is the Feynman-Kac-Itô formula.

2. The Feller semigroup

We consider the following symbol (interpreted as a classical relativistic Hamiltonian for $m = 1, c = 1$) $h : \mathbb{R}^d \to \mathbb{R}_+$ defined by $h(\xi) := \langle \xi \rangle - 1 \equiv \sqrt{1 + |\xi|^2} - 1$. Let us observe (as in [17]) that it defines a *conditional negative definite function* (see [47]) and thus has a Lévy-Khincin decomposition (see Appendix 2 to Section XIII of [47]). Computing $(\nabla h)(\xi)$ and $(\Delta h)(\xi)$ and using the general Lévy-Khincin decomposition (see for example [47]), one obtains that there exists a Lévy measure $\mathsf{n}(\mathrm{d}y)$, i.e. a non-negative, σ-finite measure on $\mathbb{R}^d$, for which $\min\{1, |y|^2\}$ is integrable on $\mathbb{R}^d$, such that

$$h(\xi) \;=\; -\int_{\mathbb{R}^d} \mathsf{n}(\mathrm{d}y) \left\{ \mathrm{e}^{\mathrm{i}y\cdot\xi} - 1 - \mathrm{i}\,(y \cdot \xi)\, I_{\{|x|<1\}}(y) \right\}, \tag{2.1}$$

where $I_{\{|x|<1\}}$ is the characteristic function of the open unit ball in $\mathbb{R}^d$. One has the following explicit formula (see [17]):

$$\mathsf{n}(\mathrm{d}y) \;=\; 2(2\pi)^{-(d+1)/2} |y|^{-(d+1)/2} K_{(d+1)/2}(|y|)\, \mathrm{d}y, \tag{2.2}$$

with K_ν the modified Bessel function of third type and order ν. We recall the following asymtotic behaviour of these functions:

$$0 \;<\; K_\nu(r) \;\leq\; C \max(r^{-\nu}, r^{-1/2}) \mathrm{e}^{-r}, \quad \forall r > 0, \quad \forall \nu > 0. \tag{2.3}$$

We shall denote by $\mathcal{H}^s(\mathbb{R}^d)$ the usual Sobolev spaces of order $s \in \mathbb{R}$ on $\mathbb{R}^d$ and by H_0 the pseudodifferential operator $h(D) \equiv \mathfrak{Op}(h)$ considered either as a continuous operator on $\mathcal{S}(\mathbb{R}^d)$ and on $\mathcal{S}^*(\mathbb{R}^d)$ or as a self-adjoint operator in $L^2(\mathbb{R}^d)$ with domain $\mathcal{H}^1(\mathbb{R}^d)$. The semigroup generated by H_0 is explicitly given by the convolution with the following function (for $t > 0$ and $x \in \mathbb{R}^d$):

$$\overset{\circ}{\wp}_t(x) \ := \ (2\pi)^{-d} \frac{t}{\sqrt{|x|^2+t^2}} \int_{\mathbb{R}^d} \mathrm{d}\xi \, \mathrm{e}^{\left(t-\sqrt{(|x|^2+t^2)(|\xi|^2+1)}\right)} \ =$$

$$= \ 2^{-(d-1)/2} \, \pi^{-(d+1)/2} \, t\mathrm{e}^t (|x|^2+t^2)^{-(d+1)/4} K_{(d+1)/2}(\sqrt{|x|^2+t^2}) \qquad (2.4)$$

(see [20], [2]). We have

$$\overset{\circ}{\wp}_t(x) \ > \ 0 \quad \text{and} \quad \int_{\mathbb{R}^d} \mathrm{d}x \, \overset{\circ}{\wp}_t(x) \ = \ 1. \qquad (2.5)$$

From (2.3) one easily can deduce the following estimation

$$\exists C > 0 \quad \text{such that} \quad \overset{\circ}{\wp}_t(0) \ \leq Ct^{-d}(1+t^{d/2}), \quad \forall t > 0. \qquad (2.6)$$

Let us set

$$C_\infty(\mathbb{R}^d) \ := \ \left\{ f \in C(\mathbb{R}^d) \ | \ \lim_{|x|\to\infty} f(x) = 0 \right\} \qquad (2.7)$$

and endow it with the Banach norm $\|f\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x)|$. Using the above properties of the function $\overset{\circ}{\wp}_t$ we can extend e^{-tH_0} to a well-defined bounded operator $P(t)$ acting in $C_\infty(\mathbb{R}^d)$.

REMARK **2.1** *One can easily verify that* $\{P(t)\}_{t\geq 0}$ *is a Feller semigroup, i.e.:*

1. $P(t)$ *is a contraction:* $\|P(t)f\|_\infty \leq \|f\|_\infty$, $\forall f \in C_\infty(\mathbb{R}^d)$;
2. $\{P(t)\}_{t\geq 0}$ *is a semigroup:* $P(t+s) = P(t)P(s)$;
3. $P(t)$ *preserves positivity:* $P(t)f \geq 0$ *for any* $f \geq 0$ *in* $C_\infty(\mathbb{R}^d)$;
4. *We have* $\lim_{t\searrow 0} \|P(t)f - f\|_\infty = 0$, $\forall f \in C_\infty(\mathbb{R}^d)$.

3. The perturbed Hamiltonian

Suppose given a magnetic field of class $\mathcal{C}^\infty_{\mathsf{pol}}(\mathbb{R}^d)$ and let us choose a potential vector A, such that $B = dA$, with components also of class $\mathcal{C}^\infty_{\mathsf{pol}}(\mathbb{R}^d)$ (this is always possible, as said before). We shall denote by H_A the operator

$\mathfrak{Op}^A(h)$, considered either as a continuous operator on $\mathcal{S}(\mathbb{R}^d)$ and on $\mathcal{S}^*(\mathbb{R}^d)$ (by duality) or as an unbounded operator on $L^2(\mathbb{R}^d)$ with domain $\mathcal{C}_0^\infty(\mathbb{R}^d)$. Using the Fourier transform one easily proves that for $u \in \mathcal{C}_0^\infty(\mathbb{R}^d)$:

$$[H_0 u](x) \;=\; -\int_{\mathbb{R}^d} n(\mathrm{d}y)\, \left[u(x+y) - u(x) - I_{\{|z|<1\}}(y)\,(y\cdot\partial_x u)\,(x)\right]. \quad (3.1)$$

Recalling the definition of $\mathfrak{Op}^A(h)$, we remark that

$$[H_A u](x) \;=\; \left[\mathfrak{Op}^A(h)u\right](x) \;=\; \left[\mathfrak{Op}(h)\left(\mathrm{e}^{\mathrm{i}(x-.)\cdot\Gamma^A(x,.)}u\right)\right](x) \;= \quad (3.2)$$

$$=\; \left[H_0\left(\mathrm{e}^{\mathrm{i}(x-.)\cdot\Gamma^A(x,.)}u\right)\right](x).$$

Combining the above two equations one gets easily

$$[H_A u](x) \;=\; -\int_{\mathbb{R}^d} n(\mathrm{d}y)\, \left[\mathrm{e}^{-\mathrm{i}y\cdot\Gamma^A(x,x+y)}u(x+y) - u(x) - \right. \quad (3.3)$$

$$\left. -I_{\{|z|<1\}}(y)\,(y\cdot(\partial_x - \mathrm{i}A(x))u)\,(x)\right].$$

Repeating the arguments in [17] with $\Gamma^A(x, x+y)$ replacing $A((x+y)/2)$ one proves the following results similar to those in [17].

PROPOSITION **3.1** *Considered as unbounded operator in $L^2(\mathbb{R}^d)$, H_A is essential self-adjoint on $\mathcal{C}_0^\infty(\mathbb{R}^d)$. Its closure, also denoted by H_A, is a positive operator.*

PROPOSITION **3.2** *For any $u \in L^2(\mathbb{R}^d)$ such that $H_A u \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$*

$$\Re\left[(\mathrm{sign}u)(H_A u)\right] \;\geq\; H_0|u|.$$

Using the method in [49] we can prove the following result.

PROPOSITION **3.3** *For any $u \in L^2(\mathbb{R}^d)$ we have:*

1. *for any $\lambda > 0$ and for any $r > 0$*

$$\left|(H_A+\lambda)^{-r}\,u\right| \;\leq\; (H_0+\lambda)^{-r}\,|u|; \quad (3.4)$$

2. *for any $t \geq 0$*

$$\left|\mathrm{e}^{-tH_A}u\right| \;\leq\; \mathrm{e}^{-tH_0}|u|. \quad (3.5)$$

We associate to H_A its sesquilinear form

$$\mathcal{D}(\mathfrak{h}_A) = \mathcal{D}(H_A^{1/2}),$$

$$\mathfrak{h}_A(u,v) := (H_A^{1/2}u, H_A^{1/2}v), \quad \forall (u,v) \in \mathcal{D}(\mathfrak{h}_A)^2. \tag{3.6}$$

Consider now a function $V \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$, $V \geq 0$ and associate to it the sesquilinear form

$$\mathcal{D}(\mathfrak{q}_V) := \{u \in L^2(\mathbb{R}^d) \mid \sqrt{V}u \in L^2(\mathbb{R}^d)\},$$

$$\mathfrak{q}_V(u,v) := \int_{\mathbb{R}^d} \mathrm{d}x\, V(x)u(x)\overline{v(x)}, \quad \forall (u,v) \in \mathcal{D}(\mathfrak{q}_V)^2. \tag{3.7}$$

Both these sesquilinear forms are symmetric, closed and positive. We shall abbreviate $\mathfrak{h}_A(u) \equiv \mathfrak{h}_A(u,u)$ and $\mathfrak{q}_V(u) \equiv \mathfrak{q}_V(u,u)$.

PROPOSITION **3.4** *Let $V : \mathbb{R}^d \to \mathbb{R}$ be a measurable function that can be decomposed as $V = V_+ - V_-$ with $V_\pm \geq 0$ and $V_\pm \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$. Moreover let us suppose that the sesquilinear form $\mathfrak{q}_{V_-}$ is small with respect to $\mathfrak{h}_0$ (i.e. it is $\mathfrak{h}_0$-relatively bounded with bound strictly less then 1). Then the sesquilinear form $\mathfrak{h}_A + \mathfrak{q}_{V_+} - \mathfrak{q}_{V_-}$, that is well defined on $\mathcal{D}(\mathfrak{h}_A) \bigcap \mathcal{D}(\mathfrak{q}_{V_+})$, is symmetric, closed and bounded from below, defining thus an inferior semibounded self-adjoint operator $H(A;V) \equiv H := H_A \dot{+} V$ (sum in sense of forms).*

Proof. The sesquilinear form $\mathfrak{h}_A + \mathfrak{q}_{V_+}$ (defined on the intersection of the form domains) is clearly positive, symmetric and closed. We shall prove now that the sesquilinear form $\mathfrak{q}_{V_-}$ is $\mathfrak{h}_A + \mathfrak{q}_{V_+}$-bounded with bound strictly less then 1, so that the conclusion of the proposition follows by standard arguments.

Let us denote by $H_+ := H_A \dot{+} V_+$ the unique positive self-adjoint operator associated to the sesquilinear form $\mathfrak{h}_A + \mathfrak{q}_{V_+}$ by the representation theorem 2.6 in §VI.2 of [29]. As $V_+ \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$, we have $\mathcal{C}_0^\infty(\mathbb{R}^d) \subset \mathcal{D}(\mathfrak{h}_A) \bigcap \mathcal{D}(\mathfrak{q}_{V_+})$ and thus we can use the form version of the Kato-Trotter formula from [30]:

$$\mathrm{e}^{-tH_+} \;=\; s - \lim_{n\to\infty} \left(\mathrm{e}^{-(t/n)H_A}\, \mathrm{e}^{-(t/n)V_+}\right)^n, \qquad \forall t \geq 0. \tag{3.8}$$

Let us recall the formula ($r > 0$ and $\lambda > 0$)

$$(H_+ + \lambda)^{-r} \;=\; \Gamma(r)^{-1} \int_0^\infty \mathrm{d}t\; t^{r-1}\, \mathrm{e}^{-t\lambda}\, \mathrm{e}^{-tH_+}. \tag{3.9}$$

Combining the above two equalities we obtain

$$\left|(H_+ + \lambda)^{-r} f\right| \;\leq\; \Gamma(r)^{-1} \int_0^\infty \mathrm{d}t\; t^{r-1}\, \mathrm{e}^{-t\lambda}\, \left|\mathrm{e}^{-tH_+} f\right| \;= \tag{3.10}$$

$$= \;\Gamma(r)^{-1} \int_0^\infty \mathrm{d}t\; t^{r-1} \left| s - \lim_{n\to\infty} \left(\mathrm{e}^{-(t/n)H_A}\, \mathrm{e}^{-(t/n)V_+}\right)^n f \right| \;\leq$$

$$\leq\; (H_0 + \lambda)^{-r} |f|,$$

by using the second point of Proposition 3.3.

Taking $u = (H_0+\lambda)^{-1/2} g$ with $g \in L^2(\mathbb{R}^d)$ arbitrary and $\lambda > 0$ large enough and using the hypothesis on V_- we deduce that there exists $a \in [0,1)$, $b \geq 0$ and $a' \in [0,1)$ such that

$$\mathfrak{q}_{V_-}(u) \leq a\|H_0^{1/2} u\|^2 + b\|u\|^2 = a\|H_0^{1/2}(H_0+\lambda)^{-1/2} g\|^2 + b\|(H_0+\lambda)^{-1/2} g\|^2 \leq$$

$$\leq (a + b/\lambda)\|g\|^2 \leq a'\|g\|^2. \tag{3.11}$$

For any $v \in \mathcal{D}(\mathfrak{h}_A) \bigcap \mathcal{D}(\mathfrak{q}_{V_+})$ let $f := (H_+ + \lambda)^{1/2} v$ and $g := |f|$. Using now (3.10) with $r = 1/2$, (3.11) and the explicit form of $\mathfrak{q}_{V_-}$ we conclude that

$$\mathfrak{q}_{V_-}(v) = \mathfrak{q}_{V_-}\left((H_+ + \lambda)^{-1/2} f\right) \leq \mathfrak{q}_{V_-}\left((H_0 + \lambda)^{-1/2} g\right) \leq \tag{3.12}$$

$$\leq a'\|g\|^2 \;=\; a' \left\|(H_+ + \lambda)^{1/2} v\right\|^2 \;=\; a'\left[\mathfrak{h}_A(v) + \mathfrak{q}_+(v) + \lambda\|v\|^2\right].$$

□

DEFINITION **3.1** *For a potential function V satisfying the hypothesis of Proposition* 3.4*, we call the operator $H = H(A;V)$ introduced in the same proposition the relativistic Hamiltonian with potential V and magnetic vector potential A.*

The spectral properties of H only depend on the magnetic field B, different choices of a gauge giving unitarly equivalent Hamiltonians, due to the gauge covariance of our quantization procedure.

PROPOSITION **3.5** *Let B be a magnetic field with $\mathcal{C}^\infty_{\mathsf{pol}}(\mathbb{R}^d)$ components and A a vector potential for B also having $\mathcal{C}^\infty_{\mathsf{pol}}(\mathbb{R}^d)$ components. Assume that $V : \mathbb{R}^d \to \mathbb{R}$ is a measurable function that can be decomposed as $V = V_+ - V_-$ with $V_\pm \geq 0$, $V_+ \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$ and $V_- \in L^p(\mathbb{R}^d)$ with $p \geq d$. Then*

1. *$\mathfrak{q}_{V_-}$ is a $\mathfrak{h}_0$-bounded sesquilinear form with relative bound 0;*

2. *the Hamiltonian H defined in Definition* 3.1 *is bounded from below and we have* $\sigma_{\mathsf{ess}}(H) = \sigma_{\mathsf{ess}}(H_A \dot{+} V_+) \subset [0, \infty)$.

Proof. 1. Using Observation 3 in §2.8.1 from [37], we conclude that for $d > 1$, the Sobolev space $\mathcal{H}^{1/2}(\mathbb{R}^d)$ (that is the domain of the sesquilinear form $\mathfrak{h}_0$) is continuously embedded in $L^r(\mathbb{R}^d)$ for $2 \leq r \leq 2d/(d-1) < \infty$. Also using Hölder inequality, we deduce that for $r = 2p/(p-1) \in [2, 2d/(d-1)]$, for $p \geq d$

$$\|V_-^{1/2}u\|_2^2 \leq \|V_-\|_p \|u\|_r^2 \leq c\|V_-\|_p \|u\|^2_{\mathcal{H}^{1/2}(\mathbb{R}^d)}, \tag{3.13}$$

$\forall u \in \mathcal{H}^{1/2}(\mathbb{R}^d) = \mathcal{D}(\mathfrak{h}_0)$. Thus $V_-^{1/2} \in \mathbb{B}(\mathcal{H}^{1/2}(\mathbb{R}^d); L^2(\mathbb{R}^d))$; now let us prove that it is even compact. Let us observe that for $d \leq p < \infty$, $\mathcal{C}_0^\infty(\mathbb{R}^d)$ is dense in $L^p(\mathbb{R}^d)$. Thus, for $d \leq p < \infty$ let $\{W_\epsilon\}_{\epsilon>0} \subset \mathcal{C}_0^\infty(\mathbb{R}^d)$ be an approximating family for $V_-^{1/2}$ in $L^{2p}(\mathbb{R}^d)$, i.e. $\|V_-^{1/2} - W_\epsilon\|_{2p} \leq \epsilon$. Moreover, for any sequence $\{u_j\} \subset \mathcal{H}^{1/2}(\mathbb{R}^d)$ contained in the unit ball (i.e. $\|u_j\|_{\mathcal{H}^{1/2}} \leq 1$) we may suppose that it converges to $u \in \mathcal{H}^{1/2}(\mathbb{R}^d)$ for the weak topology on $\mathcal{H}^{1/2}(\mathbb{R}^d)$ and thus $\|u\|_{\mathcal{H}^{1/2}} \leq 1$. It follows that $W_\epsilon u_j$ converges to $W_\epsilon u$ in $L^2(\mathbb{R}^d)$ and due to (3.13) we have:

$$\|(V_-^{1/2} - W_\epsilon)(u - u_j)\| \leq C^{1/2}\|V_-^{1/2} - W_\epsilon\|_{L^{2p}} \|u - u_j\|_{\mathcal{H}^{1/2}} \leq 2c^{1/2}\epsilon, \quad \forall j \geq 1.$$

We conclude that $V_-^{1/2}u_j$ converges in $L^2(\mathbb{R}^d)$ to $V_-^{1/2}u$ and using the duality we also get that V_- is a compact operator from $\mathcal{H}^{1/2}(\mathbb{R}^d)$ to $\mathcal{H}^{-1/2}(\mathbb{R}^d)$. Using exercise 39 in ch. XIII of [47] we deduce that $\mathfrak{q}_-$ has zero relative bound with respect to $\mathfrak{h}_0$.

2. The conclusion of point 1 implies that the operator $V_-^{1/2}(H_0 + 1)^{-1/2} \in \mathbb{B}[L^2(\mathbb{R}^d)]$ is compact. Using the first point of Proposition 3.3 with $\lambda = -1$ and $r = 1/2$, and Pitt Theorem in [45], we conclude that the operator $V_-^{1/2}(H_A \dot{+} V_+ + 1)^{-1/2} \in \mathbb{B}[L^2(\mathbb{R}^d)]$ is also compact. Thus $V_- : \mathcal{D}(\mathfrak{h}_A + \mathfrak{q}_{V_+}) \to \mathcal{D}(\mathfrak{h}_A + \mathfrak{q}_{V_+})$ is compact and the conclusion (2) follows from exercise 39 in ch. XIII of [47]. □

4. The Feynman-Kac-Itô formula

In this section we gather some probabilistic notions and results needed in the proof of Theorem 1.1. The main idea is that we obtain a Feynman-Kac-Itô formula (following [20]) for the semigroup defined by $H(A, V)$ and this

allows us to reduce the problem to the case $B = 0$. For this last one we repeat then the proof in [5] giving all the necessary details for the case of singular potentials V; here an essential point is an explicit formula for the integral kernel of the operator $\mathrm{e}^{-tH(0,V)}$ in terms of a Lévy process.

Let $(\Omega, \mathfrak{F}, \mathsf{P})$ be a probability space, i.e. $\mathfrak{F}$ is a σ-algebra of subsets of Ω and P is a non-negative σ-aditive function on $\mathfrak{F}$ with $\mathsf{P}(\Omega) = 1$. For any integrable random variable $X : \Omega \to \mathbb{R}$ we denote its expectation value by

$$\mathsf{E}(X) := \int_\Omega X(\omega)\mathsf{P}(d\omega). \tag{4.1}$$

For any sub-σ-algebra $\mathfrak{G} \subset \mathfrak{F}$ we denote its associated conditional expectation by $\mathsf{E}(X \mid \mathfrak{G})$; this is the unique $\mathfrak{G}$-measurable random variable $Y : \Omega \to \mathbb{R}$ satisfying

$$\int_B Y(\omega)\mathsf{P}(\mathrm{d}\omega) = \int_B X(\omega)\mathsf{P}(\mathrm{d}\omega), \qquad \forall B \in \mathfrak{G}. \tag{4.2}$$

Let us recall the following properties of the conditional expectation (see for example [26]):

$$\mathsf{E}\left(\mathsf{E}(X \mid \mathfrak{G})\right) = \mathsf{E}(X), \tag{4.3}$$

$$\mathsf{E}(XZ \mid \mathfrak{G}) = Z\mathsf{E}(X \mid \mathfrak{G}), \tag{4.4}$$

for any $\mathfrak{G}$-measurable random variable $Z : \Omega \to \mathbb{R}$, such that ZX is integrable.

We also recall the Jensen inequality ([48], [26]): for any convex function $\varphi : \mathbb{R} \to \mathbb{R}$, and for any lower bounded random variable $X : \Omega \to \mathbb{R}$ the following inequality is valid

$$\varphi(\mathsf{E}(X)) \leq \mathsf{E}(\varphi(X)). \tag{4.5}$$

Following [6], we can associate to our Feller semigroup $\{P(t)\}_{t\geq 0}$, defined in Section 2, a Markov process $\{(\Omega, \mathfrak{F}, \mathsf{P}_x), \{X_t\}_{t\geq 0}, \{\theta_t\}_{t\geq 0}\}$; that we briefly recall here:

- Ω is the set of *"cadlag"* functions on $[0, \infty)$, i.e. functions $\omega : [0, \infty) \to \mathbb{R}^d$ (paths) that are continuous to the right and have a limit to the left in any point of $[0, \infty)$.

- $\mathfrak{F}$ is the smallest σ-algebra for which the *coordinate functions* $\{X_t\}_{t\geq 0}$, with $X_t(\omega) := \omega(t)$, are measurable.

- P_x is a probability on Ω such that for any $n \in \mathbb{N}^*$, for any ordered set $\{0 < t_1 \leq \ldots \leq t_n\}$ and any family $\{B_1, \ldots, B_n\}$ of Borel subsets in $\mathbb{R}^d$, we have

$$\mathsf{P}_x \{X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n\} \; = \tag{4.6}$$

$$= \int_{B_1} \mathrm{d}x_1 \, \mathring{\wp}_{t_1}(x - x_1) \int_{B_2} \mathrm{d}x_2 \, \mathring{\wp}_{t_2 - t_1}(x_1 - x_2) \; \ldots \int_{B_n} \mathrm{d}x_n \, \mathring{\wp}_{t_n - t_{n-1}}(x_{n-1} - x_n).$$

  One can deduce that, if E_x denotes the expectation value with respect to P_x, then for any $f \in \mathcal{C}_\infty(\mathbb{R}^d)$ and for any $t \geq 0$ one has

$$\mathsf{E}_x(f \circ X_t) \; = \; [P(t)f](x). \tag{4.7}$$

  We also remark that P_x is the image of the probability $\mathsf{P}_0 \equiv \mathsf{P}$ under the map $S_x : \Omega \to \Omega$ defined by $[S_x\omega](t) := x + \omega(t)$.

- For any $t \geq 0$, the map $\theta_t : \Omega \to \Omega$ is defined by $[\theta_t\omega](s) := \omega(s + t)$. If we denote by $\mathfrak{F}_t$ the sub-σ-algebra of $\mathfrak{F}$ generated by the processes $\{X_s\}_{0\leq s\leq t}$, then for any $t \geq 0$ and any bounded random variable $Y : \Omega \to \mathbb{R}$

$$\mathsf{E}_x \left(Y \circ \theta_t \mid \mathfrak{F}_t\right)(\omega) \; = \; \mathsf{E}_{X_t(\omega)}(Y), \quad \mathsf{P}_x - a.e. \text{ on } \Omega. \tag{4.8}$$

We use the fact that (see [25], [20]) the probability P_x is concentrated on the set of paths X_t such that $X_0 = x$ and by the Lévy-Ito Theorem:

$$X_t = x + \int_0^{t_+} \int_{\mathbb{R}^d} y \, \tilde{N}_X(\mathrm{d}s \, \mathrm{d}y). \tag{4.9}$$

Here $\tilde{N}_X(\mathrm{d}s \, \mathrm{d}y) := N_X(\mathrm{d}s \, \mathrm{d}y) - \hat{N}_X(\mathrm{d}s \, \mathrm{d}y)$, $\hat{N}_X(\mathrm{d}s \, \mathrm{d}y) := \mathsf{E}_x(N_X(\mathrm{d}s \, \mathrm{d}y)) = \mathrm{d}s \, \mathsf{n}(\mathrm{d}y)$ with $\mathsf{n}(\mathrm{d}y)$ the Lévy measure appearing in (2.1) and N_X a 'counting measure' on $[0, \infty) \times \mathbb{R}^d$ that for $0 < t < t'$ and B a Borel subset of $\mathbb{R}^d$ is defined as $N_X((t, t'] \times B) \; :=$

$$:= \# \left\{ s \in (t, t'] \; \mid \; X_s \neq X_{s-}, \, X_s X_{s-} \in B \right\}. \tag{4.10}$$

Following the procedure developped in [20] by Ichinose and Tamura one obtains a Feynman-Kac-Itô formula for Hamiltonians of the type $H = H_A \dot{+} V$. In fact we have

PROPOSITION **4.1** *Under the same conditions as in Definition* 3.1*, for any function* $u \in L^2(\mathbb{R}^d)$ *we have*

$$\left(\mathrm{e}^{-tH}u\right)(x) \;=\; \mathsf{E}_x\left((u\circ X_t)\,\mathrm{e}^{-S(t,X)}\right), \quad t\geq 0, x\in\mathbb{R}^d \tag{4.11}$$

where

$$\begin{aligned} S(t,X) \;:=\; \mathrm{i}\int_0^{t_+}\int_{\mathbb{R}^d}\tilde{N}_X(\mathrm{d}s\,\mathrm{d}y)\left\langle\int_0^1 \mathrm{d}r\,\left(A(X_{s_-}+ry)\right),\,y\right\rangle + \\ +\mathrm{i}\int_0^t\int_{\mathbb{R}^d}\hat{N}_X(\mathrm{d}s\,\mathrm{d}y)\left\langle\left(\int_0^1\mathrm{d}r\,A(X_s+ry)-A(X_s)\right),\,y\right\rangle + \\ +\int_0^t \mathrm{d}s\,V(X_s). \end{aligned} \tag{4.12}$$

In the sequel we shall take $A=0$ and $V\in C_0^\infty(\mathbb{R}^d)$. As it is proved in [6], the operator $\mathrm{e}^{-t(H_0\dot{+}V)}$ has an integral kernel that can be described in the following way. Let us denote by $\mathfrak{F}_{t-}$ the sub-σ-algebra of $\mathfrak{F}$ generated by the random variables $\{X_s\}_{0\leq s<t}$. For any pair $(x,y)\in[\mathbb{R}^d]^2$ and any $t>0$ we define a measure $\mu_{0,x}^{t,y}$ on the Borel space $(\Omega,\mathfrak{F}_{t-})$ by the equality

$$\mu_{0,x}^{t,y}(M) \;:=\; \mathsf{E}_x\left[\chi_M\,\mathring{\wp}_{t-s}(X_s-y)\right], \tag{4.13}$$

for any $M\in\mathfrak{F}_s$ and $0\leq s<t$, where χ_M is the characteristic function of M. This measure is concentrated on the family of *'paths'* $\{\omega\in\Omega\mid X_0(\omega)=x, X_{t-}(\omega)=y\}$ and we have $\mu_{0,x}^{t,y}(\Omega)=\mathring{\wp}_t(x-y)$.

PROPOSITION **4.2** *Let* $F:\Omega\to\mathbb{R}$ *be a non-negative* $\mathfrak{F}_{t-}$*-measurable random variable and let* $f:\mathbb{R}^d\to\mathbb{R}$ *be a positive borelian function. Then the following equality holds for any* $t>0$ *and any* $x\in\mathbb{R}^d$*:*

$$\begin{aligned} \int_{\mathbb{R}^d}\mathrm{d}y\left\{\int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega)\,F(\omega)\,\mathrm{e}^{-\int_0^t\mathrm{d}s\,V(X_s)}\right\}f(y) \;= \\ =\;\mathsf{E}_x\left(F\,\mathrm{e}^{-\int_0^t\mathrm{d}s\,V(X_s)}\,f(X_t)\right). \end{aligned} \tag{4.14}$$

Proof. This is a direct consequence of relations (2.29) and (2.33) from [6]. □

Let us now take $A=0$ in Proposition 4.1 and $F=1$ in Proposition 4.2 in order to deduce that the operator $\mathrm{e}^{-t(H_0\dot{+}V)}$ is an integral operator with integral kernel given by the function

$$\wp_t(x,y) \;:=\; \int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega)\,\mathrm{e}^{-\int_0^t\mathrm{d}s\,V(X_s)}, \quad t>0,\ (x,y)\in\mathbb{R}^d\times\mathbb{R}^d. \tag{4.15}$$

Proposition 3.3 from [6] implies that the function $[0,\infty)\times\mathbb{R}^d\times\mathbb{R}^d \ni (t,x,y) \mapsto \wp_t(x,y) \in \mathbb{R}$ is non-negative, continuous and verifies $\wp_t(x,y) = \wp_t(y,x)$. We shall also need the following result.

PROPOSITION **4.3** *For any $t > 0$, any $x \in \mathbb{R}^d$ and any function $g : \Omega \to \mathbb{R}$ that is integrable with respect to the measure $\mu_{0,x}^{t,x}$ we have the equality:*

$$\int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega)\, g(\omega) \;=\; \int_\Omega \mu_{0,0}^{t,0}(\mathrm{d}\omega)\, g(x+\omega). \tag{4.16}$$

Proof. It is evidently sufficient to prove that for any $s \in [0,t)$ and any $M \in \mathfrak{F}_s$ we have

$$\mu_{0,x}^{t,x}(M) \;=\; \left(\mu_{0,0}^{t,0} \circ S_x^{-1}\right)(M)$$

where the map $S_x : \Omega \to \Omega$ is defined by $(S_x(\omega)(t) := x + \omega(t)$. We noticed previously the identity $\mathsf{P}_x = \mathsf{P}_0 \circ S_x^{-1}$; thus for any function $F : \Omega \to \mathbb{R}$ integrable with respect to P_x we have $\mathsf{E}_x(F) = \mathsf{E}_0(F \circ S_x)$. We remark that $X_s(\omega + x) = \omega(s) + x = X_s(\omega) + x$, and using the definition of the measure $\mu_{0,x}^{t,x}$ in (4.13), we obtain

$$\mu_{0,x}^{t,x}(M) = \mathsf{E}_x\left[\chi_M \mathring{\wp}_{t-s}(X_s - x)\right] = \mathsf{E}_0\left[(\chi_M \circ S_x)\, \mathring{\wp}_{t-s}(X_s)\right] = \tag{4.17}$$

$$= \mathsf{E}_0\left[(\chi_{S_x^{-1}(M)}\, \mathring{\wp}_{t-s}(X_s)\right] = \mu_{0,0}^{t,0}\left(S_x^{-1}(M)\right) = \left[\mu_{0,0}^{t,0} \circ S_x^{-1}\right](M).$$

□

5. Proof of the bound for $N(0;V)$

In this Section we will consider $A = 0$ and we shall work only with a potential $V = V_+ - V_-$ satisfying the properties:

- $V_\pm \geq 0$,
- $V_+ \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$,
- $V_- \in L^d(\mathbb{R}^d) \cap L^{d/2}(\mathbb{R}^d)$.

We shall use the notations $H := H_0 \dot{+} V$, $H_+ := H_0 \dot{+} V_+$, $H_- := H_0 \dot{+} (-V_-)$ for the operators associated to the sesquilinear forms $\mathfrak{h} = \mathfrak{h}_0 + \mathfrak{q}_V$, $\mathfrak{h}_+ = \mathfrak{h}_0 + \mathfrak{q}_{V_+}$, $\mathfrak{h}_- = \mathfrak{h}_0 - \mathfrak{q}_{V_-}$.

Due to the results of Proposition 3.5 we have $\sigma_{\mathsf{ess}}(H) = \sigma_{\mathsf{ess}}(H_+) \subset \sigma(H_+) \subset [0,\infty)$ and $\sigma_{\mathsf{ess}}(H_-) = \sigma_{\mathsf{ess}}(H_0) = \sigma(H_0) = [0,\infty)$.

For any potential function W verifying the same conditions as V above, we denote by $N(W)$ the number of strictly negative eigenvalues (counted with their multiplicity) of the operator $H_0 \dot{+} W$. The following result reduces our study to the case $V_+ = 0$.

LEMMA **5.1** *The following inequality is true:*

$$N(V) \ \leq \ N(-V_-).$$

In particular we have that $N(V) = \infty$ implies that $N(-V_-) = \infty$.

Proof. We apply the *Min-Max* principle (see Theorem XIII.2 in [47]) noticing that $\mathcal{D}(\mathfrak{h}_-) = \mathcal{D}(\mathfrak{h}_o) \supset \mathcal{D}(\mathfrak{h})$ and $\mathfrak{h}_- \leq \mathfrak{h}$ and we deduce that the operator H_- has at least $N(V)$ strictly negative eigenvalues. □

Thus we shall suppose from now on that $V_+ = 0$.

5.1. Reduction to smooth, compactly supported potentials

In this subsection we shall prove that we can suppose $V_- \in C_0^\infty(\mathbb{R}^d)$. This will be done by approximation, using a result of the type of Theorem 4.1 from [50].

LEMMA **5.2** *Let V and V_n ($n \geq 1$) functions as in Proposition* 3.4. *In addition, $V_+ = V_{n,+} = 0$ for all $n \geq 1$ and $\lim_{n\to\infty} V_{n,-} = V_-$ in $L^1_{\mathrm{loc}}(\mathbb{R}^d)$ and $V_{n,-}$ are uniformly $\mathfrak{h}_0$-bounded with relative bound < 1. We set $H_n := H_A \dot{+} V_n$. Then $H_n \to H$ when $n \to \infty$ in strong resolvent sense.*

Proof. We denote by $\mathfrak{h}_n$ the quadratic form associated to H_n, i.e. $\mathfrak{h}_n = \mathfrak{h}_A - \mathfrak{q}_{n,-}$, where $\mathfrak{q}_{n,-}$ is associated to $V_{n,-}$ by (3.7). We have $D(h_n) = D(h_A) \subset D(q_{n,-})$, and according to Proposition 3.4 there exist $\alpha \in (0,1)$ and $\beta > 0$ such that

$$\mathfrak{q}_{n,-}(v) \leq \alpha \mathfrak{h}_A(v) + \beta \parallel v \parallel, \quad \forall v \in D(\mathfrak{h}_A), \ \forall n \geq 1. \tag{5.1}$$

It follows that $\mathfrak{h}_n$ are uniformly lower bounded and the norms defined on $D(\mathfrak{h}_A)$ by $\mathfrak{h}_A$ and $\mathfrak{h}_n$ are equivalent, uniformly with respect to $n \geq 1$. Moreover, $C_0^\infty(\mathbb{R}^d)$ is a core for H_A, thus for $\mathfrak{h}_A$, $\mathfrak{h}$ and $\mathfrak{h}_n$ also.

Let $f \in L^2(\mathbb{R}^d)$ and $u_n := (H_n + i)^{-1} f \in D(H_n) \subset D(\mathfrak{h}_A)$, $n \geq 1$. We have clearly

$$\| u_n \| \leq \| f \|, \quad |\mathfrak{h}_n(u_n)| = |(H_n u_n, u_n)| \leq \| f \|, \quad \forall n \geq 1. \tag{5.2}$$

From (5.1), the subsequent comments and (5.2) it follows that the sequence $(u_n)_{n\geq 1}$ is bounded in $D(\mathfrak{h}_A)$, while the sequence $\left(V_{n,-}^{1/2} u_n\right)_{n\geq 1}$ is bounded in $L^2(\mathbb{R}^d)$. Let $u \in L^2(\mathbb{R}^d)$ be a limit point of the sequence $(u_n)_{n\geq 1}$ with respect to the weak topology on $L^2(\mathbb{R}^d)$. By restricting maybe to a subsequence, we may assume that there exist $\psi, \eta \in L^2(\mathbb{R}^d)$ such that $H_A^{1/2} u_n \underset{n\to\infty}{\to} \psi$ and $V_{n,-}^{1/2} u_n \underset{n\to\infty}{\to} \eta$ in the weak topology of $L^2(\mathbb{R}^d)$. For $g \in D\left(H_A^{1/2}\right)$ we have

$$\left(H_A^{1/2} g, u\right) = \lim_{n\to\infty} \left(H_A^{1/2} g, u_n\right) = \lim_{n\to\infty} \left(g, H_A^{1/2} u_n\right) = (g, \psi),$$

thus $u \in D(H_A^{1/2})$ and $H_A^{1/2} u = \psi$. Then $u \in D(\mathfrak{q}_-)$ and for any $g \in C_0^\infty(\mathbb{R}^d)$

$$(\eta, g) = \lim_{n\to\infty} \left(V_{n,-}^{1/2} u_n, g\right) = \lim_{n\to\infty} \left(u_n, V_{n,-}^{1/2} g\right) = \left(u, V_-^{1/2} g\right) = \left(V_-^{1/2} u, g\right),$$

implying $V_-^{1/2} u = \eta$.

It follows that for every $g \in C_0^\infty(\mathbb{R}^d)$ we have

$$(g, f) = (g, (H_n + \mathrm{i}) u_n) = \mathfrak{h}_n(g, u_n) - \mathrm{i}(g, u_n) =$$

$$= \left(H_A^{1/2} g, H_A^{1/2} u_n\right) - \left(V_{n,-}^{1/2} g, V_{n,-}^{1/2} u_n\right) - \mathrm{i}(g, u_n) \to \mathfrak{h}(g, u) - \mathrm{i}(g, u).$$

Consequently, $u \in D(H)$ and $(H + \mathrm{i})u = f$. Thus the sequence $(u_n)_{n\geq 1}$ has the single limit point $u = (H + \mathrm{i})^{-1} f$ for the weak topology of $L^2(\mathbb{R}^d)$. It follows that $(H_n \pm \mathrm{i})^{-1} f \to (H \pm \mathrm{i})^{-1} f$ weakly in $L^2(\mathbb{R}^d)$ for $n \to \infty$.

By the resolvent identity we get

$$\| (H_n \pm \mathrm{i})^{-1} f \|^2 = \frac{\mathrm{i}}{2} \left((f, (H_n - \mathrm{i})^{-1} f) - (f, (H_n + \mathrm{i})^{-1} f) \right) \to \| (H \pm \mathrm{i})^{-1} f \|^2,$$

therefore $(H_n + \mathrm{i})^{-1} f \to (H + \mathrm{i})^{-1} f$ in $L^2(\mathbb{R}^d)$. □

A direct consequence of Lemma 5.2 and Theorem VIII.20 from [47] is

COROLLARY **5.1** *Under the hypothesis of Lemma* 5.2*, for any function f bounded and continuous on $\mathbb{R}$ and any $u \in L^2(\mathbb{R}^d)$, we have $f(H_n)u \to f(H)u$.*

Approximating V_- is done by the standard procedures: cutoffs and regularization. The first of the lemmas below is obvious.

LEMMA **5.3** *Let $V_- \in L^1_{\mathsf{loc}}(\mathbb{R}^d)$ with $V_- \geq 0$ and assume that its associated sesquilinear form is $\mathfrak{h}_0$-bounded with relative bound strictly less then 1. Let $\theta \in C_0^\infty([0,\infty))$ satisfy the following: $0 \leq \theta \leq 1$, θ is a decreasing function, $\theta(t) = 1$ for $t \in [0,1]$ and $\theta(t) = 0$ for $t \in [2,\infty)$.*

If we denote by $\theta^n(x) := \theta(|x|/n)$ and $V_-^n = \theta^n V_-$, then $V_-^n \to V_-$ in $L^1_{\mathsf{loc}}(\mathbb{R}^d)$, $0 \leq V_-^n \leq V_-^{n+1}$ and the sesquilinear forms associated to V_-^n are $\mathfrak{h}_0$-bounded with relative bound strictly less then 1, uniformly in $n \in \mathbb{N}^*$.

Moreover, if we denote by $\mathfrak{h}^n$ the sesquilinear form associated to the operator $H_A \dot{+} (-V_-^n)$, we have $\mathfrak{h}^{(n)} \geq \mathfrak{h}^{(n+1)} \geq \mathfrak{h}$ and $\mathfrak{h}^{(n)}(u) \underset{n\to\infty}{\to} \mathfrak{h}(u)$ for any $u \in \mathcal{D}(\mathfrak{h}_A)$.

If, in addition, $V_- \in L^p(\mathbb{R}^d)$, $p \geq 1$, then $V_-^n \in L^p_{\mathrm{comp}}(\mathbb{R}^d)$, $\|V_-^n\|_{L^p} \leq \|V_-\|_{L^p}$ for any $n \geq 1$, and $V_-^n \to V_-$ in $L^p(\mathbb{R}^d)$.

LEMMA **5.4** (a) *Let $V_- \in L^1_{\mathrm{loc}}(\mathbb{R}^d)$, $V_- \geq 0$ and $\mathfrak{h}_0$-bounded with relative bound < 1. Let $\theta \in C_0^\infty(\mathbb{R}^d)$, $\theta \geq 0$ and $\int_{\mathbb{R}^d} \theta = 1$. We set $\theta_n(x) := n^d \theta(nx)$, $x \in \mathbb{R}^d$, $n \in \mathbb{N}^*$ and $V_{n,-} := V_- * \theta_n \in C_0^\infty$. In particular, $V_{n,-} \in C_0^\infty(\mathbb{R}^d)$ if $V_- \in L^1_{\mathrm{comp}}(\mathbb{R}^d)$.*

Then $V_{n,-} \to V_-$ in $L^1_{\mathrm{loc}}(\mathbb{R}^d)$ for $n \to \infty$ and the functions $V_{n,-}$ are non-negative and uniformly h_0-bounded, with relative bound < 1. Moreover, $\mathfrak{h}_n(u) \to \mathfrak{h}(u)$ for any $u \in D(\mathfrak{h}_A)$, where $\mathfrak{h}_n$ is the quadratic form associated to $H_n := H_A \dot{+} (-V_n)$.

(b) *If, in addition, $V_- \in L^p(\mathbb{R}^d)$ with $p \geq 1$, then $V_{n,-} \in L^p(\mathbb{R}^d) \cap C^\infty(\mathbb{R}^d)$, $\| V_{n,-} \|_{L^p} \leq \| V_- \|_{L^p}$, $\forall n \geq 1$ and $V_{n,-} \to V_-$ in $L^p(\mathbb{R}^d)$.*

Proof. (a) We have for any $x \in \mathbb{R}^d$

$$V_{n,-}(x) = \int_{\mathbb{R}^d} \mathrm{d}y\, \theta_n(y) V_-(x-y) = \int_{\mathbb{R}^d} \mathrm{d}y\, \theta(y) V_-(x - n^{-1}y). \tag{5.3}$$

By the Dominated Convergence Theorem, for any compact $K \subset \mathbb{R}^d$

$$\int_K \mathrm{d}x\, |V_{n,-}(x) - V_-(x)| \leq \int_{\mathbb{R}^d} \mathrm{d}y\, \theta(y) \int_K \mathrm{d}x\, |V_-(x - n^{-1}y) - V_-(x)| \to 0,$$

hence $V_{n,-}$ converges to V_- in $L^1_{\mathrm{loc}}(\mathbb{R}^d)$ when $n \to \infty$.

If V_- is relatively small with respect to $\mathfrak{h}_0$, we use the fact that $H_0^{1/2}$ is a convolution operator (hence it commutes with translations) and using the

comments after inequality (5.1), we deduce that for any $u \in C_0^\infty(\mathbb{R}^d)$ there exists $\alpha \in (0,1)$ and $\beta \geq 0$ such that

$$\int_{\mathbb{R}^d} \mathrm{d}x\, V_{n,-}|u|^2 = \int_{\mathbb{R}^d} \mathrm{d}y\, \theta_n(y) \int_{\mathbb{R}^d} \mathrm{d}z\, V_-(z)|u(z+y)|^2 \leq$$

$$\leq \int_{\mathbb{R}^d} \mathrm{d}y\, \theta_n(y) \left[\alpha \parallel H_0^{1/2} u(\cdot + y) \parallel^2 + \beta \parallel u(\cdot + y) \parallel^2\right] =$$

$$= \alpha \parallel H_0^{1/2} u \parallel^2 + \beta \parallel u \parallel^2 .$$

(b) From (5.3) it follows that

$$\parallel V_{n,-} \parallel_{L^p} \leq \int_{\mathbb{R}^d} \mathrm{d}y\, \theta_n(y) \parallel V_-(\cdot - y) \parallel_{L^p} \leq \parallel V_- \parallel_{L^p} .$$

Also, using the Dominated Convergence Theorem, we infer that

$$\parallel V_{n,-} - V_- \parallel_{L^p} \leq \int_{\mathbb{R}^d} \mathrm{d}y\, \theta(y) \parallel V_-(\cdot) - V_-(\cdot - n^{-1}y) \parallel_{L^p} \to 0.$$

□

Thus Lemmas 5.3 and 5.4 imply, for a potential function V_- satisfying the hypothesis of the Lemma, the existence of a sequence $(V_{n,-})_{n\geq 1} \subset C_0^\infty(\mathbb{R}^d)$ such that $V_{n,-} \geq 0$, $\parallel V_{n,-} \parallel_{L^p} \leq \parallel V_- \parallel_{L^p}$, $\forall n \geq 1$, $V_{n,-} \to V_-$ in $L^p(\mathbb{R}^d)$ (for $p = d$ and $p = d/2$) when $n \to \infty$ and the functions $V_{n,-}$ are uniformly $\mathfrak{h}_0$-bounded with relative bound < 1.

LEMMA **5.5** *Assume that there exists a constant $C > 0$, such that the inequality*

$$N(-V_{n,-}) \leq C \left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_{n,-}(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_{n,-}(x)|^{d/2} \right) \tag{5.4}$$

holds for any $n \geq 1$. Then one also has

$$N(-V_-) \leq C \left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^{d/2} \right). \tag{5.5}$$

Proof. We set $H_{n,-} := H_0 \dot{+} (-V_{n,-})$; $(E_{n,-}(\lambda))_{\lambda\in\mathbb{R}}$ will be the spectral family of $H_{n,-}$ and $(E_-(\lambda))_{\lambda\in\mathbb{R}}$ the spectral family of H_-. For $\lambda < 0$, we denote by $N_\lambda(W)$ the number of eigenvalues of $H_0 \dot{+} W$ which are strictly smaller than λ (for any potential function W satisfying the hypothesis at the

begining of this section). It suffices to show that for any $\lambda < 0$ not belonging to the spectrum of H_-, one has the inequality

$$N_\lambda(-V_-) \leq C\left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^{d/2}\right). \tag{5.6}$$

Since $V_{n,-}$ converges to V_- in $L^1_{\mathrm{loc}}(\mathbb{R}^d)$, cf. Lemma 5.2, $H_{n,-}$ will converge to H_- in strong resolvent sense. By [29], Ch. VIII, Th. 1.15, this implies the strong convergence of $E_{n,-}(\lambda)$ to $E_-(\lambda)$ for any $\lambda \notin \sigma(H_-)$. By Lemmas 1.23 and 1.24 from [29], Ch. VII, for $\lambda < 0$ such that $\lambda \notin \sigma(H_-)$, one also has $\| E_{n,-}(\lambda) - E_-(\lambda) \| \to 0$. Let us suppose that there exists some $\lambda < 0$ not belonging to $\sigma(H_-)$ and such that for it the inequality (5.6) is not verified. Thus for the given $\lambda < 0$ we have $\forall n \geq 1$:

$$N(-V_{n,-}) \leq C\left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^{d/2}\right) < N_\lambda(-V_-).$$

But for n large enough, one has $N_\lambda(-V_-) = N_\lambda(-V_{n,-})$ and thus

$$\begin{aligned} N_\lambda(-V_-) = N_\lambda(-V_{n,-}) \leq N(-V_{n,-}) \leq \\ \leq C\left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_{n,-}(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_{n,-}(x)|^{d/2}\right) \leq \\ \leq C\left(\int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^d + \int_{\mathbb{R}^d} \mathrm{d}x\, |V_-(x)|^{d/2}\right) \end{aligned}$$

that is a contradiction with our initial hypothesis. □

5.2. Proof of the Theorem 1.1 without magnetic field

We shall assume from now on that $V_+ = 0$ and $0 \leq V_- \in C_0^\infty(\mathbb{R}^d)$. We check a Birman-Schwinger principle. For $\alpha > 0$ we set $K_\alpha := V_-^{1/2}(H_0 + \alpha)^{-1}V_-^{1/2}$; it is a positive compact operator on $L^2(\mathbb{R}^d)$.

LEMMA **5.6**

$$N_{-\alpha}(-V_-) \leq \sharp\,\{\mu > 1 \mid \mu \text{ eigenvalue of } K_\alpha\}. \tag{5.7}$$

Proof. We introduce the sequence of functions $\mu_n : [0, \infty) \to (-\infty, 0]$, $n \geq 1$, where $\mu_n(\lambda)$ is the n'th eigenvalue of $H_0 - \lambda V_-$ if this operator has at least n strictly negative eigenvalues and $\mu_n(\lambda) = 0$ if not. Cf. [47], §XIII.3, μ_n is continuous and decreasing (even strictly decreasing on intervals on which it

is strictly negative). Obviously, we have $N_{-\alpha}(-V_-) \leq \# \{n \geq 1 \mid \mu_n(1) < -\alpha\}$. Now fix some n such that $\mu_n(1) < -\alpha$ and recall that $\mu_n(0) = 0$. The function μ_n is continuous and injective on the interval $[\epsilon_n, 1]$, where $\epsilon_n := \sup\{\lambda \geq 0 \mid \mu_n(\lambda) = 0\}$, therefore it exists a unique $\lambda \in (0,1)$ such that $\mu_n(\lambda) = -\alpha$. Thus

$$N_{-\alpha}(-V_-) = \# \{\lambda \in (0,1) \mid \exists n \geq 1 \ s.t. \ \mu_n(\lambda) = -\alpha\} =$$

$$= \# \{\lambda \in (0,1) \mid \exists \varphi \in D(H_0) \setminus \{0\} \ s.t. \ (H_0 - \lambda V_-)\varphi = -\alpha\varphi\} \leq$$

$$\leq \# \{\lambda \in (0,1) \mid \exists \psi \in L^2(\mathbb{R}^d) \setminus \{0\} \ s.t. \ K_\alpha \psi = \lambda^{-1}\psi\},$$

where for the last inequality we set $\psi := V_-^{1/2}\varphi$, noticing that the equality $(H_0 + \alpha)\varphi = \lambda V_- \varphi$ implies $\psi \neq 0$. □

LEMMA **5.7** *Let $F : [0,\infty) \to [0,\infty)$ be a strictly increasing continuous function with $F(0) = 0$. Then $F(K_\alpha)$ is a positive compact operator and the next inequality holds:*

$$N_{-\alpha}(-V_-) \leq F(1)^{-1} \sum_{F(\mu) \in \sigma[F(K_\alpha)], F(\mu) > F(1)} F(\mu).$$

Proof. The first part is obvious. Using (5.7) and F's monotony, we get

$$N_{-\alpha}(-V_-) \leq \sharp\{\mu > 1 \mid \mu \in \sigma(K_\alpha)\} = \# \{F(\mu) \mid \mu > 1, F(\mu) \in \sigma[F(K_\alpha)]\} =$$

$$= \sum_{\mu > 1, F(\mu) \in \sigma[F(K_\alpha)]} \frac{F(\mu)}{F(\mu)} \leq F(1)^{-1} \sum_{\mu > 1, F(\mu) \in \sigma[F(K_\alpha)]} F(\mu).$$

□

So, we shall be interested in finding functions F having the properties in the statement above, such that $F(K_\alpha) \in B_1$ (the ideal of trace-class operators in $L^2(\mathbb{R}^d)$) and such that $\mathrm{Tr}\,[\mathrm{F}(\mathrm{K}_\alpha)]$ is conveniently estimated.

Using an idea from [48], we are going to consider functions of the form

$$F(t) := t \int_0^\infty \mathrm{d}s \, \mathrm{e}^{-s} g(ts), \quad t \geq 0,$$

where $g : [0,\infty) \to [0,\infty)$ is continuous, bounded and $g \not\equiv 0$. Plainly, $F : [0,\infty) \to [0,\infty)$ is continuous, $F(0) = 0$, satisfies $F(t) \leq Ct$ for some $C > 0$ and the identity

$$F(t) = \int_0^\infty \mathrm{d}r \, \mathrm{e}^{-rt^{-1}} g(r)$$

implies that F is strictly increasing. We shall use the notations $F = \Phi(g)$, $\tilde{g}(t) := tg(t)$.

In particular, $g_\lambda(t) = \mathrm{e}^{-\lambda t}$, $\lambda > 0$ leads to $F_\lambda(t) = t(1+\lambda t)^{-1}$. In the sequel, relations valid for this particular case will be extended to the following case, that we shall be interested in:

$$g_\infty : [0,\infty) \to [0,\infty), \quad g_\infty(t) = 0 \text{ if } 0 \le t \le 1, \quad g_\infty(t) = 1 - 1/t \text{ if } t > 1, \tag{5.8}$$

by using an approximation that we now introduce. The first lemma is obvious.

LEMMA **5.8** *Let g_∞ be given by* (5.8). *For $n \ge 1$ we define $g_n : [0,\infty) \to [0,1]$, $g_n(t) = g(t)$ for $0 \le t \le n$, $g_n(t) = \frac{2n-1}{t} - 1$ for $n \le t \le 2n-1$, $g_n(t) = 0$ for $t \ge 2n-1$. Then $g_n \in C_0((0,\infty))$, $0 \le g_n \le g_{n+1} \le g_\infty$, $\forall n$ and $g_n \to g_\infty$ when $n \to \infty$ uniformly on any compact subset of $[0,\infty)$.*

LEMMA **5.9** *Let f be a nonnegative continuous function on $[0,\infty)$ such that $\lim_{t\to\infty} f(t) = 0$. There exists a sequence $(f^k)_{k\ge1}$ of real functions on $[0,\infty)$ with the properties*

(a) *Every f^k is a finite linear combination of functions of the form g_λ, $\lambda > 0$.*

(b) *$f^k \ge f^{k+1} \ge f \ge 0$ on $[0,\infty)$, $\forall k \ge 1$,*

(c) *$f^k \to f$ uniformly on $[0,\infty)$ when $k \to \infty$.*

Proof. We define the function $h : [0,1] \to [0,\infty)$, $h(s) := f(-\ln s)$ for $s \in (0,1]$, $h(0) := 0$. It follows that $h \in C([0,1])$. We can chose now two sequences of positive numbers $\{\epsilon_k\}_{k\ge1}$ and $\{\delta_k\}_{k\ge1}$ verifying the properties: $\lim\limits_{k\to\infty}(\epsilon_k + \delta_k) = 0$ and $\delta_k - \epsilon_k \ge \epsilon_{k+1} + \delta_{k+1} > 0, \forall k \ge 1$ (for example we may take $\delta_k = (k+2)^{-1}$ and $\epsilon_k = (k+2)^{-3}$). Using the Weierstrass Theorem we may find for any $k \ge 1$ a real polynomial P'_k such that $\sup\limits_{s\in[0,1]} |h(s) - P'_k(s)| \le \epsilon_k$ and let us denote by $P_k := P'_k + \delta_k$. We get:

$$\sup_{s\in[0,1]} |h(s) - P_k(s)| \le \epsilon_k + \delta_k \underset{k\to\infty}{\to} 0,$$

$$h \le h + \delta_{k+1} - \epsilon_{k+1} \le P'_{k+1} + \delta_{k+1} = P_{k+1} \le h + \delta_{k+1} + \epsilon_{k+1} \le$$

$$\le h + \delta_k - \epsilon_k \le P'_k + \delta_k = P_k$$

on $[0,1]$. Thus $f^k(t) := P_k(\mathrm{e}^{-t})$ defined on $[0,\infty)$ for $k \ge 1$ have the required properties. □

PROPOSITION **5.1** *Let $F_\infty := \Phi(g_\infty)$. The operator $F_\infty(K_\alpha)$ is self-adjoint, positive and compact on $L^2(\mathbb{R}^d)$. It admits an integral kernel of the form*

$$[F_\infty(K_\alpha)]\,(x,y) = \tag{5.9}$$

$$= V_-^{1/2}(x)V_-^{1/2}(y)\int_0^\infty \mathrm{d}t\,\mathrm{e}^{-\alpha t}\int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega)g_\infty\left(\int_0^t \mathrm{d}s\,V_-(X_s)\right),$$

which is continuous, symmetric, with $[F_\infty(K_\alpha)]\,(x,x) \geq 0$.

Proof. The first part is clear. To establish (3.27), we treat first the operator $B_\lambda := F_\lambda(K_\alpha)$, $\lambda > 0$. We have

$$B_\lambda = K_\alpha(1+\lambda K_\alpha)^{-1} \implies B_\lambda = K_\alpha - \lambda B_\lambda K_\alpha. \tag{5.10}$$

The second resolvent identity gives

$$(H_0+\alpha)^{-1} - (H_0+\lambda V_- + \alpha)^{-1} = \lambda(H_0+\lambda V_- + \alpha)^{-1}V_-(H_0+\alpha)^{-1}.$$

Multiplying by $V_-^{1/2}$ to the left and to the right and taking into account (5.10) and the definition of K_α, one gets

$$B_\lambda = V_-^{1/2}(H_0+\lambda V_- + \alpha)^{-1}V_-^{1/2} = V_-^{1/2}\left[\int_0^\infty \mathrm{d}t\,\mathrm{e}^{-\alpha t}\mathrm{e}^{-t(H_0+\lambda V_-)}\right]V_-^{1/2}.$$

By Proposition 4.2 and its consequences, for any $u \in C_0(\mathbb{R}^d)$, $u \geq 0$, we have

$$[F_\lambda(K_\alpha)u]\,(x) = \tag{5.11}$$

$$= V_-^{1/2}(x)\int_0^\infty \mathrm{d}t e^{-\alpha t}\int_{\mathbb{R}^d}\mathrm{d}y\left[\int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega)\,g_\lambda\left(\int_0^t \mathrm{d}s\,V_-(X_s)\right)\right]V_-^{1/2}(y)u(y).$$

Since Φ maps monotonous convergent sequences into monotonous convergent sequences, by applying Lemmas 5.8 and 5.9 and the Monotonous Convergence Theorem (B. Levi), we get (5.11) for $\lambda = \infty$, for the couple (g_∞, F_∞).

We introduce the notation

$$G_\lambda(t;x,y) := \int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega)\,g_\lambda\left(\int_0^t \mathrm{d}s\,V_-(X_s)\right), \tag{5.12}$$

for $t > 0$, $x, y \in \mathbb{R}^d$, $0 < \lambda \leq \infty$. By the consequences of Proposition 4.2, for any $0 < \lambda < \infty$ the function G_λ is continuous on $(0,\infty)\times\mathbb{R}^d\times\mathbb{R}^d$ and symmetric in x, y. To obtain the same properties for $\lambda = \infty$, we approximate g_∞ by using once again Lemmas 5.8 and 5.9. So it exists a sequence $(f_n)_{n\geq 1}$ of real continuous functions on $[0,\infty)$, each one being a finite linear combination

of functions of the form g_λ, such that f_n converges to g_∞ uniformly on any compact subset of $[0,\infty)$. On the other hand, if $M>0$ is an upper bound for V_-, we have

$$0 \leq \int_0^t \mathrm{d}s\, V_-(X_s) \leq Mt,$$

and $\mu_{0,x}^{t,y}(\Omega) = \mathring{\wp}_t(x-y)$. It follows that G_∞ is, uniformly on compact subsets of $[0,\infty)\times\mathbb{R}^d\times\mathbb{R}^d$, the limit of a sequence of continuous functions, which are symmetric in x,y. Thus G_∞ has the same properties. Moreover, since $0\leq g_\infty\leq 1$ and $g_\infty(t)=0$ for $0\leq t\leq 1$, we have $G_\infty(t;x,y)=0$ for $t\leq 1/M$. Using (2.4) and (2.3), there is a constant $C>0$ such that

$$0 \leq G_\infty(t;x,y) \leq C, \quad \forall t>0,\ \forall x,y\in\mathbb{R}^d. \tag{5.13}$$

From (5.11) for $\lambda=\infty$, we infer that $F_\infty(K_\alpha)$ has an integral kernel of the form

$$\left[F_\infty(K_\alpha)\right](x,y) = V_-^{1/2}(x)V_-^{1/2}(y)\int_0^\infty \mathrm{d}t\,\mathrm{e}^{-\alpha t}G_\infty(t;x,y), \tag{5.14}$$

so (3.27) is verified. The continuity of $F_\infty(K_\alpha)$ follows from the Dominated Convergence Theorem and from (5.13). The symmetry is obvious, and the last property of the statement follows from $F_\infty(K_\alpha)\geq 0$. □

REMARK **5.1** *By a lemma from* [47], *§XI.4, $F_\infty(K_\alpha)\in B_1$ if the function $\mathbb{R}^d\ni x\mapsto\left[F_\infty(K_\alpha)\right](x,x)$ is integrable and one has*

$$\mathrm{Tr}\left[F_\infty(K_\alpha)\right] = \int_{\mathbb{R}^d}\mathrm{d}x\ \left[F_\infty(K_\alpha)\right](x,x). \tag{5.15}$$

Setting $D_\infty(t;x) := V_-(x)G_\infty(t;x,x)$, $t>0, x\in\mathbb{R}^d$, we have

$$\left[F_\infty(K_\alpha)\right](x,x) = \int_0^\infty \mathrm{d}t\,\mathrm{e}^{-\alpha t}D_\infty(t;x). \tag{5.16}$$

To check the integrability of this function, one introduces

$$\Psi_\infty : (0,\infty)\times\mathbb{R}^d\to\mathbb{R}_+,$$

$$\Psi_\infty(t;x) := t^{-1}\int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega)\,\tilde{g}_\infty\left(\int_0^t \mathrm{d}s\, V_-(X_s)\right),$$

where $\tilde{g}_\infty(t) := tg_\infty(t)$. The role of this function is stressed by

LEMMA **5.10** *For $d \geq 3$ consider the following constant depending only on d:*

$$\overline{C}_d := C\left(\int_1^\infty \mathrm{d}s\, s^{-d}\, g_\infty(s) \vee \int_1^\infty \mathrm{d}s\, s^{-d/2}\, g_\infty(s)\right) = C\int_1^\infty \mathrm{d}s\, s^{-d/2}\, g_\infty(s)$$

where C is the constant verifying (2.6). One has

$$\int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t}\int_{\mathbb{R}^d} \mathrm{d}x\, \Psi_\infty(t;x) \leq \overline{C}_d\left(\int_{\mathbb{R}^d} \mathrm{d}x\, V_-^d(x) + \int_{\mathbb{R}^d} \mathrm{d}x\, V_-^{d/2}(x)\right). \tag{5.17}$$

Proof. The function $\tilde{g}_\infty$ is convex and $\frac{\mathrm{d}s}{t}$ is a probability on $(0,t)$; thus by the Jensen inequality we obtain

$$\tilde{g}_\infty\left(\int_0^t \mathrm{d}s\, V_-(X_s)\right) \leq \int_0^t \frac{\mathrm{d}s}{t}\, \tilde{g}_\infty\left(t\, V_-(X_s)\right).$$

Let us also remark that for the constant $\overline{C}_d$ to be finite we have to ask that $d \geq 3$ for the factor $s^{-d/2}$ to be integrable at infinity, because the convexity condition on $\tilde{g}_\infty$ rather implies that g_∞ cannot vanish at infinity.

Then

$$\int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t}\int_{\mathbb{R}^d} \mathrm{d}x\, \Psi_\infty(t;x) \leq$$

$$\leq \int_0^\infty \mathrm{d}t\, t^{-2}\, \mathrm{e}^{-\alpha t}\int_{\mathbb{R}^d} \mathrm{d}x\, \left[\int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega)\int_0^t \mathrm{d}s\, \tilde{g}_\infty\left(tV_-(X_s)\right)\right].$$

Using now Proposition 4.3, the last expression is equal to:

$$\int_0^\infty \mathrm{d}t\, t^{-2}\, \mathrm{e}^{-\alpha t}\int_{\mathbb{R}^d} \mathrm{d}x\, \left[\int_\Omega \mu_{0,0}^{t,0}(\mathrm{d}\omega)\int_0^t \mathrm{d}s\, \tilde{g}_\infty\left(tV_-(x+\omega(s))\right)\right] =$$

$$= \int_0^\infty \mathrm{d}t\, t^{-2}\, \mathrm{e}^{-\alpha t}\left[\int_\Omega \mu_{0,0}^{t,0}(\mathrm{d}\omega)\int_0^t \mathrm{d}s\int_{\mathbb{R}^d} \mathrm{d}x\, \tilde{g}_\infty\left(tV_-(x)\right)\right] =$$

$$= \int_0^\infty \mathrm{d}t\, t^{-1}\, \mathrm{e}^{-\alpha t}\left[\int_\Omega \mu_{0,0}^{t,0}(\mathrm{d}\omega)\right]\int_{\mathbb{R}^d} \mathrm{d}x\, \tilde{g}_\infty\left(tV_-(x)\right) =$$

$$= \int_0^\infty \mathrm{d}t\, t^{-1}\, \mathrm{e}^{-\alpha t}\, \mathring{\wp}_t(0)\int_{\mathbb{R}^d} \mathrm{d}x\, \tilde{g}_\infty\left(tV_-(x)\right) \leq$$

$$\leq C\int_{\mathbb{R}^d} \mathrm{d}x\, \left[\int_0^\infty \mathrm{d}t\, t^{-d-1}(1+t^{d/2})\tilde{g}_\infty\left(tV_-(x)\right)\right] \leq$$

$$\leq \overline{C}_d\left(\int_{\mathbb{R}^d} \mathrm{d}x\, V_-^d(x) + \int_{\mathbb{R}^d} dx\, V_-^{d/2}(x)\right),$$

where we have used the fact that $s < 1$ implies $g_\infty(s) = 0$. □

The next result gives the connection between D_∞ and Ψ_∞:

PROPOSITION **5.2**

$$\int_{\mathbb{R}^d} \mathrm{d}x\, D_\infty(t,x) \;=\; \int_{\mathbb{R}^d} \mathrm{d}x\, \Psi_\infty(t,x).$$

Proof. First let us verify the following identity for any $t > 0$:

$$\int_{\mathbb{R}^d} \mathrm{d}x\, D_\lambda(t,x) \;=\; \int_{\mathbb{R}^d} \mathrm{d}x\, \Psi_\lambda(t,x), \quad \text{for } \lambda \in (0,\infty) \tag{5.18}$$

where D_λ and Ψ_λ are defined in terms of g_λ in the same way that D_∞ and Ψ_∞ are defined in terms of g_∞. Let us point out that both D_λ and Ψ_λ are positive measurable functions on $(0,\infty)\times\mathbb{R}^d$ but only the integral on the left hand side of (5.18) is evidently finite by what we have proven so far. For simplifying the writing we shall take $\lambda = 1$. For any $r \in [0,t]$ we denote by

$$S_r := \mathrm{e}^{-r(H_0+V_-)} V_- \mathrm{e}^{-(t-r)(H_0+V_-)}.$$

Following the remarks after Proposition 4.2 above, for $r \in (0,t)$, both exponentials appearing in the above right hand side are integral operators with non-negative continuous integral kernels; thus S_r will also be an integral operator with non-negative continuous kernel that we shall denote by K_r, and we can compute it explicitely as follows. For a non-negative $u \in C_0(\mathbb{R}^d)$, using Proposition 4.1 with $A = 0$ gives

$$(S_r u)(x) = \mathsf{E}_x \left\{ \mathrm{e}^{-\int_0^r V_-(X_\rho)\mathrm{d}\rho} V_-(X_r) \mathsf{E}_{X_r} \left[\mathrm{e}^{-\int_0^{t-r} V_-(X_\sigma)\mathrm{d}\sigma} u(X_{t-r}) \right] \right\}$$

and using the Markov property (4.8) we obtain

$$\mathsf{E}_{X_r} \left[\mathrm{e}^{-\int_0^{t-r} V_-(X_\sigma)\mathrm{d}\sigma} u(X_{t-r}) \right] = \mathsf{E}_x \left[\mathrm{e}^{-\int_0^{t-r} V_-(X_\sigma\circ\theta_r)\mathrm{d}\sigma} u(X_t) \mid \mathfrak{F}_r \right] =$$

$$= \mathsf{E}_x \left[\mathrm{e}^{-\int_r^t V_-(X_\sigma)\mathrm{d}\sigma} u(X_t) \mid \mathfrak{F}_r \right].$$

As the function $\mathrm{e}^{-\int_0^r V_-(X_\rho)\mathrm{d}\rho} V_-(X_r) : \Omega \to \mathbb{R}$ is evidently $\mathfrak{F}_r$-measurable, we get (using the property (4.4) of conditional expectations)

$$(S_r u)(x) = \mathsf{E}_x \left\{ \mathsf{E}_x \left(V_-(X_r) \mathrm{e}^{-\int_0^t V_-(X_\sigma)\mathrm{d}\sigma} u(X_t) \mid \mathfrak{F}_r \right) \right\}.$$

We use now the property (4.3) and Proposition 4.2 taking $F := V_-(X_r)$ in order to get

$$(S_r u)(x) = \mathsf{E}_x \left\{ V_-(X_r) \mathrm{e}^{-\int_0^t V_-(X_\sigma)\mathrm{d}\sigma} u(X_t) \right\} =$$

$$= \int_{\mathbb{R}^d} \mathrm{d}y \left\{ \int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega) V_-(X_r) \mathrm{e}^{-\int_0^t V_-(X_\sigma)\mathrm{d}\sigma} \right\} u(y).$$

In conclusion for any $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$ we have

$$K_r(x,y) = \int_\Omega \mu_{0,x}^{t,y}(\mathrm{d}\omega) V_-(X_r) \mathrm{e}^{-\int_0^t V_-(X_\sigma)\mathrm{d}\sigma}. \tag{5.19}$$

Using Proposition 4.3 we obtain

$$\int_{\mathbb{R}^d} \mathrm{d}x\, K_r(x,x) \leq \int_{\mathbb{R}^d} \mathrm{d}x \left[\int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega) V_-(\omega(r)) \right] =$$

$$\int_{\mathbb{R}^d} \mathrm{d}x \left[\int_\Omega \mu_{0,0}^{t,x}(\mathrm{d}\omega) V_-(x+\omega(r)) \right] = \overset{\circ}{\wp}_t(0) \int_{\mathbb{R}^d} \mathrm{d}x\, V_-(x) < \infty, \quad \forall t > 0.$$

Thus, for any $r \in [0,t]$ the operator S_r is trace class. Moreover, due to the properties of the trace we have $\mathrm{Tr}S_r = \mathrm{Tr}S_0$, $\forall r \in [0,t]$. We have:

$$\mathrm{Tr}S_0 = \frac{1}{t}\int_0^t \mathrm{d}r\, (\mathrm{Tr}S_0) = \frac{1}{t}\int_0^t \mathrm{d}r\, (\mathrm{Tr}S_r) = \frac{1}{t}\int_0^t \mathrm{d}r \left[\int_{\mathbb{R}^d} \mathrm{d}x\, K_r(x,x) \right] =$$

$$= \frac{1}{t}\int_{\mathbb{R}^d} \mathrm{d}x \left[\int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega) \tilde{g}_1 \left(\int_0^t \mathrm{d}s\, V_-(X_s) \right) \right] = \int_{\mathbb{R}^d} \mathrm{d}x \Psi_1(t,x)$$

In particular, for any $t > 0$, $\Psi_1(t;\cdot)$ is integrable on $\mathbb{R}^d$.

On the other hand

$$\mathrm{Tr}S_0 = \int_{\mathbb{R}^d} K_0(x,x)\mathrm{d}x = \int_{\mathbb{R}^d} \mathrm{d}x\, V_-(x) \int_\Omega \mu_{0,x}^{t,x}(\mathrm{d}\omega) \mathrm{e}^{-\int_0^t \mathrm{d}\rho\, V_-(X_\rho)}$$

$$= \int_{\mathbb{R}^d} \mathrm{d}x\, V_-(x) G_1(t;x,x) = \int_{\mathbb{R}^d} \mathrm{d}x\, D_1(t;x).$$

One uses the approximation properties contained in Lemmas 5.8 and 5.9 as well as the Monotone Convergence Theorem. □

Proof. of Theorem 1.1 *for* $B = 0$

We can assume $V_+ = 0$ and $V_- \in C_0^\infty(\mathbb{R}^d)$. Lemma 5.7 implies that for any $\alpha > 0$ one has

$$N_{-\alpha}(-V_-) \leq F_\infty(1)^{-1} \mathrm{Tr}\left[F_\infty(K_\alpha) \right].$$

Using (5.15), (5.16), we obtain

$$\mathrm{Tr}\left[F_\infty(K_\alpha) \right] = \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t} \int_{\mathbb{R}^d} \mathrm{d}x\, D_\infty(t;x) =$$

$$= \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t} \int_{\mathbb{R}^d} \mathrm{d}x\, \Psi_\infty(t;x). \tag{5.20}$$

Inequality (6.1) for $B = 0$ follows from (5.20) and Lemma 5.10. In addition $C_d = F_\infty(1)^{-1}\overline{C}_d$. □

6. Proof of the bounds in the magnetic case

Proof. of Theorem 1.1 *for* $B \neq 0$.

Analogously to Section 5, we can assume $V_+ = 0$ and $V_- \in C_0^\infty(\mathbb{R}^d)$. For $\alpha > 0$ one sets $K_\alpha(A) := V_-^{1/2}(H_A + \alpha)^{-1}V_-^{1/2}$. By inequality (3.4) for $r = 1$ and also using Pitt's Theorem [45], $K_\alpha(A)$ is a positive compact operator, and the same can be said about $F_\infty[K_\alpha(A)]$. We show that $F_\infty[K_\alpha(A)] \in B_1$ and we estimate the trace-norm. Repeating the arguments from the beginning of the proof of Proposition 5.1,

$$F_\lambda[K_\alpha(A)] = V_-^{1/2} \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t}\mathrm{e}^{-t(H_A+\lambda V_-)}V_-^{1/2}. \tag{6.1}$$

By using Proposition 4.1, we get for any $u \in C_0(\mathbb{R}^d)$, $u \geq 0$

$$[F_\lambda[K_\alpha(A)]\, u](x) = \tag{6.2}$$

$$= V_-^{1/2}(x) \int_0^\infty \mathrm{d}t\, \mathrm{e}^{-\alpha t} E_x\left[u(X_t)V_-^{1/2}(X_t)\mathrm{e}^{-\mathrm{i}S_A(t,X)} g_\lambda\left(\int_0^t \mathrm{d}s\, V_-(X_s)\right)\right].$$

Approximating g_∞ by means of Lemmas 5.8 and 5.9 and using the Monotone Convergence Theorem, we see that (6.2) also holds for the pair (g_∞, F_∞). The next inequality follows:

$$|F_\infty[K_\alpha(A)]\, u| \leq F_\infty(K_\alpha)|u|, \quad \forall u \in L^2(\mathbb{R}^d). \tag{6.3}$$

By Lemma 15.11 from [48], we have $F_\infty[K_\alpha(A)] \in B_1$ and

$$\mathrm{Tr}\,(F_\infty[K_\alpha(A)]) \leq \mathrm{Tr}\,(F_\infty[K_\alpha]). \tag{6.4}$$

Denoting by $N_{-\alpha}(B, -V_-)$ the number of eigenvalues of $H_A - V_-$ strictly less than $-\alpha$, analogously to Lemmas 5.6 and 5.7, we deduce that

$$N_{-\alpha}(B, -V_-) \leq F_\infty(1)^{-1}\mathrm{Tr}\,(F_\infty[K_\alpha]). \tag{6.5}$$

Inequality (6.1) follows from (6.5) by using the estimations at the end of Section 5. The constant C_d is the same as for the case $B = 0$. □

Proof. of Corollary 1.1. The idea of the proof is standard (cf. [48] for instance), but one has to use parts of the arguments from the proof of Theorem 1.1 in the case $B = 0$.

1. We show that it is enough to treat the case $V_+ = 0$.

We denote by N (resp. N_-) the number of strictly negative eigenvalues of $H_A \dot{+} V$ (resp. $H_A \dot{+} (-V_-)$). We have $N, N_- \in [0, \infty]$ and the min-max principle shows that $N \leq N_-$. In addition, if $H_A \dot{+} V$ has strictly negative eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots$, then $H_A \dot{+} (-V_-)$ has strictly negative eigenvalues $\lambda_1^- \leq \lambda_2^- \leq \ldots$ and $\lambda_j^- \leq \lambda_j$, $j \geq 1$. Therefore, one has $\sum_{j\geq 1} |\lambda_j|^k \leq \sum_{j\geq 1} |\lambda_j^-|^k$.

2. We show that treating compactly supported V_- is enough (remark that this property implies that $V_- \in L^p(\mathbb{R}^d)$ for any $p \in [1, d+k]$).

We take into account the approximation sequence defined in Lemma 5.3. The sequence of forms $(\mathfrak{h}^n)_{n\geq 1}$ satisfies the hypothesis of Theorem 3.11, Ch. VIII from [29]. If we denote by $\lambda_1 \leq \lambda_2 \leq \ldots$ the strictly negative eigenvalues of $H_A \dot{+} V$ and by $\lambda_1^{(n)} \leq \lambda_2^{(n)} \leq \ldots$ the strictly negative eigenvalues of $H^{(n)} := H_A \dot{+} V^{(n)}$, once again by Theorem 3.15, Ch. VIII from [29], we have $\lambda_j^{(n)} \geq \lambda_j$, $\forall j, n \in \mathbb{N}^*$ and $\lambda_j^{(n)}$ converges to λ_j. So it will be sufficient to prove (6.1) for the operators $H^{(n)}$.

3. We assume from now on that $V = -V_-$, $V_- \in L^{d+k}(\mathbb{R}^d)$ $(k > 0)$ and that $\mathrm{supp}(V_-)$ is compact. Let $\beta_0 > 0$ and for $\beta \in (0, \beta_0]$ let

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N_{-\beta}} < -\beta$$

be the eigenvalues of $H = H_A \dot{+} (-V_-)$ strictly smaller than $-\beta$ and let

$$\overline{\lambda}_1 \leq \overline{\lambda}_2 \leq \cdots \leq \overline{\lambda}_{M(\beta)} < -\beta$$

be the distinct eigenvalues with m_j the multiplicity of $\overline{\lambda}_j$, $1 \leq j \leq M(\beta)$. We have $N_{-\alpha} := N_{-\alpha}(B, -V_-)$. Using the definition of the Stieltjes integral and integration by parts, we get

$$\sum_{j=1}^{N_{-\beta}} |\lambda_j|^k = \sum_{j=1}^{M(\beta)} m_j |\overline{\lambda}_j|^k = \sum_{j=1}^{M(\beta)} |\overline{\lambda}_j|^k \left(N_{\overline{\lambda}_{j+1}} - N_{\overline{\lambda}_j} \right) = \int_{\lambda_1}^{-\beta} |\lambda|^k \mathrm{d}N_\lambda =$$

$$= |\beta|^k N_{-\beta} + k \int_{\lambda_1}^{-\beta} |\lambda|^{k-1} N_\lambda \, \mathrm{d}\lambda. \tag{6.6}$$

We denote by I the last integral and use (6.5) and (5.20) and the arguments in the proof of Lemma 5.10 to estimate I:

$$I = \int_\beta^{-\lambda_1} \alpha^{k-1} N_{-\alpha} \mathrm{d}\alpha = [F_\infty(1)]^{-1} \int_\beta^{-\lambda_1} \alpha^{k-1} \mathrm{Tr} F_\infty(K_\alpha) \mathrm{d}\alpha =$$

$$= [F_\infty(1)]^{-1} \int_{\mathbb{R}^d} \mathrm{d}x \int_0^\infty \mathrm{d}t\, \Psi_\infty(t,x) \int_\beta^{-\lambda_1} \mathrm{d}\alpha\, \alpha^{k-1} \mathrm{e}^{-\alpha t} \leq$$

$$\leq [F_\infty(1)]^{-1} \int_{\mathbb{R}^d} \mathrm{d}x \int_0^\infty \mathrm{d}t\, t^{-1} \mathring{\wp}_t(0) \tilde{g}_\infty(tV_-(x)) \int_\beta^{-\lambda_1} \mathrm{d}\alpha\, \alpha^{k-1} \mathrm{e}^{-\alpha t} \leq$$

$$\leq C\, [F_\infty(1)]^{-1} \int_{\mathbb{R}^d} \mathrm{d}x \int_0^\infty \mathrm{d}t \left(t^{-d-1} + t^{-d/2-1} \right) \tilde{g}_\infty(tV_-(x)) \int_\beta^{-\lambda_1} \mathrm{d}\alpha\, \alpha^{k-1} \mathrm{e}^{-\alpha t}.$$

The α integral may be bounded by

$$\int_0^\infty \mathrm{d}\alpha\, \alpha^{k-1} \mathrm{e}^{-\alpha t} = t^{-k} \int_0^\infty \mathrm{d}s\, s^{k-1} \mathrm{e}^{-s} \leq C t^{-k}.$$

Recalling that $\tilde{g}_\infty(t) = 0$ for $t \leq 1$ and $\tilde{g}_\infty(t) = t - 1$ for $t > 1$, we get that $\tilde{g}_\infty(tV_-(x)) = 0$ for $V_-(x) = 0$ and for $V_-(x) > 0$

$$\int_0^\infty \mathrm{d}t\, t^{-k} \left(t^{-d-1} + t^{-d/2-1} \right) \tilde{g}_\infty(tV_-(x)) =$$

$$= [V_-(x)]^{d+k} \int_1^\infty s^{-d-k-1}(s-1)\mathrm{d}s + [V_-(x)]^{d/2+k} \int_1^\infty s^{-d/2-k-1}(s-1)\mathrm{d}s,$$

the integrals being convergent for $d \geq 2$.

Using these estimations in (6.6) we conclude that

$$\sum_{j=1}^{N_{-\beta}} \left(|\lambda_j|^k - |\beta|^k \right) \leq C \left\{ \int_{\mathbb{R}^d} [V_-(x)]^{d+k}\, \mathrm{d}x + \int_{\mathbb{R}^d} [V_-(x)]^{d/2+k}\, \mathrm{d}x \right\},$$

thus

$$\sum_{j=1}^{N_{-(\beta_0)}} \left(|\lambda_j|^k - |\beta|^k \right) \leq C \left\{ \int_{\mathbb{R}^d} [V_-(x)]^{d+k}\, \mathrm{d}x + \int_{\mathbb{R}^d} [V_-(x)]^{d/2+k}\, \mathrm{d}x \right\},$$

with the constant C not depending on β or β_0. Taking the limit $\beta \searrow 0$ ends the proof. □

References

[1] AVRON, J., HERBST, I., SIMON, B., *Schrödinger operators with magnetic fields. I General interactions*, Duke Math. J. **45**, *4* (1978), pp. 847–883.

[2] CARMONA, R., MASTERS, W.C., SIMON, B., *Relativistic Schrödinger operators: Asymptotic behaviour of eigenfunctions*, Journal of Functional Analysis **91** (1990), pp. 117–143.

[3] CYCON, H.L., FROESE, R.G., KIRSCH, W., SIMON, B., *Schrödinger Operators with Applications to Quantum Mechanics and Global Geometry*, Springer, Berlin, 1987.

[4] CWICKEL, M., *Weak type estimates for singular values and the number of bound states of Schrödinger operators*, Ann. Math. **206** (1977), pp. 93–100.

[5] DAUBECHIES, I., *An uncertainty principle for fermions with generalized kinetic energy*, Commun. Math. Phys. **90** (1983), pp. 511–520.

[6] DEMUTH, M., VAN CASTEREN, J.A., *Stochastic spectral theory for self-adjoint Feller operators*, Birkhäuser, 2000.

[7] DIMASSI, M., RAIKOV, G., *Spectral asymptotics for quantum Hamiltonians in strong magnetic fields*, Cubo Mat. Educ. **3** (2001), pp. 317–391.

[8] FRANK, R.L., LIEB, E.H., SEIRINGER, R., *Hardy-Lieb-Thirring inequalities for fractional Schrödinger operators*, J. Amer. Math. Soc. **21** (2008), pp. 925–950.

[9] GÉRARD, C., MARTINEZ, A. SJÖSTRAND, J., *A mathematical approach to the effective Hamiltonian in perturbed periodic problems*, Commun. Math. Phys. **142** (1991), pp. 217–244.

[10] HÖRMANDER, L., *The Weyl calculus of pseudo-differential operators*, Comm. Pure Appl. Math. **32** (1979), pp. 359–443.

[11] HÖRMANDER, L., *The Analysis of Linear Partial Differential Operators, III*, Springer-Verlag, New York, 1985.

[12] HÖRMANDER, L., *The Analysis of Linear Partial Differential Operators, IV*, Springer-Verlag, New York, 1985.

[13] HELFFER, B., SJÖSTRAND, J., *Equation de Schrödinger avec champ magnetique et équation de Harper*, in Springer Lecture Notes in Physics, **345** (1989), pp. 118–197.

[14] HELFFER, B., SJÖSTRAND, J., *On diamagnetism and de Haas-van Alphen effect*, Ann. I.H.P. **52** (1990), pp. 303–375.

[15] HEMPEL, R., HERBST, I., *Strong magnetic fields, Dirichlet boundaries, and spectral gaps*, Comm. Math. Phys. **169** (1995), pp. 237–259.

[16] ICHINOSE, I., *The nonrelativistic limit problem for a relativistic spinless particle in an electromagnetic field*, J. Funct. Anal. **73**, *2* (1987), pp. 233–257.

[17] ICHINOSE, I., *Essential selfadjointness of the Weyl quantized relativistic Hamiltonian*, Ann. Inst. H. Poincaré Phys. Théor. **51**, *3* (1989), pp. 265–297.

[18] ICHINOSE, T., ICHINOSE, W., *On the essential self-adjointness of the relativistic Hamiltonian with a negative scalar potential*, Rev. Math. Phys. **7**, *5* (1995), pp. 709–721.

[19] ICHINOSE, T., TAMURA, H., *Path integral for the Weyl quantized relativistic Hamiltonian*, Proc. Japan Acad. Ser. A Math. Sci. **62**, *3* (1986), pp. 91–93.

[20] ICHINOSE, T., TAMURA, H., *Imaginary-time path integral for a relativistic spinless particle in an electromagnetic field*, Comm. Math. Phys. **105**, *2* (1986), pp. 239–257.

[21] ICHINOSE, T., TSUCHIDA, T., *On Kato's inequality for the Weyl quantized relativistic Hamiltonian*, Manuscripta Math. **76**, *3-4* (1992), pp. 269–280.

[22] ICHINOSE, T., TSUCHIDA, T., *On essential selfadjointness of the Weyl quantized relativistic Hamiltonian*, Forum Math. **5**, *6* (1993), pp. 539–559.

[23] IFTIMIE, V., *Uniqueness and existence of the integrated density of states for Schrödinger operators with magnetic field and electric potential with singular negative part*, Publ. Res. Inst. Math. Sci. **41** (2005), pp. 307–327.

[24] IFTIMIE, V., MĂNTOIU, M. PURICE, R., Magnetic pseudodifferential operators, Publ. Res. Inst. Math. Sci. **43** (2007), pp. 585–623.

[25] IKEDA W., WATANABE, S., *Stochastic differential equations and diffusion processes*, North-Holland, 1981.

[26] JACOB, N., *Pseudodifferential operators and Markov processes. III Markov processes and applications*, World Scientific, 2005.

[27] KARASEV, M.V., OSBORN, T.A., *Symplectic areas, quantization and dynamics in electromagnetic fields*, J. Math. Phys. **43**, *2* (2002), pp. 756–788.

[28] KARASEV, M.V., OSBORN, T.A., *Quantum magnetic algebra and magnetic curvature*, J. Phys. A **37**, *6* (2004), pp. 2345–2363.

[29] KATO, T., *Perturbation theory for linear operators*, Springer, 1976.

[30] KATO, T., MASUDA, K., *Trotter's product formula for nonlinear semigroups generated by the subdifferentials of convex functionals*, Journal of the Mathematical Society of Japan **30** (1978), pp. 169–178.

[31] LI, P., YAU, S.T., *On the Schrödinger equation and the eigenvalue problem*, Comm. Math. Phys. **88** (1983), pp. 309–318.

[32] LIEB, E.: *Bounds on the eigenvalues of the Laplace and Schrödinger operators*, Bull. Amer. Math. Soc. **82**, *5* (1976), pp. 751–753.

[33] LIEB, E., THIRRING, W., Bounds for the kinetic energy of fermions which proves the stability of matter, Phys. Rev. Lett. **35** (1975), pp. 687–689.

[34] MÜLLER, M., *Product rule for gauge invariant Weyl symbols and its applications to the semiclassical description of guiding center motion*, J. Phys. A **32** (1999), pp. 1035–1052.

[35] MĂNTOIU, M., PURICE, R., *The algebra of observables in a magnetic field*, Mathematical Results in Quantum Mechanics (Taxco, 2001), Contemporary Mathematics **307**, Amer. Math. Soc., Providence, RI (2002), pp. 239–245.

[36] MĂNTOIU, M., PURICE, R., *The Magnetic Weyl calculus*, J. Math. Phys. **45** (2004), pp. 1394–1417.

[37] MĂNTOIU, M., PURICE, R., *Strict deformation quantization for a particle in a magnetic field*, J. Math. Phys. **46**, (2005), 15 pp.

[38] MĂNTOIU, M., PURICE, R., *The mathematical formalism of a particle in a magnetic field*, in *Mathematical physics of quantum mechanics*, pp. 417–434, Lecture Notes in Phys., **690**, Springer, 2006.

[39] MĂNTOIU, M., PURICE, R., RICHARD, S., *Twisted crossed products and magnetic pseudodifferential operators*, in *Advances in operator algebras and mathematical physics*, pp. 137–172, Theta Ser. Adv. Math., **5**, Theta, 2005.

[40] MĂNTOIU, M., PURICE, R., RICHARD, S., *Spectral and propagation results for Schrödinger magnetic operators*, J. Funct. Anal. **250** (2007), pp. 42–67. (2007).

[41] MELGAARD, M., ROZENBLUM, G.V., *Spectral estimates for magnetic operators*, Math. Scand. **79** (1996), pp. 237–254.

[42] NAGASE, M., UMEDA, T., *Weyl quantized Hamiltonians of relativistic spinless particles in magnetic fields*, J. Funct. Anal. **92** (1990), pp. 136–164.

[43] NAGASE, M., UMEDA, T., *Spectra of relativistic Schrödinger operatores with magnetic vector potentials*, Osaka J. Math. **30** (1993), pp. 839–853.

[44] PASCU, M., *On the essential spectrum of the relativistic magnetic Schrödinger operator*, Osaka J. Math. **39**, *4* (2002), pp. 963–978.

[45] PITT, L.D., *A compactness condition for linear operators on function spaces*, Journal of Operator Theory **1** (1979), 49–54.

[46] ROZENBLUM, G., *Distribution of the discrete spectrum of singular differential operators*, Izvestia Vuz, Mathematika, **20**, *2* (1976), pp. 75–86.

[47] REED, M., SIMON, B., *Methods of modern mathematical physics, I–IV*, Academic Press, 1972–1979.

[48] SIMON, B., *Functional integration and quantum physics*, Academic Press, 1979.

[49] SIMON, B., *Kato's inequality and the comparaison of semigroups*, J. Funct. Anal. **32** (1979), pp. 97–101.

[50] SIMON, B., *Maximal and minimal Schrödinger forms*, J. Oper. Th. **32** (1979), pp. 37–47.

[51] TRIEBEL, H., Interpolation theory, function spaces, differential operators, VFB Deutscher Verlag der Wiessenschaften, Berlin, 1978.

[52] UMEDA, T., *Absolutely continuous spectra of relativistic Schrödinger operators with magnetic vector potentials*, Proc. Japan Acad. **70** (1994), Ser. A, pp. 290–291.

Topics in Applied Mathematics & Mathematical Physics

Approximate inertial manifolds, induced trajectories, and approximate solutions for semilinear parabolic equations, based upon these; applications to flow and diffusion problems

by Anca-Veronica Ion [1]

Contents

[1]"Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: averionro@yahoo.com.
The paper was supported by CEEX Grant CEEX05-D11-25/2005.

1. Introduction

In the study of dissipative semi-dynamical systems generated by semilinear parabolic equations, the theory of qualitative behavior of the system at large times plays an important role. By parabolic semilinear equations we mean partial differential equations that can be written as abstract equations in a Hilbert space, of the form:

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nu Au + R(u) = f, \tag{1.1}$$

where u is a function of time with values in a Hilbert space H (whose definition comprises the boundary value conditions imposed to equation (1.1)). We attach to the above equation an initial condition

$$u(0) = u_0, \tag{1.2}$$

with u_0 in H. We assume that A is a linear operator, defined on a dense subspace $D(A)$ of H, self-adjoint, positive definite, with compact inverse, while R is a nonlinear operator defined on $D(R) \subset D(A)$. We do not insist here on the hypotheses on R, but we assume that it is such that the Cauchy problem (1)-(2) has an unique solution on $[0,T]$, for every $u_0 \in H$ and every $T > 0$. Hence a semi-dynamical system is generated by the above problem, by setting $S(t)u_0 = u(t,u_0)$, where $u(t,u_0)$ is the solution of (1.1)–(1.2).

For this presentation we assume that f is in H. We also assume that the semi-dynamical system generated by (1.1) is dissipative in the sense that there is a bounded absorbing set for it. An absorbing set is a set B having the property that, for every bounded set $M \subset H$, there is a value of t, depending on M, let us denote it by t_M, with the property that, $S(t)M \subset B$ for $t \geq t_M$. For the particular problems we consider here, there also are absorbing balls in some subspace V of H, with $D(A) \subset V \subset H$.

In the theory of qualitative behavior at large times of solutions of equations of the form (1.1), the notion of global attractor plays an important role. A global attractor [3] is a compact set of the phase space H, invariant to the semigroup $S(t)_{t\geq 0}$, that attracts the bounded sets of the phase space, when time tends to infinity. This means that the global attractor bears in its structure the properties of the behavior of the semi-dynamical system at large times. For many problems of interest the existence of an attractor was proved [37].

The study of the geometrical and topological properties of the global attractors flourished since the last two decades of the XXth century and the major

hope was that a connection between the structure of the attractor and very complex phenomena like turbulence in the flow of the fluids will be found.

In this context, another interesting notion appeared, that of *inertial manifold* (i.m.) [11]. It is a finite dimensional, invariant and at least Lipschitz manifold having the property that it exponentially attracts all the trajectories of the problem. More than that, an i.m. has the property of *asymptotic completeness* meaning that for every u_0 in H there is a v_0 on the i.m. such that the distance between the trajectories passing through the two points decreases exponentially with time.

The invariance of the i.m. implies the fact that we can construct a restriction of the problem to this manifold. The restricted problem is named *inertial form* [11], [37] and, since the i.m. is finite dimensional, is equivalent with a system of ODEs. The above defined asymptotic completeness of the i.m. implies that the asymptotic behavior at large times of the dynamical system is described by the asymptotic behavior of the inertial form. Hence the large times study of the initial semi-dynamical system (infinite dimensional since its phase space is H) can be reduced to that of a finite-dimensional one.

Another important consequence of the properties of the i.m.s is that, when a global attractor exists, it is contained in the i.m. These considerations explain the large interest shown by the scientific community in inertial manifolds. From among the great number of papers devoted to the inertial manifolds we remind: [11] (with the extended version [12]), [8], [9], [5], [36]. The important monograph [37] had a second edition in 1997.

From a theoretical point of view, the i.m.s looked very promising, but major obstacles appeared in trying to use their properties in the study of concrete problems. One is due to the fact that existence of i.m.s is in most papers proved by a fixed point theorem, and is not constructive. There is a constructive proof in [2] but it uses some integral manifolds whose construction is equivalent with solving the equation. Another problem is a restrictive hypothesis among the hypothesis of the existence theorems- the hypothesis of a spectral gap that imposes the existence of two successive eigenvalues of A situated at a "large enough" distance [1], [12], [37]. This hypothesis is not fulfilled by many problems, (e.g. is not fulfilled for the two-dimensional Navier-Stokes equations).

In this situation the approximate inertial manifolds were defined as approximations of i.m.s or as substitutes of these, when the i.m.s could not be proved to exist. An approximate inertial manifold (a.i.m.) is a finite dimensional, at least Lipschitz manifold in the space H, with the property that all the trajectories of the dynamical system enter a narrow neighborhood of the manifold

at a certain moment and never leave the neighborhood after. Even if it has not the invariance property, an a.i.m. is important because, if the problem has a global attractor, it is contained in the narrow neighborhood mentioned above.

The localization of the attractors in the space of phases was a first interesting application field of the a.i.m.s. Besides this, a.i.m.s found very interesting applications in the construction of some approximate solutions (the numerical integration) of the nonlinear evolution problems. Examples of papers devoted to a.i.m.s are: [10], [13], [23], [26], [27], [28], [33], [35], [37], [38], [39].

In Section 2 we present some methods, that use a.i.m.s, for the construction of approximate solutions for problems of the type (1.1)–(1.2), the so-called *non-linear Galerkin method* and *post-processed Galerkin method.*

We include a method conceived by us, that we named *repeatedly adjusted and post-processed Galerkin method*, that is connected to the preceding methods but brings some simplifications to these. In Section 3 we present the way these method work for the two-dimensional Navier-Stokes equations with periodic boundary conditions, and in Section 4, for a two-dimensional reaction-diffusion equation, with Von Neumann boundary conditions.

In order to settle the notations and the functional framework of our presentation, we shortly remind below the Galerkin spectral method for the abstract equation (1.1).

1.1. The Galerkin method

In the hypotheses we assumed on the operator A of equation (1.1), it follows that A has positive eigenvalues that form a tending to infinity sequence:

$$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n \leq ..., \ \lambda_n \underset{n\to\infty}{\to} \infty.$$

The eigenfunctions of A form a total (orthonormal) system for H. We consider the set, denoted Γ_m, of the first distinct m eigenvalues (in increasing order) and the eigenfunctions corresponding to these. We denote by P the orthogonal projection operator on the subspace spanned by these eigenfunctions and we set $Q = I - P$ (where I is the identity application on H). The solution u of (1.1)–(1.2) is projected by the two projectors and we set

$$p = Pu,$$
$$q = Qu.$$

It follows that the functions p and q are solutions of

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p+q) = Pf, \tag{1.3}$$

$$\frac{\mathrm{d}q}{\mathrm{d}t} + \nu Aq + QR(p+q) = Qf, \tag{1.4}$$

$$p(0) = Pu_0, \tag{1.5}$$

$$q(0) = Qu_0. \tag{1.6}$$

Usually, the component q of the solution is proved to be, at large times, "little" in the norm of H compared to the p component. That is, an inequality of the form

$$|q\,(t)| \leq C_0 \delta^a, \tag{1.7}$$

where

$$\delta = \frac{\lambda_1}{\lambda_{m+1}}, \tag{1.8}$$

and a is some positive number, is true. For the Navier-Stokes equations it is proved in [38] that a inequality of the type (1.7) holds, with $a = 1$ and C_0 depending on m. We proved in [19] that the inequality can be improved in the sense that it is true with a C_0 that does not depend on m. For the reaction-diffusion equation, $|q(t)|$ is of the order of δ for large enough times [4].

If in the equation (1.4) q is neglected in the presence of p, we find the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p) = Pf. \tag{1.9}$$

This is the Galerkin approximation of the equation (1.1). The solution of the problem (1.9) with the initial condition (1.5), that we denote by $p_G(.)$, is the Galerkin approximation of the solution of (1.1)-(1.2). For several problems it is proved in the literature that inequalities of the type

$$|u(t) - p_G(t)| \leq C\delta^\alpha,$$

where $u(t)$ is the solution of the problem (1.1)-(1.2), $\delta > 0$ is defined by (1.8), and $\alpha > 0$.

As example, for a reaction-diffusion equation with Neumann boundary values and for the two-dimensional Navier-Stokes equations, $\alpha = 1$ (in the hypothesis $f \in H$). The problem (1.9), (1.5) is equivalent to a system of ordinary differential equations for the coordinates of $p(t)$ along the eigenfunctions that span PH. The definition of δ shows that the greater will be m, (hence the dimension of PH), the smaller will be the error.

In the construction of the Galerkin equation, the q component of the solution (that is proved to be small for large times) is approximated with 0. The nonlinear Galerkin (and/or post-processed) methods of approximation are based upon the idea of approximating $q(t)$ by using a a.i.m instead of the manifold $\mathbf{q}_0$.

2. Modified Galerkin methods

The nonlinear Galerkin (and/or post-processed) methods of approximation are based upon the idea of approximating $q(t)$ by using an a.i.m instead of taking $q \simeq 0$.

2.1. Families of a.i.m.s used in the modified Galerkin methods

There are several types of a.i.m.s defined in the literature. Among them, those defined in [10], [38], [39] (for the Navier-Stokes equations – NSE) generated new numerical integration methods, based on the Galerkin method. They form a family $\{\mathcal{M}_j\}_{j\geq 0}$ and are the graphs of some functions $\Phi_j : P\mathcal{H} \to Q\mathcal{H}$. The definitions of these a.i.m.s for the NSE are presented in Section 3 while those for the RDE are given in Section 4. A.i.m.s of the type of those cited above may be (and were) defined for many particular problems of the form (1.1)–(1.2). The main property of these a.i.m.s, on which their use in the construction of the numerical methods is based, is the following: the distance (in the norm of H) between *q(t)* and the image of *p(t)* on the a.i.m. $\mathcal{M}_n$ is of the order of $\delta^{a(n)}$ that is

$$|q(t) - \Phi_n(p(t))| \leq C\delta^{a(n)}, \tag{2.1}$$

where $a(n)$ is increasing with n.

For example, for the two-dimensional NSE it is proved [38], [39] that $a(n) = (n+3)/2$. Since, for NSE, about the H norm of $q(t)$ only the fact of being of the order of δ is known, it is clear that any of the above a.i.m.s provides a better approximation of $q(t)$ than the so-called plane manifold $q = 0$, for the mentioned problem.

2.2. The nonlinear Galerkin methods

The nonlinear Galerkin method (NL Galerkin method) was first defined in [29]. The method relies on the idea that $\Phi_0(p(t))$ is a better approximation of $q(t)$ than 0, and considers, instead of the Galerkin equation (3.25), the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p + \Phi_0(p)) = Pf, \tag{2.2}$$

with initial condition (1.5). By denoting with $\widetilde{p}_0(.)$ the solution of this problem, the approximate solution of (1.1)–(1.2) is taken as

$$v_0(t) = \widetilde{p}_0(t) + \Phi_0(\widetilde{p}_0(t)).$$

As it is natural, since $\Phi_n(p(t))$ approximates $q(t)$ better and better with the increase of n, the next idea, appeared in [6], was to consider the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p + \Phi_n(p)) = Pf, \tag{2.3}$$

with the initial condition (1.5). Let $\widetilde{p}_n(.)$ the solution of this problem. The approximate solution is then defined as

$$v_n(t) = \widetilde{p}_n(t) + \Phi_n(\widetilde{p}_n(t)).$$

For the problems considered in the context of nonlinear Galerkin problems, it is proved that the error is of the order of $\delta^{b(n)}$, where $b(n)$ is increasing with n.

E.g., for the Navier-Stokes equations it is proved in [7] that $b(n) = (n+3)/2$, while for the reaction-diffusion equation it is asserted in [32] that $b(n) = n+2$ provided $f \in H$.

2.3. Post-processed Galerkin methods

In [14] the following modified Galerkin method is proposed, that also uses a.i.m.s. Let again $p_G(.)$ be the solution of (1.9), (1.5). Then the value of $\Phi_0(p_G(t))$ is computed at the right end side of the time interval $[0, T]$, that is in T. The approximate solution in T is defined as

$$w(T) = p_G(T) + \Phi_0(p_G(T)).$$

This method is named *the post-processed Galerkin method (PP Galerkin method)* because the solution of the Galerkin problem is corrected only in the final phase, after finishing the numerical integration of the Galerkin problem, by using the first a.i.m. of the family described in 2.1 (hence post-processed). The error of this approximate solution is less than that of the Galerkin method. Thus, for the two-dimensional Navier-Stokes equations, it is shown in [14] to be of the order of $\delta^{5/4}$. Another estimate is proved in [15], i.e. the error is proved to be of the order of $L^2\delta^{3/2}$, where $L = 1 + \ln(2m^2)$. This latter estimate of the error is not necessary better than the former because of the coefficient L^2.

The next idea appeared in the literature [32] was to postprocess the NL Galerkin method of the preceding section. More precisely, the equation (2.3) is considered, it is integrated on all the time interval $[0, T]$, then $\Phi_{n+1}(\widetilde{p}_n(T))$, is computed, and the approximate solution in T is defined as

$$w_n(T) = \widetilde{p}_n(T) + \Phi_{n+1}(\widetilde{p}_n(T)).$$

This method is called the nonlinear post-processed Galerkin method (NL PP Galerkin method). In [32] the use of the method is exemplified on the reaction-diffusion equation and it is proved that, if $f \in H$, then the error is of the order of $\ln m\, \delta^{n+3}$.

2.4. A new modified Galerkin method

In [38], in the context of the study of the NSE, a family of functions, $\{q_j\}_{j\geq 0}$, $q_j : \mathbb{R}^+ \to Q\mathcal{H}$, having the property

$$|q_j(t) - q(t)| \leq k_j L^{1+j/2}\delta^{(3+j)/2} \tag{2.4}$$

for large enough times is constructed. Here the coefficients k_j depend on the data of the problem (ν, $|f|$, λ_1), and $L = 1 + \ln\frac{\lambda_{m+1}}{\lambda_1}$. Actually, the function q_0 is of the form

$$q_0 = \Phi_0(p),$$

while, for $j \geq 1$, q_j are recursively defined by relations of the type

$$q_j = F_j(Qf, p, q_0, ..., q_{j-1}). \tag{2.5}$$

The functions $u_j = p + q_j$, $j \geq 0$ define the so-called *induced trajectories*, $\{u_j(t); t \geq 0\}$, associated to the trajectory $\{u(t); t \geq 0\}$ of the dynamical system. Relation (2.4) shows that the functions u_j, $j \geq 0$, are approximations

of the exact solution, of increasing with j accuracy. The definition of the a.i.m.s $\mathcal{M}_j$ used in the nonlinear Galerkin methods for the NSE are based upon the definitions of the functions q_j.

In [20], for the two-dimensional NSE with periodic boundary conditions, we defined a new type of modified Galerkin method, that uses some approximations of the induced trajectories and not the a.i.m.s. We describe here the method in the general context of equation (1.1). The purpose of the method is that of working with a very low-dimensional projection space PH, and the idea from which we started is that, however small is the dimension of PH, if we have a very good approximation for q, let us denote it by $\widetilde{q}$, then a very good approximation for p will be obtained by solving the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + g(p + \widetilde{q}) = Pf.$$

In consequence, a good approximation of u may be obtained. The method is structured on several levels. One of the ideas we followed in developing this method is that of having to integrate only differential equations of the same level of difficulty as the Galerkin equation. This was possible by using approximations of induced trajectories instead of a.i.m.s.

Level 0. This level has two stages. The first is the classical Galerkin method, i.e. we solve the problem (1.9), (1.5) and we consider its solution, $p_G(.)$.

The second stage consists in defining the function of time, with values in $Q\mathcal{H}$:

$$\widetilde{q}_0(t) = \Phi_0(p_G(t)), \tag{2.6}$$

the function Φ_0 being the one that defines the first a.i.m. of the family cited in 2.1.

Then we define the approximate solution at this first level as

$$\widetilde{u}_0 = p_G + \widetilde{q}_0.$$

Since the function $\widetilde{q}_0(t)$ will be used at the second level of our method, in the numerical implementation of this method, the function $\widetilde{q}_0$ should be computed in each point of the time mesh, unlike in the post-processing defined in [14], where it is computed only at the final point of the integration interval $[0, T]$. Besides this, Level 0 of our method is essentially the Galerkin post-processed method.

Level 1. We consider the problem

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p + \widetilde{q}_0) = Pf, \tag{2.7}$$

$$p(0) = Pu_0$$

and we denote by $\widetilde{p}_0$ its solution. This is an "adjusted" Galerkin problem.

This equation is essentially different from the corresponding one of the NL Galerkin method (see equation (2.3)) since $\widetilde{q}_0$ is known from Level 1.

Then we define

$$\widetilde{q}_1(t) = F_1(Qf, \widetilde{p}_0(t), \widetilde{q}_0(t)).$$

The approximate solution is

$$\widetilde{u}_1 = \widetilde{p}_0 + \widetilde{q}_1.$$

Level $j > 1$.

We assume that $\widetilde{q}_0,\ \widetilde{q}_1,\ \widetilde{q}_2, ..., \widetilde{q}_{j-1}$ were constructed. The problem

$$\frac{\mathrm{d}p}{\mathrm{d}t} + \nu Ap + PR(p + \widetilde{q}_{j-1}) = Pf, \qquad (2.8)$$
$$p(0) = Pu_0,$$

is considered and its solution is denoted by $\widetilde{p}_{j-1}$. Then we denote

$$\widetilde{q}_j = F_j\left(Qf,\ \widetilde{p}_{j-1},\ \widetilde{q}_0,\ \widetilde{q}_1, ..., \widetilde{q}_{j-1}\right)$$

and the approximate solution is

$$\widetilde{u}_j = \widetilde{p}_{j-1} + \widetilde{q}_j.$$

At first sight, the idea of performing several time integrations seems a bad idea, since every such integration involves a large amount of computations. However, a careful analysis shows that the amount of computations involved in the NL Galerkin method (based upon the a.i.m. $\mathcal{M}_j$) is greater than that involved in solving the problems from Level 1 to the eq. (2.8) of Level j, inclusive. Such an analysis is performed for the Navier-Stokes equations in 3.8. Hence our method, that we call *the repeatedly adjusted and post-processed Galerkin method (R-APP Galerkin method)* is an alternative to the NL Galerkin method. The final post-processing, by adding $\widetilde{q}_j$ to $\widetilde{p}_{j-1}$ is equivalent to the post-processing of NL Galerkin method and does not imply a large amount of calculi since it will be performed only in some selected moments of time (eventually only at the last moment, T). In what concerns the error, for the problems discussed below we can prove that the error of R-APP Galerkin method is of the same order of magnitude as that for NL PP Galerkin method, for the two particular problems in Sections 3 and 4.

3. Modified Galerkin methods for the Navier-Stokes equation

We present here the modified Galerkin methods for the Navier-Stokes equations: the NL, NL PP Galerkin methods already defined in the literature and our R-APP Galerkin method.

3.1. The setting of the problem

We consider the problem of the two-dimensional flow of a incompressible Newtonian fluid, modeled by the Navier-Stokes equations. We impose periodic boundary conditions and choose the periodicity cell to be a square, $\Omega = (0,l) \times (0,l)$. Thus the problem is

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\, \mathbf{u} + \nabla p = \mathbf{f}, \tag{3.1}$$

$$\operatorname{div} \mathbf{u} = 0, \tag{3.2}$$

where $\mathbf{u}(t,\mathbf{x}) \in \mathbb{R}^2$ is the velocity of the fluid, $t \geq 0$, $\mathbf{x} \in \Omega$, $p(t,\mathbf{x}) \in \mathbb{R}$ is the pressure of the fluid, ν is the kinematic viscosity, and $\mathbf{f}$ is the volume force. We add the initial condition

$$\mathbf{u}(0,\cdot) = u_0(\cdot). \tag{3.3}$$

We assume that $\mathbf{f}$ is independent of time and is an element of $\left[L^2_{per}(\Omega)\right]^2$. As is usual in the study of the Navier-Stokes equations with periodic boundary conditions, we assume that [40], [34]

$$\overline{\mathbf{f}} = \frac{1}{l^2} \int_\Omega \mathbf{f}(\mathbf{x})\, \mathrm{d}\mathbf{x} = \mathbf{0}, \tag{3.4}$$

and that the pressure is a periodic function on Ω. For simplicity we will assume also that the average $\overline{\mathbf{u}}$ of the velocity over the periodicity cell is zero.

The velocity $\mathbf{u}$ is thus looked for in the space $\mathcal{H} = \left\{\mathbf{v};\ \mathbf{v} \in \left[L^2_{per}(\Omega)\right]^2,\ \operatorname{div} \mathbf{v} = \mathbf{0}, \overline{\mathbf{v}} = 0\right\}$ with the scalar product $(\mathbf{u},\mathbf{v}) = \int_\Omega (u_1 v_1 + u_2 v_2)\, \mathrm{d}\mathbf{x}$, (where $\mathbf{u} = (u_1,u_2)$, $\mathbf{v} = (v_1,v_2)$) and the induced norm is denoted by $|\cdot|$. Let us also consider the space $\mathcal{V} = \left\{\mathbf{u} \in \left[H^1_{per}(\Omega)\right]^2,\ \operatorname{div} \mathbf{u} = \mathbf{0}, \overline{\mathbf{u}} = 0\right\}$, with the scalar product $((\mathbf{u},\mathbf{v})) = \sum_{i,j=1}^2 \left(\frac{\partial u_i}{\partial x_j}, \frac{\partial v_i}{\partial x_j}\right)$, and the induced norm, denoted by $\|\cdot\|$.

The variational formulation of the Navier-Stokes equations [40] leads, for the periodic boundary conditions, to the Cauchy problem

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\,\mathbf{u} = \mathbf{f} \quad \text{in } \mathcal{V}', \tag{3.5}$$

$$\mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_0 \in \mathcal{H}. \tag{3.6}$$

The notations

$$\mathbf{B}(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\nabla)\mathbf{v}, \tag{3.7}$$

$$\mathbf{B}(\mathbf{u}) = \mathbf{B}(\mathbf{u},\mathbf{u}), \tag{3.8}$$

will be used below.

We remind here the classical existence and uniqueness results for the Navier-Stokes equations in $\mathbb{R}^2$, with periodic boundary conditions.

We denote $\mathbf{A} = -\boldsymbol{\Delta}$. The definition domain of the linear operator $\mathbf{A}$ is $D(\mathbf{A}) = \mathcal{V} \cap H^2_{per}(\Omega)$.

THEOREM **3.1** [40]. *a) If* $\mathbf{u}_0 \in \mathcal{H}$, $\mathbf{f} \in \mathcal{H}$, *then the problem* (3.5), (3.6) *has an unique solution* $\mathbf{u} \in C^0([0,T];\mathcal{H}) \cap L^2(0,T;\mathcal{V})$. *b) If, in addition,* $\mathbf{u}_0 \in \mathcal{V}$, *then* $\mathbf{u} \in C^0([0,T];\mathcal{V}) \cap L^2(0,T;D(\mathbf{A}))$. *The solution is, in this latter case, analytic in time on the positive real axis.*

The semi-dynamical system $\{S(t)\}_{t\geq 0}$ generated by problem (3.5) is dissipative [37]. More precisely, there is a $\rho_0 > 0$ such that for every $R > 0$, there is a $t_0(R) > 0$ with the property that for every $\mathbf{u}_0 \in \mathcal{H}$ with $|\mathbf{u}_0| \leq R$, we have $|S(t)\,\mathbf{u}_0| \leq \rho_0$ for $t > t_0(R)$. In addition, there are absorbing balls in $\mathcal{V}$ and $\mathbf{D}(\mathbf{A})$ for $\{S(t)\}_{t\geq 0}$, [34] i.e. there are $\rho_1 > 0$, $\rho_2 > 0$ and, for every $R > 0$, there are $t_1(R)$, $t_2(R)$ with $t_2(R) \geq t_1(R) \geq t_0(R)$ such that $|\mathbf{u}_0| \leq R$ implies $\|S(t)\,\mathbf{u}_0\| \leq \rho_1$ for $t > t_1(R)$ and $|\mathbf{A}S(t)\,\mathbf{u}_0| \leq \rho_2$ for $t > t_2(R)$.

3.2. The decomposition of the space, the projected equations

The eigenvalues of $\mathbf{A}$ are $\lambda_{j_1,j_2} = \frac{4\pi^2}{l^2}\left(j_1^2 + j_2^2\right)$, $(j_1,\ j_2) \in \mathbb{N}^2 \setminus \{(0,0)\}$, and the corresponding eigenfunctions are

$$\mathbf{w}^{s\pm}_{j_1,j_2} = \frac{\sqrt{2}}{l}\frac{(j_2, \mp j_1)}{|\mathbf{j}|}\sin\left(2\pi\frac{j_1x_1 \pm j_2x_2}{l}\right),$$

$$\mathbf{w}^{c\pm}_{j_1,j_2} = \frac{\sqrt{2}}{l}\frac{(j_2, \mp j_1)}{|\mathbf{j}|}\cos\left(2\pi\frac{j_1x_1 \pm j_2x_2}{l}\right),$$

where $|\mathbf{j}| = \left(j_1^2 + j_2^2\right)^{\frac{1}{2}}$ [38]. These eigenfunctions form a total system for $\mathcal{H}$. For a fixed $m \in \mathbb{N}$ we consider the set Γ_m of eigenvalues λ_{j_1,j_2} having $0 \leq j_1, j_2 \leq m$. We define

$$\lambda := \lambda_{1,0} = \lambda_{0,1} = \frac{4\pi^2}{l^2},$$
$$\Lambda := \lambda_{m+1,0} = \lambda_{0,m+1} = \frac{4\pi^2}{l^2}\left(m+1\right)^2,$$
$$\delta = \delta\left(m\right) := \frac{\lambda}{\Lambda} = \frac{1}{(m+1)^2}.$$

Λ is the least eigenvalue not belonging to Γ_m. The eigenfunctions corresponding to the eigenvalues of Γ_m span a finite-dimensional subspace of $\mathcal{H}$. We denote by $\mathbf{P}$ the orthogonal projection operator on this subspace and by $\mathbf{Q}$ the orthogonal projection operator on the complementary subspace. We write for the solution $\mathbf{u}$ of (3.5), (3.6), $\mathbf{u} = \mathbf{p} + \mathbf{q}$, where $\mathbf{p} = \mathbf{P}\mathbf{u}$, $\mathbf{q} = \mathbf{Q}\mathbf{u}$.

By projecting equation (3.5) on the above constructed spaces, we obtain

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{P}\mathbf{B}(\mathbf{p}+\mathbf{q}) = \mathbf{P}\mathbf{f}, \tag{3.9}$$
$$\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} - \nu\Delta\mathbf{q} + \mathbf{Q}\mathbf{B}(\mathbf{p}+\mathbf{q}) = \mathbf{Q}\mathbf{f}. \tag{3.10}$$

In [10] is proved that for every $R > 0$, there is a moment $t_3\left(R\right) \geq t_2(R)$ such that for every $|\mathbf{u}_0| \leq R$,

$$\begin{aligned} &|\mathbf{q}\left(t\right)| \leq K_0 L^{\frac{1}{2}}\delta, \quad \|\mathbf{q}\left(t\right)\| \leq K_1 L^{\frac{1}{2}}\delta^{\frac{1}{2}}, \\ &\left|\mathbf{q}'\left(t\right)\right| \leq K_0' L^{\frac{1}{2}}\delta, \quad |\Delta\mathbf{q}\left(t\right)| \leq K_2 L^{\frac{1}{2}}, \qquad t \geq t_3\left(R\right), \end{aligned} \tag{3.11}$$

where, for our choice of the set of eigenvalues Γ_m, $L = 1 + \ln(2m^2)$. In [19] we proved that estimates of the same order are true for the various norms of $\mathbf{q}\left(t\right)$ above, but with coefficients of the powers of δ not depending on m.

3.3. Induced trajectories for the Navier-Stokes problem

In [38] the notion of induced trajectory is defined and a family of induced trajectories is constructed for this problem. The asymptotic expansions that rely behind this construction are not made explicit there.

A family of functions, $\{\mathbf{q}_j;\ j \in \mathbb{N}\}$, that satisfy the equations

$$-\nu\Delta\mathbf{q}_0 + \mathbf{QB}(\mathbf{p}) = \mathbf{Qf}, \tag{3.12}$$

$$-\nu\boldsymbol{\Delta}\mathbf{q}_1 + \mathbf{QB}(\mathbf{p}) + \mathbf{QB}(\mathbf{p},\mathbf{q}_0) + \mathbf{QB}(\mathbf{q}_0,\mathbf{p}) = \mathbf{Qf}, \tag{3.13}$$

$$-\nu\Delta\mathbf{q}_2 + \mathbf{QB}(\mathbf{p}) + \mathbf{QB}(\mathbf{p},\mathbf{q}_1) + \mathbf{QB}(\mathbf{q}_1,\mathbf{p}) + \mathbf{QB}(\mathbf{q}_0,\mathbf{q}_0) + \mathbf{q}_0' = \mathbf{Qf}, \tag{3.14}$$

$$-\nu\boldsymbol{\Delta}\mathbf{q}_j + \mathbf{q}_{j-2}' + \mathbf{QB}(\mathbf{p}) + \mathbf{QB}(\mathbf{p},\mathbf{q}_{j-1}) + \tag{3.15}$$

$$+\mathbf{QB}(\mathbf{q}_{j-1},\mathbf{p}) + \mathbf{QB}(\mathbf{q}_{j-2},\mathbf{q}_{j-2}) = \mathbf{Qf},\ \ j \geq 2,$$

is defined.

If $\mathbf{p}(t)$ is, as above, the $\mathbf{P}$ projection of the solution $\mathbf{u}(t)$ of the NSE, the sets $\{\mathbf{u}_j(t) = \mathbf{p}(t) + \mathbf{q}_j(t);t \geq 0\}$ are called *induced trajectories* associated to the trajectory $\{\mathbf{u}(t) = \mathbf{p}(t) + \mathbf{q}(t);t \geq 0\}$. The inequalities

$$|\mathbf{q}_j| \leq \kappa_j\delta L^{1/2},\ \ \|\mathbf{q}_j\| \leq \kappa_j\delta^{1/2}L^{1/2},\ |\mathbf{q}_j'| \leq \kappa_j\delta L^{1/2},$$

are proved in [38], as well as the following

$$|\mathbf{q}(t) - \mathbf{q}_j(t)| \leq \bar{\kappa}_j L^{(1+j)/2}\delta^{(3+j)/2}. \tag{3.16}$$

3.4. A family of approximate inertial manifolds for the Navier-Stokes equations

The family of induced trajectories above, more precisely the functions $\mathbf{q}_j$, $j \geq 0$, form the starting point for the construction of a family of approximate inertial manifolds defined in the literature, the first one in [10] and the following in [38] and [39]. The first a.i.m. of this family is the graph $\mathcal{M}_0$ of the function $\Phi_0 : \mathbf{P}\mathcal{H} \to \mathbf{Q}\mathcal{H}$, that satisfies the relation

$$-\nu\boldsymbol{\Delta}\Phi_0(\mathbf{X}) + \mathbf{QB}(\mathbf{X}) = \mathbf{Qf},$$

where $\mathbf{X} \in \mathbf{P}\mathcal{H}$. Thus $\Phi_0(\mathbf{X})$ is explicitly given by

$$\Phi_0(\mathbf{X}) = (-\nu\boldsymbol{\Delta})^{-1}(\mathbf{Qf} - \mathbf{QB}(\mathbf{X})). \tag{3.17}$$

The connection between this definition and the definition (3.12) of $\mathbf{q}_0$ is obvious: the set of points $\{\mathbf{p}(t)+\mathbf{q}_0(t);\ t\geq 0\}$ lies on $\mathcal{M}_0$. The next a.i.m. defined in [38] is $\mathcal{M}_1$, the graph of the function $\Phi_1 : \mathbf{P}\mathcal{H} \to \mathbf{Q}\mathcal{H}$, given by the solution of the problem

$$-\nu\mathbf{\Delta}\Phi_1(\mathbf{X}) + \mathbf{QB}(\mathbf{X}) + \mathbf{QB}(\mathbf{X},\Phi_0(\mathbf{X})) + \mathbf{QB}(\Phi_0(\mathbf{X}),\mathbf{X}) = \mathbf{Qf},$$

that is

$$\Phi_1(\mathbf{X}) = -(\nu\mathbf{\Delta})^{-1}\left[\mathbf{Qf} - \mathbf{QB}(\mathbf{X}) - \mathbf{QB}(\mathbf{X},\Phi_0(\mathbf{X})) - \mathbf{QB}(\Phi_0(\mathbf{X}),\mathbf{X})\right]. \tag{3.18}$$

The relation with the definition (3.13) of the corresponding function $\mathbf{q}_1$ is clear.

For $j\geq 2$, inspired by the definition (3.15) of $\mathbf{q}_j$, the a.i.m. $\mathcal{M}_j$ is defined as the graph of $\Phi_j : \mathbf{P}\mathcal{H} \to \mathbf{Q}\mathcal{H}$, with $\Phi_j(\mathbf{X})$ the solution of

$$\begin{aligned}-\nu\mathbf{\Delta}\Phi_j(\mathbf{X}) + \mathbf{QB}(\mathbf{X}) + \mathbf{QB}(\mathbf{X},\Phi_{j-1}(\mathbf{X})) + \mathbf{QB}(\Phi_{j-1}(\mathbf{X}),\mathbf{X}) + \\ + \mathbf{QB}(\mathbf{\Phi}_{j-2}(\mathbf{X})) + \mathbf{D\Phi}_{j-2}(\mathbf{X})\,\Gamma_{j-2}(\mathbf{X}) = \mathbf{Qf},\end{aligned}$$

where $\mathbf{D\Phi}_{j-2}(\mathbf{X})\,\Gamma_{j-2}(\mathbf{X})$ is the Fréchet differential of $\mathbf{\Phi}_{j-2}(\mathbf{X})$, applied to

$$\Gamma_{j-2}(\mathbf{X}) = \nu\mathbf{\Delta X} - \mathbf{PB}(\mathbf{X} + \mathbf{\Phi}_{j-2}(\mathbf{X})) + \mathbf{Pf}. \tag{3.19}$$

Hence

$$\begin{aligned}\mathbf{\Phi}_j(\mathbf{X}) = -(\nu\mathbf{\Delta})^{-1}\left[\mathbf{Qf} - \mathbf{QB}(\mathbf{X}) - \mathbf{QB}(\mathbf{X},\mathbf{\Phi}_{j-1}(\mathbf{X})) - \right. \\ \left. - \mathbf{QB}(\mathbf{\Phi}_{j-1}(\mathbf{X}),\mathbf{X}) - \mathbf{QB}(\mathbf{\Phi}_{j-2}(\mathbf{X})) - \mathbf{D\Phi}_{j-2}(\mathbf{X})\,\Gamma_{j-2}(\mathbf{X})\right].\end{aligned} \tag{3.20}$$

The inequalities (3.16) allow us to estimate the distance between the trajectories of the problem and the a.i.m.s. This is immediate for the first two a.i.m.s, since for $j=0,1$, we have $\mathbf{u}_j(t)\in\mathcal{M}_j$, and thus

$$\mathrm{dist}_{\mathcal{H}}(\mathbf{u}(t),\ \mathcal{M}_j) \leq \mathrm{dist}(\mathbf{u}(t),\mathbf{u}_j(t)) = |\mathbf{q}(t)-\mathbf{q}_j(t)|.$$

For the a.i.m.s $\mathcal{M}_j$ with $j>1$, some extra work is necessary, since $\mathbf{D\Phi}_{j-2}(\mathbf{p}(t))\,\Gamma_{j-2}(\mathbf{p}(t))$ is only an approximation of $[\mathbf{q}_{j-2}(\mathbf{p}(t))]'$. However, in [38] and [39] it is proved that

$$\mathrm{dist}_{\mathcal{H}}(\mathbf{u}(t),\mathcal{M}_j) \leq \bar{\kappa}_j L^{(1+j)/2}\delta^{(3+j)/2}.$$

3.5. Nonlinear Galerkin method for the Navier Stokes equations

The nonlinear Galerkin method was first presented in [29]. It is defined for a class of equations that contains the Navier-Stokes equations as a particular case, i.e. an equation of the type (1.1) with

$$R(u) = B(u) + Cu,$$

where $B(u) = B(u,u)$, $B(.,.)$ is a bilinear operator having essentially the properties of $\mathbf{B}$ and C is a linear operator. It is assumed that $A + C$ is positive in H and C is bounded from $V = D(A^{1/2})$ to H.

We write the method for the Navier-Stokes problem we considered here (that is we take $A = -\nu\boldsymbol{\Delta}$, $B = \mathbf{B}$, $C = 0$). It consists in approximating in the $\mathbf{P}$ projection of the equation, the function $\mathbf{q}$ with help of the first a.i.m. of the family described above. That is, instead of the Galerkin equation, the equation

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{P}\left[\mathbf{B}(\mathbf{p}) + \mathbf{B}(\mathbf{p}, \boldsymbol{\Phi}_0(\mathbf{p})) + \mathbf{B}(\boldsymbol{\Phi}_0(\mathbf{p}), \mathbf{p})\right] = \mathbf{P}\mathbf{f}, \tag{3.21}$$

with the initial condition

$$\mathbf{p}(0) = \mathbf{P}\mathbf{u}_0,$$

is considered, where $\boldsymbol{\Phi}_0$ is given by (3.17) (the notations are adapted to ours). We see from the term $\mathbf{PB}(\mathbf{p} + \boldsymbol{\Phi}_0(\mathbf{p}))$ the term $\mathbf{PB}(\boldsymbol{\Phi}_0(\mathbf{p}), \boldsymbol{\Phi}_0(\mathbf{p}))$ is missing. This is because it is of lower order than the preceding terms.

As for the equation of $\boldsymbol{\Phi}_0$, this is taken in [29] as

$$\Phi_0(\mathbf{p}) = (-\nu\boldsymbol{\Delta})^{-1}\mathbf{Q}_{2m}\left[\mathbf{f} - \mathbf{B}(\mathbf{p})\right], \tag{3.22}$$

where $\mathbf{Q}_{2m}$ is the projection operator defined as $\mathbf{Q}_{2m} = \mathbf{Q}\mathbf{P}_{2m}$, where $\mathbf{P}_{2m}$ is the projector on the space spanned by the eigenfunctions corresponding to the eigenvalues in Γ_{2m} (of λ_{j_1,j_2} having $0 \leq j_1, j_2 \leq 2m$). This is because the space $\mathbf{Q}\mathcal{H}$ is infinite dimensional and a truncation must be made (at least for $\mathbf{f}$, since for periodic boundary conditions, if $\mathbf{X} \in \mathbf{P}\mathcal{H}$ then $\mathbf{B}(\mathbf{X})$ is anyway in $\mathbf{P}_{2m}\mathcal{H}$).

Let us denote, together with the authors of [29], the solution of (3.21) by $\mathbf{u}_m$. It is proved in the paper we refer at, that, if $\mathbf{u}_0 \in \mathcal{H}$ then $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^2(0,T;\mathcal{V})$, $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^p(0,T;\mathcal{H})$, strongly (for any $T > 0$, $p \geq 1$) and $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^\infty(\mathbb{R}^+;\mathcal{H})$ weak-star.

If $\mathbf{u}_0 \in \mathcal{V}$ then $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^2(0,T;D(A))$, $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^p(0,T;\mathcal{V})$, strongly (for any $T>0$, $p\geq 1$) and $\mathbf{u}_m \underset{m\to\infty}{\to} \mathbf{u}$ in $L^\infty(\mathbb{R}^+;\mathcal{V})$ weak-star.

As an alternative nonlinear Galerkin method, that starting from the equation (similar to (2.2))

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{PB}(\mathbf{p}+\mathbf{\Phi}_0(\mathbf{p})) = \mathbf{Pf}, \tag{3.23}$$

with Φ_0 defined by (3.22) is also given in [29]. Convergence results similar to those asserted above are proved.

In [7] an estimate of the error of the method is given

$$|\mathbf{u}(t) - [\mathbf{u}_m(t) + \mathbf{\Phi}_0(\mathbf{u}_m(t))]| \leq C(t)\delta^{3/2}.$$

In [6] the NL Galerkin method is improved by using more accurate a.i.m.s. The equation that provides the approximate solution is (we write it here also for the N-S equations)

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{PB}(\mathbf{p}+\mathbf{\Phi}_j(\mathbf{p}))] = \mathbf{Pf}, \tag{3.24}$$

where $\mathbf{\Phi}_j$ is the the function whose graph is the corresponding a.i.m. (similar to that defined in (3.20), but slightly different). Let us denote by $\mathbf{u}_{m,j}$ the solution of (3.24) and by $\mathbf{v}_{m,j} = \mathbf{u}_{m,j} + \mathbf{\Phi}_j(\mathbf{u}_{m,j})$. It is proved in [6] that if $\mathbf{u}_0 \in \mathcal{V}$, both $\mathbf{u}_{m,j}$ and $\mathbf{v}_{m,j}$ converge to $\mathbf{u}$ (when $m\to\infty$) in $L^2(0,T;D(A))$ and in $L^p(0,T;\mathcal{V})$, strongly (for all $p\geq 1$ and all $T>0$), and in $L^\infty(\mathbb{R}^+;\mathcal{V})$ weak-star. It is also proved that, for a fix j, $\mathbf{z}_{m,j} = \mathbf{\Phi}_j(\mathbf{u}_m)$ converges (when $m\to\infty$) to 0 in $L^\infty(\mathbb{R}^+;\mathcal{V})$ and $L^2(0,T;D(A))$ strongly for any $T>0$. In [7] some estimates for the error are obtained. More precisely, for the NSE, it is shown that (with our numbering of the a.i.m.s)

$$|\mathbf{u}(t) - \mathbf{v}_{m,j}(t)| \leq K_j L_m^{(j+3)/2}\delta^{(j+3)/2}.$$

3.6. Post-processed Galerkin method for the Navier-Stokes equations

The ideas on which the post-processed Galerkin method relies are exposed in 2.3. In [14] a general equation is considered and the Navier-Stokes equation

is treated as a particular case. The solution $\mathbf{p}_G$ of the Galerkin equation,

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{PB}(\mathbf{p}) = \mathbf{Pf}, \tag{3.25}$$
$$\mathbf{p}(\mathbf{0}) = \mathbf{Pu}_0,$$

is post-processed. This means, at a certain moment T (the end of the time interval on which the integration of (3.25) was performed) the image of $\mathbf{p}_G$ on the first a.i.m. $\mathcal{M}_0$, that is $\Phi_0(\mathbf{p}_G(T))$, is computed and is added to $\mathbf{p}_G(T)$. It is proved that, if $\mathbf{f} \in \mathcal{H}$, then

$$|\mathbf{u}(T) - (\mathbf{p}_G(T) + \Phi_0(\mathbf{p}_G(T)))| \leq C\delta^{5/4}. \tag{3.26}$$

In a subsequent paper, [15], the same authors prove another estimate for the Navier-Stokes problem. More exactly, they prove that, for $\mathbf{f} \in \mathcal{H}$,

$$|\mathbf{u}(T) - (\mathbf{p}_G(T) + \Phi_0(\mathbf{p}_G(T)))| \leq CL^2\delta^{3/2}. \tag{3.27}$$

Estimate (3.27) is not necessarily better than (3.26), since the coefficient L^2 appears (as before, $L = 1 + \ln(2m^2)$). In [32] the method is improved. That paper considers a reaction-diffusion equation, but the algorithm works for the Navier-Stokes equations as well. Instead of the Galerkin equations, the NL Galerkin equations (3.24) are considered. The solution $\mathbf{u}_{m,j}(t)$ of these equations is post-processed, i.e. the sum

$$\mathbf{u}_{m,j}(T) + \Phi_{j+1}(\mathbf{u}_{m,j}(T))$$

is considered and proposed as an approximate solution. The estimate of the error is made in [32] for the reaction-diffusion equation, hence is not relevant for the Navier-Stokes equation.

3.7. The repeatedly adjusted and post-processed Galerkin method for the Navier-Stokes equation

We adapt the general method presented in 2.4 to the Navier-Stokes equations.

Level 0. We define the first step of this level as the classical Galerkin method. Let us consider the Cauchy problem

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\boldsymbol{\Delta}\mathbf{p} + \mathbf{PB}(\mathbf{p}) = \mathbf{Pf}, \tag{3.28}$$
$$\mathbf{p}(0) = \mathbf{Pu}_0.$$

We denote by $\mathbf{p}_G(t)$ its solution and define

$$\widetilde{\mathbf{q}}_0(t) = \mathbf{\Phi}_0(\mathbf{p}_G(t)).$$

In the implementation of the method, the equation (3.28) must be numerically integrated. We remark that the values of $\widetilde{\mathbf{q}}_0(t)$ must be computed in every point of the time mesh used in the course of the numerical integration, since they will be used at the next level of the method.

We define the function

$$\widetilde{\mathbf{u}}_0(t) = \mathbf{p}_G(t) + \widetilde{\mathbf{q}}_0(t). \tag{3.29}$$

This preliminary level differs from the PP Galerkin method only in the post-processing part, in the fact that we compute $\widetilde{\mathbf{q}}_0(t)$ at any moment of time and not only at the end of the time interval on which (3.28) is integrated.

Level 1. Now we consider the problem

$$\begin{aligned} \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\mathbf{\Delta p} + \mathbf{PB}(\mathbf{p}+\widetilde{\mathbf{q}}_0) &= \mathbf{Pf}, \\ \mathbf{p}(0) &= \mathbf{Pu}_0, \end{aligned} \tag{3.30}$$

with $\widetilde{\mathbf{q}}_0(t)$ computed at the preceding step. Since $\widetilde{\mathbf{q}}_0(t)$ is already known, this equation is not more difficult to integrate than the simple Galerkin equation attached to the Navier-Stokes equation. It is an adjusted Galerkin equation since the nonlinear term is adjusted by adding to $\mathbf{p}(t)$ the term $\widetilde{\mathbf{q}}_0(t)$ that approximates $\mathbf{q}(t)$ better than $\mathbf{0}$ does. We denote by $\widetilde{\mathbf{p}}_0(t)$ the solution of problem (3.30). The computation of the error showed that $\widetilde{\mathbf{p}}_0$ is a better approximation of $\mathbf{p}$ than $\mathbf{p}_G$ (see the comments in 3.8).

Then we define

$$\begin{aligned} \widetilde{\mathbf{q}}_1(t) = \; & -(\nu\triangle)^{-1}[\mathbf{Qf} - \mathbf{QB}(\widetilde{\mathbf{p}}_0(t)) - \mathbf{QB}(\widetilde{\mathbf{p}}_0(t), \widetilde{\mathbf{q}}_0(t)) - \\ & -\mathbf{QB}(\widetilde{\mathbf{q}}_0(t), \widetilde{\mathbf{p}}_0(t))] \end{aligned}$$

The approximate solution will be defined at this level as

$$\widetilde{\mathbf{u}}_1(t) = \widetilde{\mathbf{p}}_0(t) + \widetilde{\mathbf{q}}_1(t). \tag{3.31}$$

This function is an approximation of $\mathbf{u}_1$ that defines the second induced trajectories.

Level j ($j \geq 2$). We assume that we constructed the functions $\widetilde{\mathbf{q}}_{j-2},\ \widetilde{\mathbf{q}}_{j-1}(t)$. We consider the adjusted Galerkin problem

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\boldsymbol{\Delta}\mathbf{p} + \mathbf{PB}\left(\mathbf{p}+\widetilde{\mathbf{q}}_{j-1}\right) = \mathbf{Pf}, \tag{3.32}$$

$$\mathbf{p}(0) = \mathbf{Pu}_0,$$

and denote by $\widetilde{\mathbf{p}}_{j-1}\left(t\right)$ its solution. Then we set

$$\widetilde{\mathbf{q}}_j\left(t\right) = \left(-\nu\boldsymbol{\Delta}\right)^{-1}\Big[\mathbf{Qf} - \mathbf{QB}(\widetilde{\mathbf{p}}_{j-1}\left(t\right)) - \mathbf{QB}(\widetilde{\mathbf{p}}_{j-1}\left(t\right),\widetilde{\mathbf{q}}_{j-1}\left(t\right)) - \tag{3.33}$$

$$-\mathbf{QB}(\widetilde{\mathbf{q}}_{j-1}\left(t\right),\widetilde{\mathbf{p}}_{j-1}\left(t\right)) - \mathbf{QB}(\widetilde{\mathbf{q}}_{j-2}\left(t\right),\widetilde{\mathbf{q}}_{j-2}\left(t\right)) - \widetilde{\mathbf{q}}'_{j-2}\left(t\right)\Big].$$

We define the approximate solution at this level as

$$\widetilde{\mathbf{u}}_j(t) = \widetilde{\mathbf{p}}_{j-1}\left(t\right) + \widetilde{\mathbf{q}}_j(t).$$

We remark that $\widetilde{\mathbf{u}}_j\left(t\right)$ is an approximation of $\mathbf{u}_j\left(t\right)$ (that defines a induced trajectory of the family constructed in [38]).

We must say that, at the last level, as in the NL PP Galerkin method, we may correct $\widetilde{\mathbf{p}}_{j-1}$ by adding $\widetilde{\mathbf{q}}_j$ only at some moments of interest (the final postprocessing step).

We also must remark that, when the method is numerically implemented, the projector $\mathbf{Q}$ must be replaced by a finite dimensional projector as, e.g. $\mathbf{Q}_{2m}$ defined in Section 3.5.

3.8. The error of the R-APP Galerkin method

It is not the purpose of this work to present the explicit calculus of the error of the methods presented. We proved in [20] that

$$|\mathbf{p}(t) - \widetilde{\mathbf{p}}_j(t)| \leq C\delta^{5/4+j/2}$$

and

$$|\mathbf{q}(t) - \widetilde{\mathbf{q}}_j(t)| \leq C\delta^{3/2+j/2},$$

where C depends on the data of the problem: $\Omega,\ \mathbf{f},\ \nu,\ \lambda_1$, and on t but not on m.

With other methods, other estimates may be obtained. If we start from estimates of [15] of $\left|\mathbf{p}\left(t\right) - \mathbf{p}_G\left(t\right)\right|$, where $\mathbf{p}_G\left(t\right)$ is, as before, the classical Galerkin approximation of the solution, that is

$$\left|\mathbf{p}\left(t\right) - \mathbf{p}_G\left(t\right)\right| \leq C'L^2\delta^{3/2},$$

an improvement of the estimate of the error of the successive solutions by a factor of $\delta^{1/4}$ seems to be obtained. However, the appearance of the factor L^2 ($L = 1 + \ln(2m^2)$) diminishes this success. A very careful analysis of the constants C, C' should be performed in order to see what approach is better.

Anyway, the R-APP Galerkin provides approximates solutions as accurate as those provided by the NL PP Galerkin method.

3.9. R-APP Galerkin method compared to the high-order accuracy NLPP Galerkin method

The R-APP Galerkin method is intended to bring some simplifications to the NL Galerkin methods that use high accuracy approximate inertial manifolds. Hence this method makes sense only if more of its levels are passed through.

The simplifications come from the following facts:

a) the use of some already known functions (the $\widetilde{\mathbf{q}}_j$s) for the adjustment of the Galerkin equation, makes the equations for the approximations of $\mathbf{p}$ to have essentially the same structure as the Galerkin equation; this imply simplifications of the algorithms for the numerical integration of these equations, compared to the corresponding equations of the NL Galerkin equations;

b) the use of the "approximate induced trajectories" instead of the approximate inertial manifolds makes some computations easier, because, in the function $\widetilde{\mathbf{q}}_j$ the term $\widetilde{\mathbf{q}}'_{j-2}$ appears instead of the corresponding term $\mathbf{D}\mathbf{\Phi}_{j-2}(\mathbf{X})\,\Gamma_{j-2}(\mathbf{X})$ of the a.i.m. Φ_j; the term $\widetilde{\mathbf{q}}'_{j-2}$ can be approximated by the numerical derivative (since we know its values in the points of the time mesh);

c) when we proceed to Level j of the method, all we need are the values of $\widetilde{\mathbf{q}}_{j-2}$ and $\widetilde{\mathbf{q}}_{j-1}$, while all values of $\widetilde{\mathbf{p}}_k$, $k < j-1$ and $\widetilde{\mathbf{q}}_k$, $k < j-2$ may be erased from the memory of the computer; this must be compared to the NL Galerkin method that uses $\mathcal{M}_j$, where in the course of a single numerical integration one must handle the values of all functions Φ_k, $k \leq j$, and all these must be stored in the memory of the computer.

In order to compare the R-APP Galerkin method with the NL PP Galerkin method, we must look at the levels j with $j \geq 2$.

Let us analyze in parallel the first stage of Level 3 (that delivers us the function $\widetilde{\mathbf{p}}_2$) of our method and the corresponding NL Galerkin method (that uses the a.i.m. $\mathcal{M}_2$). It is easier to follow our reasoning on this particular case than than on the general one.

In order to make the following as clear as possible, we describe the computations necessary for the simple Euler integration method. Of course, more elaborated algorithms must be used, but the difficulties added by these should be evaluated for each specific algorithm individually.

In order to proceed, we consider a time-mesh $0 = t_0 < t_1 < t_2 < ... < t_k < ... < t_N = T$ on the time integration interval $[0, T]$.

Let us make the notations

$$\Gamma_G(\mathbf{p}) = \mathbf{Pf} + \nu\boldsymbol{\Delta}\mathbf{p} - \mathbf{PB}(\mathbf{p})$$

$$\widetilde{\Gamma}_j(\mathbf{p}) = \mathbf{Pf} + \nu\boldsymbol{\Delta}\mathbf{p} - \mathbf{PB}(\mathbf{p}+\widetilde{\mathbf{q}}_j).$$

R-APP Galerkin method, at the third level, requires the following computations for the determination of $\widetilde{\mathbf{p}}_2(t)$:

at Level 0 – computation of $\mathbf{p}_G(t_k)$, $k = 1, ..., N$, by numerical integration of eq. (3.28) (this is equivalent with the computation of $\Gamma_G(\mathbf{p_G}(t_{k-1}))$); then computation of $\widetilde{\mathbf{q}}_0(t_k)$;

at Level 1 – computation of $\widetilde{\mathbf{p}}_0(t_k)$, $k = 1, ..., N$, by numerical integration of eq. (3.30) (this is equivalent with the computation of $\widetilde{\Gamma}_0(\widetilde{\mathbf{p}}(t_{k-1}))$); then computation of $\widetilde{\mathbf{q}}_1(t_k)$;

at Level 2 – computation of $\widetilde{\mathbf{p}}_1(t_k)$, $k = 1, ..., N$, by numerical integration of eq. (3.32) with $j = 2$, (this is equivalent with the computation of $\widetilde{\Gamma}_1(\widetilde{\mathbf{p}}_1(t_{k-1}))$), then computation of $\widetilde{\mathbf{q}}_2(t_k)$;

at Level 3 – computation of $\widetilde{\mathbf{p}}_2(t_k)$, by numerical integration of eq. (3.32) with $j = 3$ (this is equivalent with the computation of $\widetilde{\Gamma}_2(\widetilde{\mathbf{p}}_2(t_{k-1}))$).

NL Galerkin method that uses $\mathcal{M}_2$, presented in [29], consists in the integration of the system of ODEs

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} - \nu\Delta\mathbf{p} + \mathbf{P}[\mathbf{B}(\mathbf{p} + \boldsymbol{\Phi}_2(\mathbf{p}))] = \mathbf{Pf}, \qquad (3.34)$$

$$\mathbf{p}(0) = \mathbf{Pu}_0,$$

where the function Φ_2 is given by

$$-\nu\Delta\boldsymbol{\Phi}_2(\mathbf{p}) + \mathbf{Q}_{2m}\mathbf{B}(\mathbf{p} + \boldsymbol{\Phi}_\mathbf{1}(\mathbf{p})) + \mathbf{q}_1^1 = \mathbf{Q}_{2m}\mathbf{f}, \quad (3.35)$$

$$-\nu\Delta\mathbf{q}_1^1 + \mathbf{Q}_{2m}\left[\mathbf{B}(\left(\mathbf{p}_0^1,\ \mathbf{p} + \boldsymbol{\Phi}_\mathbf{1}(\mathbf{p})\right) + \mathbf{B}\left(\mathbf{p} + \boldsymbol{\Phi}_\mathbf{1}(\mathbf{p}), \mathbf{p}_\mathbf{0}^\mathbf{1}\right)\right] = \mathbf{0},$$

$$\mathbf{p}_\mathbf{0}^\mathbf{1} - \nu\boldsymbol{\Delta}\mathbf{p} + \mathbf{P}[\mathbf{B}(\mathbf{p} + \boldsymbol{\Phi}_\mathbf{1}(\mathbf{p}))] = \mathbf{Pf},$$

$$-\nu\Delta\boldsymbol{\Phi}_1(\mathbf{p}) + \mathbf{Q}_{2m}\mathbf{B}(\mathbf{p} + \boldsymbol{\Phi}_\mathbf{0}(\mathbf{p})) = \mathbf{Q}_{2m}\mathbf{f},$$

$$-\nu\Delta\boldsymbol{\Phi}_0(\mathbf{p}) + \mathbf{Q}_{2m}\mathbf{B}(\mathbf{p}) = \mathbf{Q}_{2m}\mathbf{f}.$$

We reproduced here the definition of $\mathcal{M}_2$ from [29], but we adapted the notations from [29] to our notations and we started counting a.i.m.s with 0, as in [38], while in [29] this count begins with 1.

In the course of the numerical integration, with $\mathbf{p}(t_{k-1})$, $k = 1, ..., N$, already determined, in order to find $\mathbf{p}(t_k)$, we have to compute:

$\mathbf{\Phi}_0(\mathbf{p}(t_{k-1}))$, $\mathbf{\Phi}_1(\mathbf{p}(t_{k-1}))$, $\Gamma_1(\mathbf{p}(t_{k-1}))$ (for the calculation of $\mathbf{p_0^1}(t_k)$, with Γ_1 given by (3.19), $j = 3$), $\mathbf{q}_1^1(t_{k-1})$, $\mathbf{\Phi}_2(\mathbf{p}(t_{k-1}))$, and finally $\Gamma_2(\mathbf{p}(t_{k-1}))$. This will yield $\mathbf{p}(t_k)$.

Now we can compare the two methods from the point of view of the computations involved. We have the following:

– computation of $\widetilde{\mathbf{q}}_0(t_j)$ is equivalent to that of $\mathbf{\Phi}_0(\mathbf{p}(t_j))$;

– computation of $\widetilde{\mathbf{q}}_1(t_j)$ is equivalent to that of $\mathbf{\Phi}_1(\mathbf{p}(t_j))$;

– computation of $\widetilde{\Gamma}_1(\widetilde{\mathbf{p}}_1(t_j))$ is equivalent to that of $\Gamma_1(\mathbf{p}(t_j))$;

– computation of $\widetilde{\mathbf{q}}_2(t_j)$ is equivalent to that of $\mathbf{\Phi}_2(\mathbf{p}(t_j))$, assuming that $\mathbf{q}_1^1(t_j)$ is already computed;

– finally we observe that the computation of $\mathbf{p}_G(t_j)$ and $\widetilde{\mathbf{p}}_0(t_j)$ (from R-APP Galerkin method) together, involve less computations than that of $\mathbf{q}_1^1(t_j)$ (from the NL Galerkin method).

This is because in computing $\mathbf{p}_G(t_j)$ we have to compute a number of $4m^2 + 2m$ projections of the term $\Gamma_G(\mathbf{p_G}(t_{j-1}))$ and in computing $\widetilde{\mathbf{p}}_0(t_j)$ we have to compute $4m^2 + 2m$ projections of the term $\widetilde{\Gamma}_0(\widetilde{\mathbf{p}}(t_{j-1}))$, while in computing $\mathbf{q}_1^1(t_j)$ we have to compute $12m^2 + 6m$ projections.

At the following level, induced trajectories, respectively a.i.m.s, of higher order are used. The definition of these involves approximations of the derivatives similar to the above. Hence, the difference in the amounts of computations between the two methods increases with the order of the method. It follows that the R-APP Galerkin method involves a smaller amount of computations than the NL Galerkin method.

The computational effort involved in the final post-processing part is eased in the R-APP Galerkin method by the fact that, by using approximations of the induced trajectories we can approximate directly (by numerical derivative) the function $\mathbf{q}'$, while in the NL PP Galerkin method it is approximated by the differential $\mathbf{D\Phi}_{j-2}(\mathbf{X})\Gamma_{j-2}(\mathbf{X})$. In conclusion, the R-APP Galerkin method brings simplifications to the NL PP Galerkin method relying on higher accuracy a.i.m.s.

4. Modified Galerkin methods for a reaction-diffusion problem

We consider a reaction-diffusion (RD) equation of the form

$$\frac{\partial u}{\partial t} - D\left(\Delta u - u\right) + g(u) = f, \tag{4.1}$$

where u is a real-valued function, $u = u(t, \mathbf{x}),\ \mathbf{x} \in \Omega = (0,\ l) \times (0,\ l),\ l > 0$, D is the diffusion coefficient and the function g is a polynomial function of odd degree. In order to simplify the following considerations we take here a polynomial function of degree 3,

$$g(u) = b_0 + b_1 u + b_2 u^2 + b_3 u^3,\ \ b_i \in \mathbb{R},\ b_3 > 0.$$

We take $f \in L^2(\Omega)$. To the equation (4.1) we associate an initial condition

$$u(0) = u_0 \tag{4.2}$$

and the boundary condition

$$\left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = 0. \tag{4.3}$$

The phase space is here $\mathcal{H} = L^2(\Omega)$. We consider also the space $\mathcal{V} = H^1(\Omega)$ with the usual norm.

The operator $A = -\Delta + I$ is a positive-definite, self–adjoint, with compact inverse operator with definition domain $D(A) = H^2(\Omega)$. The following existence result may be obtained by the Galerkin-Faedo method [37], [34]

THEOREM **4.1** *If $u_0 \in \mathcal{H}$, then there exists a unique solution $u \in C(\mathbb{R}^+;\ \mathcal{H})$, $u \in L^2(0,T;\ \mathcal{V}) \cap L^{2p}(0,T;L^{2p}(\Omega))$ where $p > 1,\ T > 0$. If, more than that, $u_0 \in \mathcal{V}$, then $u \in C([0,T);\ \mathcal{V}) \cap L^2(0,T;\ H^2(\Omega))$.*

The semi-dynamical system $\{S(t)\}_{t\geq 0}$, generated by (4.1) is proved to be dissipative in $\mathcal{H}$ and $\mathcal{V}$ [37], [34]. Hence there is a $\rho_0 > 0$ (respectively a $\rho_1 > 0$), such that for every $R > 0$, there is a moment $t_0(R)$ (respectively $t_1(R) > t_0(R)$) with the property that for every $u_0 \in \mathcal{H}$ with $|u_0| \leq R$, we have $|S(t)u_0| < \rho_0$, for $t \geq t_0(R)$ (respectively $\|S(t)u_0\| < \rho_1$, for $t \geq t_1(R)$).

4.1. The splitting of the space

The eigenvalues of A are

$$\lambda_{j,k} = \frac{\pi^2}{l^2}[j^2 + k^2] + 1$$

and the corresponding eigenfunctions are

$$w_{j,k} = \frac{\sqrt{\alpha_j \alpha_k}}{l} \cos\frac{j\pi x}{l} \cos\frac{k\pi y}{l},$$

where $\alpha_j = 1$ for $j = 0$ and $\alpha_j = 2$ for $j \neq 0$.

As for the Navier-Stokes equations, we consider the set Γ_m of eigenvalues λ_{j_1,j_2} with $0 \leq j_1, j_2 \leq m$. We make the notations

$$\Lambda = \lambda_{m+1,0} = \lambda_{0,m+1},$$
$$\delta = \frac{1}{\Lambda}.$$

We also consider the space spanned by the eigenfunctions corresponding to these eigenvalues and we denote by P the projector on this space. We set $Q = I - P$, where I is the identity on $\mathcal{H}$, $p = Pu,\ q = Qu$.

We project the equation (4.1) by using these projectors, to obtain

$$\frac{\mathrm{d}p}{\mathrm{d}t} - D\,(\Delta p - p) + Pg(p+q) = Pf,$$
$$\frac{\mathrm{d}q}{\mathrm{d}t} - D\,(\Delta q - q) + Qg(p+q) = Qf.$$

It can be proved (e.g. [4]) that

$$|q| \leq C\delta$$

for t great enough, where the coefficient C depends on the data of the problem.

4.2. Induced trajectories for the reaction-diffusion problem

In constructing a family of induced trajectories for the reaction-diffusion problem, we try an asymptotic analysis of the RD equations. We develop the function q in series of powers of δ

$$q = \delta\left(k_0 + \delta k_1 + \delta^2 k_2 + \delta^3 k_3 + ...\right). \tag{4.4}$$

We have

$$\begin{aligned} g(p+q) &= g(p) + g'(p)q + \frac{1}{2}g''(p)q^2 + \frac{1}{6}g'''(p)q^3 = \\ &= g(p) + g'(p)\delta\left(k_0 + \delta k_1 + \delta^2 k_2 + \delta^3 k_3 + ...\right) + \\ &+ \frac{1}{2}g''(p)\left[\delta\left(k_0 + \delta k_1 + \delta^2 k_2 + \delta^3 k_3 + ...\right)\right]^2 + \\ &+ \frac{1}{6}g'''(p)\left[\delta\left(k_0 + \delta k_1 + \delta^2 k_2 + \delta^3 k_3 + ...\right)\right]^3, \end{aligned}$$

hence, by ordering the terms after the powers of δ,

$$\begin{aligned} g(p+q) &= g(p) + \delta g'(p)k_0 + \\ &+ \delta^2\left[g'(p)k_1 + \frac{1}{2}g''(p)k_0^2\right] + \\ &+ \delta^3\left[g'(p)k_2 + \frac{1}{2}g''(p)2k_0k_1 + \frac{1}{6}g'''(p)k_0^3\right] + \\ &+ \delta^4\left[g'(p)k_3 + \frac{1}{2}g''(p)\left(k_1^2 + 2k_0k_2\right) + \frac{1}{6}g'''(p)3k_0^2k_1\right] + ... \,. \end{aligned} \tag{4.5}$$

Then, by substituting (4.4) in the equation for q, we obtain

$$\begin{aligned} &\delta k_0' + \delta^2 k_1' + \delta^3 k_2' + \delta^4 k_3' + ... \\ &- D\left[\delta\Delta k_0 + \delta^2\Delta k_1 + \delta^3\Delta k_2 + \delta^4\Delta k_3 + \delta^5\Delta k_4 + ...\right] + \\ &+ D\left[\delta k_0 + \delta^2 k_1 + \delta^3 k_2 + \delta^4 k_3 + \delta^5 k_4 + ...\right] + \\ &+ Qg(p) + \delta Qg'(p)k_0 + \delta^2 Q\left[g'(p)k_1 + \frac{1}{2}g''(p)k_0^2\right] + \\ &+ \delta^3 Q\left[g'(p)k_2 + \frac{1}{2}g''(p)2k_0k_1 + \frac{1}{6}g'''(p)k_0^3\right] + \\ &+ \delta^4 Q\left[g'(p)k_3 + \frac{1}{2}g''(p)\left(k_1^2 + 2k_0k_2\right) + \frac{1}{6}g'''(p)3k_0^2k_1\right] + ... = Qf. \end{aligned}$$

In ordering the terms in (4.5) we simply performed an algebraic calculus, and treated the right-hand side as a polynomial in δ, but when we look for the terms of the same order of magnitude, a careful analysis should be performed. Since $k_j(t) \in Q\mathcal{H}$, we have

$$|\Delta k_j| \geq \Lambda\,|k_j| = \frac{1}{\delta}\,|k_j| \tag{4.6}$$

and it follows that the term $\delta^{j+1}\nu\Delta k_j$ is of the order of j. We also must evaluate carefully the terms containing products or powers of $k_j s$. E.g., for the term $\frac{1}{2}g''(p)k_0^2$ we have the estimates

$$\left|\frac{1}{2}g''(p)k_0^2\right| = \left(\int_\Omega \left(g''(p)\right)^2 k_0^4 \mathrm{d}x\right)^{1/2} \leq \left(\int_\Omega \left(g''(p)\right)^4 dx\right)^{1/4} \left(\int_\Omega k_0^8 \mathrm{d}x\right)^{1/4}.$$

Sobolev embedding theorem gives

$$\|u\|_{L^p(\Omega)} \leq C(p,s)\,\|u\|_s\,,$$

with $1/p = 1/2 - s/2$, $s < 1$, and, since

$$\|u\|_s \leq C\,\|u\|_1\,,$$

we obtain

$$\left(\int_\Omega k_0^8 dx\right)^{1/4} = \|k_0\|^2_{L^8(\Omega)} \leq C^2(8,\frac{3}{4})\,\|k_0\|^2_{3/4} \leq C^2(8,\frac{3}{4})\,\|k_0\|^2_1\,.$$

In a similar way we see that $\left(\int_\Omega \left(g''(p)\right)^4 \mathrm{d}x\right)^{1/4}$ is a function of ρ_0 and ρ_1. This together with inequality $\|k_0\|_1 \geq \left(\frac{1}{\delta}\right)^{1/2} |k_0|$ show that all we can say about the term $\frac{1}{2}\delta^2 g''(p)k_0^2$ is that it is of order δ and we have to consider it together with the terms of the same order. Similar reasonings will be considered implicit for the other terms containing products or powers of k_js. Thus we obtain the relations:

$$\begin{aligned}
-\delta D\Delta k_0 + Qg(p) &= Qf,\\
\delta k_0' - \delta^2 D\Delta k_1 + \delta D k_0 + \delta Q g'(p)k_0 + \frac{1}{2}\delta^2 Q g''(p)k_0^2 &= 0,\\
\delta^2 k_1' - \delta^3 D\Delta k_2 + \delta^2 D k_1 + \delta^2 Q g'(p)k_1 + &\\
+\frac{1}{2}\delta^3 Q g''(p)2k_0k_1 + \frac{1}{6}\delta^3 Q g'''(p)k_0^3 &= 0,\\
\delta^3 k_2' - \delta^4 D\Delta k_3 + \delta^3 D k_2 + \delta^3 Q g'(p)k_2 + &\\
\frac{1}{2}\delta^4 Q g''(p)\left(k_1^2 + 2k_0k_2\right) + \frac{1}{6}\delta^4 Q g'''(p)3k_0^2k_1 &= 0,\\
&\vdots
\end{aligned}$$

Now we define the functions

$$q_j = \delta k_0 + \delta^2 k_1 + \delta^3 k_2 + \delta^4 k_3 + ... + \delta^{j+1} k_j.$$

By summing the equations for k_j, we obtain equations for q_j:

$$\begin{aligned} -D\Delta q_0 + Qg(p) &= Qf, \qquad (4.7)\\ q_0' - D\Delta q_1 + Dq_0 + Qg\,(p) + Qg'(p)q_0 + \frac{1}{2}Qg''(p)q_0^2 &= Qf,\\ q_1' - D\Delta q_2 + Dq_1 + Qg\,(p) + Qg'\,(p)\,q_1 +&\\ \frac{1}{2}Qg''(p)q_0^2 + \frac{1}{2}Qg''(p)2q_0\,(q_1 - q_0) + \frac{1}{6}Qg'''(p)q_0^3 &= Qf,\\ q_2' - D\Delta q_3 + Dq_2 + Qg\,(p) + Qg'\,(p)\,q_2 +&\\ \frac{1}{2}Qg''(p)q_1^2 + \frac{1}{6}Qg'''(p)3q_0^2\,(q_1 - q_0) &= Qf,\\ \vdots& \end{aligned}$$

We see that the nonlinearity of the polynomial makes the equations neither "beautiful", nor with a clear structure. However, we consider the functions

$$u_j(t) = p(t) + q_j(t),$$

and define the *induced trajectories* of the problem as the sets $\{u_j(t);\ t \geq 0\}$. These will be used to define the R-APP method for the reaction-diffusion equations.

4.3. Approximate inertial manifolds for the reaction-diffusion equation

In the NL Galerkin method and in the NL PP Galerkin method described in literature [32], the following a.i.m.s are defined for the RD equation: for any $j \geq 0$, $\mathcal{M}_j$ is the graph of the function $\Phi_j : P\mathcal{H} \to Q\mathcal{H}$, described below

$$DA\Phi_0(p) + Qg(p) = Qf, \qquad (4.8)$$

$$q_{j-1}^1 + DA\Phi_j(p) + Qg\,(p + \Phi_{j-1}(p)) = Qf, \quad j \geq 1. \qquad (4.9)$$

Here $q_{j-1}^1 = \mathbf{D}\Phi_{j-1}(p)\Gamma_{j-1}(p)$, with $\mathbf{D}\Phi_{j-1}(p)$ the Fréchet differential of Φ_{j-1} computed in p and applied to $\Gamma_{j-1}(p) = Pf - DAp - Pg(p + \Phi_{j-1}(p))$.

If we would want to construct a family of a.i.m.s $\widetilde{\mathcal{M}}_j$ starting from the induced trajectories we defined above (as is done in [38] for the Navier-Stokes equation), the first a.i.m. of the family, $\widetilde{\mathcal{M}}_0$, would be identical with $\mathcal{M}_0$ since the function $\widetilde{\Phi}_0$ defining it would be identical to Φ_0 of (4.8), as the equation for $q_0(t)$ shows.

The second a.i.m., $\widetilde{\mathcal{M}}_1$, would be quite different from $\mathcal{M}_1$ above. That is, it would be the graph of the function $\widetilde{\Phi}_1$ defined by the equation

$$\begin{array}{rl} DQ\widetilde{\Phi}_0(p)\Gamma_0(p) - D\Delta\widetilde{\Phi}_1(p) & +D\widetilde{\Phi}_0(p) + Qg(p) + \\ +Qg'(p)\widetilde{\Phi}_0(p) + & \frac{1}{2}Qg''(p)\widetilde{\Phi}_0(p)^2 = Qf, \end{array} \tag{4.10}$$

with $\Gamma_0(p) = Pf + D(\Delta p - p) - Pg(p + \widetilde{\Phi}_0(p))$. We see that the difference between this equation and that for Φ_1, that we write explicitly below

$$\mathbf{D}\Phi_0(p)\Gamma_0(p) - D\Delta\Phi_1(p) + D\Phi_1(p) + Qg(p + \Phi_0(p)) = Qf, \tag{4.11}$$

consists essentially in the presence of the term $\frac{1}{6}g'''(p)\Phi_0(p)^3$ in this latter equation. If the polynomial g would be of higher degree, the difference between the two families of a.i.m.s, that defined starting from the induced trajectories and the one defined by the relations (4.8) and (4.9) would increase. However, for the sake of the elegance of the definitions, (4.11) may be taken as the equation for $\Phi_1(p)$ even if it does not spring from an accurate asymptotic analysis. The presence of the higher order terms does not affect the order of magnitude of the distance between the exact solution of the R-D equation and the first a.i.m. [21].

4.4. "Induced trajectories" inspired by a.i.m.s

For the sake of the simplicity of the definitions and having in mind some simplifications of the computations in the R-APP Galerkin method below, we can choose an alternate definition for the induced trajectories of the R-D problem, inspired from the definitions of the a.i.m. of [32]. That is, we define the functions $\widetilde{q}_j$ through the relations

$$\begin{array}{c} DA\widetilde{q}_0 + Qg(p) = Qf, \\ \widetilde{q}'_{j-1} + DA\widetilde{q}_j + Qg(p + \widetilde{q}_{j-1}) = Qf, \quad j \geq 1, \end{array} \tag{4.12}$$

where $p(t) = Pu(t)$. The functions

$$\widetilde{u}_j = p + \widetilde{q}_j$$

define the new "induced trajectories" $\{\widetilde{u}_j(t);\ t \geq 0\}$.

4.5. The NL Galerkin method for the RDE

The NL Galerkin method for RDE consists in integrating the differential equation:

$$\frac{\mathrm{d}p}{\mathrm{d}t} + DAu + g(p + \Phi_0(p)) = Pf, \tag{4.13}$$

with the initial condition

$$p(0) = Pu_0. \tag{4.14}$$

If we denote by y_m its solution, the approximate solution is taken as

$$y_m(t) + \Phi_0(y_m(t)).$$

In [32] it is asserted that, for large enough t,

$$|u(t) - (y_m(t) + \Phi_0(y_m(t)))| \leq C\delta^2.$$

Improved NL Galerkin methods make use of the higher accuracy a.i.m.s, $\mathcal{M}_j$, $j \geq 1$. That is an equation of the type

$$\frac{\mathrm{d}p}{\mathrm{d}t} + DAu + g(p + \Phi_j(p)) = Pf, \tag{4.15}$$

with the initial condition (4.14) is solved, let $y_{m,j}$ be its solution. The approximate solution of the RDE is taken as:

$$y_{m,j}(t) + \Phi_j(y_{m,j}(t)).$$

In [32] it is proved that the $\mathcal{H}$ norm of the error of this approximate solution is of the order of $C(t)\delta^{j+2}$.

4.6. The PP NL Galerkin method for the RDE

Also in [32] the NL Galerkin method is post-processed, i.e. to the solution $y_{m,j}$ of the NL Galerkin problem, considered in T, the quantity $\Phi_{j+1}(y_{m,j}(T))$ is added and

$$y_{m,j}(T) + \Phi_{j+1}(y_{m,j}(T))$$

is taken as the approximate solution in T. It is proved in [32] that

$$|u(t) - (y_{m,j}(t) + \Phi_{j+1}(y_{m,j}(t)))| \leq C \ln m\, \delta^{j+3}.$$

4.7. The R-APP Galerkin method for the RDE

We describe the R-APP Galerkin method for the reaction-diffusion equation. In [21] we presented a variant of our method that has as initial level a NL Galerkin method (this was meant to skip a numerical integration - that of the Galerkin problem). Let us denote generically

$$q_j = F_j(Qf, p, q_0, q_1, ..., q_{j-1}),$$

either the functions given by the set of relations (4.7) or the functions $\widetilde{q}_j$ given by (4.12). We see that in this latter case, F_j, $j \geq 1$ actually depends only on Qf, p, q_{j-1}, q'_{j-1}.

Level 0. We consider the NL Galerkin problem

$$\begin{aligned} \frac{\mathrm{d}p}{\mathrm{d}t} - D\left(\Delta p - p\right) + Pg(p) &= Pf, \\ p(0) &= Pu_0 \end{aligned} \tag{4.16}$$

and denote it's solution by p_G.

Then we compute, at every moment of time

$$\widetilde{q}_0\left(t\right) = F_0(Qf, p_G\left(t\right)).$$

When the numerical implementation of the method is actually done, this is equivalent to the computation of q_1 at the nodes of the time mesh, and $q'_0\left(t_i\right)$ is approximated by $(q_0(t_i) - q_0(t_{i-1}))/(t_i - t_{i-1})$. The approximate solution is

$$u_0 = p_G + \widetilde{q}_0.$$

Level 1. We consider the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} - D\left(\Delta p - p\right) + Pg(p + \widetilde{q}_0) = Pf,$$

and denote its solution by $\widetilde{p}_0$. Then we compute

$$\widetilde{q}_1\left(t\right) = F_1(Qf, \widetilde{p}_0\left(t\right), \widetilde{q}_0\left(t\right)).$$

The approximate solution at this level is defined as

$$\widetilde{u}_j = \widetilde{p}_{j-1} + \widetilde{q}_j.$$

Level j > 1. We assume $\widetilde{q}_0, \widetilde{q}_1, ..., \widetilde{q}_{j-1}$ were successively constructed. We consider the equation

$$\frac{\mathrm{d}p}{\mathrm{d}t} - D\left(\Delta p - p\right) + Pg(p + \widetilde{q}_{j-1}) = Pf,$$

and denote its solution by $\widetilde{p}_{j-1}$. Then we compute

$$\widetilde{q}_j\left(t\right) = F_j(Qf, \widetilde{p}_{j-1}\left(t\right), \widetilde{q}_0\left(t\right), \widetilde{q}_1(t), ..., \widetilde{q}_{j-1}(t),).$$

The approximate solution at this level is defined as

$$\widetilde{u}_j = \widetilde{p}_{j-1} + \widetilde{q}_j.$$

Remarks: 1. While the equations for p_j are equivalent to a finite, constant number, of (differential) equations, the equations for q_j are equivalent to a system of equations having (if Qf admits non-null projections on an infinite number of eigenfunctions) a infinite number of equations.

Hence a truncation must be done. In [6] the truncation is made by using a projector, denoted P_{2m}, that is the analogous of P but with $2m$ instead of m. If Qf would have nonzero projections only on a finite number of eigenfunctions, then q_j would also be finite dimensional. In this situation, we could also compute the dimension of q_j, by using the consequences of the trigonometrical relation $2\cos\alpha\cos\beta = \cos\left(\alpha+\beta\right) + \cos\left(\alpha-\beta\right)$, on the products of eigenfunctions. Then, in order to not affect the estimate of the error predicted by our method, we could take a truncation of Qf, let us denote it by $Q_j f$ such that $\left|\Delta^{-1}\left(Qf - Q_j f\right)\right|$ is less that the error of the level j.

2. Both families of $\{q_j\}_{j\geq 0}$ defined above present advantages and disadvantages one relative to the other. The first family, defined in (4.7), has the advantage of demanding a smaller amount of computations since in (4.7) fewer terms than in (4.12) are taken into account at a certain level. It presents the disadvantage of recalling all q_i with $i < j$, at a certain level j. The second family of approximations of q, given by (4.12), recalls at a certain level j, only the values of q_{j-1}. This is important from the point of view of organizing the memory of the computer in the numerical implementation of the method. However, this second family takes into account more terms in the polynomial g. This increases a lot the computations when g has a high degree.

4.8. Estimates of the error

By using the method of [32], we can prove that both families of induced trajectories defined above lead to the same orders of error, for every level of the R-APP method, as the corresponding NL PP Galerkin method. That is, we can prove [22] that at the level $j+1$ of our method

$$|p - \widetilde{p}_j| \leq C_j \, (\ln m \,) \, \delta^{j+3}$$

and

$$|q - \widetilde{q}_{j+1}| \leq K_j \delta^{j+3},$$

and thus

$$|u - \widetilde{u}_{j+1}| \leq [C_j \, (\ln m \,) + K_j] \, \delta^{j+3}.$$

4.9. Comments on the method

The comparison of the computational cost of the R-APP Galerkin method to that of the NL Galerkin method is similar to that we performed for the Navier-Stokes equations. The conclusions are the same: the R-APP Galerkin method is more economic than the NL PP Galerkin method. The difference in the computational cost between the two methods increases with their level.

References

[1] P. Constantin, C. Foiaş, *Navier-Stokes Equations,* Chicago Lectures in Math., Univ. of Chicago Press, IL, 1988.

[2] P. Constantin, C. Foias, B. Nikolaenko, R. Temam, *Spectral barriers and inertial manifolds for dissipative partial differential equations*, J. Dynamics Differential Equations, **1** (1989), 45–73.

[3] P. Constantin, C. Foiaş, R. Temam, *Attractors representing turbulent flows,* Mem. of AMS, **53** (1985), 314, AMS, Providence,USA.

[4] A. Debussche, M. Marion, *On the construction of families of approximate inertial manifolds*, J. Diff. Eqns., **100** (1992), 173–201.

[5] F. Demengel, J.M. Ghidaglia, *Some remarks of the smoothness of the inertial manifolds*, J. Math. Anal. Appl., **155** (1991), 177–225.

[6] C. Devulder, Martine Marion, *A class of numerical algorithms for large time integration: the nonlinear Galerkin methods*, SIAM J. Numer. Anal., **29** (1992), 462–483.

[7] C. Devulder, M. Marion, E.S.Titi, *On the rate of convergence of Nonlinear Galerkin methods,* Math. Comp. **60** (1993), 495–515.

[8] A. Debussche, R. Temam, *Inertial manifolds and the slow manifolds in meteorology,* Diff. Int. Eqns., **4** (1991), 897–931.

[9] A. Debussche, R. Temam, *Inertial manifolds and their dimensions*, Dynamical Systems, Theory and Applications, S.I. Andersson, A. E. Andersson, O. Ottoson (Eds), World Scientific, Singapore, 1993.

[10] C. Foiaş, O. Manley, R. Temam, *Modelling of the interactions of the small and large eddies in two dimensional turbulent flows,* Math. Modelling and Num. Anal., **22** (1988), 93–114.

[11] C. Foiaş, G. R. Sell, R. Temam, *Variétés inertielles des équations differentielles,* C.R. Acad. Sci., Ser. I, **301** (1985), 139–141.

[12] C. Foias, G. R. Sell, R. Temam, *Inertial manifolds for nonlinear evolutionary equations,* J. Differential Equations, **73** (1988), 309–353.

[13] C. Foiaş, G. R. Sell, E. S. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations,* Journal of Differential Equations, **1** (1989), 199–244.

[14] B. Garcia-Archilla, Julia Novo, E. S. Titi, *Postprocessing the Galerkin method: a novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., **35** (1998), 941–972.

[15] B. Garcia-Archilla, Julia Novo, E. S. Titi, *An approximate inertial manifolds approach to postprocessing the Galerkin method for the Navier-Stokes equation*, Mathematics of Computation, **68**(1999), 893–911.

[16] B. Garcia-Archilla, E. S. Titi, *Postprocessing Galerkin methods: The finite element case*, SIAM J. Numer. Anal., **37** (2000), 477–490.

[17] Adelina Georgescu, *Hydrodynamic stability theory*, Martinus Nijhoff Publ., Dordrecht, 1985.

[18] D. Henry, *Geometric theory of semilinear parabolic equations,* Springer, Berlin, 1991.

[19] Anca-Veronica Ion, *An improvement of an inequality concerning the solution of the two-dimensional Navier-Stokes equations*, to be published in ROMAI Journal, **3**, 2 (2007).

[20] Anca-Veronica Ion, *A new modified Galerkin method for the two-dimensional Navier-Stokes equations*, to be published.

[21] Anca-Veronica Ion, *A new modified Galerkin method in the study of a reaction-diffusion equation*, Works of the Middle-Volga Mathemetical Society, 1, **X** (2008), 96–106.

[22] Anca-Veronica Ion, *On the accuracy of the repeatedly adjusted and post-processed Galerkin method for a reaction-diffusion equation*, to be published.

[23] D.A. Jones, L. G. Margolin, E. S. Titi, *On the effectiveness of the inertial manifold-a computational study*, Theoret. Comput. Fluid Dynamics, **7** (1995), 243–260.

[24] M.S. Jolly, R. Rosa, R. Temam, *Accurate computations on inertial manifolds,* SIAM J. Sci. Comp., **22** (2001), 2216–2238.

[25] G. J. Lord, *Attractors and inertial manifolds for finite-difference approximations of the complex Guinsburg-Landau equations,* SIAM J. Numer. Anal., **34** (1997), 1483–1512.

[26] M. Luskin, G. R. Sell, *Approximation theories for inertial manifolds*, J. Model. And Numer. Anal., **3** (1989), 445–461.

[27] M. Marion, *Approximate inertial manifolds for reaction-diffusion equations in high space dimension*, J. Dynamics Differential Equations, **1** (1989), 245–267.

[28] M. Marion, *Approximate inertial manifolds for a pattern formation Cahn-Hilliard equation*, Math. Model. And Numer. Anal., **23** (1989), 463–488.

[29] M. Marion, R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., **26** (5) (1989), 1139–1157.

[30] M. Marion, R. Temam, *Nonlinear Galerkin methods: the finite element case*, Numer. Math., **57** (1990), 1–22.

[31] L. Margolin, E. S. Titi, S. Wynne, *The postprocessing Galerkin and nonlinear Galerkin methods – a truncation analysis point of view,* SIAM J. Numer. Anal., **41** (2003), 695–714.

[32] Julia Novo, E.S. Titi, S. Wynne, *Efficient methods using high accuracy approximate inertial manifolds*, Numer. Math., **87** (3) (2001), 523–554.

[33] K. Promislow, R. Temam, M. Marion, *Localization and approximation for attractors for the Guinzburg-Landau equation,* J, Dynamics Differential Equations, **3** (1991), 491–514.

[34] J. C.Robinson, *Infinite-dimensional dynamical systems; An introduction to dissipative parabolic PDEs and the theory of global attractors,* Cambridge University Press, 2001.

[35] R. Rosa, *Approximate inertial manifolds of exponential order,* Discrete and Continuous Dynamical Systems, **1** (1995), 421–449.

[36] R. Rosa, R. Temam, *Inertial manifolds and normal hyperbolicity*, Acta Applicandae Mathematicae, **45** (1996), 1–50.

[37] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics*, Applied Mathematical Sciences, **68**, Springer, Berlin, 1988.

[38] R. Temam, *Induced trajectories and approximate inertial manifolds*, Math. Mod. Num. Anal., **23** (1989), 541–561.

[39] R. Temam, *Attractors for the Navier-Stokes equations, localization and approximation,* J. Fac. Sci. Univ. Tokyo, Soc. IA, Math., **36** (1989), 629–647.

[40] R. Temam, *Navier-Stokes equations and nonlinear functional analysis,* CBMS-NSF Reg. Conf. Ser. in Appl. Math., SIAM, 1995.

Topics in Applied Mathematics & Mathematical Physics

Diffusion Processes. Physical Models and Numerical Approximation

by Stelian Ion [1]

Contents

[1]"Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: istelian@ima.ro.
The paper was supported by CEEX Grant CEEX05-D11-25/2003.

1. Introduction

We report some mathematical results on the numerical approximation of a class of nonlinear diffusion problems. We are concerned with the convection-diffusion-reaction equation (CDRE)

$$\frac{\partial b(u)}{\partial t} - \operatorname{div}\left(\kappa(u)\nabla u + \boldsymbol{f}(u)\right) = g(t,x,u), \tag{1.1}$$

and generalized porous medium equation (GPME),

$$\frac{\partial u}{\partial t} - \triangle\phi(u) = r(u), \tag{1.2}$$

where div and ∇ are taken with respect to $x \in \mathbb{R}^n$; $\triangle = \operatorname{div}\nabla$ is the Laplace operator and $u(t,x)$ is the scalar unknown function.

There are some reasons to work with two different equations. The both equations quantify diffusion phenomena but in different manner. The diffusion flux is modeled by $\kappa(u)\nabla u$ in the CDRE and by $\operatorname{grad}\phi(u)$ in the GPME. In some cases the two forms can be interchanged but in other cases is not possible. For example, if $\kappa(\cdot)$ is an integrable function one can put $\phi(u) = \int^u \kappa(s)\mathrm{d}s$. Although in almost any physicaly interesting cases this transformation can be done the calculation of the function ϕ, especially when one deals with numerical approximation, can be a hard problem. In such a case is recomandable to use the CDRE form. On the other hand if $\phi(\cdot)$ is a differentiable function one has $\kappa(u) = \phi'(u)$. If $\phi(\cdot)$ is onlya continuous function it is not posible to evaluate the diffusion coefficient.

The outline of the paper follows.

In Section 2 we delineate some mechanical problems and we will make comments on the constitutive functions.

In Section 3 we present the essential facts relative to solvability of the Cauchy problem. We revise the concepts of weak solution and weak entropy solution and we will present a comparison criterion.

Section 4 is devoted to the numerical approximation.

The numerical solution of the Cauchy problem is obtained in two steps. In the first step a system of ordinary differential equation is set up and in the second step this ODE system is numericaly integrated.

The mathematical properties of the ODE model are strongly determined by the numerical diffusion flux and the numerical convective flux. We will define a numerical approximation of the diffusion flux and a numerical approximation of the convective flux that lead to a quasimonotone ODE system. Using

this property we will show that there exists a comparison principle and we will provide the bounds for the solutions of the discrete model that are independent of the mesh size of triangulation.

In Section 5 we give two numerical algorithms to solve GPME equation and Richards' equation respectively. To integrate the ODE system which approximate the GPME equation we will use implicit Euler method and we we will setup an iterative algorithm to solve the system of nonlinear algebraic equation that results.

To solve Richards' equation we use an adaptive time marching scheme and an inexact Newton type method to solve nonlinear equation.

2. Physical Models

The mathematical models (1.1) and (1.2) cover a wide range of physical phenomena such that: heat transfer, infiltration of water through porous media, transport of contaminant in porous media, the flow of the gas through porous media, plasma radiation, to remaind a few.

The simplest example of the model problem (1.1) is the linear caloric equation:

$$\frac{\partial u}{\partial t} = \operatorname{div}(\kappa \nabla u), \tag{2.1}$$

where u models the temperature and $\kappa > 0$ represents the thermal conductivity. Here it is supposed that the caloric flux obeys the Fourier law $q = -\kappa \nabla T$ and that the thermal conductivity is independent of temperature. The condition $\kappa > 0$ reflects the fact that heat propagates from high to lower temperature.

If the temperature of the body is high enough one must consider the radiation effects and the temperature dependence of thermal conductivity. For example, if the power radiated by a body to environment follows the Stefan-Boltzmann law of the forth powers, for both the body and the medium, the heat equation becomes [8]

$$\frac{\partial u}{\partial t} = \operatorname{div}(\kappa(u) \nabla u) - k_r(u^4 - u_e^4). \tag{2.2}$$

The unsaturated water flow through porous media is described by the well known Richards' equations [7]

$$\frac{\partial \theta(h)}{\partial t} - \operatorname{div}(K(h) \nabla h + \boldsymbol{e}_3 K(h)) = 0, \tag{2.3}$$

where θ represents the relative volumetric water content, h represents the pressure head, K is the hydraulic conductivity and $\boldsymbol{e}_3$ is the upward vertical versor. The function $\theta(h)$ is a continuous positive function and it is strictly increasing function on the interval $(-\infty, 0]$ and a constant function on $h > 0$. Also the hydraulic conductivity is a continuous positive function strictly increasing on $(-\infty, 0]$ and a constant function on the set $h > 0$. The hydraulic conductivity becomes zero as h approaches $-\infty$.

The transport of contaminant in porous media is governed by an equation of the form [9], [10]

$$\frac{\partial \left(C + \lambda C^p\right)}{\partial t} + \boldsymbol{v} \cdot \nabla C = \operatorname{div}(D \nabla C) + g(x, C), \tag{2.4}$$

where C represents the mass concentration of the contaminant, $\boldsymbol{v}$ denotes the velocity of the fluid flow, supposed to be constant. The term $\lambda C^p,\ \lambda \geq 0$ takes into account the adsorption reaction by means of Freundlich isotherm. The absorption reaction is described by the term $g(x, C)$ that usually is given by

$$g = -\alpha C^q \tag{2.5}$$

with $\alpha > 0$, $q > 0$ (the order of the reaction).

An extremely used form of the GPME is given by the

$$\frac{\partial u}{\partial t} = \triangle u^m + \lambda u^r. \tag{2.6}$$

For $m > 1$ (slow diffusion) the equation models the flow of the gas through porous medium for $m < 1$ (fast diffusion) the model is encountered in plasma physics, kinetic theory and solid state.

The Stefan problem can be written as a GPME equation with

$$\phi(u) = \lambda \begin{cases} \max\{0, (u-1)\}, & \text{if } u \geq 0, \\ u, & \text{if } u < 0. \end{cases}$$

3. Mathematical Settings

In this section we review some results concerning the solution of the nonlinear diffusion equations.

The constitutive functions are supposed to satisfy:

A1
$$\left|\left| \begin{array}{ll} b:\mathbb{R}\to\mathbb{R}, & \text{is a continuous and nondecreasing function,} \\ \kappa:\mathbb{R}\to\mathbb{R}_+, & \text{is a continuous and nondecreasing function,} \\ f:\mathbb{R}\to\mathbb{R}^n, & \text{is a local Lipschitz vector function,} \\ g:\mathbb{R}_+\times\Omega\times\mathbb{R}\to\mathbb{R}, & \text{is a Caratheodory function.} \end{array} \right.$$

A2
$$\left|\left| \begin{array}{l} \phi\in C([0,\infty))\cup C^1((0,\infty)),\ \phi(0)=0,\ \text{nodecreasing function,} \\ r\in C([0,\infty)),\ r(0)=0. \end{array} \right.$$

We consider the Cauchy problem for both equations. The domain Ω on which the problem is considered satisfies:

A3 $\left|\left| \ \Omega\in\mathbb{R}^n,\ \text{is an open, bounded and connected set.}\right.$

The initial conditions and boundary data are written as

$$\left\{ \begin{array}{ll} u(0,x)=u_0(x), & x\in\Omega. \\ u=u_D, & t>0, x\in\partial\Omega. \end{array} \right. \tag{3.1}$$

We assume that

A4
$$\left|\left| \begin{array}{l} u_0\in L^\infty(\Omega), \\ u_D\in L^2((0,T):W^{1,2}(\Omega))\cap L^\infty((0,T)\times\Omega). \end{array} \right.$$

Cauchy problem for CDRE. The Cauchy problem is defined by the equation (1.1) in a domain Ω in $\mathbb{R}^n$, the initial condition and boundary data (3.1).

Due to the nonlinear parabolic term $b(u)$ and nonlinear diffusion coefficient $\kappa(u)$ the problem (1.1) can be a degenerate problem and consequently there exists no classical solutions.

The notion of *weak solution* for the problem of the type (1.1) was introduced by Alt and Luckhaus in [1]. By imposing some proper conditions on the constitutive functions, boundary data and initial conditions, the authors were able to prove the existence of the weak solution in the case of the parabolic-elliptic degeneration, $b(u)$ is a constant function on some interval of positive measure and the diffusion coefficient is a strict positive function.

DEFINITION **3.1** (Weak Solution (H. W. Alt and S. Luckhaus)) *A measurable function u is a weak solution of the Cauchy problem* (1.1) *and* (3.1) *if it satisfies:*

1) $u-u_D\in L^2((0,T):W_0^{1,2}(\Omega))$,

2) $b(u)\in L^\infty((0,T):L^1(\Omega))$ *and* $\dfrac{\partial b(u)}{\partial t}\in L^2((0,T):W^{-1,2}(\Omega))$ *with initial*

values $b(u_0)$, that is,

$$\int_0^T \left\langle \frac{\partial b(u)}{\partial t}, v \right\rangle \mathrm{d}t + \int_0^T \int_\Omega (b(u) - b(u_0)) \frac{\partial v}{\partial t} \mathrm{d}x\mathrm{d}t = 0 \tag{3.2}$$

for every $v \in L^2((0,T) : W_0^{1,2}(\Omega)) \cap W^{1,1}((0,T) : L^1(\Omega))$, $v(T,\cdot) \equiv 0$

3) $\kappa(u)\nabla u, g(\cdot,\cdot,u(\cdot,\cdot)) \in L^2((0,T)\times\Omega), f(u) \in \left(L^2((0,T)\times\Omega)\right)^n$ and u satisfies the differential equation, that is,

$$\int_0^T \left\langle \frac{\partial b(u)}{\partial t}, v \right\rangle \mathrm{d}t + \int_0^T \int_\Omega (\kappa(u)\nabla u + f(u)) \cdot \nabla v \mathrm{d}x\mathrm{d}t = \int_0^T \int_\Omega g(t,x,u) v \mathrm{d}x\mathrm{d}t \tag{3.3}$$

for every test function $v \in L^2(0,T : W_0^{1,2}(\Omega))$.

In the paper [7] Carrillo extrapolates the concept of entropy solution introduced by Kruzhkov in theory of hyperbolic PDE [14]. He showed that there exists a unique *weak entropy solution* of the Cauchy problem with homogeneous boundary data, $u_D = 0$, even in the case of parabolic-hyperbolic degeneration. Such kind of degeneration appears when the diffusion coefficient is a null function on some interval with the positive measure.

The weak entropy solution is a weak solution that in addition satisfies an integral entropy inequality.

Let us introduce the function

$$K(u) = \int_0^u \kappa(s)\mathrm{d}s,$$

DEFINITION **3.2** (Weak entropy solution. Homogeneous case (Carrillo)) *An weak entropy solution of the Cauchy problem* (1.1) *and* (3.1) *with $u_D = 0$, is*

a weak solution which in addition satisfies the entropy inequality

$$\begin{aligned}
&\int_0^T\int_\Omega H_0(u-s)\Big((\nabla K(u)+f(u)-f(s))\cdot\nabla v-\\
&\qquad -(b(u)-b(s))\frac{\partial v}{\partial t}-gv\Big)\mathrm{d}x\mathrm{d}t-\int_\Omega (b(u_0)-b(s))^+v(0)\mathrm{d}x\le 0,\\
&\int_0^T\int_\Omega H_0(-s-u)\Big((\nabla K(u)+f(u)-f(-s))\cdot\nabla v-\\
&\qquad -(b(u)-b(-s))\frac{\partial v}{\partial t}-gv\Big)\mathrm{d}x\mathrm{d}t-\int_\Omega (b(u_0)-b(-s))^-v(0)\mathrm{d}x\ge 0,
\end{aligned}\tag{3.4}$$

for any $(s,v)\in\mathbb{R}\times\big(L^2((0,T):W^{1,2}(\Omega))\cap W^{1,1}((0,t):L^\infty(\Omega))\big)$ *such that* $s\ge 0, v\ge 0$ *and* $v(T)=0$.

In the entropy conditions the following notations:

$$H_0(s)=\begin{cases}1, \text{ if } s>0\\ 0, \text{ if } s\le 0\end{cases}\qquad s^+=\begin{cases}s, \text{ if } s>0\\ 0, \text{ if } s\le 0\end{cases}$$

were used. If $\kappa>0$ then the two definitions of the weak solution coincide and any weak solution is an entropy solution [7].

To deal with nonhomogeneous Dirichlet conditions for degenerate problem one supplementary difficulty is to give a sense to boundary conditions. In the paper [18] C. Mascia, A. Porreta and A. Terracina proved the existence of the weak entropy solution of the Cauchy problem with nonhomogeneous Dirichlet data. Their definition is as follows. Denote by Q_T the direct product $Q_T=(0,T)\times\Omega$. Also we use the notations:

$$\begin{aligned}
\mathcal{E}(u,v)&=\nabla\,|K(u)-K(v)|+\operatorname{sgn}(u-v)(f(u)-f(v)),\\
\mathcal{B}(u,v,w)&=\mathcal{E}(u,v)+\mathcal{E}(u,w)-\mathcal{E}(v,w).
\end{aligned}$$

The domain Ω is such that there exists a C^2–covering of $\partial\Omega$, $\mathcal{A}=\{U_i\}_{i=1,m}$, of open sets such that $\partial\Omega\subset\cup\overline{U}_i$ and, in some local coordinates $x=(x',x_n)$, there exists a C^2 function $x_n=\alpha_i(x')$ such that $U_i\cap\partial\Omega=\{x_n=\alpha_i(x')\}$, $U_i\cap\Omega=\{x_n<\alpha_i(x'\}$.

A sequence $\{\vartheta_\delta\}$ of $C^2(\Omega)\cap C^0(\overline{\Omega})$ functions is named a boundary layer sequence if

$$\lim_{\delta\to 0^+}\vartheta_\delta=1,\ \text{pointwise in }\Omega,\ 0\le\vartheta_\delta\le 1,\ \vartheta_\delta=0\text{ on }\partial\Omega.$$

DEFINITION **3.3** (Weak Entropy Solution. Nonhomogeneous case (Mascia et al.)) *A function $u \in L^\infty((0,T)\times\Omega)$ is an entropy solution of Cauchy problem* (1.1) *and* (3.1) *if*

1) (regularity)

$$K(u) \in L^2((0,T) : W^{1,2}(\Omega))$$

and for any $U \in \mathcal{A}$, and any positive $\psi \in C_0^\infty(U)$ we have

$$\Big(- |u - u_D|\psi, \mathcal{E}(u, u_D)\psi \Big) \in \mathcal{DM}(Q)_2,$$

where $\mathcal{DM}(Q)_2$ is the set of divergence-measure vector fields in Q.

2) (entropy condition in interior of Q_T)

$$\int\limits_{Q_T} \left\{ |b(u) - b(s)| \, \frac{\partial v}{\partial t} - \mathcal{E}(u,s)\nabla v + gv \right\} \mathrm{d}x\mathrm{d}t \geq 0$$

for any $v \in W_0^{1,2}(Q_T)$ and $v \geq 0$ and $s \in \mathbb{R}$.

3) (initial condition)

$$\lim_{t\to 0^+} \int\limits_\Omega |u(t,x) - u_0(x)| \, \mathrm{d}x = 0$$

4) (boundary conditions) *in sense of trace in $L^2((0,T) : W^{1,2}(\Omega))$ we have*

$$K(u) = K(u_D), \quad t > 0, \ x \in \partial\Omega,$$

and for any boundary layer sequence ϑ_δ, and for any $U \in \mathcal{A}$, and any positive $\psi \in C_0^\infty(U)$ we have

$$\liminf_{\delta\to 0} \int\limits_{Q_T} \mathcal{B}(u,s,u_D)\nabla\vartheta_\delta \xi\psi \mathrm{d}x\mathrm{d}t \geq 0, \ \forall s \in \mathbb{R},$$

for any $\xi \in L^2((0,T) : W^{1,2}(\Omega)), \xi \geq 0$.

Cauchy problem for GPME. The Cauchy problem consists in the equation (1.2) and the data (3.1).

The existence of the weak solution was proved by many authors see for example, [4], [25].

DEFINITION **3.4** (M. Borelli and M. Ughi) *A nonnegative function u defined on the $\overline{\Omega} \times [0,T]$ is said to be a* weak solution *of the Cauchy problem* (1.2) *and* (3.1) *if*

1) $u \in C\left([0,T]; L^1(\Omega)\right) \cap L^\infty([0,T] \times \Omega)$,

2) *for any test function $\eta \in C^{1,0}([0,T] \times \overline{\Omega}) \cap C^{2,1}((0,T] \times \Omega)$ such that $\eta \geq 0$ on $(0,T] \times \Omega$ and $\eta = 0$ on $(0,T] \times \partial\Omega$ u satisfies the integral identity:*

$$\begin{aligned} \int_\Omega u(t,x)\eta(t,x)\mathrm{d}x = & \int_\Omega u_0(x)\eta(0,x)\mathrm{d}x - \int_0^t \int_{\partial\Omega} \phi(u_D)\frac{\partial\eta}{\partial n} + \\ & + \int_0^t \int_\Omega \left[u\partial_t\eta + \phi(u)\triangle\eta + r(u)\eta\right] \mathrm{d}t\mathrm{d}x \end{aligned} \tag{3.5}$$

for any $0 \leq t \leq T$.

The presence of the reaction term and nonlinearity in the equation (1.2) generate interesting phenomena namely, extinction time or blow up of the solution and the finite speed of propagation of disturbance [25].

Such problems have been studied by several authors: Borelli-Ughi [4], Ferreira-Vasquez [13], Leoni [16], Levin-Sacks [17], Peletier and Z. Junninig [23]. In the case $r(u) = 0$ and $\phi(s) = s^m, 0 < m < 1, u_D = 0$ there exists an extinction time T_e such that the problem (1.2 has a unique classical solution, positive on $\Omega \times [0,T_e]$ and null for $t \geq T_e$, see [17].

For generalized fast diffusion with strong absorption and $\Omega = \mathbb{R}^2$ there also exists an extinction time and the support of the solution is bounded for any time $t > 0$, [4].

In the power case, $\phi(s) = s^m, r(s) = \lambda p^s, \lambda > 0$, the numerical methods to compute the solution of the similar problem (1.2) have been proposed by M.-N. Le Roux, [21] the case $m > 1$, M.-N. Le Roux and P.-E. Mainge, [22].

Pointwise comparison principle. For both Cauchy problems CDRE and GPME there exists several comparison criteria [1], [10], [25]. We will give here a result that allows one to compare two solutions with respect to their boundary and initial conditions.

For any two real functions $f(x)$ and $g(x)$ we write $f \leq g$ if $f(x) \leq g(x), \forall x \in \Omega$. In addition to assumptions **A1** the constitutive functions in CDRE problem satisfy

$$\mathbf{A1'}\left\|\begin{array}{l}(1)\,\kappa:\mathbb{R}\to\mathbb{R}_+,\ \kappa(u)\geq\eta,\\ (2)\,|\kappa(u_1)-\kappa(u_2)|<C|u_1-u_2|^{\gamma_1},\ \gamma_1>\dfrac{1}{2},\ \forall u_1,u_2\in\mathbb{R},\\ (3)\,|\boldsymbol{f}(u_1)-\boldsymbol{f}(u_2)|<C|u_1-u_2|^{\gamma_2},\ \gamma_2>\dfrac{1}{2},\ \forall u_1,u_2\in\mathbb{R},\\ (4)\,g(u_1)-g(u_2)<C(b(u_1)-b(u_2)),\ \text{for } u_1>u_2.\end{array}\right.$$

THEOREM **3.1** (Comparison Theorem) *Let (u_D,u_0), $(\widehat{u}_D,\widehat{u}_0)$ be such that $u_D\leq\widehat{u}_D, u_0\leq\widehat{u}_0$. Let u and $\widehat{u}$) be two bounded weak solutions of the Cauchy problem* (1.1), (3.1) *associated to (u_D,u_0) and $(\widehat{u}_D,\widehat{u}_0)$ respectively. Assume, in addition, that*

$$b(u)_t, b(\widehat{u})_t\in L^1((0,T)\times\Omega).$$

Then

$$u\leq\widehat{u}$$

on $(0,T)\times\Omega$.

Proof. We follow the main ideas from [1]. As in [1] for any $\delta>0$ let $\Psi_\delta(\alpha)=\min(1,\max(0,\alpha/\delta))$. The function $w=\Psi_\delta(u-\widehat{u})$ belongs to $L^2(0,T:W_0^{1,2}(\Omega))$ and its gradient is given by

$$\nabla w=\left\{\begin{array}{ll}\dfrac{1}{\delta}\left(\nabla u-\nabla\widehat{u}\right), & \text{if } 0<u-\widehat{u}<\delta\\ 0, & \text{otherwise}\end{array}\right.$$

Set w as test function in (3.3). Then

$$\begin{aligned}&\int_0^t\int_\Omega\left(b(u)_t-b(\widehat{u})_t\right)w\mathrm{d}x\mathrm{d}t+\underbrace{\frac{1}{\delta}\int_0^t\int_{\Omega_\delta}\left(\kappa(u)\nabla u-\kappa(\widehat{u})\nabla\widehat{u}\right)\nabla(u-\widehat{u})\mathrm{d}x\mathrm{d}t}_{I_1}+\\ &+\underbrace{\frac{1}{\delta}\int_0^t\int_{\Omega_\delta}\left(\boldsymbol{f}(u)-\boldsymbol{f}(\widehat{u})\right)\cdot\nabla(u-\widehat{u})\mathrm{d}x\mathrm{d}t}_{I_2}=\int_0^t\int_\Omega\left(g(u)-g(\widehat{u})\right)w\mathrm{d}x\mathrm{d}t,\end{aligned}\tag{3.6}$$

where $\Omega_\delta:=\{x|0<h-\widehat{h}<\delta\}$. The integral I_1 can be rewritten as

$$I_1=\int_0^t\int_{\Omega_\delta}\kappa(u)||\nabla(u-\widehat{u})||^2\mathrm{d}x\mathrm{d}t+\int_0^t\int_{\Omega_\delta}(\kappa(u)-\kappa(\widehat{u}))\nabla\widetilde{u}\cdot\nabla(u-\widehat{u})\mathrm{d}x\mathrm{d}t.$$

Using Young inequality, $ab \leq C(\epsilon)p^{-1}a^p + \epsilon q^{-1}b^q$, and **A1′**-(1) we obtain

$$I_1 \geq \left(\eta - \frac{\epsilon}{2}\right) \int_0^t \int_{\Omega_\delta} ||\nabla(u - \widehat{u})||^2 \mathrm{d}x\mathrm{d}t - \frac{C(\epsilon)}{2} \int_0^t \int_{\Omega_\delta} (\kappa(u) - \kappa(\widehat{u}))^2 ||\nabla \widetilde{u}||^2 \mathrm{d}x\mathrm{d}t$$

and

$$I_2 \geq -\frac{\epsilon}{2} \int_0^t \int_{\Omega_\delta} ||\nabla(u - \widehat{u})||^2 \mathrm{d}x\mathrm{d}t - \frac{C(\epsilon)}{2} \int_0^t \int_{\Omega_\delta} ||\boldsymbol{f}(u) - \boldsymbol{f}(\widehat{u})||^2 \mathrm{d}x\mathrm{d}t.$$

Then

$$I_1 + I_2 \geq (\eta - \epsilon) \int_0^t \int_{\Omega_\delta} ||\nabla(u - \widehat{u})||^2 \mathrm{d}x\mathrm{d}t - C\delta^{2\gamma} \int_0^T \int_{\Omega_\delta} (||\nabla \widetilde{u}||^2 + 1) \mathrm{d}x\mathrm{d}t.$$

From **A1′**(4) the production can be estimate as

$$\int_0^t \int_{\Omega} (g(u) - g(\widehat{u}))\, w \mathrm{d}x\mathrm{d}t \leq \int_0^t \int_{\Omega} 1_{\{u - \widehat{u} > 0\}} \max\{0, g(u) - g(\widehat{u}\} \mathrm{d}x\mathrm{d}t \leq$$

$$\leq C \int_0^t \int_{\Omega} \max\{0, b(u) - b(\widehat{u})\} \mathrm{d}x\mathrm{d}t.$$

Taking $\epsilon < \eta$ we obtain

$$\begin{aligned} \int_0^t \int_{\Omega} \quad & (b(u)_t - b(\widehat{u})_t)\, w \mathrm{d}x\mathrm{d}t + \frac{c}{\delta} \int_0^t \int_{\Omega_\delta} ||\nabla(u - \widehat{u})||^2 \mathrm{d}x\mathrm{d}t \leq \\ & \leq C\delta^{2\gamma - 1} \int_0^T \int_{\Omega_\delta} (||\nabla \widetilde{u}||^2 + 1) \mathrm{d}x\mathrm{d}t + \int_0^t \int_{\Omega} \max\{0, b(u) - b(\widehat{u})\} \mathrm{d}x\mathrm{d}t. \end{aligned} \tag{3.7}$$

For $\delta \to 0$ the first term on the right converge to 0 and the first term on left becomes

$$\lim_{\delta \to 0} \int_0^t \int_{\Omega} (b(u)_t - b(\widehat{u})_t)\, w \mathrm{d}x\mathrm{d}t = \int_0^t \int_{\Omega} 1_{\{u - \widehat{u} > 0\}} (b(u)_t - b(\widehat{u})_t) \mathrm{d}x\mathrm{d}t =$$

$$= \int_0^t \int_\Omega \partial_t \max\{b(u) - b(\widehat{u}), 0\} \mathrm{d}x\mathrm{d}t = \int_\Omega \max\{b(u) - b(\widehat{u}), 0\}(t,x)\mathrm{d}x.$$

One obtains

$$\int_\Omega \max\{0, b(u) - b(\widehat{u})\}\mathrm{d}x\mathrm{d}t \leq \int_0^t \int_\Omega \max\{0, b(u) - b(\widehat{u})\}\mathrm{d}x\mathrm{d}t,$$

and using Gronwall's inequality we get

$$b(u) \leq b(\widehat{u}),$$

and using this inequality in (3.7) we have $\nabla(u - \widehat{u}) = 0$ on the set $\{0 < u - \widehat{u}\}$. So, we have $u - \widehat{u}$=const. which implies $u - \widehat{u} \leq 0$ since on boundary $u \leq \widehat{u}$. As a corollary of the comparison principle one can obtain an upper bound for the solution of Cauchy problems in the both case CDRE and GPME equations.

COROLLARY **3.1** *Assume that* **A1** *and* **A1′** *are fullfiled and* $g(t,x,u) = g(u)$, $g(0) = 0$. *Let* u *be the solution of the problem* (1.1), (3.1) *on some interval* $[0,T]$. *Then*

1) *if* $u_D \geq 0$ *and* $u_0 \geq 0$ *so is* $u \geq 0$,

2) *Let* $\alpha = ||u_D||_{L^\infty([0,T]\times\partial\Omega)}$, $\beta = \max\{||u_0||_\infty, \alpha\}$. *If* $\alpha > 0$ *we assume that* $g(w) \geq 0$. *Let* $w(t)$ *be the solution of the differential equation*

$$\begin{aligned} &\partial_t b(w) = g(w) \\ &w(0) = \beta. \end{aligned}$$

on the same interval $t \in [0,T]$. *Then the solution* u *satisfies*

$$u < w \text{ on } [0,T].$$

Proof. 1). One compares the solution u with the trivial solution $v = 0$.

2). Define the function $v(t,x) = w(t), \forall x \in \overline{\Omega}$. The function $v(t,x)$ verifies the equation (1.1), at the time $t = 0$ $v(0,x) = \beta > u_0$ and on boundary $v(t,x)|_{x\in\partial\Omega} = w(t) \geq \beta > u_D$ that implies $u < v$.

COROLLARY **3.2** *In the GPME the diffusion function and production function are given by* $\phi(u) = u^m$, $r(u) = -\lambda u^s$ *respectively* $\lambda > 0, m > 0, s > 0$. *The initial conditions satisfy* **A4**, $u_0 > 0$ *and* $u_D = 0$. *Let* $\beta = ||u_0||_\infty$.
1) *If* $s > 1$ *then the solution* u *of the problem* 1.2, 3.1 *satisfies*

$$||u||_\infty < \beta \left(1 - \lambda(1-s)\beta^{s-1}t\right)^{\frac{1}{1-s}}.$$

2) *If $s < 1$ then there exists a time T^*, extinction time, given by*

$$T^* = \frac{1}{\lambda}\frac{\beta^{1-s}}{1-s}$$

such that the solution exists on the interval $[0, T^*]$ and it satisfies

$$||u||_\infty < \beta\left(1 - \frac{t}{T^*}\right)^{\frac{1}{1-s}}.$$

Proof. In the generalized porous medium equation

$$\partial_t u = \triangle u^m - \lambda u^s$$

we make the substitution $u^m = v$ and we obtain

$$\partial_t v^p = \triangle v - \lambda v^r,$$

$$v_{t=0} = u_0^m, v|_{x\in\partial\Omega} = 0,$$

where $p = 1/m, r = s/m$. By using the corollary 1 one obtain that the function v is bounded from above by the solution of differential equation

$$\begin{aligned} &pw^{p-1}w' = -\lambda w^r, \\ &w(0) = \beta^m, \end{aligned}$$

which has the solution

$$w = \beta^m(1 - \lambda(1-s)\beta^{s-1}t)^{\frac{m}{1-s}}.$$

4. Quasimonotone ODE Approximation

4.1. Discrete Approximation

By the method of lines (MOL), one can associate an ordinary differential system of equations (ODE) to a parabolic partial differential equation. The MOL consists in the discretization of the space variable using one of the standard methods as finite element, finite differences or finite-volume method (FVM). The FVM fits very well to conservative equations and there exists a large literature devoted to the method, we recall here the papers that deal with Dirichlet problem, [6] for hyperbolic PDE, [11], [12], [19] for nonlinear parabolic PDE.

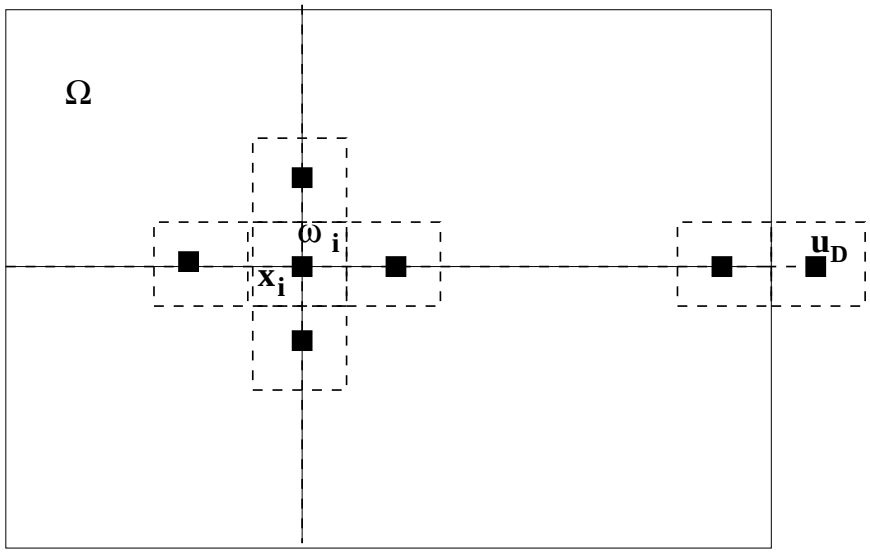


Fig. 1: Triangulation of polygonal domain in $\mathbb{R}^2$.

The FVM deals with a decomposition of the domain Ω into small polygonal domains ω_i and a net of the inner knots x_i. The assembly $\{\omega_i, x_i\}$ defines a triangulation of the domain and it is an admissible mesh if it satisfies the following conditions, [12].

DEFINITION **4.1** (Admissible mesh) *The triangulation* $\mathcal{T} = \{(\omega_i, x_i)\}_{i\in I}$ *is called an admissible mesh if it satisfies:*

A5 ω_i *is open poligonal set* $\subseteq \Omega$, $x_i \in \overline{\omega}_i$
(1) $\overline{\cup_{i\in I}\omega_i} = \overline{\Omega}$
(2) $\forall i \neq j \in I$ *and* $\overline{\omega}_i \cap \overline{\omega}_j \neq \Phi$, *either* $\mathcal{H}_{n-1}(\overline{\omega}_i \cap \overline{\omega}_j) = 0$ *or* $\sigma_{ij} := \overline{\omega}_i \cap \overline{\omega}_j$ *is a common* $(n-1)$-face of ω_i and ω_j
(3) *for all* σ_{ij}, $\quad [x_i, x_j] \perp \sigma_{ij}$

Here $\mathcal{H}_{n-1}$ is the $(n-1)$-dimensional Hausdorff measure. For each volume ω_i that has a common $(n-1)$-face with the boundary $\partial\Omega$ one defines an external volume $\omega_{i_b} \in C\Omega$ by the reflection of the ω_i with respect to the face $\sigma_{i_b} = \omega_i \cap \partial\Omega$. Denotes by $\mathcal{T}^b$ the collection of all external volumes $\{(\omega_{i_b}, x_{i_b})\}$ and by I^b the set of their indices. Let $\mathcal{T}^e = \mathcal{T} \cup \mathcal{T}^b$ and $I^E = I \cup I^b$. We say that the volumes ω_i, $\omega_j \in \mathcal{T}^e$ are neighbours if they share a common $n-1$–face and we denote by $\boldsymbol{n}_{i,j}$ the unit normal vector to the face σ_{ij} that point to ω_j.

Discrete form of CDRE. The space discretized equations are derived from the integral form of (1.1) for each control volume ω_i

$$\int_{\omega_i} \frac{\partial b(u)}{\partial t} \mathrm{d}x - \int_{\partial\omega_i} (\kappa(u)\nabla u + \boldsymbol{f}(u)) \cdot \boldsymbol{n} \mathrm{d}a = \int_{\omega_i} g(t,x,u)\mathrm{d}x, \ \forall i \in I. \quad (4.1)$$

By a proper approximation of the volume integrals and surface integrals one obtains discrete form of CDRE.

We define the numerical diffusion coefficient $\widetilde{\kappa} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ by

$$\widetilde{\kappa}(u,v) = \max(\kappa(u), \kappa(v)). \quad (4.2)$$

It is easy to show that numerical diffusion coefficient satisfies

A6
$$\left\| \begin{array}{ll} \widetilde{\kappa}(u,v) = \widetilde{\kappa}(v,u), & \text{symmetry}, \\ (\widetilde{\kappa}(u_1,v) - \widetilde{\kappa}(u_2,v))(u_1-u_2) > 0, & \text{monotonicity}, \\ \widetilde{\kappa}(u,u) = \kappa(u), & \text{consistency}. \end{array} \right.$$

Corresponding to each face σ_{ij} we admit that there exists a numerical flux function $\widetilde{f} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with the following properties:

A7
$$\left\| \begin{array}{ll} \widetilde{f}_{i,j}(u,v) = -\widetilde{f}_{j,i}(v,u), & \text{conservation}, \\ (\widetilde{f}_{i,j}(u_1,v) - \widetilde{f}_{i,j}(u_2,v))(u_1-u_2) \le 0, & \text{monotonicity}, \\ (\widetilde{f}_{i,j}(u,v_1) - \widetilde{f}_{i,j}(u,v_2))(v_1-v_2) \ge 0, & \\ \widetilde{f}_{i,j}(u,u) = \boldsymbol{f}(u) \cdot \boldsymbol{n}_{i,j}, & \text{consistency}. \end{array} \right.$$

A numerical convective flux which satisfies **A7** is systematicaly used in the approximation of hyperbolic equation see, for example [6]. The integrals in (4.1) will be approximated as follows:

$$\begin{aligned} &\int_{\omega_i} \frac{\partial b(u)}{\partial t} \mathrm{d}x \approx \mathrm{m}(\omega_i) \frac{\partial b(u_i)}{\partial t}, \\ &\int_{\partial\omega_i} \kappa(u)\nabla u \cdot \boldsymbol{n} \mathrm{d}a \approx \sum_{j \in \mathcal{N}(i)} \mathrm{m}(\sigma_{ij}) \widetilde{\kappa}(u_i,u_j) \frac{u_j - u_i}{d_{ij}}, \\ &\int_{\partial\omega_i} \boldsymbol{f}(u) \cdot \boldsymbol{n} \mathrm{d}a \approx \sum_{j \in \mathcal{N}(i)} \widetilde{f}_{i,j}(u_i,u_j), \\ &\int_{\omega_i} g(t,x,u)\mathrm{d}x \approx \int_{\omega_i} g(t,x,u_i)\mathrm{d}x := g_i(t,u_i). \end{aligned}$$

$\mathcal{N}(i)$ denotes all neighbours in $\mathcal{T}^e$ of ω_i, $\mathrm{m}(\omega_i)$ represents the volume of polyhedron ω_i and $\mathrm{m}(\sigma_{ij})$ represents the $n-1$-dimensional measure of the face σ_{ij} and $d_{i,j} = |x_i - x_j|$.

The initial data and boundary conditions are approximated by:

$$u_{0i} = \frac{1}{\mathrm{m}(\omega_i)} \int\limits_{\omega_i} u_0(x)\mathrm{d}x, \tag{4.3}$$

$$u_{i_b} = \frac{1}{\mathrm{m}(\sigma_{i_b})} \int\limits_{\sigma_{i_b}} u_D \mathrm{d}a, \tag{4.4}$$

respectively.

As a result one can define a Cauchy problem for a system of ordinary differentila equations whose solution gives an approximation of the Cauchy problem (1.1), (3.1).

$$\begin{cases} \dfrac{\mathrm{d}b(u_i)}{\mathrm{d}t} = \displaystyle\sum_{j\in\mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \left[\widetilde{\kappa}(u_i,u_j)\frac{u_j-u_i}{d_{ij}} + \widetilde{f}_{i,j}(u_i,u_j)\right] + g_i(t,u_i) \\ u_i|_{t=0} = u_{0i}, \end{cases} \tag{4.5}$$

for $t > 0$ and for any $i \in I$.

Let us introduce the numerical diffusion-convection flux functions

$$\mathcal{F}_i(\boldsymbol{u};\boldsymbol{u}_D) = \sum_{j\in\mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \left[\widetilde{\kappa}(u_i,u_j)\frac{u_j-u_i}{d_{ij}} + \widetilde{f}_{i,j}(u_i,u_j)\right] \tag{4.6}$$

then the ODE approximation reads as

$$\frac{\mathrm{d}b(u_i)}{\mathrm{d}t} = \mathcal{F}_i(\boldsymbol{u};\boldsymbol{u}_D) + g_i(t,u_i). \tag{4.7}$$

The boundary conditions are taken into account by the volume elements next to boundary $\partial\Omega$. For such element the contribution of the boundary values to the $\mathcal{F}_i$ is given by

$$\frac{\mathrm{m}(\sigma_{i_b})}{\mathrm{m}(\omega_i)} \left[\widetilde{\kappa}(u_{i_b},u_j)\frac{u_{i_b}-u_i}{d_{i_b}} + \widetilde{f}_{i,i_b}(u_i,u_{i_b})\right]. \tag{4.8}$$

Infiltration model. Here is an example of a numerical convective flux that satisfies **A7** with $\boldsymbol{f}(u) = \boldsymbol{e}_3 K(u)$ that appears in the Richards' equation (2.3).

$$\widetilde{f}_{i,j}(u,v) = \frac{1}{2}\left(\boldsymbol{e}_3\cdot\boldsymbol{n}_{i,j} + |\boldsymbol{e}_3\cdot\boldsymbol{n}_{i,j}|\right)K(v) + \frac{1}{2}\left(\boldsymbol{e}_3\cdot\boldsymbol{n}_{i,j} - |\boldsymbol{e}_3\cdot\boldsymbol{n}_{i,j}|\right)K(u). \tag{4.9}$$

Discrete form of GPME. For each control volume ω_i we write

$$\int\limits_{\omega_i} \frac{\partial u}{\partial t} \mathrm{d}x - \int\limits_{\partial\omega_i} \frac{\partial \phi(u)}{\partial n} \mathrm{d}a = \int\limits_{\omega_i} r(u) \mathrm{d}x, \ \forall i \in I. \tag{4.10}$$

To approximate (4.10) we use the same schemes as in previous paragraph. The new integral that contains the diffusion function ϕ will be approximated by

$$\int\limits_{\partial\omega_i} \frac{\partial \phi(u)}{\partial n} \mathrm{d}a \approx \sum_{j \in \mathcal{N}(i)} \mathrm{m}(\sigma_{ij}) \frac{\phi(u_j) - \phi(u_i)}{d_{ij}}. \tag{4.11}$$

The ODE approximation of (4.10) is given by

$$\frac{\partial u_i}{\partial t} = \sum_{j \in \mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \frac{\phi(u_j) - \phi(u_i)}{d_{ij}} + r(u_i). \tag{4.12}$$

The boundary conditions are taken into account by the volume elements next to boundary $\partial\Omega$. For such an element the boundary values enters into the play by a term of the form

$$\frac{\mathrm{m}(\sigma_{i^e})}{\mathrm{m}(\omega_i)} \frac{\phi(u_D^{i_e}) - \phi(u_i)}{d_{ij}^e}. \tag{4.13}$$

For shortness we introduce the notation

$$\mathcal{G}_i = \sum_{j \in \mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \frac{\phi(u_j) - \phi(u_i)}{d_{ij}}.$$

4.2. ODE Model

As in the continuum case we want to prove that the solutions of ODE (4.5) and (4.10) obey a comparison criterion.

For that, we firstly prove that $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$ satisfy Kamke conditions.

LEMMA **4.1** *Assume* **A2**, **A6** *and* **A7**. *Then*

$$\mathcal{F}_i(\boldsymbol{u}^e) = 0, \ \mathcal{G}_i(\boldsymbol{u}^e) = 0 \tag{4.14}$$

for any constant state $u_i = u, \forall i \in I^e$.

$\boldsymbol{\mathcal{F}}$ *and* $\boldsymbol{\mathcal{G}}$ *satisfy Kamke conditions, that is*

$$\mathcal{F}_i(\boldsymbol{v}^e) \geq \mathcal{F}_i(\boldsymbol{w}^e), \ \mathcal{G}_i(\boldsymbol{v}^e) \geq \mathcal{G}_i(\boldsymbol{w}^e), \ \forall i \in I, \tag{4.15}$$

for any two vectors that satisfy $v_k \geq w_k$, $\forall k \in I^e$, *and* $v_i = w_i$.

Proof. To prove (4.14) we have

$$\mathcal{F}_i(\boldsymbol{u}^e) = \sum_{j\in\mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \boldsymbol{f}(u)\cdot \boldsymbol{n}_{ij} = 0.$$

We only prouve the counterpart relativ to $\boldsymbol{\mathcal{F}}$. To prove the Kamke conditions we have

$$\mathcal{F}_i(\boldsymbol{v}^e) - \mathcal{F}_i(\boldsymbol{w}^e) =$$

$$\sum_{j\in\mathcal{N}(i)} \frac{\mathrm{m}(\sigma_{ij})}{\mathrm{m}(\omega_i)} \left[\widetilde{\kappa}(u,v_j)\frac{v_j-u}{d_{ij}} + \widetilde{f}_{i,j}(u,v_j) - \widetilde{\kappa}(u,w_j)\frac{w_j-u}{d_{ij}} - \widetilde{f}_{i,j}(u,w_j)\right]$$

and from (4.2) and the monotonicity property of **A7** the affirmation results.

As $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{G}}$ are both quasimonotone and nondecreasing with respect to boundary data vectorial functions the next two results are equaly true for discrete ODE (4.12).

Assumptions on source term

A1″ There exists the real numbers $\underline{\alpha} < \alpha < \beta < \overline{\beta}$ such that
(1) $b \in C^1((\underline{\alpha},\overline{\beta}))$ and $b' > 0$ on $(\underline{\alpha},\overline{\beta})$.
There exists two Lipschitz functions $\overline{g}, \underline{g} : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ such that
(2) $\underline{g}(t,u) \le g(t,x,u) \le \overline{g}(t,u),\ \forall u \in (\underline{\alpha},\overline{\beta})$,
(3) $\underline{g}(t,\alpha) \le 0,\ \overline{g}(t,\beta \ge 0$.

THEOREM **4.1** (Boundedness of discrete solutions) *Consider the Cauchy problem* (4.5). *Assume* **A1**, **A1″**, **A4**, **A6**, **A7**. *We suppose also that initial conditions and boundary data satisfy*

$$\alpha \le u_0(x) \le \beta, \text{a.e } x \in \Omega, \alpha \le u_D(t,x) \le \beta, \text{ a.e } (t,x) \in (0,T)\times\Omega. \quad (4.16)$$

Let $\underline{u}(t)$ *be the solution of the problem*

$$\begin{cases} \dfrac{\partial b(u)}{\partial t} = \underline{g}(t,u) \\ \\ |u|_{t=0} = \alpha, \end{cases} \quad (4.17)$$

$\overline{u}(t)$ *be the solution of the problem*

$$\begin{cases} \dfrac{\partial b(u)}{\partial t} = \overline{g}(t,u) \\ \\ |u|_{t=0} = \beta \end{cases} \quad (4.18)$$

and $T_{\sup} = \inf(\sup\{t|\underline{u}(t) > \underline{\alpha}, \overline{u}(t) < \overline{\beta}\}, T)$ *Then the solution* $\boldsymbol{u}(t)$ *of the Cauchy problem is bounded by* $\underline{u}$ *and* $\overline{u}$ *on the interval* $[0, T_{\sup}]$ *i.e.,*

$$\underline{u}(t) \leq u_i(t) \leq \overline{u}(t) \, \forall i \in I, \forall t \in [0, T_{\sup}] \tag{4.19}$$

Proof. The essential tool in the proof is the Nickel's theorem that provide the monotony of the solution of the quasimonotone ODE. The Kamke conditions ensure us that we deal with quasimonotone system.

Observe that the conditions **A1**$''$-3 guaranties that

$$\underline{\alpha} \leq \underline{u}(t) \leq \alpha, \; \beta \leq \overline{u}(t) \leq \overline{\beta}. \tag{4.20}$$

Define

$$\underline{\mathcal{F}}_i(\boldsymbol{u}) = \mathcal{F}_i(\boldsymbol{u}; \underline{\boldsymbol{u}}), \overline{\mathcal{F}}_i(\boldsymbol{u}) = \mathcal{F}_i(\boldsymbol{u}; \overline{\boldsymbol{u}}).$$

From (4.4), (4.8), (4.15), (4.20) and the conditions **A1**$'$-2 one obtains

$$\underline{\mathcal{F}}_i(\boldsymbol{u}) + \underline{g}(t,u) \leq \mathcal{F}_i(\boldsymbol{u}; \boldsymbol{u}_D) + g_i(t,u) \leq \overline{\mathcal{F}}_i(\boldsymbol{u}) + \overline{g}(t,u).$$

Since $u_i^{\sup}(t) = \overline{u}(t), \forall i \in I$ is a solution of the problem

$$\begin{cases} \dfrac{\mathrm{d}b(u_i)}{\mathrm{d}t} = \overline{\mathcal{F}}_i(\boldsymbol{u}) + \overline{g}(t, u_i) \\ u_i|_{t=0} = \beta, \end{cases} \tag{4.21}$$

$u_i^{\inf}(t) = \underline{u}(t), \forall i \in I$ is a solution of the problem

$$\begin{cases} \dfrac{\mathrm{d}b(u_i)}{\mathrm{d}t} = \underline{\mathcal{F}}_i(\boldsymbol{u}) + \underline{g}(t, u_i) \\ u_i|_{t=0} = \alpha, \end{cases} \tag{4.22}$$

and $\alpha \leq u_{0i} < \beta$ one can apply the Nickel's theorem and one obtains

$$u_i^{\inf}(t) \leq u_i(t) \leq u_i^{\sup}(t),$$

which is (4.19).

THEOREM **4.2** (Comparison theorem. Discrete case) *Assume we are as in the boundedness theorem. Let* $\boldsymbol{u}(t)$ *and* $\widehat{\boldsymbol{u}}(t)$, $t \in (0,T)$, *be the solutions of the problem* (4.5) *associated to* $(\boldsymbol{u}_D, \boldsymbol{u}_0)$ *and* $(\widehat{\boldsymbol{u}}_D, \widehat{\boldsymbol{u}}_0)$ *respectively. Suppose that*

$$\boldsymbol{u}_D \leq \widehat{\boldsymbol{u}}_D < 0, \; \boldsymbol{u}_0 \leq \widehat{\boldsymbol{u}}_0 < 0.$$

Then

$$\boldsymbol{u} \leq \widehat{\boldsymbol{u}}$$

on $(0,T)$.

Proof. The same as in the boundedness theorem.

5. Numerical Algorithms and Numerical Results

In this section we give two numerical algorithms to solve GPME equation and Richards' equations respectively.

5.1. Fast Diffusion with Strong Absorption

We will present here an algorithm to solve numerically (4.12) in the case of the fast diffusion with strong absorption. In addition to assumptions **A2** the constitutive functions ϕ and r satisfy

A2′ $\left\|\begin{array}{l}\phi \text{ is increasing function and } \lim_{s\to 0}\phi(x)/x=\infty,\\ r(s)\le 0, \text{ for } s>0,\end{array}\right.$

The ODE can be rewritten as

$$\frac{\partial u_i}{\partial t}=A_{ij}\phi(u_j)+r(u_i). \tag{5.1}$$

We use the classical full implicit Euler time integration scheme to integrate the system. It follows

$$u^{n+1}=u^n+\triangle t\left(A\phi(\mathbf{u}^{n+1})+r(u^{n+1})\right), \tag{5.2}$$

where $\triangle t$ represents the time step. Depending on the initial data u_0 and the type of nonlinearity of the functions ϕ and r to solve the arising system can be a very hard problem, in the vicinity of the zero the derivative of the function ϕ in the case of fast diffusion become infinite. We propose here an algorithm suggested by the Gauss-Sidel iterative method. The method uses the very special structure of the matrix A generated by finite volume method. One writes the matrix A as

$$A=\widetilde{A}+\Gamma,$$

where Γ is a diagonal matrix containing the diagonal entries of the matrix A. We point the following properties of the two matrices

$$\widetilde{A}_{ij}\ge 0,\ \Gamma_{ii}<0,\ \sum_j \widetilde{A}_{ij}\le -\Gamma_{ii}. \tag{5.3}$$

We rewrite also the functions ϕ and r as

$$\phi(x)=\widetilde{\phi}(x)\cdot x,\ r(x)=-\widetilde{r}(x)\cdot x. \tag{5.4}$$

The equation (5.2) can be written now as

$$\left(I + \triangle t\left(-\Gamma\widetilde{\phi}(u^{n+1}) + \widetilde{r}(u^{n+1})\right)\right)u^{n+1} = u^n + \triangle t\widetilde{A}\phi(u^{n+1}). \tag{5.5}$$

The next theorem gives the main properties of the solution of implicit Euler method.

THEOREM **5.1** *In addition of the conditions* **A2** *and* **A2′** *we assume that* $\widetilde{\phi}$ *is a nonincreasing function and* $\widetilde{r} \geq 0$. *If the initial data and boundary conditions are positive and upper bounded functions, i.e.*

$$0 \leq u_0 \leq \rho,\ 0 \leq u_D \leq \rho,$$

then for any time step $\triangle t$ *there exists a solution of the equation* (5.2) *that satisfies*

$$0 \leq u^n \leq \rho,\ \forall n. \tag{5.6}$$

Proof. Let us assume that for a time level n there exists a solution u^n that satisfies (5.6). We will use the Browder fixed point theorem to demonstrate the existence of u^{n+1} with the same properties. Define the $\mathbb{R}^N$-values function Ψ by

$$\Psi_i(y) = \frac{u_i^n + \triangle t\sum_j \widetilde{A}_{ij}\phi(y_j)}{1 + \triangle t\left(-\Gamma_{ii}\widetilde{\phi}(y_i) + \widetilde{r}(y_i)\right)}.$$

We claim that the function Ψ is a continuous function on the set $[0,\rho]^N$ and take values in the same set. So, it has a fixed point.

Since $\widetilde{\phi}$ and $\widetilde{r}$ are continuous functions on $(0,\infty)$ and let us assume that their limits in 0 are finite we can prolong by continuity the function Ψ in 0. It is obviously that $\Psi_i > 0$. For the upper bound we have

$$\Psi_i(y) - \rho \leq \frac{u_i^n + \triangle t\sum_j \widetilde{A}_{ij}\phi(y_j)}{1 - \triangle t\Gamma_{ii}\widetilde{\phi}(y_i)} - \rho =$$

$$= \frac{u_i^n - \rho + \triangle t\left(\sum_j \widetilde{A}_{ij}\phi(y_j) + \rho\Gamma_{ii}\widetilde{\phi}(y_i)\right)}{1 - \triangle t\Gamma_{ii}\widetilde{\phi}(y_i)}.$$

For any $y \in [0,\rho]^N$ we have

$$\sum_j \widetilde{A}_{ij}\phi(y_j) + \rho\Gamma_{ii}\widetilde{\phi}(y_i) \leq \phi(\rho)\sum_j \widetilde{A}_{ij} + \rho\Gamma_{ii}\widetilde{\phi}(y_i) \leq$$

$$\leq -\phi(\rho)\Gamma_{ii} + \rho\Gamma_{ii}\widetilde{\phi}(y_i) = -\rho\Gamma_{ii}(\widetilde{\phi}(\rho) - \widetilde{\phi}(y_i)) \leq 0.$$

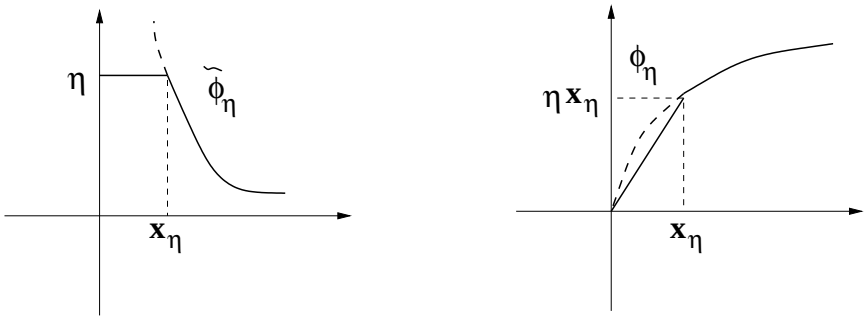


Fig. 2: The regularization of the flux function.

To obtain the first inequality one uses: assumptions **A2′** (ϕ is a nondecreasing function), boundary data is bounded from above by ρ and $\widetilde{A}_{ij} > 0$, the second inequality results from the structure of the matrix A and the last inequality from the constitutive assumption on the $\widetilde{\phi}$.

So, we have

$$0 \leq \Psi_i(y) \leq \rho$$

and for it results the existences of a fixed point, say u. Since for any i one has

$$1 + \triangle t \left(-\Gamma_{ii} \widetilde{\phi}(y_i) + \widetilde{r}(y_i) \right) < \infty, \text{ on } [0, \rho],$$

it follows that the fix point u is a solution of the of the nonlinear equation (5.6).

Let us analyse the case in which the functions $\widetilde{\phi}$ and $\widetilde{r}$ have infinite limits in 0. One regularises the function $\widetilde{\phi}$ by

$$\widetilde{\phi}_\eta(x) = \begin{cases} \eta, & \text{if } \widetilde{\phi}(x) > \eta \\ \widetilde{\phi}(x), & \text{if } \widetilde{\phi}(x) \leq \eta \end{cases} \tag{5.7}$$

and from it one has

$$\phi_\eta(x) = \begin{cases} x\eta, & \text{if } \phi(x) > x\eta \\ \phi(x), & \text{if } \phi(x) \leq x\eta. \end{cases} \tag{5.8}$$

Obviously

$$\phi_\eta(x) \leq \phi(x), \lim_{\eta \to \infty} \phi_\eta(x) = \phi(x).$$

In a similar manner we define r_η.

With the functions ϕ_η and r_η we are in the previous case and then results that there exists a solution $u_\eta \in [0, \rho]^N$ of the equation

$$u_\eta = u^n + \triangle t \left(A\phi_\eta(u_\eta) + r_\eta(u_\eta) \right). \tag{5.9}$$

As the sequence u_η is bounded we can extract a subsequence u_{η_n} that converges to a point $u \in [0,\rho]^N$. The problem is to demonstrate that the limit point u is a solution of the original equation, i.e.

$$u = u^n + \triangle t\,(A\phi(u) + r(u))\,.$$

Let us denote by $F_\eta(u)$ and F r.h.s., of the preceding equations, respectively. We have

$$\begin{aligned} ||u - F(u)||_\infty = \quad & ||u - u_{\eta_n} + (F_{\eta_n}(u_{\eta_n}) - F_{\eta_n}(u)) + (F_{\eta_n}(u) - F(u)||_\infty \leq \\ & \leq ||u - u_{\eta_n}||_\infty + ||F_{\eta_n}(u_{\eta_n}) - F_{\eta_n}(u)||_\infty + \\ & + ||F_{\eta_n}(u) - F(u)||_\infty\,. \end{aligned}$$

We will show that, for any $\varepsilon > 0$,

$$||u - F(u)||_\infty \leq \varepsilon.$$

Observe that the first term and the last term can be made arbitrary small,

$$||u - u_{\eta_n}||_\infty + ||F_{\eta_n}(u) - F(u)||_\infty < \frac{\varepsilon}{2}$$

for any $n > n^\varepsilon$. The middle term can be evaluate as $||\cdot||_\infty$

$$\begin{aligned} ||F_{\eta_n}(u_{\eta_n}) - F_{\eta_n}(u)||_\infty \leq \quad & \triangle t\big(\,||A(\phi_{\eta_n}(u_{\eta_n}) - \phi_{\eta_n}(u))||_\infty + \\ & + ||r_{\eta_n}(u_{\eta_n}) - r_{\eta_n}(u)||_\infty\,\big) \leq \\ \leq \quad & \triangle t(||A||\,||\phi_{\eta_n}(u_{\eta_n}) - \phi_{\eta_n}(u)||_\infty + \\ & + ||r_{\eta_n}(u_{\eta_n}) - r_{\eta_n}(u)||_\infty). \end{aligned}$$

For each component i we look at

$$|\phi_{\eta_n}(u_{\eta_n i}) - \phi_{\eta_n}(u_i)|$$

and note that if u_i is not equal with zero then for a great enough number n one has

$$|\phi_{\eta_n}(u_{\eta_n i}) - \phi_{\eta_n}(u_i)| = |\phi(u_{\eta_n i}) - \phi(u_i)|\,,$$

if u_i equals zero then

$$|\phi_{\eta_n}(u_{\eta_n i}) - \phi_{\eta_n}(u_i)| = \phi_{\eta_n}(u_{\eta_n i}) \leq \phi(u_{\eta_n i}).$$

Using the continuity of the function ϕ we can find a number n_1^ε such that

$$||\phi_{\eta_n}(u_{\eta_n}) - \phi_{\eta_n}(u)||_\infty \leq \frac{\varepsilon}{4||A||\triangle t}$$

for any $n > n_1^{\varepsilon}$. Using the same arguments we can prove that

$$||r_{\eta_n}(u_{\eta_n}) - r_{\eta_n}(u)||_{\infty}) < \frac{\varepsilon}{4\triangle t}$$

for any $n > n_2^{\varepsilon}$. Hence, there exists a n^{ε} such that

$$||F_{\eta_n}(u_{\eta_n}) - F_{\eta_n}(u)||_{\infty} \leq \frac{\epsilon}{2}$$

for any $n > n_{\epsilon}$.

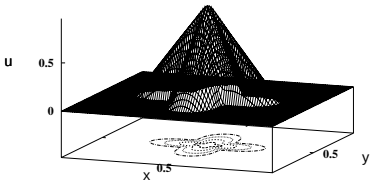
This end the proof of the theorem.

In our implementation we calculate the solution of the Euler scheme by the following iterative solver

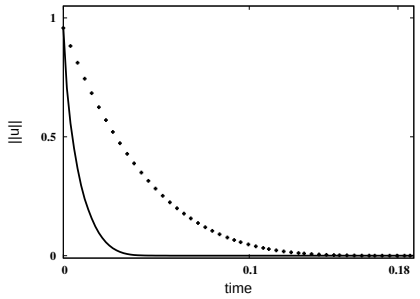
$$\begin{gathered}\left(I + \triangle t\left(-\Gamma\widetilde{\phi}_{\eta}(u^{n+1,k}) + \widetilde{r}_{\eta}(u^{n+1,k})\right)\right)u^{n+1,k+1} = \\ u^n + \triangle t\widetilde{A}\phi_{\eta}(u^{n+1,k}).\end{gathered} \tag{5.10}$$

Numerical Simulation. For the numerical simulation we chose a very simple domain $\Omega = [0,1] \times [0,1]$. The fast diffusion with absorption is modeled by $\phi(s) = s^m, r(s) = -\lambda \cdot s^p$.

Table 1: Extinction phenomenon, extinction time $T^e = 0.18$. $\phi(s) = s^{0.75}$, $r(s) = -21. \times s^{0.5}$, $u_D = 0$



Initial Profile



Comparison of the numerical solution (solid line) with a theoretical estimation (points drawing).

5.2. Water Infiltration through Stratified Soil. Richard's Equation

We consider stratified soil. Hereafter the stratified soil means a block-wise homogeneous soil with horizontal parallel homogeneous strata, see figure (3).

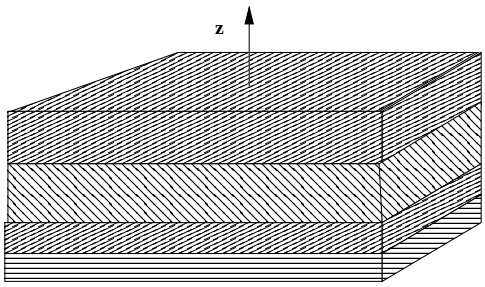


Fig. 3: Stratified porous soil. Each layer is modelled by different constitutive function.

In the case of stratified soil the different mechanical properties of the soils require different constitutive functions which in turn lead to a partial differential equation with discontinuous coefficient. On an interface of two different strata one must impose some compatible conditions to have a well defined problem. Physical considerations require the continuity of the pressure head and normal components of the velocity. So, we have

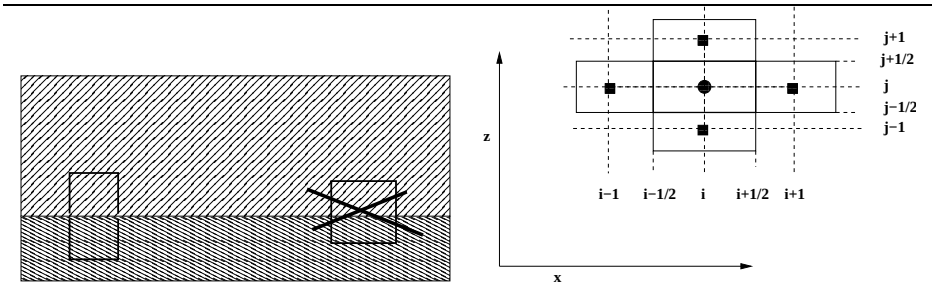
$$\begin{aligned} h|_- &= h|_+, \\ \boldsymbol{v}\cdot\mathbf{n}|_- &= \boldsymbol{v}\cdot\mathbf{n}|_+. \end{aligned} \tag{5.11}$$

Taking into account the compatibility relations (5.11) appear that it is more convenient to work with the $\theta - h$ form of Richards' equation, i.e.,

$$\begin{aligned} \partial_t \int_V \theta \mathrm{d}x &= \int_{\partial V} K(\theta)\frac{\partial(h+z)}{\partial n}\mathrm{d}s, \\ \theta &= \theta(h) \end{aligned} \tag{5.12}$$

We assume that the flow domain is the 2D rectangle $\Omega = [0,a]\times[0,b]$ which is stratified in N_s strata $[0,a]\times[Z_{i-1},Z_i]$ with $Z_0 = 0, Z_{N_s} = b$.

Let $0 = x_{1/2} < x_{1+1/2} < \cdots < x_{N+1/2} = a$, $0 = z_{1/2} < z_{1+1/2} < \cdots < z_{M+1/2} = b$ be two partitions of the intervals $[0,a]$ and $[0,b]$ respectively. We define the control volumes $\omega_{i,j} = \left[x_{i-1/2}, x_{i+1/2}\right]\times\left[z_{j-1/2}, z_{j+1/2}\right]$, $i = \overline{1,N}$, $j = \overline{1,M}$ and the net inner knots $\boldsymbol{r}_{i,j} = (x_i, z_j)$, $x_i = \dfrac{x_{i-1/2}+x_{i+1/2}}{2}$, $z_j = \dfrac{y_{j-1/2}+y_{j+1/2}}{2}$, $i = \overline{1,N}$, $j = \overline{1,M}$. We assume that the partition $\{\omega_{i,j}\}$ is a *conform partition* with respect to stratification of the domain Ω,



Conformal Mesh with respect to strata

Volumes of the control.

Fig. 4: 2D mesh.

i.e for any j the line $z = Z_j$ does not intersect the interior of any control volume $\omega_{i,j}$.

On each volume $\omega_{i,j}$ one approximates the pressure by a constant value $h_{i,j}$ and water content by a constant value $\theta_{i,j}$. On the common boundary $\sigma_{i+1/2,j} = \omega_{i,j} \cup \omega_{i+1,j}$ of two neighbors we approximate the flux by

$$\int\limits_{\sigma_{i+1/2,j}} K(\theta)\frac{\partial(h+z)}{\partial n}\mathrm{ds} \approx K_{i+1/2,j}\frac{h_{i+1,j}-h_{i,j}}{\triangle x_{i+1}} \tag{5.13}$$

where the numerical hydraulic conductivity $K_{i+1/2,j}$ is an approximation of the hydraulic conductivity $K(\theta)$,

$$K_{i+1/2,j} = \widetilde{K}(\theta_{i,j}, \theta_{i+1,j}). \tag{5.14}$$

We assume that the function $\widetilde{K}(\cdot,\cdot)$ is a symmetric and continuous function with respect to its arguments. As result, we obtain a differential algebraic system of equation (DAE), $\theta - h$ form of Richards' equation,

$$\begin{cases} m_{i,j}\dfrac{\mathrm{d}\theta_{i,j}}{\mathrm{d}t} = K_{i+1/2,j}\dfrac{h_{i+1,j}-h_{i,j}}{\triangle x_{i+1}} - K_{i-1/2,j}\dfrac{h_{i,j}-h_{i-1,j}}{\triangle x_i} + \\ \qquad + K_{i,j+1/2}\left(\dfrac{h_{i,j+1}-h_{i,j}}{\triangle z_{j+1}}+1\right) - K_{i,j-1/2}\left(\dfrac{h_{i,j}-h_{i,j-1}}{\triangle z_j}+1\right), \\ \theta_{i,j} = \theta(h_{i,j}). \end{cases} \tag{5.15}$$

To integrate the DAE (5.15) we use an implicit multi-step method, [5]. Let $\{t_{n-k}, t_{n-k+1}, ..., t_n\}$ be a sequence of moments of time and denotes by $\theta^m = \theta(t_m) \in \mathbb{R}^{NM}$, $NM = N \times M$. Supposing that one knows the values $\{\theta^{n-k}, \theta^{n-k+1}, ..., \theta^n\}$, the values θ^{n+1} and h^{n+1} at the next moment of time

t_{n+1} are calculated as follows. Define a predictor polynomial $\omega^P(t)$ and a corrector polynomial $\omega^C(t)$. The predictor polynomial interpolates the values $\{\theta^{n-k}, \theta^{n-k+1}, ..., \theta^n\}$ at moments of time $\{t_{n-k}, t_{n-k+1}, ..., t_n\}$, Lagrange interpolation,

$$\omega^P(t) = \sum_{j=0}^{k} q_j(t)\theta^{n-j}. \tag{5.16}$$

For each $j = \overline{0,k}$ the polynomial $q_j(t)$ is given by

$$q_j(t) = \prod_{i=0, i\neq j}^{k} \frac{t - t_{n-i}}{t_{n-j} - t_{n-i}}.$$

The corrector polynomial $\omega^C(t)$ interpolates the unknowns θ^{n+1} and the values of $\omega^P(t)$ at the moments of time t_{n+1} and $\{t_{n+1} - j\triangle t_n; j = \overline{1,k}\}$, respectively. The unknowns θ^{n+1} and h^{n+1} are determined by imposing to the corrector polynomial $\omega^C(t)$ and to h^{n+1} to satisfies the DAE. Then a system of nonlinear equation results. By denoting

$$\begin{aligned}
&\mathcal{F}_{i,j}(\boldsymbol{\theta}^{n+1}, \boldsymbol{h}^{n+1}) := \\
&K_{i+1/2,j}(\theta^{n+1})\frac{h^{n+1}_{i+1,j} - h^{n+1}_{i,j}}{\triangle x_{i+1}} - K_{i-1/2,j}(\theta^{n+1})\frac{h^{n+1}_{i,j} - h^{n+1}_{i-1,j}}{\triangle x_i} + \\
&K_{i,j+1/2}(\theta^{n+1})\left(\frac{h^{n+1}_{i,j+1} - h^{n+1}_{i,j}}{\triangle z_{j+1}} + 1\right) - K_{i,j-1/2}(\theta^{n+1})\left(\frac{h_{i,j} - h_{i,j-1}}{\triangle z_j} + 1\right)
\end{aligned} \tag{5.17}$$

one obtains

$$\begin{cases} m_{i,j}\left(\dfrac{a}{\triangle t^n}\theta^{n+1}_{i,j} - w^{P,n}_{i,j}\right) = & \mathcal{F}_{i,j}\left(\boldsymbol{\theta}^{n+1}, \boldsymbol{h}^{n+1}\right), \\ \theta^{n+1}_{i,j} = & \theta(h^{n+1}_{i,j}), \end{cases} \tag{5.18}$$

where $w^{P,n}_{i,j}$ are known quantities as functions of the preceding values of θ.

The nonlinear system (5.18) is solved iteratively using an inexact Newton step followed by a Broyden step until a desired accuracy is obtained. Let $\mathcal{R}$ be given by

$$\boldsymbol{\mathcal{R}}(\boldsymbol{\theta}, \boldsymbol{h}) = \boldsymbol{m}\left(\frac{a}{\triangle t^n}\boldsymbol{\theta} - \boldsymbol{w}^{P,n}\right) - \boldsymbol{\mathcal{F}}\left(\boldsymbol{\theta}, \boldsymbol{h}\right). \tag{5.19}$$

The matrix $\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{h})$ of the iterative process in INS is an approximation of the full Jacobian of the function $\mathcal{R}$, the product of it with a vector $\boldsymbol{w}$ read

as

$$\mathcal{J}(\boldsymbol{\theta},\boldsymbol{h})\boldsymbol{w}=\boldsymbol{m}\frac{a}{\triangle t^n}C(\boldsymbol{h})\boldsymbol{w}-\widetilde{\boldsymbol{\mathcal{F}}}\left(\boldsymbol{\theta},\boldsymbol{w}\right), \tag{5.20}$$

where

$$\widetilde{\boldsymbol{\mathcal{F}}}\left(\boldsymbol{\theta},\boldsymbol{w}\right)=\partial_h\boldsymbol{\mathcal{F}}\left(\boldsymbol{\theta},\boldsymbol{w}\right) \tag{5.21}$$

and

$$C(\cdot)=\frac{\mathrm{d}\theta(\cdot)}{\mathrm{d}h}.$$

The nonlinear solver is:

$$\left\|\begin{array}{ll}
\text{Inexact Newton step} & \\
\mathcal{J}(\boldsymbol{\theta}^{n+1,k},\boldsymbol{h}^{n+1,k})\boldsymbol{\delta}_h^{NS}=-\boldsymbol{\mathcal{R}}\left(\boldsymbol{\theta}^{n+1,k},\boldsymbol{h}^{n+1,k}\right), & (\mathrm{s}\,1)\\
\overline{\boldsymbol{h}}^{n+1,k+1}=\boldsymbol{h}^{n+1,k}+\boldsymbol{\delta}_h^{NS}, & (\mathrm{s}\,2)\\
\overline{\boldsymbol{\theta}}^{n+1,k+1}=\theta(\overline{\boldsymbol{h}}^{n+1,k+1}), & (\mathrm{s}\,2)\\
\text{Broyden step} & \\
\mathcal{J}(\boldsymbol{\theta}^{n+1,k},\boldsymbol{h}^{n+1,k})\boldsymbol{\delta}_h^{BS}=-\boldsymbol{\mathcal{R}}\left(\overline{\boldsymbol{\theta}}^{n+1,k},\overline{\boldsymbol{h}}^{n+1,k}\right), & (\mathrm{s}\,3)\\
\boldsymbol{\delta}_h^{k+1}=\boldsymbol{\delta}_h^{BS}\dfrac{\langle\boldsymbol{\delta}_h^{NS},\boldsymbol{\delta}_h^{NS}\rangle}{\langle\boldsymbol{\delta}_h^{NS},\boldsymbol{\delta}_h^{NS}\rangle-\langle\boldsymbol{\delta}_h^{NS},\boldsymbol{\delta}_h^{BS}\rangle}, & (\mathrm{s}\,4)\\
\boldsymbol{h}^{n+1,k+1}=\overline{\boldsymbol{h}}^{n+1,k}+\boldsymbol{\delta}_h^{k+1}, & (\mathrm{s}\,5)\\
\boldsymbol{\theta}^{n+1,k+1}=\theta(\boldsymbol{h}^{n+1,k+1}). & (\mathrm{s}\,5)
\end{array}\right. \tag{5.22}$$

The linear equations in the steps s1 and s3 are solved by Conjugate Gradient Method for linear system with symmetric and positive definite matrix. We present some numerical tests obtained using the above algorithm. As empirical models for water content $\theta(h)$ and hydraulic conductivity $K(\theta)$ we use the van Genuchten model,

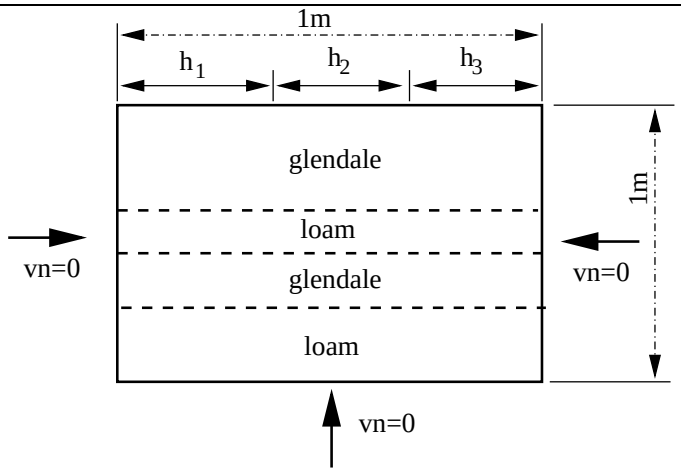
$$S(h)=\left\{\begin{array}{ll}(1+(\alpha h)^n)^{-m}, & h<0,\\ 1, & h\geq 0,\end{array}\right. \tag{5.23}$$

$$K(S)=\left\{\begin{array}{ll}K_sS^l\left(1-\left(1-S^{1/m}\right)^m\right)^2, & 0<S<1,\\ K_s, & S\geq 1,\end{array}\right. \tag{5.24}$$

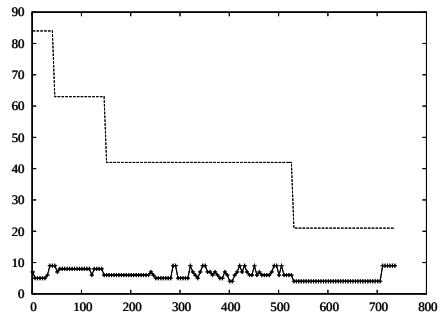
where S represents the relative water content

$$S = \frac{\theta - \theta_r}{\theta_s - \theta_r}.$$

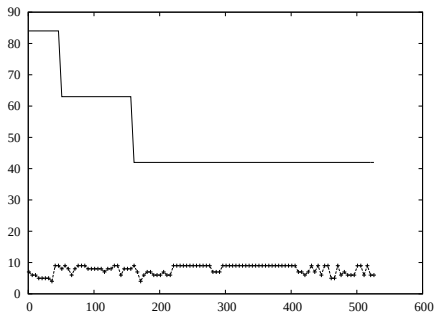
The soil in the test is a layered soil with two alternate strata.



Physical configuration. The parameters for the *loam soil* in the van Genuthen model are: $n = 2$, $\alpha = 3.35\,\mathrm{m}^{-1}$, $l = 0.5$, $K_s = 0.3318\,\mathrm{mh}^{-1}$, $\theta_r = 0.012$, $\theta_s = 0.368$ and for the *Glendale soil* are: $n = 1.3954$, $\alpha = 1.04\,\mathrm{m}^{-1}$, $l = 0.5$, $K_s = 0.545 \times 10^{-2}\,\mathrm{mh}^{-1}$, $\theta_r = 0.106$, $\theta_s = 0.4686$. The initial datum is $h^0 = -1.0\,\mathrm{m}$ in the whole domain. The boundary conditions are of the mixt type.

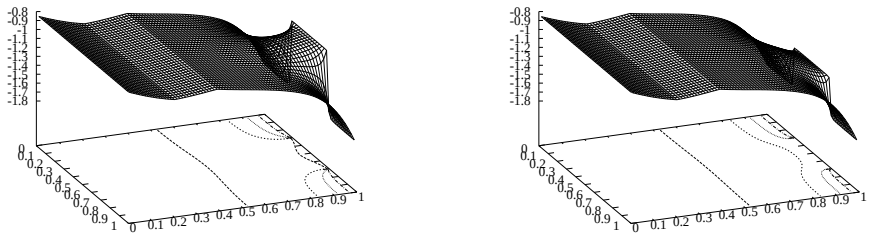


$h_1 = -0.75\,\mathrm{m}$, $h_2 = -0.0\,\mathrm{m}$, $h_3 = -0.75\,\mathrm{m}$.

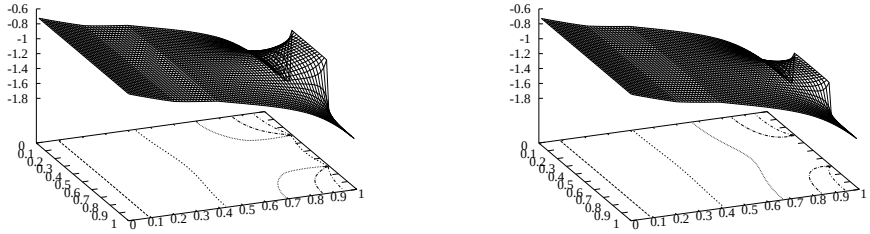


$h_1 = -0.75\,\mathrm{m}$, $h_2 = -0.3\,\mathrm{m}$, $h_3 = -0.75\,\mathrm{m}$.

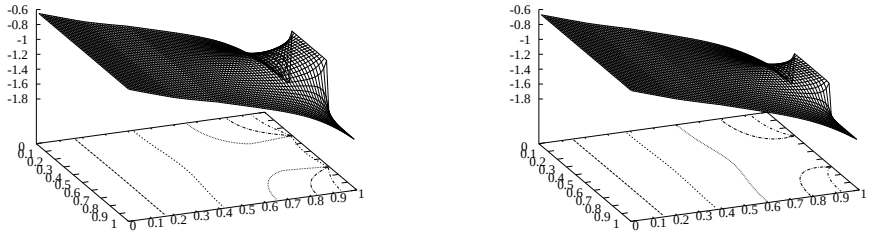
The number of the iterations versus time step. Numbers of iteration in nonlinear solver (line-point) and the total numbers of iteration time step in CGM method The time simulation was 48h.



$t = 6\text{h}$



$t = 24\text{h}$



$t = 48\text{h}$

The comparative profiles of the pressure head for two different Dirichlet datum on the top boundary at different moments of time. $h_1 = -0.75\,\text{m}$, $h_2 = -0.\,\text{m}$, $h_3 = -0.75\,\text{m}$.(left), $h_1 = -0.75\,\text{m}$, $h_2 = -0.3\,\text{m}$, $h_3 = -0.75\,\text{m}$.(right).

References

[1] H. W. Alt, S. Luckhaus, *Quasilinear eliptic-parabolic differential equations*, Math. Z., **183** (1983), pp. 311–341.

[2] V. Barbu, *Ecuaţii diferenţiale*, Editura Junimea, Iaşi (1985).

[3] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, 1988.

[4] M. Borelli and M. Ughi, *The fast diffusion equation with strong absorbtion: the instantaneous shrinking phenomenon*, Rendi. Instit. Mat. Univ. Trieste, **26** (1994), pp. 109–140.

[5] K.E. Brenam, S.L. Campbell and L.R. Petzald, *Numetical Solution of Initial Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics, SIAM, 1996.

[6] B. Cockburn, F. Coquel and P. G. Lefloch, *Convergence of the Finite Volume Method For Multidimensional Conservation Laws*, SIAM J. Numer. Anal., **32** (1995), No. 3, pp. 687–705.

[7] J. Carrillo, *Entropy Solutions for Nonlinear Degenerate Problems*, Arch. Rational Mech. Anal., **147** (1999), pp. 269–361.

[8] Carslaw H. S., Jaeger J. C., *Conduction of heat in solids*, Oxford, 1959.

[9] C.N. Dawson, C. J. van Duijn, and R.E. Grundy, *Large Time Asymptotics in Contaminant Transport in Porous Media*, SIAM J. Appl. Math., **56** (1996), No. 4, pp. 965–993.

[10] J.I. Diaz, *Qualitative Study of Nonlinear Parabolic Equations: an Introduction*, Extracta Matematicae, **16**(3) (2001), pp. 303–341.

[11] R. Eymard, Th. Gallouet and R. Herbin, *Error Estimate for Approximate Solutions of a Nonlinear Convection-Diffusion Problem*, Advances in Differential Equations, **7** (2002), No. 4, pp. 419–440.

[12] R. Eymard, Th. Gallouet and R. Herbin, *Finite Volume Method*, in Handbook of Numerical Analysis, G. Ciarlet and J.L. Lions eds., North Holland, 2000.

[13] R. Ferreira and J.L. Vasquez, *Extiction behavior for fast diffusion equation with absorption*, Nonlinear Anal., **43** (2001), pp. 551–563.

[14] S. N. Kruskov, *First order quasilinear equations with several independent variables*, Math. USSR Sb., **10** (1970), pp. 217–243.

[15] V. Lakshmikantham and S. Leela *Differential and Integral Inequalities: Theory and Applications.* Academic Press, New York (1969).

[16] G. Leoni, *A very singular solution for the porous media equation* $u_t = \triangle u^m - u^p$ *when* $0 < m < 1$, J. Differential Equations, **132** (1996), pp. 353–376.

[17] H.E. Levin and P.E. Sacks, *Some existence and nonexistence theorems for solution of degenerate parabolic equation*, J. Diff. Equat., **52** (1984), pp. 135–161.

[18] C. Mascia, A. Porretta and A. Terracina, *Nonhomogeneous Dirichlet problems for degenerate parabolic-hyperbolic equations*, Arch. Ratio. Mech. anal., **163** (2002), pp. 87–124.

[19] A. Michel and J. Vovelle, *Entropy Formulation for Parabolic Degenerate Equations with General Dirichlet Boundary Conditions and application to the Convergence of FV Methods*, SIAM J. Numer. Anal., **41** (2003), No. 6, pp. 2262–2293.

[20] K. Nickel, *Bounds for the Set of Solutions of Functional-Differential Equations*, MRC Tech. Summary Report No. 1782, Univ. of Wisconsin, Madison (1977).

[21] M.-N. Le Roux, *Semidiscretization in time of a fast diffussion*, J. Math. Anal. Appl., **137**, no. 2, (1989), pp. 354–370.

[22] M.-N. Le Roux and P.-E. Mainge, *Numerical solution of a fast diffusion equation*, Mathematics of Computation, **68**, no. 226, (1999), pp. 461–485.

[23] L.A. Peletier and Zhao Junning, *Large time behavior of solution of the porous media equation with absorption: The fast diffusion case*, Nonlinear Anal., **17** (1990), pp. 107–121.

[24] F. Otto, L^1*-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Dirential Equations, **131**(1) (1986), pp. 20–38.

[25] J.L. Vasquez, *Smoothing and Decay Estimates for Nonlinear Diffusion Equations of Porous Medium Type*, Oxford University Press, 2006.

Topics in Applied Mathematics & Mathematical Physics

On the Numerical Simulation of a Class of Reactive Boltzmann Type Equation

by Dorin Marinescu [1]

Contents

[1]"Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: marinescu.dorin@gmail.com

The paper was supported by CEEX Grant CEX05-D11-06/03.10.2005 and CEEX05-D11-25/03.10.2005

1. Introduction

It is known that the classical Boltzmann equation describes the evolution of the simple gas. The Boltzmann equation represents the connecting bridge between the microscopic and macroscopic description of the simple fluid evolution. The kinetics of the simple gas is essentially governed by elastic binary collisions between structureless particles belonging to a unique species, the multiple collisions being very improbable Ref. [1]. However, this equation is not able to describe the evolution of the real gas with chemical reactions and/or ionization processes. Then inelastic collisions must be considered by the kinetic models. Boltzmann himself was aware of the importance of the inelastic collisions in the real fluid evolution Ref. [9].

The classical Boltzmann equation is almost unanimously considered as appropriate for the kinetics of the rarefied simple gas. A kinetic theory for the reactive (real) gas is a more difficult task Ref. [30, 21]. As compared to the classical Boltzmann equation for the simple gas, kinetic reactive models exhibit new mathematical difficulties due the contribution of the particle internal states to the gas evolution (in particular the presence or reaction thresholds) and the existence of collision channels with multiple reaction participants Ref. [8, 25, 24, 29]. In the case of the reacting gas mixtures the mass balance does not hold for a given species. Then, the mass conservation for a specie must be replaced by the total mass balance. In the reactive models is present a transfer between the kinetic energy and the internal molecular energy. Consequently, the kinetic energy balance must be replaced by the total energy balance (i.e. kinetic energy + internal molecular energy). Then, the transport properties of the reacting gas mixtures differ from the properties of the simple gas.

Various models have been introduced to describe the kinetics of the real (reactive) gas. An important role is played by the Boltzmann-like semi-quantum equations. A known example is the Wang-Chang-Uhlenbeck-de Boer system of kinetic equations [32] for the real gas with binary collisions. This model refers to a gas of particles with classical translational motion, but with quantum internal structure. Essentially, the difference from the Boltzmann model Ref. [11] for the simple gas is to associate to each internal state a distribution function, and to relate each transition from one quantum internal state (of some chemical species) to another with a cross-section matrix.

A more general model introduced by Ludwig and Heil [25] extends Wang-Chang-Uhlenbeck-de Boer model. This model describes reactions in a diatomic gas without emission or absorption of radiation. It includes processes

of recombinations by triple collisions, as well as three post-collisional products like dissociation and ionization induced by collisions Ref. [8, 25, 24].

In some Wang-Chang-Uhlenbeck-de Boer or Ludwig and Heil model the number of equations depends on the number of distribution functions, i.e. on the number of different quantum internal states owned by the gas particles during the gas evolution. It is known that, there exists only at most a countable set of internal states. However, only a finite number of internal states will significantly contribute to the gas kinetics. Consequently, the Wang-Chang-Uhlenbeck-de Boer and Ludwig and Heil models are described by a finite number of equations.

For analytical purposes, in Ref. [16, 17, 18], the Wang-Chang-Uhlenbeck-de Boer and Ludwig and Heil equations corresponding to the model with finite number of internal states have been transcribed in abstract form, revealing the mathematical structure of the equations. In Ref. [17] was proved the existence and uniqueness of the solutions for the Cauchy problem. It was shown that the solutions verify the conservation of the total mass, momentum and energy respectively. Moreover, it was proved the existence of equilibrium solutions. H-theorem and a generalized law of the mass action have been rigorously proved under extended balance conditions.

The interest for reactive kinetics is not only intrinsic, but also of practical nature, in plasma physics, nuclear physics, physical chemistry of the high atmosphere, combustion theory, modeling of missiles flight.

Accurate numerical modeling of nonlinear processes in dilute, flows is critical for solving transport problems both in fundamental and applied science. In this respect Babovsky and Illner [4, 5] have proposed an efficient numerical scheme consistent with the classical Boltzmann equation. Using Nambu's ideas [26], by time discretization and local space-homogenization, Babovsky and Illner have obtained a convenient approximate form of the equation. At this point, the nonlinear character of the collision operators involve a power-like growth of the numerical complexity. To provide an algorithm, with small numerical effort, they have introduced an additonal stochastic approximation. Finally, they have proved the convergence almost sure, in some sense, of the approximation scheme. The techniques developed by Nambu [26], Babovsky and Illner of [4, 5] were also applied Ref. [6] to Pullin's equation [27] with Larsen-Borgnakke [10] scattering cross section for the one-component diatomic gas with classical internal degrees of freedom.

For the abstract model Ref. [16, 17, 18] describing the real reacting gas, in Ref. [19] was introduced a rigorous and efficient approximation scheme. This method represents a nontrivial extension of the techniques of Ref. [4, 5] for

solving space-homogeneous Boltzmann-like models of reacting gas mixtures Ref. [32, 8, 25, 24, 16, 17].

The methods of this chapter have been tested Ref. [14, 13] on the Krook-Wu [22] two-component Boltzmann equation as well as on the reactive Boltzmann models with three and four components Ref. [12, 20].

This review presents the theoretical approximation method for the solutions of the Boltzmann model introduced in Ref. [17] following the line of Ref. [19] and adding some improvements sketched in Ref. [12].

The present chapter is organized as follows.

In the next section one first recalls the main features of the Boltzmann-like equations introduced in Ref. [17]. Then, one formulates the approximation problem. In Section 3 one investigates the initial value problem for the space-homogeneous kinetic equations of Section 2, formulated in a suitable space of functions. In Section 4 one obtains a convergent, time-discretized version of the aforementioned Boltzmann-like equations. Section 5 is devoted to the generalizations of certain probabilistic selection results of Ref. [4, 5]. This is possible due to some clarifications with respect to the nature of the convergence introduced by Babovsky and Illner. More precisely, the probabilistic part of the convergence proof of Ref. [4, 5] is based on the central limit theorem for row-wise i.i.d. random variables and the Borel-Cantelli Lemma. Our argument follows from a simple version of the strong law of large numbers for arrays of (not necessarily identically distributed) row-wise independent, random variables. (Which results from the Chebyshev inequality and the Borel-Cantelli Lemma.) In Section 6, the results of Section 5 are applied to the discretized scheme obtained in Section 4. Consequently, one obtains the numerical algorithm for the original Cauchy problem. This represents our main result, namely the convergence of the numerical scheme. Finally, we discuss the limitations and possible generalizations of the model.

2. The Kinetic Model and the Approximation Procedure

Here, we briefly recall the features of the model presented in Ref. [17, 18] (see also Ref. [16]).

The leading idea behind the model is that, unequal internal states of a gas particle with internal structure can be considered as describing structure-less particles belonging to distinct chemical species. Then, a real gas mixture

of particles with internal structure can be thought as a mixture of several chemical species of mass points with unique internal states.

Specifically, the model refers to a gas consisting of N distinct species of point masses, with one-state internal energy, evolving without external forces. The following assumptions are general: (i) gas particles have free classical motion in space, between (in)elastic, instant, local collisions, without emission or absorption of photons; (ii) collision (reactions) may change momenta, as well as the chemical nature (in particular mass and internal energy) of the gas particles; any collision occurs with conservation of total mass, momentum and (kinetic+internal) energy, according to the laws of classical mechanics. (iii) in each collision (reaction) channel, the number of identical partners cannot exceed some number, say $M \geq 2$ and any collision (reaction) channel contains, at least, two particles.

Denote by $\mathcal{M}$ the folowing multi-index set

$$\mathcal{M} := \{\boldsymbol{\gamma} = (\gamma_k)_{k=1,\ldots,N} | \gamma_k \in \{0, 1, \ldots, M\}\}. \tag{2.1}$$

A gas collision (reaction) process is specified by a couple $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$. Here, the multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ represents the pre-collision (in) channel, with $\alpha_n \in \{0, 1, \ldots, M\}$ identical participants of the $n-th$ species. The multi-index $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$ represents the post-collision (out) channel, with $\beta_n \in \{0, 1, \ldots, M\}$ identical participants of the $n-th$ species.

The pair of multi-indexes $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponds to a reaction of the following type

$$\alpha_1 X_1 +, \ldots, +\alpha_N X_N \to \beta_1 X_1 +, \ldots, +\beta_N X_N, \tag{2.2}$$

between the species X_1,...,X_N, with stoichiometric coefficients $\alpha_1, \ldots, \alpha_N$, $\beta_1, .., \beta_N$. Note that if $\alpha = \beta$, the collision is elastic and if $\alpha \neq \beta$, the collision is inelastic.

For each channel $\boldsymbol{\gamma} \in \mathcal{M}$ the family $\mathcal{N}(\boldsymbol{\gamma}) := \{k \mid \gamma_k > 0 \text{ for } k = 1, \ldots, N\}$ represents the species existing in that channel. Obviously, if $k \notin \mathcal{N}(\boldsymbol{\gamma})$ the species k is not present inside the channel $\boldsymbol{\gamma}$. If $k \in \mathcal{N}(\boldsymbol{\gamma})$, then there are γ_k identical particles of the species k in the channel $\boldsymbol{\gamma}$. We denote the total number of particles in the channel $\boldsymbol{\gamma}$ by

$$|\boldsymbol{\gamma}| := \sum_{k=1}^{N} \gamma_k. \tag{2.3}$$

Their velocities are denoted by $\mathbf{w}_{k,1}, \ldots, \mathbf{w}_{k,\gamma_k} \in \mathbb{R}^3$. Also set $\mathbf{w} := ((\mathbf{w}_{k,i})_{i=1,\ldots,\gamma_k})_{k \in \mathcal{N}(\boldsymbol{\gamma})}$, understanding that $\mathbf{w} \in \mathbb{R}^{3|\boldsymbol{\gamma}|}$. We denote by

$m_k > 0$ and $E_k \in \mathbb{R}$, the mass and the internal energy, respectively of a mass-point of the species $k = 1, \ldots, N$.

Let

$$V_{\boldsymbol{\gamma}}(\mathbf{w}) := (\sum_{k=1}^{N} \gamma_k m_k)^{-1} \sum_{k \in \mathcal{N}(\boldsymbol{\gamma})} \sum_{i=1}^{\gamma_k} m_k \mathbf{w}_{k,i}, \tag{2.4}$$

and

$$W_{\boldsymbol{\gamma}}(\mathbf{w}) := \sum_{k \in \mathcal{N}(\boldsymbol{\gamma})} \sum_{i=1}^{\gamma_k} (2^{-1} m_k \mathbf{w}_{k,i}^2 + E_k). \tag{2.5}$$

be the classical mass center velocity and the total energy, respectively, for the particles in the channel $\boldsymbol{\gamma}$. According to the conservation assumptions, in the description of the gas kinetics, for each couple $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$ we consider only the collisions satisfying the relations

$$\sum_{k=1}^{N} m_k(\alpha_k - \beta_k) = 0, \tag{2.6}$$

$$V_{\boldsymbol{\alpha}}(\mathbf{w}) = V_{\boldsymbol{\beta}}(\mathbf{u}), \quad W_{\boldsymbol{\alpha}}(\mathbf{w}) = W_{\boldsymbol{\beta}}(\mathbf{u}), \tag{2.7}$$

In (2.7) $\mathbf{w} = ((\mathbf{w}_{k,i})_{i=1,\ldots,\gamma_k})_{k \in \mathcal{N}(\boldsymbol{\alpha})}$ and $\mathbf{u} = ((\mathbf{u}_{k,i})_{i=1,\ldots,\beta_k})_{k \in \mathcal{N}(\boldsymbol{\beta})}$ are the velocities of the particles in the channels $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

Note that reactions with at most one particle in some collision channel are excluded by (2.6) and (2.7), because in the absence of radiative processes, the conservation laws (2.6) and (2.7) cannot be simultaneously fulfilled. Therefore, $|\boldsymbol{\gamma}| \geq 2$. This inequality explains the restriction $M \geq 2$ in the definition (2.1) of $\mathcal{M}$. Remark that, the conservation of the total energy stated in (2.7) implies the existence of reaction thresholds and shows what happens with the internal energies of the particles participating in reactions. For instance in the case of endothermic collisions $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, i.e.

$$\sum_{k \in \mathcal{N}(\boldsymbol{\alpha})} \alpha_k E_k < \sum_{k \in \mathcal{N}(\boldsymbol{\beta})} \beta_k E_k, \tag{2.8}$$

the kinetic energy of the resulting products is lost as binding energy. In such a case the collision can be forbidden if the kinetic energy in the channel $\boldsymbol{\alpha}$ is bellow the reaction threshold. Note that, the model accepts also reaction thresholds for exothermic collisions $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

$$\sum_{k \in \mathcal{N}(\boldsymbol{\alpha})} \alpha_k E_k > \sum_{k \in \mathcal{N}(\boldsymbol{\beta})} \beta_k E_k. \tag{2.9}$$

Following the standard Boltzmann procedure (based on the molecular chaos assumption) we introduce the system of kinetic equations

$$\partial_t f_k + \mathbf{v} \cdot \nabla_x f_k = P_k(\mathbf{f}) - S_k(\mathbf{f}), \text{ for } k = 1, \dots, N, \tag{2.10}$$

as an abstract transcription of the Wang-Chang-Uhlenbeck-de Boer and Ludwig and Heil equations. Here $f_k : \mathbb{R}_+ \times \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}_+$ are the unknowns for $k = 1, \dots, N$, (with $\mathbb{R}_+ := [0, \infty)$) and $\mathbf{f} := (f_1, \dots, f_N)$. Each $f_k = f_k(t, \mathbf{v}, \mathbf{x})$ (t-time, $\mathbf{v}$ -velocity, $\mathbf{x}$ -position) is the one-particle distribution function for species $k = 1, \dots, N$ of particles. In (2.10) the gain operators P_k and the loss operators $S_k(\mathbf{f})$ describe the collision processes.

For $\mathbf{g} = (g_1, \dots, g_N)$ (with $g_1, \dots, g_N : \mathbb{R}^3 \to \mathbb{R}$) define,

$$\mathbf{g}_{\boldsymbol{\gamma}}(\mathbf{w}) := \prod_{k \in \mathcal{N}(\boldsymbol{\gamma})} \prod_{i=1}^{\gamma_k} g_k(\mathbf{w}_{k,i}), \quad \boldsymbol{\gamma} \in \mathcal{M}. \tag{2.11}$$

Formally the gain and the loss operators are defined by

$$P_k(\mathbf{g})(\mathbf{v}) = \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}} \alpha_k \int_{\mathbb{R}^{3|\boldsymbol{\beta}|} \times \mathbb{R}^{3|\boldsymbol{\alpha}|}} \sigma_{\boldsymbol{\beta}, \boldsymbol{\alpha}, k}(\mathbf{u}, \mathbf{w}, \mathbf{v}) \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{u}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w}, \tag{2.12}$$

and

$$S_k(\mathbf{g})(\mathbf{v}) = \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}} \alpha_k \int_{\mathbb{R}^{3|\boldsymbol{\beta}|} \times \mathbb{R}^{3|\boldsymbol{\alpha}|}} \sigma_{\boldsymbol{\alpha}, \boldsymbol{\beta}, k}(\mathbf{w}, \mathbf{u}, \mathbf{v}) \mathbf{g}_{\boldsymbol{\alpha}}(\mathbf{w}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w}. \tag{2.13}$$

Here, for each $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$ and $k = 1, \dots, N$,

$$\begin{aligned} &\sigma_{\boldsymbol{\alpha}, \boldsymbol{\beta}, k}(\mathbf{w}, \mathbf{u}, \mathbf{v}) := K_{\boldsymbol{\alpha}, \boldsymbol{\beta}}(\mathbf{w}, \mathbf{u}) \cdot \\ &\delta(\mathbf{w}_{k, \alpha_k} - \mathbf{v}) \cdot \delta(V_{\boldsymbol{\beta}}(\mathbf{u}) - V_{\boldsymbol{\alpha}}(\mathbf{w})) \cdot \delta(W_{\boldsymbol{\beta}}(\mathbf{u}) - W_{\boldsymbol{\alpha}}(\mathbf{w})), \end{aligned} \tag{2.14}$$

where $K_{\boldsymbol{\alpha}, \boldsymbol{\beta}} : \mathbb{R}^{3|\boldsymbol{\alpha}|} \times \mathbb{R}^{3|\boldsymbol{\beta}|} \to \mathbb{R}_+$ are given functions related to the probability of the reaction $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$. The following general properties are assumed:

1^o $K_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \equiv 0$ if $|\boldsymbol{\alpha}| < 0$, or $|\boldsymbol{\beta}| < 0$.

2^o $K_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \equiv 0$ when the probability of the collision $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is zero.

3^o $K_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \equiv 0$ if for some $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$, the condition (2.6) does not hold.

4^o $K_{\boldsymbol{\alpha}, \boldsymbol{\beta}}(\mathbf{w}, \mathbf{u})$ is invariant at the permutation of the components $\mathbf{w}_{n,1}, \dots, \mathbf{w}_{n, \alpha_n}$ of $\mathbf{w}$ for each fixed $\mathbf{u} \in \mathbb{R}^{3|\boldsymbol{\alpha}|}$, $\mathbf{w} \in \mathbb{R}^{3|\boldsymbol{\beta}|}$ and $n \in \mathcal{N}(\boldsymbol{\alpha})$; a similar

statement holds for the components of $\mathbf{u}$. (This condition expresses the "indistinguishability" of identical collision partners.)

5^o For all $\mathbf{a} \in \mathbb{R}^3$ $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$,

$$K_{\alpha,\beta}(T(\mathbf{a})\mathbf{w}, T(\mathbf{a})\mathbf{u}) \equiv K_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{w}, \mathbf{u}), \tag{2.15}$$

where $T(\mathbf{a})\mathbf{w}$ is defined on components by $(T(\mathbf{a})\mathbf{w})_{k,i} = \mathbf{w}_{k,i} + \mathbf{a}$ for $k \in \mathcal{N}(\boldsymbol{\alpha})$ and $i = 1, \ldots, \alpha_k$.

6^o There exist some given constants $C_1, \ldots, C_N > 0$, such that

$$C^{\boldsymbol{\beta}} K_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{w}, \mathbf{u}) \equiv C^{\boldsymbol{\alpha}} K_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{u}, \mathbf{w}). \tag{2.16}$$

are verified for all $(\mathbf{w}, \mathbf{u}) \in \mathbb{R}^{3|\boldsymbol{\alpha}|} \times \mathbb{R}^{3|\boldsymbol{\beta}|}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{M} \times \mathcal{M}$, where

$$C^{\boldsymbol{\gamma}} := C_1^{\gamma_1} \cdot \ldots \cdot C_N^{\gamma_N}, \tag{2.17}$$

for all $\boldsymbol{\gamma} \in \mathcal{M}$.

Note that assumption 1^o excludes the "spontaneous dissociation" as well as the "total fussion". The condition 3^o refers to the microscopic conservation of the mass. The form of $\sigma_{\boldsymbol{\alpha},\boldsymbol{\beta},k}$ in (2.14) takes into account the microscopic conservation laws of the total energy and momentum. The explicit use of only one variable, $\mathbf{w}_{k,\alpha_k}$ in $\delta(\mathbf{w}_{k,\alpha_k} - \mathbf{v})$, is possible due to "indistinguishability" of identical collision partners (condition 4^o). Assumption 5^o expresses the absence of the external fields. The generalization of the classical collision reversibility is given by the condition 6^o.

As announced before, we refer only to the space-homogeneous version of (2.10), i.e.

$$\partial_t f_k = P_k(\mathbf{f}) - S_k(\mathbf{f}), \qquad k = 1, \ldots, N. \tag{2.18}$$

Several properties (also valid in the space-inhomogeneous case [17, 18]) can be formally established as for the Ludwig and Heil equations [25], and rigorously proved by giving a meaning to (2.18) and finding classes of solutions with convenient regularity properties. Thus, formally,

$$\sum_{k=1}^{N} \int_{\mathbb{R}^3} \Phi_k^i(\mathbf{v}) \left[P_k(\mathbf{f})(\mathbf{v}) - S_k(\mathbf{f})(\mathbf{v})\right] \mathrm{d}\mathbf{v} = 0, \quad i = 0, \ldots, 4, \tag{2.19}$$

provided that all integrals involved are convergent, where $\Phi_n^0(\mathbf{v}) := m_n$, $\Phi_n^i(\mathbf{v}) = m_n v_i$, for the component v_i, $i = 1, 2, 3$, of $\mathbf{v}$, and $\Phi_n^4(\mathbf{v}) := m_n \mathbf{v}^2/2 + E_n$. By (2.19) the solutions of (2.18) are formally compatible

with the conservation of the mass ($i = 0$), bulk momentum ($i = 1, 2, 3$) and energy ($i = 4$), respectively.

One can define the H-function

$$H(\mathbf{f})(t) = \sum_{k=1}^{N} \int_{\mathbb{R}^3} \left[\log C_k f_k(t, \mathbf{v}) - 1\right] f_k(t, \mathbf{v}) \mathrm{d}\mathbf{v}, \tag{2.20}$$

for those solutions $\mathbf{f}(t, \mathbf{v})$ of (2.18), with positive components, provided that the integrals exist. In (2.20) the constants C_k are the same to the constants from the assumption 6^o. Formally, by a few algebraic manipulations, one obtains

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} H(\mathbf{f})(t) &= \sum_{k=1}^{N} \int_{\mathbb{R}^3} \left[P_k(\mathbf{f})(t, \mathbf{v}) - S_k(\mathbf{f})(t, \mathbf{v})\right] \log C_k f_k(t, \mathbf{v}) \mathrm{d}\mathbf{v} = \\ &= \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}} \int_{\mathbb{R}^{3|\boldsymbol{\beta}|} \times \mathbb{R}^{3|\boldsymbol{\alpha}|}} K_{\boldsymbol{\beta}, \boldsymbol{\alpha}}(\mathbf{u}, \mathbf{w}) \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{u}) F\left[\frac{C^{\boldsymbol{\alpha}} \mathbf{f}_{\boldsymbol{\alpha}}(t, \mathbf{w})}{C^{\boldsymbol{\beta}} \mathbf{f}_{\boldsymbol{\beta}}(t, \mathbf{u})}\right] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w} \leq 0, \end{aligned} \tag{2.21}$$

where $F(x) := \frac{1}{2}(1 - x) \log x \leq 0$ for $x \geq 0$.

The equilibrium solutions of (2.18) are Maxwellian (Gaussian) functions with determining constants (concentration, bulk velocity and temperature) related to the internal energies E_n and the constants C_n of (2.16), by the law of the mass action (for more details see e.g. Ref. [25, 17]).

We distinguish the following particular cases:

1. If $M = 3$ in (2.10-2.13), and the conditions of (2.16) are verified, then (2.10) essentially reduces to the Ludwig and Heil system of equations with discrete internal energies.

2. If $M = 2$ and the conditions of (2.16) are fulfilled with $C_1 = C_2 = 1$, then we obtain the Wang-Chang-Uhlenbeck-de Boer system of equations.

3. If $M = 2$, $N = 1$, the condition (2.16) are fulfilled and the transition functions depend only on the relative velocities of the encounters in each collision channel, then one gets the classical Boltzmann equation.

In order to introduce the numerical scheme associated to the equations (2.18), in the next section we solve a Cauchy problem for (2.18) formulated in a product of $\mathbb{L}^1$ spaces. Besides the uniqueness and global existence of the

solution, we also need the positivity of the solution and the macroscopic mass conservation. Note that, other conservation properties, as well as the existence of a H-theorem play no role in this numerical scheme. In particular, property (2.16) is not needed. However, we will state without proof a general result concerning the conservation relations and a H-theorem (only for the sake of completeness).

Roughly speaking, we would like to approximate the measures $d\mu_k^t(\mathbf{v}) := f_k(t,\mathbf{v})d\mathbf{v}$ induced by the solutions $f_k(t,\mathbf{v})$ of (2.18), $k = 1,\dots,N$, by convenient homogeneous sums of point measures, defined as follows.

Let μ be a finite positive measure on $\mathbb{R}^m$. For $a_n > 0$, where $n \in \mathbb{N}^* := \{1,2,\dots\}$, let

$$\sigma_n = \frac{a_n}{n}\sum_{i=1}^{n}\delta_{x_{i,n}}, \quad n \in \mathbb{N}^*. \tag{2.22}$$

Here $\delta_{x_{i,n}}$ is the Dirac measure on $\mathbb{R}^m$ concentrated at point $x_{i,n}$ for $i = 1,\dots,n$. The sequence of measures $(\sigma_n)_{n\in\mathbb{N}^*}$ is called a *homogeneous sum of point measures* (HSPM) approximating the measure μ, if it converges weakly to μ (in the weak sens of the measures) i.e. $\sigma_n \rightharpoonup \mu$ as $n \to \infty$.

We call a sequence $(\sigma_n)_{n\in\mathbb{N}^*}$ of the form

$$\sigma_n = \sum_{i=1}^{n}\frac{a_{i,n}}{n}\delta_{x_{i,n}}, \quad n \in \mathbb{N}^*, \tag{2.23}$$

(where $a_{i,n} > 0$ for $i \in \{1,\dots,n\}$ and $n \in \mathbb{N}^*$) a *weighted sum of point measures* (WSPM) approximating the measure μ, if it converges weakly to μ, i.e. $\sigma_n \rightharpoonup \mu$ as $n \to \infty$. Obviously, if $a_{i,n} = a_{j,n}$ for $i,j \in \{1,\dots,n\}$ and $n \in \mathbb{N}^*$, the WSPM approximation becomes a HSPM approximation.

The HSPM approximation is convenient for numerical solving of equations where the solutions are finite (probability) measures on $\mathbb{R}^m$, and where one also wishes to approximate moments of some (random) variables with respect to solutions. In this case, the control of the approximation can be made by means of the Koksma-Hlavka inequality Ref. [23], in terms of discrepancy.

We recall that, by definition Ref. [5, 15, 23], the discrepancy between the nonnegative measures μ and ν on $\mathbb{R}^m$ is given by the following formula,

$$D(\mu,\nu) := \sup_{\mathbf{a}\in\mathbb{R}^m}\left|\mu(\Lambda(\mathbf{a})) - \nu(\Lambda(\mathbf{a}))\right|, \tag{2.24}$$

where $\Lambda(\mathbf{a}) := \{\mathbf{x} \in \mathbb{R}^m \,|\, x_l \le a_l,\ l = 1,\dots,m\}$.

We also recall, Ref. [5], that a sequence of measures μ_n is said to converge to μ with respect to discrepancy if, $D(\mu_n, \mu) \to 0$ as $n \to \infty$.

It is known, Ref. [5], that if μ is a measure absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, then the convergence of μ_n to μ with respect to discrepancy is equivalent to the weak convergence in the sense of measures.

Starting with HSPM approximation for each μ_k^0 induced by the initial data in (2.18), with $k = 1, \ldots, N$, our purpose is to provide a convergent algorithm generating HSPM approximations for the measures μ_k^t, where $k = 1, \ldots, N$, at any $t > 0$.

In this respect, one chooses some fixed timestep $\Delta t < T$. Let

$$T_\Delta := \left[\left[\frac{T}{\Delta t}\right]\right], \tag{2.25}$$

where $[[x]]$ denotes the integer part of $x \in \mathbb{R}$. One associates a time-discretized version of equations to (2.18). Starting with an initial data, $f_k^0 = f_k^0(\mathbf{v})$, $k = 1, \ldots, N$, one obtains a family of functions $f_k^j(\mathbf{v})$, $j = 1, \ldots, T_\Delta$ verifying the discretized form of (2.18). The discretized version of (2.18) can be formulated in the weak form for the measures $d\bar{\mu}_k^j(\mathbf{v}) := f_k^j(\mathbf{v})d\mathbf{v}$, where $k = 1, \ldots, N$. We shall prove that if, each $\bar{\mu}_k^0$ is close, to μ_k^0, in some sense, then (for Δt sufficiently small), $\bar{\mu}_k^j$ is close to μ_k^t on the interval $((j-1)\Delta t, j\Delta t]$, with an error of order Δt, for all $j = 1, \ldots, T_\Delta$ and $k = 1, \ldots, N$.

The scheme is initialized for $k = 1, \ldots, N$ by approximating for the measures $\bar{\mu}_k^0$ by a HSPM approximation of the form:

$$\mu_{k,n}^0 := \frac{a_{k,n}}{n} \sum_{i=1}^{n} \delta_{\mathbf{v}_{k,n}} \rightharpoonup \bar{\mu}_k^0, \quad \text{as } n \to \infty. \tag{2.26}$$

The above approximation provides for all $j = 1, \ldots, T_\Delta$ and $k = 1, \ldots, N$ approximations by discrete measures $\mu_{k,n}^j \rightharpoonup \bar{\mu}_k^j$ as $n \to \infty$.

Because of the nonlinearity of the initial problem, each step of the iteration produces a power-like growing number of terms in the sums of point measures expressing $\mu_{k,n}^j$. In computations, the numerical effort would also be power-like increasing, so that the algorithm could not be effective at this level.

To approximate $\bar{\mu}_k^j$ by sums of Dirac measures with a non-increasing number of terms, for technical reasons, it is necessary to have a HSPM approximation. However, in general, $\mu_{k,n}^j$ appears as a WSPM of the form (2.23). For this reason we introduce a *homogenization procedure* of approximation to obtain measures of the form (2.22). At this level, one can reduce the numerical

effort by using probabilistic techniques of selection. Then, the convergence of the numerical scheme is proved in probabilistic terms.

3. The Existence of the Solution

Define the space $\mathbb{X} := \underbrace{\mathbb{L}^1(\mathbb{R}^3) \times \ldots \times \mathbb{L}^1(\mathbb{R}^3)}_{N \text{ times}}$ — *real*, equipped with the norm

$$\|\mathbf{g}\|_{\mathbb{X}} := \sum_{k=1}^{N} m_k \|g_k\|_{\mathbb{L}^1}, \tag{3.1}$$

where $\mathbf{g} = (g_1, \ldots, g_N)$ and $g_k \in \mathbb{L}^1(\mathbb{R}^3)$, $k = 1, \ldots, N$. We recall that $m_k > 0$ denotes the mass of a particle of species k for each $k = 1, \ldots, N$.

Note that if $\mathbf{g} \geq 0$ (i.e. $g_k \geq 0$ a.e. for all $k = 1, \ldots, N$) then the norm $\|g\|_{\mathbb{X}}$ is equal to the mass of the gas in the state described by the distribution functions given by the components of $\mathbf{g}$.

For approximation purposes, we suppose that the functions of the family $\{K_{\boldsymbol{\alpha},\boldsymbol{\beta}}\}_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}}$ are *continuous*. We formulate the Cauchy problem for (2.18) in the space $\mathbb{X}$.

Before, we must give a meaning to the collision operators P_k and S_k as operators acting in the space $\mathbb{X}$. This can be performed, using regularization as in Ref. [16, 17] to define $\sigma_{\boldsymbol{\alpha},\boldsymbol{\beta},k}$ as distributions for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M} \times \mathcal{M}$ and $k = 1, \ldots, N$.

For $m \in \mathbb{N}^*$ denote by $C_b(\mathbb{R}^m)$ the space of the bounded functions of $C(\mathbb{R}^m; \mathbb{R})$, endowed with the usual sup norm. Let $C_c(\mathbb{R}^m)$ be the subset of the functions of $C_b(\mathbb{R}^m)$ with compact support.

Let $J \in C_c(\mathbb{R})$ be positive and even function, such that $\text{supp}(J) = [-1, 1]$ and $\|J\|_{\mathbb{L}^1} = 1$. For $\varepsilon > 0$ denote by $\delta_\varepsilon(t) =: \varepsilon^{-1} J(\varepsilon^{-1} \cdot t)$ and $\delta_\varepsilon^3(y) := \delta_\varepsilon(y_1) \cdot \delta_\varepsilon(y_2) \cdot \delta_\varepsilon(y_3)$, where $y = (y_1, y_2, y_3) \in \mathbb{R}^3$. Define

$$\sigma^{\varepsilon,\eta}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{u}, \mathbf{w}) := K_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{w}, \mathbf{u}) \delta_\varepsilon^3(V_{\boldsymbol{\beta}}(\mathbf{u}) - V_{\boldsymbol{\alpha}}(\mathbf{w})) \delta_\eta(W_{\boldsymbol{\beta}}(\mathbf{u})) - W_{\boldsymbol{\alpha}}(\mathbf{w})), \tag{3.2}$$

$$P_{k\varepsilon\eta}(\mathbf{g})(\mathbf{v}) := \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \boldsymbol{\alpha}_k \left[\int_{\mathbb{R}^{3|\boldsymbol{\beta}|} \times \mathbb{R}^{3|\boldsymbol{\alpha}|-3}} \sigma^{\varepsilon,\eta}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{u}, \mathbf{w}) \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{u}) \mathrm{d}\mathbf{u} \mathrm{d}\tilde{\mathbf{w}}_k \right]_{\mathbf{w}_{k,\alpha_k} = \mathbf{v}} \tag{3.3}$$

and

$$S_{k\varepsilon\eta}(\mathbf{g})(\mathbf{v}) := \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \alpha_k \left[\int_{\mathbb{R}^{3|\boldsymbol{\beta}|}\times\mathbb{R}^{3|\boldsymbol{\alpha}|-3}} \sigma^{\varepsilon,\eta}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{w},\mathbf{u})\mathbf{g}_{\boldsymbol{\alpha}}(\mathbf{w})\mathrm{d}\mathbf{u}\mathrm{d}\tilde{\mathbf{w}}_k\right]_{\mathbf{w}_{k,\alpha_k}=\mathbf{v}}, \tag{3.4}$$

with $\mathbf{g}_{\boldsymbol{\alpha}}$ and $\mathbf{g}_{\boldsymbol{\beta}}$ as in (2.11), for all $\mathbf{g}\in C_c(\mathbb{R}^3)^N := \underbrace{C_c(\mathbb{R}^3)\times\ldots\times C_c(\mathbb{R}^3)}_{N \text{ times}}$; $\mathbf{v}\in\mathbb{R}^3$, $k\in 1,\ldots,N$. In (3.3) and (3.4), the terms with $\alpha_k = 0$, vanish, by definition, and $d\tilde{\mathbf{w}}_k$ is the Euclidean element of area on the manifold $\left\{\mathbf{w}\in\mathbb{R}^{3|\boldsymbol{\alpha}|} | \mathbf{w}_{k,\alpha_k} = \mathbf{v}\right\}$.

Let $\boldsymbol{\Omega}_{\boldsymbol{\gamma}}$ be the unit sphere in $\mathbb{R}^{3|\boldsymbol{\gamma}|-3}$, where $\boldsymbol{\gamma}\in\mathcal{M}$. The operators P_k and S_k can be defined by means of the following result.

LEMMA **3.1** *For each $\mathbf{g}\in C_c^N(\mathbb{R}^3)$, there exist the limits*

$$\dot{P}_k(\mathbf{g})(\mathbf{v}) := \lim_{\eta\to 0}\lim_{\varepsilon\to 0} P_{k\varepsilon\eta}(\mathbf{g})(\mathbf{v}), \qquad \dot{S}_k(\mathbf{g})(\mathbf{v}) := \lim_{\eta\to 0}\lim_{\varepsilon\to 0} S_{k\varepsilon\eta}(\mathbf{g})(\mathbf{v}). \tag{3.5}$$

There are the families of functions $\{r_{\boldsymbol{\beta},\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}}$, $\{p_{\boldsymbol{\beta},\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \subset C(\mathbb{R}^{3|\boldsymbol{\alpha}|}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}};\mathbb{R}_+)$ and $\{\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \subset C(\mathbb{R}^{3|\boldsymbol{\alpha}|}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}};\mathbb{R}^{3|\boldsymbol{\beta}|})$ such that

$$\dot{P}_k(\mathbf{g})(\mathbf{v}) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \alpha_k \left[\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|-3}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}}} p_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}))\mathrm{d}\tilde{\mathbf{w}}_k\mathrm{d}\mathbf{n}\right]_{\mathbf{w}_{k,\alpha_k}=\mathbf{v}}, \tag{3.6}$$

$$\dot{S}_k(\mathbf{g})(\mathbf{v}) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \alpha_k \left[\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|-3}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\mathbf{g}_{\boldsymbol{\alpha}}(\mathbf{w})\mathrm{d}\tilde{\mathbf{w}}_k\mathrm{d}\mathbf{n}\right]_{\mathbf{w}_{k,\alpha_k}=\mathbf{v}}, \tag{3.7}$$

for all $\mathbf{g}\in C_c^N(\mathbb{R}^3)$, and the following properties are verified:

i) *there are some constants $c,d>0$ such that $|\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})| \geq c|\mathbf{w}|$ for all $|\mathbf{w}|\geq d$ and $\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}$.*

ii) *if $W_{\boldsymbol{\alpha}}(\mathbf{w}) - 2^{-1}(\sum_{n=1}^N \alpha_n m_n)V_{\boldsymbol{\alpha}}(\mathbf{w})^2 - \sum_{n=1}^N \beta_n E_n \leq 0$ for some $\mathbf{w}\in\mathbb{R}^{3|\boldsymbol{\alpha}|}$, then*

$$r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}) = p_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}) = 0, \quad \text{for all } \mathbf{n}\in\boldsymbol{\Omega}_{\boldsymbol{\beta}} \text{ and } \boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}. \tag{3.8}$$

iii) *for each $\varphi\in C(\mathbb{R}^{3|\boldsymbol{\alpha}|})$ and $f\in C_c(\mathbb{R}^{3|\boldsymbol{\beta}|})$ and $\forall\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}$*

$$\begin{aligned}&\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}}} \varphi(\mathbf{w})\cdot p_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\cdot\mathbf{f}(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}))\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n} = \\ &= \int_{\mathbb{R}^{3|\boldsymbol{\beta}|}\times\boldsymbol{\Omega}_{\boldsymbol{\alpha}}} \varphi(\mathbf{u}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{u},\mathbf{n}))\cdot r_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{u},\mathbf{n})\cdot\mathbf{f}(\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{n}.\end{aligned} \tag{3.9}$$

The results of the above Lemma were obtained in Ref. [17]. However, for the sake of completeness, the proof is outlined in Appendix[2].

Property (3.8) follows by the presence of reaction thresholds (in the frame of the conservation relations (2.6) and (2.7)). Moreover, (3.6) and (3.7) are well defined, because of property i) in Lemma 3.1.

From (3.7), we can write

$$\dot{S}_k(\mathbf{g})(\mathbf{v}) = \dot{R}_k(\mathbf{g})(\mathbf{v})g_k(\mathbf{v}), \tag{3.10}$$

where

$$\begin{aligned} &\dot{R}_k(\mathbf{g})(\mathbf{v}) := \\ &:= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}} \boldsymbol{\alpha}_k \left[\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|-3}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\mathbf{g}_{\boldsymbol{\gamma};k}(\mathbf{w}_{s,i})\mathrm{d}\tilde{\mathbf{w}}_k\mathrm{d}\mathbf{n} \right]_{\mathbf{w}_{k,\alpha_k}=\mathbf{v}} . \end{aligned} \tag{3.11}$$

In (3.11), for $\boldsymbol{\gamma}\in\mathcal{N}(\boldsymbol{\gamma})$ we assumed the convention

$$\mathbf{g}_{\boldsymbol{\gamma};k}(\mathbf{w}) := \mathbf{g}_{\boldsymbol{\gamma}}(\mathbf{w})/g_k(\mathbf{w}_{k,\alpha_k}), \tag{3.12}$$

where the r.h.s. makes sense and $\mathbf{g}_{\boldsymbol{\gamma};k}(\mathbf{w}) := 0$ otherwise.

Our results are based on the following

Assumption

There is a constant $K > 0$, such that

$$\int_{\boldsymbol{\Omega}_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\mathrm{d}\mathbf{n} < K, \tag{3.13}$$

for all $\mathbf{w}\in\mathbb{R}^{3|\boldsymbol{\alpha}|}$ and $\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}$.

From (3.13), it is immediate that the maps

$$\begin{aligned} &\mathbb{X}\supset C_c(\mathbb{R}^3)^N \ni \mathbf{g}\to \dot{S}_k(\mathbf{g})\in\mathbb{L}^1(\mathbb{R}^3), \\ &\mathbb{X}\supset C_c(\mathbb{R}^3)^N \ni \mathbf{g}\to \dot{R}_k(\mathbf{g})\in C_b(\mathbb{R}^3) \end{aligned} \tag{3.14}$$

are continuous for each $k = 1,\ldots,N$. Moreover, using property (3.9) (with $\varphi = 1$, $f = \mathbf{g}_{\boldsymbol{\beta}}$) combined with Fubini's theorem, it also follows that the map

$$\mathbb{X}\supset C_c(\mathbb{R}^3)^N \ni \mathbf{g}\to \dot{P}_k(\mathbf{g})\in\mathbb{L}^1(\mathbb{R}^3) \tag{3.15}$$

[2]Note that the functions $r_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and $p_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ appear in explicit form in the proof of Lemma 3.1 (see the Appendix).

is continuous for each $k = 1, \ldots, N$.

Since $C_c(\mathbb{R}^3)^N$ is dense in $\mathbb{X}$, the maps given by (3.14-3.15) have continuous extensions to $\mathbb{X}$. These extensions will be also denoted S_k, R_k and P_k, respectively.

Note that (3.10) can be extended to all $\mathbf{g} \in \mathbb{X}$, in the sense that a.e.,

$$S_k(\mathbf{g})(\mathbf{v}) = R_k(\mathbf{g})(\mathbf{v}) g_k(\mathbf{v}), \tag{3.16}$$

for all $k = 1, \ldots, N$.

Define $\mathbf{P}, \mathbf{S} : \mathbb{X} \to \mathbb{X}$ by

$$\begin{aligned} \mathbf{P}(\mathbf{g}) &= (P_1(\mathbf{g}), \ldots, P_N(\mathbf{g})), \\ \mathbf{S}(\mathbf{g}) &= (S_1(\mathbf{g}), \ldots, S_N(\mathbf{g})), \end{aligned} \tag{3.17}$$

for all $\mathbf{g} \in \mathbb{X}$.

We consider the Cauchy problem for equation (2.18) in $\mathbb{X}$.

$$d_t f(\mathbf{f}) = \mathbf{P}(\mathbf{f}(t)) - \mathbf{S}(\mathbf{f}(t)), \quad \mathbf{f}(0) = \mathbf{f}_0. \tag{3.18}$$

THEOREM **3.1** *Let $\mathbf{f}_0 > 0$. For each $T > 0$, equation (3.18) has a unique solution $\mathbf{f}(t)$ in $\mathbb{X}$ on $[0, T]$. Moreover, for all $t \in [0, T]$ one has $\mathbf{f}(t) > 0$ and*

$$\sum_{k=1}^{N} m_k \int_{\mathbb{R}^3} f_k(t, \mathbf{v}) \mathrm{d}\mathbf{v} = \sum_{k=1}^{N} m_k \int_{\mathbb{R}^3} f_{0,k}(\mathbf{v}) \mathrm{d}\mathbf{v}. \tag{3.19}$$

Proof. One applies the Banach fixed point theorem to (3.18) written in convenient form.

Consider the cone $C_T^+ := \{\mathbf{f} \in C(0, T; \mathbb{X}) | \mathbf{f}(t) \geq 0, \text{for all } t \in [0, T]\}$ with the norm

$$\|\mathbf{f}\| := \sup_{t \in [0,T]} \|\mathbf{f}(t)\|_{\mathbb{X}}. \tag{3.20}$$

Observe that for all $k = 1, \ldots, N$, if $\mathbf{f} \in C_T^+$ then $R_k(\mathbf{f}), P_k(\mathbf{f}) \geq 0$ (since $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}$, $p_{\boldsymbol{\beta},\boldsymbol{\alpha}} \geq 0$, for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}$). Moreover, if $\mathbf{f} \in C_T^+$, then $R_k(\mathbf{f}) \in C(0, T; C_b(\mathbb{R}^3))$. Consequently the Riemann integral $\int_s^t R_k(\mathbf{f}(\tau)) \mathrm{d}\tau$ is well defined in $C_b(\mathbb{R}^3)$ for all $s, t \in [0, T]$ and $k \in \{1, \ldots, N\}$.

Let $\mathbf{f} \in C_T^+$. We define the map $[0, T] \ni t \to \mathbf{I}(\mathbf{f})(t) \in \mathbb{X}$ by the components of $\mathbf{I}(\mathbf{f})(t)$, as:

$$I_k(\mathbf{f})(t) = \exp\left[-\int_0^t R_k(\mathbf{f}(\tau))\mathrm{d}\tau\right]\cdot \mathbf{f}_{0,k}+ \\ +\int_0^t \exp\left[-\int_s^t R_k(\mathbf{f}(\tau))\mathrm{d}\tau\right]\cdot P_k(\mathbf{f}(s))\mathrm{d}s, \tag{3.21}$$

where $t\in[0,T]$. Here, the integration with respect to ds is in the sense of Riemann in $\mathbb{L}^1(\mathbb{R}^3)$.

Obviously $I_k(\mathbf{f})(t)\geq 0$ for all $t\in[0,T]$, $k=1,\ldots,N$.

The problem (3.18) can be rewritten in C_T^+, as it follows.

$$\mathbf{f}=\mathbf{I}(\mathbf{f}) \tag{3.22}$$

Let $R>\|\mathbf{f}_0\|_{\mathbb{X}}$. Define

$$\mathcal{B}(R):=\left\{\mathbf{f}\in C_T^+ \,|\, \|\mathbf{f}\|\leq R, \quad \mathbf{f}\,(0)=\mathbf{f}_0\right\}. \tag{3.23}$$

Using (3.11), (3.6) and (3.13), one can find some positive numbers $k_1(R)$ and $k_2(R)$, such that

$$\|\mathbf{I}(\mathbf{f})\|\leq \|\mathbf{f}_0\|_{\mathbb{X}}+T\cdot k_1(R), \tag{3.24}$$

and

$$\|\mathbf{I}(\mathbf{f})-\mathbf{I}(\mathbf{h})\|\leq T\cdot k_2(R)\cdot\|\mathbf{f}-\mathbf{h}\|, \tag{3.25}$$

for all $\mathbf{f},\mathbf{h}\in\mathcal{B}(R)$. Obviously, from (3.24) and (3.25), for T small enough, the map $\mathbf{I}$ becomes a strict contraction on $\mathcal{B}(R)$. Consequently $\mathbf{I}:\mathcal{B}(R)\to\mathcal{B}(R)$ and has a unique fixed point. This proves that (3.18) has a unique positive solution on $[0,T]$.

The positivity of f_k, implies that

$$\|\mathbf{f}(t)\|_{\mathbb{X}}=\sum_{k=1}^N m_k\int_{\mathbb{R}^3} f_k(t,\mathbf{v})\mathrm{d}\mathbf{v}, 0\leq t\leq T. \tag{3.26}$$

By (3.18) and using (2.6), (3.11), (3.6) and (3.9) (applied to $\varphi\equiv 1$) one obtains

$$d_t\|\mathbf{f}(t)\|_{\mathbb{X}}=\sum_{k=1}^N m_k\int_{\mathbb{R}^3}\left[P_k(\mathbf{f})-S_k(\mathbf{f})\right]\mathrm{d}\mathbf{v}=0, \tag{3.27}$$

which proves (3.19). Moreover,

$$\|\mathbf{f}\|=\sup_{0\leq t\leq T}\|\mathbf{f}(t)\|_{\mathbb{X}}=\|\mathbf{f}_0\|_{\mathbb{X}}. \tag{3.28}$$

By continuation, and uniqueness, the local solution $\mathbf{f}(t)$ can be made time-global. This ends the proof. □

For the sake of completeness we state the following result.

Let Φ_n^i be as in (2.19) for $i = 1, \ldots, 4$. With the remark that the mass conservation (3.19) has been already proved, the solution of (3.18) has the following properties.

PROPOSITION **3.1** *Let* $\mathbf{f}(t)$ *be the solution of* (3.18) *given by Theorem* 3.1.

a) *If*

$$f_{0,k},\ (1+\mathbf{v}^2)f_{0,k} \in \mathbb{L}^1(\mathbb{R}^3) \tag{3.29}$$

for each $k = 1, \ldots, N$, *then*

$$(1+\mathbf{v}^2)f_k(t) \in \mathbb{L}^1(\mathbb{R}^3) \tag{3.30}$$

and

$$\sum_{n=1}^{N} \int_{\mathbb{R}^3} \Phi_n^i(\mathbf{v}) f_n(t,\mathbf{v}) \mathrm{d}\mathbf{v} = \sum_{n=1}^{N} \int_{\mathbb{R}^3} \Phi_n^i(\mathbf{v}) f_{0,n}(\mathbf{v}) d\mathbf{v} = 0, \tag{3.31}$$

for each $k = 1, \ldots, N$ *and* $i = 1, \ldots, 4$ *and all* $t \geq 0$.

b) *In addition to the conditions (3.29), suppose that there are some constants* $C_1, \ldots, C_N > 0$ *such that conditions (2.16) hold. If*

$$f_{0,k} \log f_{0,k} \in \mathbb{L}^1(\mathbb{R}^3) \tag{3.32}$$

for each $k = 1, \ldots, N$, *then*

$$f_k(t) \log f_k(t) \in \mathbb{L}^1(\mathbb{R}^3; \mathrm{d}\mathbf{v}) \tag{3.33}$$

and

$$H(\mathbf{f})(t) := \sum_{n=1}^{N} \int_{\mathbb{R}^3} \left[\log C_n f_n(t,\mathbf{v}) - 1\right] f_n(t,\mathbf{v}) \mathrm{d}\mathbf{v} \tag{3.34}$$

is non-increasing as a function of t, *for each* $k = 1, \ldots, N$ *and all* $t \geq 0$.

The proof of this proposition is beyond the present purposes. Though, we mention that the proof uses Lemma 3.1 and the ideas introduced by of Arkeryd [2, 3] to prove results of the same nature in the case of the classical space-homogeneous Boltzmann equation.

4. Time Discretization

Let $\Delta t \in (0,T)$ be a fixed timestep. We consider the following discretized version of (3.18).

$$\begin{aligned} &\mathbf{f}^j = \mathbf{f}^{j-1} + \Delta t \cdot \left[\mathbf{P}(\mathbf{f}^{j-1}) - \mathbf{S}(\mathbf{f}^{j-1})\right], \\ &\mathbf{f}^0 = \mathbf{f}_0 \geq 0, \ a.e., \quad j = 1, \ldots, T_\Delta, \end{aligned} \tag{4.1}$$

where $\mathbf{f}^j = (f_1^j, \ldots, f_N^j)$ and $f_k^j = f_k^j(\mathbf{v})$.

The discretized scheme (4.1) may destroy the positivity of the functions $\mathbf{f}^j$ for $j \geq 1$. However, one can prove that for Δt small enough, $\mathbf{f}^j$ is positive and close, in some sense, to the solution $\mathbf{f}$ provided by Theorem 3.1.

PROPOSITION **4.1**

a) *If Δt is sufficiently small, then $\mathbf{f}^j \geq 0$ for all $j = 1, \ldots, T_\Delta$. Moreover,*

$$\left\|\mathbf{f}^j\right\| = \left\|\mathbf{f}_0\right\|, \tag{4.2}$$

for all $j = 1, \ldots, T_\Delta$.

b) *There exists some number $C = C(\|\mathbf{f}_0\|_{\mathbb{X}}) > 0$, depending only on $\|\mathbf{f}_0\|_{\mathbb{X}}$, such that*

$$\left\|\mathbf{f}(t) - \mathbf{f}^j\right\|_{\mathbb{X}} \leq C \cdot \Delta t, \tag{4.3}$$

for all $j = 1, \ldots, T_\Delta$ and $t \in ((j-1)\Delta t, j\Delta t]$.

Proof. a) First we write (4.1) more conveniently.

Let

$$\mathcal{U} := \{\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N) \,|\gamma_k \in \{0, 1, \ldots, NM\}, |\boldsymbol{\gamma}| \geq 2\}. \tag{4.4}$$

For any $\boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N$ for $k = 1, ..., N$ and $\boldsymbol{\alpha} \in \mathcal{M}$, denote

$$\boldsymbol{\xi}_{\boldsymbol{\alpha},k} := \begin{cases} \dfrac{1}{\xi_k} \displaystyle\prod_{n \in \mathcal{N}(\boldsymbol{\alpha})} \xi_n^{\alpha_n} & \text{if } \alpha_k \geq 1 \text{ and } \xi_k \neq 0, \\ \\ 0 & \text{if } \alpha_k = 0 \text{ or } \quad \xi_k = 0. \end{cases} \tag{4.5}$$

For $k = 1, \ldots, N$ and $\boldsymbol{\alpha} \in \mathcal{M}$, using the multinomial formula, we get

$$\sum_{p=2}^{NM} (\xi_1 + \ldots + \xi_N)^{p-1} = \sum_{p=2}^{NM} p^{-1} \partial_{\xi_k} (\xi_1 + \ldots + \xi_N)^p = \sum_{\boldsymbol{\alpha} \in \mathcal{U}} c^{\boldsymbol{\alpha}} \alpha_k \boldsymbol{\xi}_{\boldsymbol{\alpha},k}, \tag{4.6}$$

where

$$c^{\boldsymbol{\alpha}} := (|\boldsymbol{\alpha}| - 1)! \left(\prod_{k=1}^{N} \alpha_k! \right)^{-1}. \tag{4.7}$$

If

$$\xi_1 + \ldots + \xi_N = 1, \tag{4.8}$$

then, by (4.6) we get

$$MN - 1 = \left[\frac{1}{(M+1)^N - N - 1} \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \alpha_k c_{\boldsymbol{\alpha}} \boldsymbol{\xi}_{\boldsymbol{\alpha},k} + \sum_{\boldsymbol{\alpha} \in \mathcal{U} \setminus \mathcal{M}} \alpha_k c_{\boldsymbol{\alpha}} \boldsymbol{\xi}_{\boldsymbol{\alpha},k} \right]. \tag{4.9}$$

For each $k = 1, \ldots, N$, put

$$\xi_k = \mu_k \mathfrak{I}_k, \tag{4.10}$$

where

$$\mu_k = m_k \left(\sum_{n=1}^{N} m_n \int_{\mathbb{R}^3} f_{0,n}(\mathbf{v}) \mathrm{d}\mathbf{v} \right)^{-1} \tag{4.11}$$

and

$$\mathfrak{I}_k = \int_{\mathbb{R}^3} f_k^j(\mathbf{v}) \mathrm{d}\mathbf{v}. \tag{4.12}$$

It follows that (4.8) is satisfied, due to (4.19). Consequently, by (4.9),

$$1 = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \alpha_k \cdot \Gamma^{\boldsymbol{\alpha},k} \cdot \mathbf{I}_{\boldsymbol{\alpha},k} + \sum_{\boldsymbol{\alpha} \in \mathcal{U} \setminus \mathcal{M}} \Lambda^{\boldsymbol{\alpha},k} \cdot \mathbf{I}_{\boldsymbol{\alpha},k}, \tag{4.13}$$

where the notation $\mathbf{I}_{\boldsymbol{\alpha},k}$ is given by (4.5) for $\mathbf{I} = (\mathfrak{I}_1, \ldots, \mathfrak{I}_N)$. In (4.13),

$$\Lambda^{\boldsymbol{\alpha},k} := \frac{\alpha_k c^{\boldsymbol{\alpha}} \mu_1^{\alpha_1} \cdot \ldots \mu_{k-1}^{\alpha_{k-1}} \cdot \mu_k^{\alpha_k - 1} \cdot \mu_{k+1}^{\alpha_{k+1}} \cdot \ldots \cdot \mu_N^{\alpha_N}}{MN - 1} \tag{4.14}$$

and

$$\Gamma^{\boldsymbol{\alpha},k} := \frac{c^{\boldsymbol{\alpha}} \mu_1^{\alpha_1} \cdot \ldots \mu_{k-1}^{\alpha_{k-1}} \cdot \mu_k^{\alpha_k - 1} \cdot \mu_{k+1}^{\alpha_{k+1}} \cdot \ldots \cdot \mu_N^{\alpha_N}}{(MN - 1)\left[(M+1)^N - N - 1\right]}. \tag{4.15}$$

Multiplying on components ($k = 1, \ldots, N$), the first term of the right side of (4.1) by (4.13) and using (3.11), equation (4.1) becomes

$$f_k^j = Q_k(\mathbf{f}^{j-1}) + L_k(\mathbf{f}^{j-1}) + \Delta t \cdot P_k(\mathbf{f}^{j-1}), \tag{4.16}$$

for $k=1,\dots,N$. Here

$$Q_k(\mathbf{f}^j)(\mathbf{v}):=$$

$$=\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{M}}\alpha_k\left[\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|-3}}\left(\Gamma^{\boldsymbol{\alpha},k}-\Delta t\int_{\boldsymbol{\Omega}_{\boldsymbol{\beta}}}r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})\mathrm{d}\mathbf{n}\right)\mathbf{f}^j_{\boldsymbol{\alpha}}(\mathbf{w})\mathrm{d}\mathbf{w}_{(k)}\right]_{\mathbf{w}_{\boldsymbol{\alpha},k}=\mathbf{v}},\tag{4.17}$$

$$L_k(\mathbf{f}^j)(\mathbf{v}):=\sum_{\boldsymbol{\alpha}\in\mathcal{U}\setminus\mathcal{M}}\Lambda^{\boldsymbol{\alpha},k}\left[\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|-3}}\mathrm{d}\mathbf{w}_{(k)}\mathbf{f}^j_{\boldsymbol{\alpha}}(\mathbf{w})\right]_{\mathbf{w}_{\boldsymbol{\alpha},k}=\mathbf{v}}.\tag{4.18}$$

If K is the constant introduced in (3.13), we can choose Δt such that $\Delta t\cdot K\leq\inf_{\boldsymbol{\alpha},k}\Gamma^{\boldsymbol{\alpha},k}$.

Then, the positivity of $\mathbf{f}^j$, for all $j=1,\dots,T_\Delta$, follows by induction, using Assumption (3.13). As $\mathbf{f}^j\geq 0$ for all $j=1,\dots,T_\Delta$, then the mass conservation is always fulfilled. Indeed, by induction and using the same argument as in (3.27) we have

$$\sum_{k=1}^N m_k\int_{\mathbb{R}^3}f_k^j(\mathbf{v})\mathrm{d}\mathbf{v}=\sum_{k=1}^N m_k\int_{\mathbb{R}^3}f_{k,0}(\mathbf{v})\mathrm{d}\mathbf{v}\tag{4.19}$$

for all $j=1,\dots,T_\Delta$.

b) Combining (3.18) and (4.1), for all $j=1,\dots,T_\Delta$ we can write

$$\begin{aligned}\left\|\mathbf{f}(j\cdot\Delta t)-\mathbf{f}^j\right\|_{\mathbb{X}}&\leq\left\|\mathbf{f}(j-1)\cdot\Delta t)-\mathbf{f}^{j-1}\right\|_{\mathbb{X}}+\\&+\int_{(j-1)\cdot\Delta t}^{j\cdot\Delta t}\left\|P(\mathbf{f}(s))-P(\mathbf{f}^{j-1})\right\|_{\mathbb{X}}\mathrm{d}s+\\&+\int_{(j-1)\cdot\Delta t}^{j\cdot\Delta t}\left\|S(\mathbf{f}(s))-S(\mathbf{f}^{j-1})\right\|_{\mathbb{X}}\mathrm{d}s.\end{aligned}\tag{4.20}$$

Denote by $\mathcal{O}_j:=\left\|\mathbf{f}(j\Delta t)-\mathbf{f}^j\right\|_{\mathbb{X}}$. Using the explicit forms of $\mathbf{P}$ and $\mathbf{S}$, taking account of the conservation relations (3.19) and (4.19), we find that there is some number $c_0>0$, depending on $\|\mathbf{f}_0\|_{\mathbb{X}}$ such that $\mathcal{O}_j<\mathcal{O}_{j-1}(1+c_0\Delta t)$ for all $j=2,\dots,T_\Delta$ and $\mathcal{O}_1\leq c_0\Delta t$. Then

$$\mathcal{O}_j\leq\mathcal{O}_1(1+c_0\Delta t)^{T_\Delta}\leq c_1\cdot\Delta t,\tag{4.21}$$

with $c_1>0$ depending only on $\|\mathbf{f}_0\|_{\mathbb{X}}$. Suppose that $t\in((j-1)\Delta t,j\Delta t]$.

The explicit forms of $\mathbf{P}$ and $\mathbf{S}$ together with (3.18) and (3.19) lead to

$$\begin{aligned} &\|\mathbf{f}(t) - \mathbf{f}((j-1)\Delta t)\|_{\mathbb{X}} \leq \\ &\leq \int_{(j-1)\Delta t}^{j\Delta t} (\|\mathbf{P}(\mathbf{f}(s))\|_{\mathbb{X}} + \|\mathbf{S}(\mathbf{f}(s))\|_{\mathbb{X}}) \mathrm{d}s \leq c_2 \cdot \Delta t, \end{aligned} \tag{4.22}$$

where c_2 depends only on $\|f_0\|_{\mathbb{X}}$. Now estimation (4.3) is an immediate consequence of (4.21) and (4.22). □

For numerical purposes, one has to write the equation (4.1) in the weak form for measures. In this respect, we associate the the following measures to the solutions $\mathbf{f}(t)$ and $\mathbf{f}^j$ appearing in Proposition 4.1. For $k = 1, \ldots, N$ define

$$\mathrm{d}\mu_k^t(\mathbf{v}) := f_k(t, \mathbf{v}) \mathrm{d}\mathbf{v}, \tag{4.23}$$

where $t \geq 0$, and

$$\mathrm{d}\bar{\mu}_k^j(\mathbf{v}) := f_k^j(\mathbf{v}) \mathrm{d}\mathbf{v}, \tag{4.24}$$

for $j = 1, \ldots, T_\Delta$.

Proposition 4.1 has the following consequence expressed in terms of the discrepancy defined by (2.24).

COROLLARY **4.1** *If the conditions of Proposition* 4.1 *are fulfilled, then*

$$\max_{k=1,\ldots,N} \max_{j=1,\ldots,T_\Delta} D(\mu_k^{j\Delta t}, \bar{\mu}_k^j) \to 0 \quad \textit{as } \Delta t \to 0. \tag{4.25}$$

5. The Probabilistic Frame

The central result of this section extends, in some sense, the probabilistic methods of selection used by Babovsky and Illner [4, 5] (see e.g. Lemma 2 of Ref. [4]).

We start with a simple generalization (to row-wise independent random variables) of the strong law of large numbers for independent random variables with bounded fourth momentum (see, e.g., Theorem IV.§3-1 in Ref. [28], p.363).

Let (Ω, β, P) be a probability space. For some real random variable X, by $\langle X \rangle$ we denote its mean with respect to P.

Let $\mathbb{N}^* \ni n \to q_n \in \mathbb{N}^*$. We call the family $((X_{n,i})_{i \in \{1,\ldots,q_n\}})_{n \in \mathbb{N}^*}$ of real valued random variables on Ω an *array of row-wise independent random variables*, if for each fixed $n \in \mathbb{N}^*$ the random variables $(X_{n,i})_{i \in \{1,\ldots,q_n\}}$ are independent.

PROPOSITION **5.1** *Let $((X_{n,i})_{i\in\{1,\dots,q_n\}})_{n\in\mathbb{N}^*}$ be an array of row-wise independent random variables with zero mean. Denote $A_n := \sup\limits_{i\in\{1,\dots,q_n\}} \langle X_{n,i}^4\rangle$.*

If

$$\sum_{n=1}^{\infty} \frac{A_n}{q_n^2} < \infty, \tag{5.1}$$

then, with probability one,

$$\frac{1}{q_n}\sum_{i=1}^{q_n} X_{n,i} \to 0, \ \ as\ n\to\infty. \tag{5.2}$$

Proof. According to a version of the Borel-Cantelli Lemma, it is sufficient to show that for each $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\left(\left|\frac{1}{q_n}\sum_{i=1}^{q_n} X_{n,i}\right| > \varepsilon\right) < \infty. \tag{5.3}$$

To this end, by Chebyshev's inequality, we obtain

$$P\left(\left|\sum_{i=1}^{q_n} X_{n,i}\right| > \varepsilon\cdot q_n\right) \le \frac{1}{\varepsilon^4 q_n^4}\left\langle\left|\sum_{i=1}^{q_n} X_{n,i}\right|^4\right\rangle. \tag{5.4}$$

Expanding the fourth power, we invoke the independence of $X_{n,i}$ and use the fact that $\langle X_{n,i}\rangle = 0$. Then a simple computation shows that for all $\varepsilon > 0$,

$$0 \le \sum_{n=1}^{\infty} P\left(\frac{1}{q_n}\left|\sum_{i=1}^{q_n} X_{n,i}\right| > \varepsilon\right) \le \frac{3}{\varepsilon^4}\sum_{n=1}^{\infty}\frac{A_n}{q_n^2} < \infty. \tag{5.5}$$

This concludes the proof. □

Consider $\mathbb{N}^* \ni n \to m_n \in \mathbb{N}^*$ a sequence, such that $m_n \to \infty$ as $n\to\infty$.

For each $n\in\mathbb{N}^*$, let $\mathcal{I}_n := \{1,2,\dots,m_n\}$ be an index set and let $\mathcal{I}_n^p := \underbrace{\mathcal{I}_n\times\ldots\times\mathcal{I}_n}_{p \text{ times}}$ for a fixed $p\in\mathbb{N}^*$.

Consider some given set $\mathbf{X}\subset\mathbb{R}^m$ and a given sequence $(F_n)_{n\in\mathbb{N}^*}$ of functions $F_n : \mathbf{X}\times\mathcal{I}_n^p \to \mathbb{R}$. Define $S_n : \mathbf{X}\to\mathbb{R}$ by

$$S_n(x) := \begin{cases} \dfrac{1}{m_n^p}\displaystyle\sum_{\mathbf{j}\in\mathcal{I}_n} F_n(x,\mathbf{j}) & \text{if} \quad p\ge 2, \\ \\ \displaystyle\sum_{j=1}^{m_n} a_{n,j}F_n(x,j) & \text{if} \quad p=1, \end{cases} \tag{5.6}$$

where $((a_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ is a family of nonnegative numbers, such that

$$\begin{aligned}&\sup_{n\in\mathbb{N}^*}\sum_{l=1}^{m_n}a_{n,j}<\infty,\\&\sum_{l=1}^{m_n}a_{n,l}>0,\ \text{for all } n\in\mathbb{N}^*.\end{aligned}\tag{5.7}$$

Suppose that there is some function $F:\mathbf{X}\to\mathbb{R}$ such that, for each $x\in\mathbf{X}$,

$$F(x)=\lim_{n\to\infty}S_n(x).\tag{5.8}$$

In general, for a given n, the sum S_n contains m_n^p terms. Roughly speaking, our problem is to conveniently diminish the numbers of terms in S_n, by random selection of the terms in (5.6) and "renormalize" the resulting sum such that the convergence to $F(x)$ be kept, in some sense. In this respect, we define some special families of random variables.

Let (Ω,β,P) be a probability space, where $\Omega:=[0,1)^\infty$ (in the countable sense) is endowed with the usual product Borel $\sigma-$algebra β and P the usual product probability induced on Ω by the uniform distribution of $[0,1)$.

For each $n\in\mathbb{N}^*$ and $j\in\mathcal{I}_n$, define the weights

$$p_{n,j}:=\frac{a_{n,j}}{\sum\limits_{l=1}^{m_n}a_{n,l}},\tag{5.9}$$

where $((a_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ is the family with properties (5.7). For each $n\in\mathbb{N}^*$, let

$$q_{n,s}:=\begin{cases}0 & \text{if}\quad s=0,\\ \displaystyle\sum_{j=1}^{s}p_{n,j} & \text{if}\quad s\in\mathcal{I}_n.\end{cases}\tag{5.10}$$

For each $n\in\mathbb{N}^*$ and $l\in\mathcal{I}_n$ we consider the random variables $c_{n,l}\,,\tilde{c}_{n,l}:\Omega\to\mathcal{I}_n$ given by

$$c_{n,l}(\boldsymbol{\omega}):=[[\omega_l\cdot m_n]]+1,\tag{5.11}$$

and

$$\tilde{c}_{n,l}(\boldsymbol{\omega}):=s\ \text{ if }\omega_l\in[q_{n,s-1},q_{n,s})\,,\tag{5.12}$$

where ω_l is the l^{th} component of $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots) \in \Omega$. In (5.12) we make the convention that $[x, x) := \phi$ (the void set) for any $x \in \mathbb{R}$. Obviously, for each $j \in \mathcal{I}_n$

$$P\left(c_{n,l}(\boldsymbol{\omega}) = j\right) = \frac{1}{m_n}, \tag{5.13}$$

and

$$P\left(\tilde{c}_{n,l}(\boldsymbol{\omega}) = j\right) = p_{n,j}. \tag{5.14}$$

Consequently, $((c_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ and $((\tilde{c}_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$, are arrays of row-wise independent random variables.

Remark that the random variables $c_{n,l}$ are particular forms of $\tilde{c}_{n,l}$, with $p_{n,j} = m_n^{-1}$ in (5.9).

Let $p \geq 2$. For $n \in \mathbb{N}^*$ and $l \in \mathcal{I}_n$, define the random variables $\mathbf{J}_{n,l} : \Omega \to \mathcal{I}_n^p$ by

$$\mathbf{J}_{n,l}(\boldsymbol{\omega}) := (i, c_{n,(l-1)p+1}(\boldsymbol{\omega}), c_{n,(l-1)p+2}(\boldsymbol{\omega}), \ldots, c_{n,lp-1}(\boldsymbol{\omega})), \tag{5.15}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots) \in \Omega$.

Observe that $ip + j = i'p + j'$ if and only if $i = i'$ and $j = j'$, for all $i, i' \in N^*$ and $j, j' \in \{1, 2, \ldots, p\}$. Then, using the row-wise independence of $((c_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$, we conclude the row-wise independence of $((\mathbf{J}_{n,l})_{l\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$.

Suppose that one of the following conditions is fulfilled:

1. $\mathbf{X}$ is at most countable.

2. $\mathbf{X}$ is the whole $\mathbb{R}^m$, the function F is continuous and each $F_n(\cdot, \mathbf{j})$ is increasing with respect to the order of $\mathbb{R}^m$ for each fixed $n \in \mathbb{N}^*$ and $\mathbf{j} \in \mathcal{I}_n^p$. Define for each $n \in \mathbb{N}^*$ and $x \in \mathbf{X}$ by

$$a_n(x) := \max_{\mathbf{j}\in\mathcal{I}_n^p} \left|F_n(x, \mathbf{j})\right|. \tag{5.16}$$

PROPOSITION **5.2** 1. *Let $p \geq 2$. If*

$$\sum_{n=1}^{\infty} \frac{a_n(x)^4}{m_n^2} < \infty \tag{5.17}$$

for all $x \in \mathbf{X}$, then for each $x \in \mathbf{X}$, with probability one,

$$\lim_{n\to\infty} \frac{1}{m_n} \sum_{i=1}^{m_n} F_n(x, \cdot) \circ \mathbf{J}_{n,i} = F(x). \tag{5.18}$$

2. *Let $p = 1$. Consider $\mathbb{N}^* \ni n \to k_n \in \mathbb{N}^*$ a sequence such that, $k_n \to \infty$ as $n \to \infty$. If $k_n \leq m_n$ for all $n \in \mathbb{N}^*$, and*

$$\sum_{n=1}^{\infty} \frac{a_n(x)^4}{k_n^2} < \infty, \tag{5.19}$$

for all $x \in \mathbf{X}$, then for all $x \in \mathbf{X}$, with probability one,

$$\lim_{n\to\infty} \left(\sum_{j-1}^{m_n} a_{n,j} \right) \frac{1}{k_n} \sum_{i=1}^{k_n} F_n(x, \cdot) \circ \tilde{c}_{n,i} = F(x). \tag{5.20}$$

Proof. Remark that it is sufficient to consider the case in which all functions F_n are positive.

Case $\mathbf{X}$ *countable*

1. Let $x \in \mathbf{X}$ be fixed. For each $n \in \mathbb{N}^*$ and $i \in \mathcal{I}_n$, define

$$Y_{n,i} := F_n(x, \cdot) \circ \mathbf{J}_{n,i}. \tag{5.21}$$

The row-wise independence of $((\mathbf{J}_{n,i})_{i\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ implies that $((Y_{n,i})_{i\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ is an array of row-wise independent random variables. Let $\mathbf{j} = (j_1, \ldots, j_p) \in \mathcal{I}_n^p$. Using (5.13) and the definition (5.15) of $\mathbf{J}_{n,i}$, we get

$$P(\{\mathbf{J}_{n,i}(\boldsymbol{\omega}) = \mathbf{j}\}) = \begin{cases} m_n^{1-p} & \text{if} \quad i = j_1, \\ \\ 0 & \text{if} \quad i \neq j_1, \end{cases} \tag{5.22}$$

for all $n \in \mathbb{N}^*$ and $\mathbf{j} \in \mathcal{I}_n$. Consequently,

$$\langle Y_{n,i} \rangle = \frac{1}{m_n^{p-1}} \sum_{j_2,\ldots,j_p=1}^{m_n} F_n(x, (i, j_2, \ldots, j_p)), \tag{5.23}$$

so that

$$\frac{1}{m_n} \sum_{i=1}^{m_n} \langle Y_{n,i} \rangle = \frac{1}{m_n^p} \sum_{\mathbf{j}\in\mathcal{I}_n^p} F_n(x, \mathbf{j}) = S_n(x). \tag{5.24}$$

Put $X_{n,i} := Y_{n,i} - \langle Y_{n,i} \rangle$. Then, the family $((X_{n,i})_{i\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ satisfies the conditions of Proposition 5.1, with $A_n \leq (2a_n(x))^4$. Therefore, for each fixed x, by (5.24) and (5.6) one obtains (5.18). For each $x \in \mathbf{X}$, let Ω_x be the subset of Ω where the limit (5.18) holds. Define $\Omega_{\mathbf{X}} := \bigcap_{x\in\mathbf{X}} \Omega_x$. Since $\mathbf{X}$ is countable, we have $P(\Omega_{\mathbf{X}}) = 1$, so that the argument is complete.

2. Let $x \in \mathbf{X}$ be fixed. For each $n \in \mathbb{N}^*$ and $i \in \mathcal{I}_n$ define

$$Y_{n,i} := \left(\sum_{l=1}^{m_n} a_{n,l} \right) F_n(x, \cdot) \circ \tilde{c}_{n,i}. \tag{5.25}$$

The row-wise independence of $((\tilde{c}_{n,i})_{i \in \mathcal{I}_n})_{n \in \mathbb{N}^*}$ ensures that $((Y_{n,i})_{i \in \mathcal{I}_n})_{n \in \mathbb{N}^*}$ is an array row-wise independent family of random variables. From (5.14), we get

$$\langle Y_{n,i} \rangle = \sum_{l=1}^{m_n} a_{n,l} F_n(x, l), \tag{5.26}$$

for all $i \in \mathcal{I}_n$ and $n \in \mathbb{N}^*$. Consequently,

$$\frac{1}{k_n} \sum_{i=1}^{k_n} \langle Y_{n,i} \rangle = S_n(x). \tag{5.27}$$

Define $X_{n,i} := Y_{n,i} - \langle Y_{n,i} \rangle$. From here the argument works similarly as in 1.

Case $\mathbf{X} = \mathbb{R}^m$

1. Observe that the argument with **X** *countable* is valid on the countable set $\mathbb{Q}^m$ of the vectors of $\mathbb{R}^m$ with rational components. Further, remark that for any $x \in \mathbb{R}^m \setminus \mathbb{Q}^m$ and $\varepsilon > 0$, by the continuity of F and the monotonicity of F_n, we can find two elements $x^-, x^+ \in \mathbb{Q}^m$, with $x^- \leq x \leq x^+$ such that

$$\begin{aligned} &F(x^+) - \frac{1}{m_n} \sum_{i=1}^{m_n} F_n(x^+, \cdot) \circ \mathbf{J}_{n,i}(\boldsymbol{\omega}) - \varepsilon \leq \\ &\leq F(x) - \frac{1}{m_n} \sum_{i=1}^{m_n} F_n(x, \cdot) \circ \mathbf{J}_{n,i}(\boldsymbol{\omega}) \leq \\ &\leq F(x^-) - \frac{1}{m_n} \sum_{i=1}^{m_n} F_n(x^-, \cdot) \circ \mathbf{J}_{n,i}(\boldsymbol{\omega}) + \varepsilon, \end{aligned} \tag{5.28}$$

for all $\boldsymbol{\omega} \in \Omega$. Now we approximate x by two sequences $\left\{x_p^+\right\}_{p \in \mathbb{N}}, \left\{x_p^-\right\}_{p \in \mathbb{N}} \subset \mathbb{Q}^m$, with $x_p^- \leq x \leq x_p^+$. Then, to conclude the proof in the *case* $\mathbf{X} = \mathbb{R}^m$, we refer to the result in the *case* **X** *countable*.

2. Replacing only (5.28) with

$$F(x^+) - \frac{1}{k_n}\sum_{i=1}^{k_n} F_n(x^+,\cdot)\circ\tilde{c}_{n,i}(\boldsymbol{\omega}) - \varepsilon \leq$$

$$\leq F(x) - \frac{1}{k_n}\sum_{i=1}^{k_n} F_n(x,\cdot)\circ\tilde{c}_{n,i}(\boldsymbol{\omega}) \leq \tag{5.29}$$

$$\leq F(x^-) - \frac{1}{k_n}\sum_{i=1}^{k_n} F_n(x^-,\cdot)\circ\tilde{c}_{n,i}(\boldsymbol{\omega}) + \varepsilon,$$

one repeats step by step the arguments of the part 1 to conclude the proof of the part 2. □

The index set $\mathcal{I}_n^p$ being defined as before, let $((\mu_{n,\mathbf{j}})_{\mathbf{j}\in\mathcal{I}_n^p})_{n\in\mathbb{N}^*}$ be a bounded family of positive measures on $\mathbb{R}^m$, i.e. there exists some constant $a>0$, such that $|\mu_{n,\mathbf{j}}| \leq a$ for all $\mathbf{j}\in\mathcal{I}_n^p$ and $n\in\mathbb{N}^*$ (we recall the notation $|\mu| := \mu(\mathbb{R}^m)$ for some finite measure μ on $\mathbb{R}^m$).

Let (Ω,β,P) be the probability space be as in Proposition 5.2 and the arrays of row-wise random variables $((\mathbf{J}_{n,i})_{i\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ and $((\tilde{c}_{n,i})_{i\in\mathcal{I}_n})_{n\in\mathbb{N}^*}$ defined by (5.15) and (5.12) respectively.

THEOREM **5.1** 1. *Let $p\geq 2$. Suppose that there is a positive measure μ on $\mathbb{R}^m$, absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, such that*

$$\frac{1}{m_n^p}\sum_{\mathbf{j}\in\mathcal{I}_n^p}\mu_{n,\mathbf{j}} \rightharpoonup \mu, \ \text{as } n\to\infty. \tag{5.30}$$

Define $\mu_{n,i}(\boldsymbol{\omega}) := \mu_{n,\mathbf{j}}|_{\mathbf{j}=\mathbf{J}_{n,i}(\boldsymbol{\omega})}$ for all $\boldsymbol{\omega}\in\Omega$, all $i\in\mathcal{I}_n$ and $n\in\mathbb{N}^*$. If

$$\sum_{n=1}^{\infty}\frac{1}{m_n^2} < \infty, \tag{5.31}$$

then for $P-$almost all $\boldsymbol{\omega}$,

$$\sigma_{1,n}(\boldsymbol{\omega}) := \frac{1}{m_n}\sum_{i=1}^{m_n}\mu_{n,i}(\boldsymbol{\omega}) \rightharpoonup \mu \ \text{as } n\to\infty. \tag{5.32}$$

2. *Let $p=1$. Suppose that there is a positive measure μ on $\mathbb{R}^m$, absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, such that*

$$\sum_{l=1}^{m_n} a_{n,l}\cdot\mu_{n,l} \rightharpoonup \mu, \ \text{as } n\to\infty. \tag{5.33}$$

Define $\mu_{n,i}(\boldsymbol{\omega}) := \mu_{n,l}|_{l=\tilde{c}_{n,i}(\boldsymbol{\omega})}$ *for all* $\boldsymbol{\omega} \in \Omega$, *all* $i \in \mathcal{I}_n$ *and* $n \in \mathbb{N}^*$. *Let* $\mathbb{N}^* \ni n \to k_n \in \mathbb{N}^*$ *be a sequence such that* $k_n \leq m_n$, *for all* $n \in \mathbb{N}^*$ *and*

$$\sum_{n=1}^{\infty} \frac{1}{k_n^2} < \infty. \tag{5.34}$$

Then, for $P-$*almost all* $\boldsymbol{\omega}$,

$$\sigma_{2,n}(\boldsymbol{\omega}) := \frac{1}{k_n} \sum_{i=1}^{k_n} \mu_{n,i}(\boldsymbol{\omega}) \rightharpoonup \mu \text{ as } n \to \infty. \tag{5.35}$$

Proof. Define for each $x \in \mathbb{R}^m$

$$F_n(x, \mathbf{j}) := \int_{y \leq x} \mathrm{d}\mu_{n,\mathbf{j}}(y), \tag{5.36}$$

and

$$F(x) := \int_{y \leq x} \mathrm{d}\mu(y). \tag{5.37}$$

Then it is sufficient to observe that F and $F_n(x, \mathbf{j})$ satisfy the conditions of Proposition 5.2, (with $a_n(x) = a$) and the family $\{y \in \mathbb{R}^m \,|y \leq x\}_{x \in \mathbb{R}^m}$ is determining, Ref. [28], for the weak convergence of the measures $\mu_{n,\mathbf{j}}$. □

REMARK **5.1** *It can be easily seen that Babovsky Lemma (see Lemma* 2 *of Ref.* [4]*) is a particular case of Theorem* 5.1.1 *with* $m_n = n^2$, *for all* $n \in \mathbb{N}^*$ *and with* $\mu_{n,\mathbf{j}}$ *given by a product of two point measures.*

REMARK **5.2** *As we have mentioned in Section 1, our purpose is to approximate the solutions of* (2.18) *by sums of Dirac measures of the form* (2.22).

Due to the nonlinear character of the collision operators $\mathbf{P}$ *and* $\mathbf{S}$, *at each timestep, the numerical complexity increases dramatically (power-like). Although, we are able to reduce the computational effort using repeatedly the Theorem* 5.1.1.

However, except the case of (2.18) *modelling the one component gas with purely elastic collisions, a certain step of the numerical scheme destroys the homogeneity of the sums of Dirac measures, i.e. instead of HSPM approximations one obtains WSPM approximations. This difficulty will be surmounted by using Theorem* 5.1.2, *which converts the WSPM approximations into HSPM approximations.*

Theorem 5.1 will be the basic point of the probabilistic part of our numerical scheme for the solutions of (2.18) in the next section.

6. The Main Result

For our numerical scheme, we need a weak form of (4.16), where the functions f_k^j are replaced by the measures $\bar{\mu}_k^j$ given by (4.24). Denote

$$(\varphi, h) := \int_{\mathbb{R}^3} \varphi(\mathbf{v}) h(\mathbf{v}) \mathrm{d}\mathbf{v}, \tag{6.1}$$

for $\varphi \in C_b(\mathbb{R}^3)$ and $h \in \mathbb{L}^1(\mathbb{R}^3)$. From (4.16) using (6.1) we get

$$\left(\varphi, f_k^j\right) = (\varphi, Q_k(\mathbf{f}^{j-1})) + (\varphi, L_k(\mathbf{f}^{j-1})) + \Delta t \cdot (\varphi, P_k(\mathbf{f}^{j-1})) \tag{6.2}$$

for all $\varphi \in C_b(\mathbb{R}^3)$, all $j = 1, \ldots, T_\Delta$ and $k = 1, \ldots, N$. Denoting by

$$V(\mathbf{\Omega}_{\boldsymbol{\beta}}) := \int_{\mathbf{\Omega}_{\boldsymbol{\beta}}} \mathrm{d}\mathbf{n}, \tag{6.3}$$

in (6.2),

$$\begin{aligned} (\varphi, Q_k(\mathbf{f}^j)) := \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \alpha_k \int_{\mathbb{R}^{3|\boldsymbol{\alpha}|} \times \mathbf{\Omega}_{\boldsymbol{\beta}}} (\varphi \circ i_{k,\boldsymbol{\alpha}})(\mathbf{w}) \times \\ \times \left(\frac{\Gamma^{\boldsymbol{\alpha},k}}{V(\mathbf{\Omega}_{\boldsymbol{\beta}})} - \Delta t \cdot r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}) \right) \mathbf{f}_{\boldsymbol{\alpha}}^j(\mathbf{w}) \mathrm{d}\mathbf{w} \mathrm{d}\mathbf{n}, \end{aligned} \tag{6.4}$$

and

$$(\varphi, L_k(\mathbf{f}^j)(\mathbf{v})) := \sum_{\boldsymbol{\alpha} \in \mathcal{U} \setminus \mathcal{M}} \Lambda^{\boldsymbol{\alpha},k} \int_{\mathbb{R}^{3|\boldsymbol{\alpha}|}} (\varphi \circ i_{k,\boldsymbol{\alpha}})(\mathbf{w}) \mathbf{f}_{\boldsymbol{\alpha}}^j(\mathbf{w}) \mathrm{d}\mathbf{w}. \tag{6.5}$$

In the formulas (6.4) and (6.5), the projection application $i_{k,\boldsymbol{\gamma}} : \mathbb{R}^{3|\boldsymbol{\gamma}|} \to \mathbb{R}^3$ is defined by $i_{k,\boldsymbol{\gamma}}(\mathbf{w}) = \mathbf{w}_{k,\gamma_k}$, for $\boldsymbol{\gamma} \in \mathcal{M}$ and $k = 1, \ldots, N$. Using (3.6) and (3.9) we get

$$\begin{aligned} &(\varphi, P_k(\mathbf{f}^j)) = \\ &= \sum_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}} \beta_k \int_{\mathbb{R}^{3|\boldsymbol{\alpha}|} \times \mathbf{\Omega}_{\boldsymbol{\beta}}} \varphi \circ i_{k,\boldsymbol{\beta}}(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n})) r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}) \mathbf{f}_{\boldsymbol{\alpha}}^j(\mathbf{w}) \mathrm{d}\mathbf{w} \mathrm{d}\mathbf{n}, \end{aligned} \tag{6.6}$$

for all $\varphi \in C_b(\mathbb{R}^3)$, all $j = 0, 1, \ldots, T_\Delta$ and $k = 1, \ldots, N$.

Now, we are able to formulate (6.2) as an equation for measures. For some $\boldsymbol{\gamma} \in \mathcal{M}$ and $j = 0, 1, \ldots, T_\Delta$, define the measure $\bar{\mu}_{\boldsymbol{\gamma}}^j$ on $\mathbb{R}^{3|\boldsymbol{\gamma}|}$ by

$$\mathrm{d}\bar{\mu}_{\boldsymbol{\gamma}}^j(\mathbf{w}) = \bigotimes_{k \in \mathcal{N}_\gamma} \bigotimes_{i=1}^{\gamma_k} \mathrm{d}\bar{\mu}_k^j(\mathbf{w}_{k,i}). \tag{6.7}$$

From (6.2-6.6), using spherical coordinates

$$[0,\pi)^{3|\boldsymbol{\beta}|-5}\times[0,2\pi)\ni(\theta,\varphi)\to\mathbf{n}(\theta,\varphi)\in\boldsymbol{\Omega}_{\boldsymbol{\beta}}, \tag{6.8}$$

to integrate on each unit sphere $\boldsymbol{\Omega}_{\boldsymbol{\beta}}$, it follows that there are some sets $\mathcal{A}\subset\mathcal{U}$, $\mathcal{B}\subset\mathcal{M}$, the functions $q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}\in C(\mathbb{R}^{3|\boldsymbol{\alpha}|}\times[0,\pi)^{3|\boldsymbol{\beta}|-5}\times[0,2\pi)\,;\mathbb{R}_+)$ and $H_{\boldsymbol{\alpha},\boldsymbol{\beta},k}\in C(\mathbb{R}^{3|\boldsymbol{\alpha}|}\times[0,\pi)^{3|\boldsymbol{\beta}|-5}\times[0,2\pi)\,;\mathbb{R}^3)$ such that we can write (6.2) in the compressed form

$$\begin{aligned}\int_{\mathbb{R}^3}\varphi(\mathbf{v})\mathrm{d}\bar{\mu}_k^j(\mathbf{v})=&\sum_{\boldsymbol{\alpha}\in\mathcal{A},\boldsymbol{\beta}\in\mathcal{B}}\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|}}\mathrm{d}\bar{\mu}_{\boldsymbol{\alpha}}^{j-1}(\mathbf{w})\times\\ &\times\int_{[0,\pi)^{3|\boldsymbol{\beta}|-5}}\mathrm{d}\theta\int_0^{2\pi}(\varphi\circ H_{\boldsymbol{\alpha},\boldsymbol{\beta},k})(\mathbf{w},\theta,\phi)q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,\phi)\mathrm{d}\phi,\end{aligned} \tag{6.9}$$

for $\varphi\in C_b(\mathbb{R}^3)$ and $k\in 1,\dots,N$.

First, we consider $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}$ verifying the properties of Lemma 3.1 and we construct the algorithm starting from (6.9). Then, we show how the numerical scheme can be improved, if one introduces additional conditions on $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}$.

Now, we write (6.9) in a more convenient form. Note that, we can find some $L\in\mathbb{N}^*$ and

1. a family $\{\boldsymbol{\alpha}(l)\}_{l=1,\dots,L}\subset\mathcal{U}$ of multi-indexes,
2. a family $\{q(l)\}_{l=1,\dots,L}\subset\mathbb{N}^*$,
3. a family $\{\pi_l\}_{l=1,\dots,L}$ of measures absolute continuous with respect to the Lebesgue measure on $\mathbb{R}^{q(l)}$,
4. a family $\{R_{k,l}\}_{k=1,\dots,N;l=1,\dots,L}\subset C(\mathbb{R}^{3|\boldsymbol{\alpha}(l)|+q(l)};\mathbb{R}_+)$ of functions,
5. a family $\{h_{k,l}\}_{k=1,\dots,N;l=1,\dots,L}\subset C(\mathbb{R}^{3|\boldsymbol{\alpha}(l)|+q(l)};\mathbb{R}^3)$ of functions,

such that (6.9) can be written

$$\int_{\mathbb{R}^3}\varphi(\mathbf{v})\mathrm{d}\bar{\mu}_k^j(\mathbf{v})=\sum_{l=1}^L\int_{\mathbb{R}^{3|\boldsymbol{\alpha}(l)|+q(l)}}R_{k,l}(\mathbf{z})(\varphi\circ h_{k,l})(\mathbf{z})\mathrm{d}(\bar{\mu}_{\boldsymbol{\alpha}(l)}^{j-1}\otimes\pi_l)(\mathbf{z}). \tag{6.10}$$

Let (Ω,β,P) be as in Theorem 5.1.

a) For each $l = 1, \ldots, L$, we approximate π_l by a convenient HSPM of the form (2.22), containing n-terms, $\pi_{l,n} \rightharpoonup \pi_l$ as $n \to \infty$ (this can be done, e.g. by means of low discrepancy, well distributed sequences Ref. [6, 27]).

b) The initialization of the scheme is done by giving n-terms HSPM approximations $\nu^0_{k,n}$ of the initial data $\bar{\mu}^0_k$, where $k = 1, \ldots, N$.

c) The n-terms HSPM approximations $\nu^1_{k,n}$ of $\bar{\mu}^1_k$, with $k = 1, \ldots, N$, resulting from the scheme, can be obtained as follows:

Step 1 *(first selection).* For each $l = 1, \ldots, L$ and $k = 1, \ldots, N$ we replace $\bar{\mu}^0_k$ by $\nu^0_{k,n}$ in (6.7) (for $\boldsymbol{\gamma} = \boldsymbol{\alpha}(l)$, $j = 0$). Then for each $l = 1, \ldots, L$, we obtain a sequence of finite measures $\nu^0_{\boldsymbol{\alpha}(l),n} \rightharpoonup \bar{\mu}^0_{\boldsymbol{\alpha}(l)}$ as $n \to \infty$, implying $\nu^0_{\boldsymbol{\alpha}(l),n} \otimes \pi_{l,n} \rightharpoonup \bar{\mu}^0_{\boldsymbol{\alpha}(l)} \otimes \pi_l$ as $n \to \infty$. Obviously, each $\nu^0_{\boldsymbol{\alpha}(l),n} \otimes \pi_{l,n}$ is a sum of the form (5.30), containing $n^{|\boldsymbol{\alpha}(l)|+1}$ terms. We apply the selection algorithm cf. Theorem 5.1.1 (with $m_n = n$ and $p = |\boldsymbol{\alpha}(l)|+1$) to construct n - -terms HSPM approximations for all $\nu^0_{\boldsymbol{\alpha}(l),n} \otimes \pi_{l,n}$. Thus, by Theorem 5.1.1, for each $l = 1, \ldots, L$, there exists some set $\Omega_l \subset \Omega$, with $P(\Omega_l) = 1$, such that from $\nu^0_{\boldsymbol{\alpha}(l),n} \otimes \pi_{l,n}$, one can extract a n-terms HSPM approximation (of the form (5.32)) $\sigma_{1,l,n}(\boldsymbol{\omega}^l) \rightharpoonup \bar{\mu}^0_{\boldsymbol{\alpha}(l)} \otimes \pi_l$ as $n \to \infty$, for almost all $\boldsymbol{\omega}^l \in \Omega_l$.

Step 2 *(second selection).* In the right side of (6.10), written for $j = 1$, replace each $\bar{\mu}^0_{\boldsymbol{\alpha}(l)} \otimes \pi_l$ by the corresponding $\sigma_{1,l,n}$. Then the right side of (6.10) defines the measures $M_{k,n}$ on $\mathbb{R}^3$, for $k = 1, \ldots, N$ and $n \in \mathbb{N}^*$,

$$M_{k,n} = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} a_l R_{k,l}(\mathbf{z}_{l,i,n}(\boldsymbol{\omega}^l)) \delta_{h_{k,l}(\mathbf{z}_{l,i,n}(\boldsymbol{\omega}^l))}, \tag{6.11}$$

concentrated at the points $h_{k,l}(\mathbf{z}_{l,i,n}(\boldsymbol{\omega}^l))$, where $\mathbf{z}_{l,i,n}(\boldsymbol{\omega}^l) \in \mathbb{R}^{3|\boldsymbol{\alpha}(l)|+q(l)}$ and $a_l \geq 0$ are some constants (for $l = 1, \ldots, L$ and $i = 1, \ldots, n$). By Step 1, it follows that

$$M_{k,n} \rightharpoonup \bar{\mu}^1_k \text{ as } n \to \infty, \tag{6.12}$$

for all $\boldsymbol{\omega}^1 \in \Omega_1$, $\boldsymbol{\omega}^2 \in \Omega_2$, ..., $\boldsymbol{\omega}^L \in \Omega_L$ and for $k = 1, \ldots, N$. Now, it can be easily seen that (6.11) can be written as WSPM, containing, at most $L \cdot n$ terms.

As we mentioned before, we want to obtain HSPM approximations at the end of each step of time. We fix, for the moment, some $\boldsymbol{\omega}^1 \in \Omega_1, \ldots, \boldsymbol{\omega}^L \in \Omega_L$, so that (6.12) holds. We apply the selection algorithm formulated Theorem 5.1.2

for each fixed $k = 1, \ldots, N$, as follows. For $l = 1, \ldots, L \cdot n$ defining

$$\iota(l) := \left[\left[\frac{l-1}{L}\right]\right] + 1,$$
$$\lambda(l) := \left[\left[\frac{l-1}{n}\right]\right] + 1, \tag{6.13}$$

put

$$a_{n,l} = \frac{1}{n} a_{\lambda(l)} R_{k,\lambda(l)}(\mathbf{z}_{\lambda(l),\iota(l),n}(\boldsymbol{\omega}^{\lambda(l)})). \tag{6.14}$$

We choose $m_n = L \cdot n$ and $k_n = n$. Then, for each $k = 1, \ldots, N$, there exists some $\Omega_{L+k} \subset \Omega$, with $P(\Omega_{L+k}) = 1$, such that from $M_{k,n}$, we obtain a n-terms HSPM approximation (of the form (5.35)) $\sigma_{2,k,n}(\boldsymbol{\omega}^{L+k}; \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^L) \rightharpoonup \bar{\mu}_k^1$ as $n \to \infty$, for all $\boldsymbol{\omega}^{L+k} \in \Omega_{L+k}$. Set $\bar{\nu}_{k,n}^1(\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{L+k}) := \sigma_{2,k,n}(\boldsymbol{\omega}^{L+k}; \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^L)$. Therefore for each $\bar{\mu}_k^1$ in (6.10), we obtain a corresponding n-terms HSPM approximation $\bar{\nu}_{k,n}^1 \rightharpoonup \bar{\mu}_k^1$ as $n \to \infty$, for all $\boldsymbol{\omega}^1 \in \Omega_1, \ldots, \boldsymbol{\omega}^{L+k} \in \Omega_{L+k}$ and for all $k = 1, \ldots, N$.

e) The procedure can be repeated, with the entering data $\bar{\nu}_{k,n}^1$, to obtain HSPM approximations $\bar{\nu}_{k,n}^2(\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{2L+N+k})$ of $\bar{\mu}_k^2$ for $k = 1, \ldots, N$.

f) Repeating this procedure over and over, after j timesteps, we provide the n-terms HSPM approximations $\bar{\nu}_{k,n}^j(\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{jL+(j-1)N+k}) \rightharpoonup \bar{\mu}_k^j$ for all $\boldsymbol{\omega}^1 \in \Omega_1$, $\boldsymbol{\omega}^2 \in \Omega_2, \ldots, \boldsymbol{\omega}^{jL+(j-1)N+k} \in \Omega_{jL+(j-1)N+k}$, all $j = 1, \ldots, T_\Delta$ and all $k = 1, \ldots, N$, where $\Omega_l \subset \Omega$ with $P(\Omega_l) = 1$, for $l = 1, \ldots, T_\Delta(L+N)$.

Now, observe that we can find a family $\{Q_l\}_{l \in \mathbb{N}^*}$ of measurable maps $Q_l : \Omega \to \Omega$, with $P(Q_l^{-1}(A)) = 1$, for all $A \subset \Omega$ with $P(A) = 1$. For instance, we can consider $U, V : \Omega \to \Omega$, given by

$$U(\boldsymbol{\omega}) = U(\omega_1, \omega_2, \ldots, \omega_{2n-1}, \omega_{2n}, \ldots) := (\omega_1, \omega_3, \ldots \omega_{2n-1}, \omega_{2n+1}, \ldots), \tag{6.15}$$

$$V(\boldsymbol{\omega}) = V(\omega_1, \omega_2, \ldots, \omega_{2n-1}, \omega_{2n}, \ldots) := (\omega_2, \omega_4, \ldots, \omega_{2n}, \omega_{2n+2}, \ldots), \tag{6.16}$$

respectively, for all $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_{2n-1}, \omega_{2n}, \ldots) \in \Omega$. Then it is sufficient to put $Q_1 = U$ and $Q_l := U \circ V^{l-1}$, $l = 2, 3, \ldots$ Let

$$\Omega_{\Delta t} := \bigcap_{l=1}^{T_\Delta(L+N)} Q_l^{-1}(\Omega_l). \tag{6.17}$$

Since $P(Q_l^{-1}(\Omega_l)) = 1$ for all $l = 1, \ldots, T_\Delta(L+N)$, clearly $P(\Omega_{\Delta t}) = 1$. Defining $\nu_{k,n}^j(\boldsymbol{\omega}) := \bar{\nu}_{k,n}^j(Q_1(\boldsymbol{\omega}), \ldots, Q_{jL+(j-1)N+k}(\boldsymbol{\omega}))$ for all $\boldsymbol{\omega} \in \Omega$, $j =$

$1, \ldots, T_\Delta$, $k = 1, \ldots, N$, it follows that $\nu_{k,n}^j(\boldsymbol{\omega}) \rightharpoonup \bar{\mu}_k^j$ as $n \to \infty$, for all $\boldsymbol{\omega} \in \Omega_{\Delta t}$, $j = 1, \ldots, T_\Delta$, $k = 1, \ldots, N$.

In particular, if $D(\cdot, \cdot)$ is the discrepancy introduced in Section 2., then

$$\lim_{n\to\infty} \max_{k=1,\ldots,N} \max_{j=1,\ldots,T_\Delta} D\left(\nu_{k,n}^j(\boldsymbol{\omega}), \bar{\mu}_k^j\right) = 0, \tag{6.18}$$

for almost all $\boldsymbol{\omega} \in \Omega$.

All these and Corollary 4.1 lead to our main result.

Let $\mathbf{f}(t)$ be the solution of equation (3.18), provided by Theorem 3.1 and let μ_k^t be given by $\mathrm{d}\mu_k^t(\mathbf{v}) := f_k(t, \mathbf{v})\mathrm{d}\mathbf{v}$, for all $t \geq 0$ and $k = 1, \ldots, N$. Consider some family $\{\Delta t_p\}_{p\in\mathbb{N}}$ of discretization timesteps as in Section 4.. For each Δt_p and for the initial data $\bar{\mu}_k^0$, consider the solutions $\bar{\mu}_{k,p}^j$ of (6.10), with $j = 1, \ldots, T_\Delta$ and $k = 1, \ldots, N$. For each $\bar{\mu}_{k,p}^j$, denote by $\nu_{k,p,n}^j$ the corresponding n-terms HSPM approximation obtained by the above scheme.

Similar to (2.25), we introduce the following notation $T_{\Delta p} := [[T/\Delta t_p]]$, for all $p \in \mathbb{N}$.

THEOREM **6.1** *For each sequence of timesteps $\Delta t_p \to 0$ as $p \to \infty$, there is a sequence of positive integers $n(p) \to \infty$ as $p \to \infty$, such that*

$$\lim_{p\to\infty} \max_{k=1,\ldots,N} \max_{j=1,\ldots,T_{\Delta p}} D\left(\nu_{k,p,n(p)}^j(\boldsymbol{\omega}), \mu_k^{j\cdot\Delta t_p}\right) = 0, \tag{6.19}$$

for almost all $\boldsymbol{\omega} \in \Omega$.

Proof. Let

$$d_{p,n}(\boldsymbol{\omega}) := \max_{k=1,\ldots,N} \max_{j=1,\ldots,T_{\Delta p}} D\left(\nu_{k,p,n}^j(\boldsymbol{\omega}), \bar{\mu}_{k,p}^j\right). \tag{6.20}$$

Consider some positive sequence $\varepsilon_p \downarrow 0$ as $p \to \infty$. Using (6.18), for each p, we obtain that

$$\lim_{n\to\infty} P(d_{p,n} > \varepsilon_p) = 0. \tag{6.21}$$

Then, for each p, we can choose $n = n(p)$, such that

$$P(d_{p,n(p)} > \varepsilon_p) \leq \frac{1}{p^2}. \tag{6.22}$$

Consequently,

$$\sum_{p=1}^{\infty} P(d_{p,n(p)} > \varepsilon_p) < \infty. \tag{6.23}$$

Then, for almost all $\boldsymbol{\omega} \in \Omega$,

$$\lim_{n\to\infty} d_{p,n(p)}(\boldsymbol{\omega}) = 0. \tag{6.24}$$

Now, by Corollary 4.1, we conclude the proof of the Theorem. □

This theorem represents a space homogeneous reactive correspondent to the main result in the Babovsky-Illner simulation scheme for the classical Boltzmann equation (Theorem 7.1 of Ref. [5]).

Note that the numerical effort of the method is at most, $O(n \log n)$ (the dominant contribution being introduced by the random selections of Theorem 5.1.2, i.e. *(second selection) Step* 2). However, under additional conditions on $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}$, the sum (6.10) the numerical effort can be improved.

We consider the following simple case. Denote $\mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}} := \left\{\mathbf{w}' \in \mathbb{R}^{3|\boldsymbol{\alpha}|} | 0 < W_{\boldsymbol{\alpha}}(\mathbf{w}') - 2^{-1}(\sum_{n=1}^{N} \alpha_n m_n) V_{\boldsymbol{\alpha}}(\mathbf{w}')^2 - \sum_{n=1}^{N} \beta_n E_n\right\}$ (we recall that $W_{\boldsymbol{\alpha}}(\mathbf{w})$ is the energy defined in Section 2). By Lemma 3.1, $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}) \geq 0$ on $\mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \times \Omega_{\boldsymbol{\beta}}$. Suppose that in (6.2 -6.6), we have $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}) > 0$ on $\mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \times \Omega_{\boldsymbol{\beta}}$ for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}$. Taking into account the form of the element $\mathrm{d}\mathbf{n}$ on $\Omega_{\boldsymbol{\beta}}$ in spherical coordinates (when (6.9) is obtained from (6.2 -6.6)) it follows easily that in (6.9), each function $q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,\phi)$ can be constructed such that the set $\{\theta | q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,\phi) = 0\}$ is finite and does not depend on the choice of $(\mathbf{w},\phi) \in \mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \times [0,2\pi)$. Consequently, for each $\boldsymbol{\beta} \in \mathcal{B}$, there is a measurable set $\Theta_{\boldsymbol{\beta}} \subset [0,\pi)^{3|\boldsymbol{\beta}|-5}$ such that $q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,\phi) > 0$, for all $\mathbf{w} \in \mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, $\theta \in \Theta_{\boldsymbol{\beta}}$, $\phi \in [0,2\pi)$, $\boldsymbol{\alpha} \in \mathcal{A}$. Denote

$$I_k(\phi;\mathbf{w},\theta) := \int_0^{\phi} q_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,\rho)\mathrm{d}\rho, \qquad \phi \in [0,2\pi)\,. \tag{6.25}$$

Then, for all $\mathbf{w} \in \mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, $\theta \in \Theta_{\boldsymbol{\beta}}$, fixed, (6.25) defines an invertible map

$$[0,2\pi) \ni \phi \to I_k(\phi;\mathbf{w},\theta) \in [0, I_k(2\pi;\mathbf{w},\theta))\,, \tag{6.26}$$

with the inverse I_k^{-1}. In each integral of (6.9), with respect to $\mathrm{d}\phi$, we perform the change of variable $\phi = I_k^{-1}(y;\mathbf{w},\theta)$. Define

$$\tilde{H}_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,y) = H_{\boldsymbol{\alpha},\boldsymbol{\beta},k}(\mathbf{w},\theta,I_k^{-1}(y;\mathbf{w},\theta)). \tag{6.27}$$

We can choose some measurable sets

$$\mathcal{C}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \subseteq \mathbb{R}^{3|\boldsymbol{\alpha}|} \times [0,\pi)^{3|\boldsymbol{\beta}|-5} \times \mathbb{R}_+, \text{ for } \boldsymbol{\alpha} \in \mathcal{A}, \boldsymbol{\beta} \in \mathcal{B},$$

such that, (6.9) takes the following form

$$\int_{\mathbb{R}^3} \varphi(\mathbf{v})\mathrm{d}\bar{\mu}_k^j(\mathbf{v}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A},\boldsymbol{\beta}\in\mathcal{B}} \int_{\mathcal{C}_{\boldsymbol{\alpha}\boldsymbol{\beta}}} (\varphi \circ \tilde{H}_{\boldsymbol{\alpha},\boldsymbol{\beta},k})(\mathbf{w},\theta,y)\mathrm{d}\bar{\mu}_{\boldsymbol{\alpha}}^{j-1}(\mathbf{w})\mathrm{d}\theta dy. \tag{6.28}$$

For applications it is important to observe that the conclusion remains the same if weaker conditions are imposed on $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}$, e. g. if one supposes that for each $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{M}$, $r_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}) > 0$ on $\mathcal{D}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \times \Omega_{\boldsymbol{\beta}}$ except a countable set, etc.

Obviously, (6.28) has the form (6.10), but has the important property that if $\bar{\mu}_k^{j-1}$, for $k = 1, \ldots, N$ are HSPM, after *Step* 1 *(first selection)* the output measures are also a HSPM.

In order to obtain $\bar{\mu}_k^j$, for $k = 1, \ldots, N$ as HSPM with the same number of terms as $\bar{\mu}_k^{j-1}$, we can apply the following immediate corollary of Theorem 5.1.2, which introduces a numerical complexity of only $O(n)$.

COROLLARY **6.1** *Suppose that there is a positive measure μ on $\mathbb{R}^m$, absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$, such that*

$$\frac{1}{m_n} \sum_{l=1}^{m_n} \mu_{n,l} \rightharpoonup \mu, \text{ as } n \to \infty. \tag{6.29}$$

Define $\mu_{n,i}(\boldsymbol{\omega}) := \mu_{n,l}|_{l=\tilde{c}_{n,i}(\boldsymbol{\omega})}$ for all $\boldsymbol{\omega} \in \Omega$, all $i \in \mathcal{I}_n$ and $n \in \mathbb{N}^*$. Let $\mathbb{N}^* \ni n \to k_n \in \mathbb{N}^*$ be a sequence such that $k_n \leq m_n$, for all $n \in \mathbb{N}^*$ and

$$\sum_{n=1}^{\infty} \frac{1}{k_n^2} < \infty. \tag{6.30}$$

Then, for $P-$almost all $\boldsymbol{\omega}$,

$$\sigma_{3,n}(\boldsymbol{\omega}) := \frac{1}{k_n} \sum_{i=1}^{k_n} \mu_{n,i}(\boldsymbol{\omega}) \rightharpoonup \mu \text{ as } n \to \infty. \tag{6.31}$$

Further we can proceed as in the scheme constructed before, but without applying Theorem 5.1.2. Instead we apply Corollary 6.1. The scheme reduces to iterations alternating with selections, and the conclusion of Theorem 6.1 remains valid. The numerical effort becomes $O(n)$.

Finally remark that if (2.18) reduces the classical Boltzmann equation, for the one-component simple gas, then the sum in the r.h.s of (6.28) can be compressed to a unique term as in Ref. [4]. In general, this is not possible in the case of gas mixtures.

7. Concluding Remarks

From the above analysis, it follows that besides a convenient existence theory, only the conservation of the total mass is needed to introduce the numer-

ical scheme described here. The other properties (e.g. detailed balance, H-Theorem) of the Wang-Chang-Uhlenbeck-de Boer and Ludwig and Heil system of equations play no role in this algorithm. Note that, the numerical scheme can also be used and when the detailed balance does not hold, e.g., for models where we ignore some recombination processes (as in the situation when we consider the collisional dissociation, but neglect the recombination by triple collisions Ref. [24]).

We discuss possible generalizations as well as some limitations of the results.

1^0 In the case of non-reacting gas mixtures one can obtain similar numerical schemes for the space-dependent equation (2.10), in the frame of the theory of existence of solutions of Ref. [17]. This can be done by adapting directly the spatial cell homogenization method of Ref. [5].

2^0 In the case of reacting gas mixtures, one can also obtain similar numerical schemes for the space-dependent equation (2.10). To this end, the adaptation of the spatial cell homogenization method of Ref. [5] is not as straightforward as it appears. This is due to the collisions that produce new particles in a given spacial cell. For this purpose, we need "to establish" the space position in the cell for each "new born" particle and at the same time, to keep the control on convergence.

3^0 Assumption (3.13) replaces in the reactive model the boundedness condition on the collision law used in Ref. [4, 5]. This condition is essential for the control of the positivity of the solutions in the time-discretized equation (4.1). Indeed, Assumption (3.13) is restrictive from an analytical point of view. Nevertheless, for practical purposes, it is satisfactory for those models where the high energy-tail of the gas consists of very few molecules (see Ref. [7]).

The existence of unique positive solutions to (2.10) and (2.18) can be proved for more general transition functions $K_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ (see Ref. [18]). The simulation scheme can be also extended in this respect, but the (possible) singularities of $K_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ must not destroy the continuity of the functions $r_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and $p_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ (necessary for the convergence in the weak sense of the measures).

4^0 One can improve the approximation algorithm as follows. Instead of assigning to each species the same number of terms in HSPM, one can fix a given number of terms n for all the species. Then, when we apply the selection algorithm given by Theorem 5.1.2 (or Corollary 6.1), we can allocate to each species a number of terms "proportional" to its mass, such that the total number of terms for all the species to be (approximative) n. The same is also valid for the approximation of the initial data. By example if we designate by n_k the number of terms corresponding to the species $k = 1, \ldots, N$, then

we define

$$n_k := \left[\left[\left[n \cdot \frac{m_k \int_{\mathbb{R}^3} f_k^0(\mathbf{v})\mathrm{d}\mathbf{v}}{\sum_{l=1}^{N} m_l \int_{\mathbb{R}^3} f_l^0(\mathbf{v})\mathrm{d}\mathbf{v}} \right]\right]\right]. \tag{7.1}$$

5^0 In this numerical scheme there are three essential sources of approximation errors.

1. The errors from the approximation of the initial data.
2. The errors produced by the time discretization.
3. The errors introduced by stochastic selections.

The contribution of the stochastic errors over the time discretized scheme can be illustrated as it follows. Giving, for the chemical species $k = 1, \ldots, N$, an initial data, say $\nu_k^{0,0}$ of the form (2.22) the algorithm follows the computational chain

$$\nu_k^{0,0} \to \nu_k^{1,1} \to \nu_k^{2,2} \to \ldots \to \nu_k^{T_\Delta - 1, T_\Delta - 1} \to \nu_k^{T_\Delta, T_\Delta} \tag{7.2}$$

corresponding to the diagonal of the scheme

$$\begin{array}{ccccccccccc}
\nu_k^{0,0} & \longrightarrow & \nu_k^{0,1} & \longrightarrow & \nu_k^{0,2} & \longrightarrow & \ldots & \longrightarrow & \nu_k^{0,T_\Delta-1} & \longrightarrow & \nu_k^{0,T_\Delta} \\
 & & \wr\!\!\!\wr & & & & & & & & \\
 & & \nu_k^{1,1} & \longrightarrow & \nu_k^{1,2} & \longrightarrow & \ldots & \longrightarrow & \nu_k^{1,T_\Delta-1} & \longrightarrow & \nu_k^{1,T_\Delta} \\
 & & & & \wr\!\!\!\wr & & & & & & \\
 & & & & \nu_k^{2,2} & \longrightarrow & \ldots & \longrightarrow & \nu_k^{2,T_\Delta-1} & \longrightarrow & \nu_k^{2,T_\Delta} \\
 & & & & & & & & \vdots & & \vdots \\
 & & & & & & & & \wr\!\!\!\wr & & \\
 & & & & & & & & \nu_k^{T_\Delta-1,T_\Delta-1} & \longrightarrow & \nu_k^{T_\Delta-1,T_\Delta} \\
 & & & & & & & & & & \wr\!\!\!\wr \\
 & & & & & & & & & & \nu_k^{T_\Delta,T_\Delta}
\end{array} \tag{7.3}$$

Here, the horizontal chains represent the exact iterations of the time discretized equations, such that for each $j = 0, \ldots, T_\Delta - 1$ and $p = j+1, \ldots, T_\Delta$ the measure $\nu_k^{j,p}$ is given as $(p-j)$ - th iteration for the input data $\nu_k^{j,j}$. In addition, $\nu_k^{j,j}$ is provided by random selection form $\nu_k^{j-1,j}$, for $j = 1, \ldots, T_\Delta$.

The above computational chain shows that one can expect that the errors due to the random selections increase when the timestep Δt decreases. Indeed,

such a behavior was observed in numerical applications Ref. [13, 12]. Some theoretical estimations on the errors Ref. [12] prove that the probabilistic errors ε behave like

$$\varepsilon \sim \frac{1}{\Delta t \cdot \sqrt{n}}. \tag{7.4}$$

Consequently, when we decrease the timestep (to improve the errors for the time discretization, Proposition 4.1.b) we shall increase the number of terms for the initial approximation, in order to keep the stochastic errors in acceptable limits.

8. Appendix

Proof of Lemma 3.1.

Let $n \in \mathbb{N}^*$ and let $a_1, \ldots, a_n > 0$, be some constants. Consider the positive quadratic form defined on $\mathbb{R}^{3n}$ by

$$T := T(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{i=1}^{n} a_i \mathbf{v}_i^2, \tag{8.1}$$

where $\mathbf{v}_i \in \mathbb{R}^3$, for all $i = 1, \ldots, n$. One introduces the Jacobi-type transformation

$$\mathbb{R}^{3n} \ni (\mathbf{v}_1, \ldots, \mathbf{v}_n) \to (\underline{V}, \xi) \in \mathbb{R}^3 \times \mathbb{R}^{3n-3}, \tag{8.2}$$

where

$$\underline{V} := (\sum_{i=1}^{n} a_i)^{-1} \sum_{i=1}^{n} a_i \mathbf{v}_i, \tag{8.3}$$

and $\xi := (\xi_1, \ldots, \xi_{n-1})$, with

$$\xi_i := \left[\frac{1}{a_{i+1}} + \frac{1}{\sum\limits_{j=1}^{i} a_j} \right]^{-\frac{1}{2}} \left[\mathbf{v}_{i+1} - \frac{\sum\limits_{j=1}^{i} a_j \mathbf{v}_j}{\sum\limits_{j=1}^{i} a_j} \right], \tag{8.4}$$

for $i = 1, \ldots, n-1$.

By (8.2), the form T takes the form

$$T = T(\underline{V}, \xi) = \left(\sum_{i=1}^{n} a_i \right) \cdot \underline{V}^2 + \xi^2. \tag{8.5}$$

Define

$$W_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}) := W_{\boldsymbol{\alpha}}(\mathbf{w}) - \frac{1}{2}\left(\sum_{n=1}^{N}\alpha_n m_n\right)\cdot V_{\boldsymbol{\alpha}}(\mathbf{w})^2 - \sum_{n=1}^{N}\beta_n E_n, \tag{8.6}$$

and

$$t_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}) := \begin{cases} [W_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w})]^{1/2} & \text{if} \quad W_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{8.7}$$

Now, consider the form on $\mathbb{R}^{3|\boldsymbol{\beta}|}$,

$$T_{\boldsymbol{\beta}}(\mathbf{u}) := W_{\boldsymbol{\beta}}(\mathbf{u}) - \sum_{n=1}^{N}\beta_n E_n \tag{8.8}$$

and a corresponding Jacobi-type transformation as in (8.2),

$$\mathbb{R}^{3|\boldsymbol{\beta}|} \ni \mathbf{u} \to (\underline{V}, \xi) \in \mathbb{R}^3 \times \mathbb{R}^{3|\boldsymbol{\beta}|-3}, \tag{8.9}$$

with $\xi := (\xi_1, \ldots, \xi_{|\boldsymbol{\beta}|-1})$, where $\xi_i \in \mathbb{R}^3$, for all $i = 1, \ldots, |\boldsymbol{\beta}| - 1$. Denote by $\Delta_{\boldsymbol{\beta}}$ the Jacobian determinant of the transformation. Let ξ be represented in spherical coordinates on $\mathbb{R}^{3|\boldsymbol{\beta}|-3}$, $\xi = r\mathbf{n}$, with $(r, \mathbf{n}) \in [0, \infty) \times \Omega_{\boldsymbol{\beta}}$. Consider the inverse map

$$\mathbb{R}^3 \times \mathbb{R}_+ \times \Omega_{\boldsymbol{\beta}} \ni (\underline{V}, r, \mathbf{n}) \to \mathbf{u}(\underline{V}, r, \mathbf{n}) \in \mathbb{R}^{3|\boldsymbol{\beta}|} \tag{8.10}$$

of the transformation $\mathbf{u} \to (\underline{V}, r, \mathbf{n})$ and set

$$\mathbf{u}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}) := \mathbf{u}(\underline{V}, r, \mathbf{n})|_{\,\underline{V}=V_{\boldsymbol{\alpha}}(\mathbf{w}), r=t_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w})}. \tag{8.11}$$

Obviously, for all $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{M}$ such that (2.6) is satisfied, we have

$$V_{\boldsymbol{\beta}}(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n})) = V_{\boldsymbol{\alpha}}(\mathbf{w}) \qquad W_{\boldsymbol{\beta}}(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n})) = W_{\boldsymbol{\alpha}}(\mathbf{w}). \tag{8.12}$$

Define

$$\begin{aligned} p_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}) &:= 2^{-1}\Delta_{\boldsymbol{\beta}} \cdot t_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w})^{3|\boldsymbol{\beta}|-5} K_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{u}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}), \mathbf{w}), \\ r_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n}) &:= 2^{-1}\Delta_{\boldsymbol{\beta}} \cdot t_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w})^{3|\boldsymbol{\beta}|-5} K_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{w}, \mathbf{u}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\mathbf{w}, \mathbf{n})). \end{aligned} \tag{8.13}$$

From (8.12), one obtains property i) of the Lemma 3.1. Property ii) follows from the definitions introduced in (8.7) and (8.13).

The limits (3.6) and (3.7), can be obtained from (3.3) and (3.4). We start the computation with the integral upon d$\mathbf{u}$, by choosing $(\underline{V}, r, \mathbf{n})$ as new integration variables such that $\mathbf{u} = \mathbf{u}(\underline{V}, r, \mathbf{n})$. Since $\mathbf{f}_{\boldsymbol{\alpha}} \in C_c(\mathbb{R}^{3|\boldsymbol{\alpha}|})$ and $\mathbf{f}_{\boldsymbol{\beta}} \in C_c(\mathbb{R}^{3|\boldsymbol{\beta}|})$, using the properties of $K_{\boldsymbol{\alpha},\boldsymbol{\beta}}$, δ_ε^3, δ_η and $\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}$, we obtain (3.6) and (3.7) by repeated application of Lebesgue's dominated convergence theorem.

Using a similar argument as in the proof (3.6), for all $f \in C_c(\mathbb{R}^{3|\boldsymbol{\beta}|})$ and $\varphi \in C_b(\mathbb{R}^{3|\boldsymbol{\alpha}|})$, we get

$$\begin{aligned} &\lim_{\eta\to 0}\lim_{\varepsilon\to 0}\int_{\mathbb{R}^{3|\boldsymbol{\alpha}|}\times\mathbb{R}^{3|\boldsymbol{\beta}|}} \varphi(\mathbf{w})\sigma^{\varepsilon,\eta}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{u},\mathbf{w})f(\mathbf{u})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{u} \\ &= \int_{\mathbb{R}^{3|\boldsymbol{\alpha}|}\times\boldsymbol{\Omega}_{\boldsymbol{\beta}}} \varphi(\mathbf{w})p_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n})f(\mathbf{u}_{\boldsymbol{\beta},\boldsymbol{\alpha}}(\mathbf{w},\mathbf{n}))\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{n}, \end{aligned} \tag{8.14}$$

giving the left side of (3.9). To obtain the right side of (3.9), we repeat the procedure, but first we perform the integral upon d$\mathbf{w}$ in the left side of (8.14) (using the change of variables induced by the Jacobi-type transformation $\mathbb{R}^{3|\boldsymbol{\alpha}|} \ni \mathbf{w} \to (\underline{V}, \xi) \in \mathbb{R}^3 \times \mathbb{R}^{3|\boldsymbol{\alpha}|-3}$, associated to the form $T_{\boldsymbol{\alpha}}(\mathbf{w}) = W_{\boldsymbol{\alpha}}(\mathbf{w}) - \sum_{n=1}^N \boldsymbol{\alpha}_n E_n$, and then taking the representation of $\xi \in \mathbb{R}^{3|\boldsymbol{\alpha}|-3}$ in spherical coordinates). □

References

[1] AIZENMAN I., Duke Math. J. **45**, 809 (1978).

[2] ARKERYD L., Achiv for Rat. Mech. and Anal. **45**, 1 (1972).

[3] ARKERYD L., Achiv for Rat. Mech. and Anal. **45**, 17 (1972).

[4] BABOVSKY H., European Journal of Mechanics B/Fluids **8**, 41 (1989).

[5] BABOVSKY H., ILLNER R., SIAM J. Num. Anal. **26**, 45 (1989).

[6] BABOVSKY H., GROPENGIESSER F., NEUNZERT H., STRUCKMEIER J., WIESEN B., *Low Discrepancy Methods for the Boltzmann Equation*, Bericht Nr. 32, Fachbereich Mathematik, Kaiserslautern Univ. (1988).

[7] BABOVSKY H., ILLNER R., *The Essence of Particle Simulation of the Boltzmann Equation*, in Collection "Multidimensional Hyperbolic Problems and Computations", p. 13, J. Glimm and A. Majda Eds. (The IMA Vol. in Math.and Appl., vol. 29) Springer, New York (1991).

[8] Bärwinkel K., Wolters H., *Reaktionskinetik und Transport in Relaxierenden Gasen*, Report BMFT-FB W 75-28, Dornier, Friedrichshafen, 1975.

[9] Boltzmann L., *Further studies on the thermal equilibrium of gas molecules*, in Brush S.G., Kinetic Theory Volume 2, Irreversible Processes, Pergamon, Oxford (1966).

[10] Borgnakke C., Larsen P., J. Comp. Phys. **18**, 405 (1975).

[11] Cercignani C., *The Boltzmann Equation and Its Applications*, in Applied Mathematical Science, No. 67, Springer Berlin (1988).

[12] Espesset A., Grünfeld C.P., Marinescu D., *Tests of a Simulation Method for Boltzmann-like Models with Chemical Reactions*, Journal of Computational Physics, Vol. **175**, pp. 225–248 (2002).

[13] Espesset A., Marinescu D., *Tests of a Simulation Method for a System of Boltzmann Equations*, Computers & Mathematics with Applications, Vol. **40** (6-7) pp. 805–812 (2000).

[14] Georgescu E., Grünfeld C.P., Marinescu D., *Tests of a Simulation Method for the Multicomponent Boltzmann Equation*, Balkan Phys. Lett. **4**, 94 (1996).

[15] Gropengiesser F., Neunzert H., Struckmeier J, *Computational Methods for the Boltzmann Equation*, Bericht Nr. 43, Fachbereich Mathematik, Kaiserslautern Univ. (1990).

[16] Grünfeld C.P., C.R. Acad. Sci. Paris Tome I, **316**, 953 (1993).

[17] Grünfeld C.P., Georgescu E., *On a Class of Kinetic Equations for Reacting, Gas Mixtures*, Mat. Fiz., Analiz, Geom., **2**, 408 (1995).

[18] Grünfeld C.P., *Nonlinear Kinetic Models with Chemical Reactions*, in N. Bellomo and M. Pulvirenti, editors, *Modeling in Applied Sciences, A Kinetic Theory Approach* (Birkhäuser, Boston, 2000).

[19] Grünfeld C.P., Marinescu D., *On the Numerical Simulation of a Class of Reactive Boltzmann Type Equations*, Transp. Theory Stat. Phys., **26**, 287 (1997).

[20] Grünfeld C.P., Marinescu D., *Tests of a Convergent Numerical Scheme for Nonlinear Boltzmann-like Models with Chemical Reactions: Two, Three and Four Species*, Proceedings of The 3-rd International Colloquium "Mathematics in Engineering and Numerical Physics" October

7-9, 2004, Bucharest, Geometry Balkan Press, Bucharest, Romania, pp. 60–72, (2005).

[21] HOFFMAN D.K., DAHLER J.S., J. Stat Phys. **1**, 521 (1969).

[22] KROOK M., WU T.T., Phys. Rev Lett. **38**, 991 (1977).

[23] KUIPERS L., NIEDERREITER H., *Uniform Distribution of Sequences*, John Wiley & Sons (1974).

[24] KUŠČER I., Physica A **176**, 542 (1991).

[25] LUDWIG G., HEIL M., *Boundary-Layer Theory with Dissociation and Ionization*, in "Advances in Applied Mechanics" vol. **6**, p. 39, Academic Press, New York (1960).

[26] NAMBU K., J. Phys. Soc. Japan **49** 2042 (1980).

[27] PULLIN D.I., Phys. Fluids **21**, 209 (1978)

[28] Shiryayev S.N., *Probability*, in Graduate Texts in Mathematics, vol. **95**, Springer, Berlin (1984).

[29] SMITH F.T., *Triple Collisions and Termolecular Reaction Rates*, in "Kinetic processes in gases and plasmas", p. 321, A.R. Hochstim, Ed., Academic Press (1969).

[30] SNIDER R.F., *Transport Properties of Dilute Gases with Internal Structure*, in "Transport Phenomena", L.N. Phys. Vol. **31**, p. 469, Springer, Berlin (1974).

[31] WANG CHANG C.S., UHLENBECK E., *Transport phenomena in polyatomic gases*, Engineering Research Report CM-681, University of Michigan (1951).

[32] WANG CHANG C.S., UHLENBECK E., DE BOER J., *The Heat Conductivity and Viscosity of Polyatomic Gases*, in "Studies in Statistical Mechanics", J. De Boer and G.E. Uhlenbeck Eds, vol. **2**, Part C, p. 241, North Holland, Amsterdam (1964).

Topics in Applied Mathematics & Mathematical Physics

Mathematical Models of Diffusion in Nonhomogeneous Porous Media

Gabriela Marinoschi [1]

Contents

1. Physical context and mathematical hypotheses

From the hydraulic point of view, the problems we shall study are related to a Darcian flow of an incompressible fluid in an isotropic, nonhomogeneous nondeformable porous medium with a variable porosity and with no hysteresis development.

[1]"Gheorghe Mihoc–Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, e-mail: gabimarinoschi@yahoo.com.

This work was elaborated under the contract CEEX05-D11-25/2005.

The general boundary value problem. Assume that the flow domain Ω is an open bounded subset of $\mathbf{R}^N$ $(N = 1, 2, 3)$, and the time runs within the finite time interval $(0, T)$. The boundary of Ω is denoted by Γ and it is considered piecewise smooth. The vector of space variables is denoted by $x = (x_1, x_2, x_3) \in \Omega$ and the time by $t \in (0, T)$.

We consider the Richards' equation describing the water infiltration into an isotropic, nonhomogeneous, unsaturated porous medium with a variable porosity, with initial data and various boundary conditions (see [7])

$$\frac{\partial(m(x)S_w(h))}{\partial t} - \nabla \cdot (k(h)\nabla h) + \frac{\partial k(h)}{\partial x_3} = f \text{ in } Q = \Omega \times (0, T), \tag{1.1}$$

$$h(x, 0) = h_0(x) \text{ in } \Omega, \tag{1.2}$$

$$\textit{boundary conditions for } h \text{ on } \Sigma = \Gamma \times (0, T). \tag{1.3}$$

The unknown in Richards' equation is the *capillary pressure* $h(x, t)$ (or *pressure head*, or *water pressure* in the unsaturated soil), S_w is the *water saturation* in pores, m is the medium porosity and $\theta = m(x)S_w$ is the *volumetric water content* or *soil moisture*. In this work the dependence of m on x models the nonhomogeneity of the medium. The function k is the *hydraulic conductivity*, $f(x, t)$ is a source (or sink) in the flow domain and h_0 is the initial pressure distribution in the domain, f and h_0 being given. In general $m \in (0, 1)$ but a limit case with m tending to 0 may have a physical relevance. The properties of the dependence of S_w and k on h will be specified.

In particular, we shall exemplify the theory for the case of the medium having a part of the boundary, Γ_α semipermeable, allowing a water flux across it and the other part Γ_u at which the pressure will be given. Here, Γ_u and Γ_α are disjoint and $\Gamma = \Gamma_u \cup \overline{\Gamma_\alpha}$. In infiltration problems, we can often meet the situation in which water ponds on the soil surface (let it be Γ_u). This happens when the rainfall rate is greater than the soil conductivity at saturation and the soil begins to saturate from the surface, or when the soil surface is in contact with an open water body, for example the bottom of a lake. In consequence the boundary conditions we shall consider are

$$h(x, t) = h_u(x, t) \geq 0 \text{ on } \Sigma_u = \Gamma_u \times (0, T), \tag{1.4}$$

$$q \cdot \nu = f_\alpha \text{ on } \Sigma_\alpha = \Gamma_\alpha \times (0, T), \tag{1.5}$$

where q is the water flux defined by

$$q(x, t) = k(h)i_3 - k(h)\nabla h, \tag{1.6}$$

ν is the outer normal vector at the boundary and i_3 is the unit vector of the Ox_3 axis, downwards directed.

We can reverse the boundary conditions by considering that Γ_α is the soil surface and Γ_u is the underground boundary. Thus we can interpret that the flux through the soil surface, is provided by a water supply as a rain or irrigation and that the lower part of the porous medium is in contact with the phreatic aquifer.

Description of the hydraulic model. The behaviour of an unsaturated soil, i.e., partially filled with water, is completely known from the hydraulic point of view if two functions are given: one is the *retention curve*

$$S_w = \widetilde{C}^*(h), \tag{1.7}$$

linking the *water saturation* S_w, to the pressure head h, and the other is the *hydraulic conductivity*

$$k = k(h), \tag{1.8}$$

both depending nonlinearly on h. For an isotropic soil the latter is a scalar function.

Since we study the nonhysteretic case, the retention curve and the hydraulic conductivity are assumed single-valued functions of the pressure.

In soil sciences, the unsaturated pressure is considered negative ($h < 0$) and the saturation is characterized by $h = 0$. Also, it is considered that the process of infiltration-drainage (opposite to infiltration) takes place between two limits of h. The lowest limit is denoted h_r and at this pressure head the soil is considered dry even if some water still resides in the pores and the hydraulic conductivity is still positive. The corresponding water saturation is denoted S_r and the volumetric water content θ_r is called *residual moisture* (see [7]). The upper limit is $h = 0$ where saturation is reached and water saturation becomes equal to 1. However, we shall denote this value by S_s. At saturation, moisture attains its *saturation value* θ_s equal to the medium porosity at this point (if the porosity is not constant). The parts of the medium where $h > 0$ are completely saturated. We define the derivative of the water saturation with respect to the pressure

$$\widetilde{C}(h) = \frac{\mathrm{d}S_w}{\mathrm{d}h}(h). \tag{1.9}$$

For the saturated flow, when $h \geq 0$, the previously functions take constant values.

Generally, the hydraulic models raise a difficult mathematical problem. When the pressure head in the unsaturated soil comes close to the saturation value, $\widetilde{C}$ vanishes and Richards' equation degenerates. Correspondingly, the diffusion coefficient expressed as a function of moisture exhibits a blow-up development around saturation. In soil sciences the model which reflects this behaviour is the strongly nonlinear Green-Ampt limit model, see [10]. The situation in which $\widetilde{C}(0) > 0$ corresponds to a less nonlinear hydraulic behaviour, the typical model for this class being the Burgers' model, see [10], too. Depending on the particularities of the hydraulic functions which are determined by the soil pore structure, the models of water infiltration range between these two limit models (see [44]).

Previous theoretical results. In the most mathematical literature devoted to this subject the blow-up of the diffusivity in the diffusive form of Richards' equation was avoided, by considering a finite-valued diffusivity, or studying the problem only in the pressure form (see [2], [4], [12], [19], [20], [25], [26], [27], [37], [38]). More recently, in the paper [9] a model of the saturated-unsaturated flow lying on a special definition of the boundary conditions that changes during the phenomenon evolution, has been developed also for a finite value of the diffusivity at saturation (which was implied by the assumption that $\widetilde{C}(0) > 0$). Following the technique presented in [20] the model was reduced to systems in class of Stefan-like problems of high-order, see [19].

However, apart from specific infiltration problems, previous existence and uniqueness studies for solutions to the elliptic-parabolic equation

$$\frac{\partial(b(u))}{\partial t} + \nabla \cdot (a(\nabla u, b(u))) + f(b(u)) = 0 \text{ in } \Omega \times (0,T)$$

have been presented in the literature especially using a technique inspired by the method of entropy solutions introduced by S.N. Krushkov in [28]. Originally, this method was devoted to prove L^1-contraction for entropy solutions for scalar conservation laws, i.e., generalized solutions in the sense of distributions satisfying admissibility conditions similar to those of entropy growth in gas dynamics (see also [8]). J. Carillo applied Krushkov's method to second order equations (see [13], [14], [15], [16]). F. Otto (see [35], [36]) proved a L^1-contraction principle and uniqueness of solutions for this type of equation by applying Krushkov's technique only to the time variable. H.W. Alt and S. Luckhaus showed in [1] that the natural solution space for this

equation is given by all functions u of finite energy in the sense that

$$\sup_{t\in(0,T)}\int_\Omega \Psi(b(u(t)))\mathrm{d}x + \int_Q |\nabla u|^r \,\mathrm{d}x\mathrm{d}t < \infty,$$

where Ψ is the Legendre transform of the primitive of b.

We also mention the results of J.L. Vázquez regarding the fast diffusion equations (see [18], [40], [41], [42], [17] and the book [43]).

Concerning the degenerate evolution equations, extensive studies have been performed for linear operators, relying on the properties of the resolvent of an appropriate multivalued linear operator accounting for the multiplication by the function m (see [21], [23] and the monograph [22]). We mention also the paper [24] related to a similar topic in which a degenerate model with homogeneous Dirichlet boundary conditions and no transport was studied.

The analysis of the well-posedness of the diffusive form of Richards' equation in the unsaturated case ($\theta < \theta_s$) with the porosity m constant, was developed in the papers [6], [29], [30], [31] within a functional approach. The existence results which were deduced showed that solutions reaching saturation can be obtained but only on zero-measure subsets of Q. Somehow, this was expected because the unsaturated model reflects a behaviour of a particular soil only and not the general feature of the process which includes the possible soil saturation.

In the paper [32] a rigorous mathematical model able to describe the saturation occurrence (with the blow-up of the diffusivity) was introduced for a homogeneous porous medium (with m constant) in the diffusive form and developed then in [33].

In the first part of this chapter we introduce the diffusive models of water flow in saturated-unsaturated media characterized by a space variation of the porosity. Then we analyze a model with mixed boundary conditions involving a flux on a part of the boundary and a nonhomogeneous Dirichlet condition corresponding to a singular situation on another part of the domain boundary. The model will be degenerate because we shall assume that porosity can vanish on a subset of Ω.

2. Diffusion models in nonhomogeneous porous media

We intend first to reveal how the particular character of the hydraulic models is determined by the behaviour of the functions $\widetilde{C}^*$ and k around 0.

Mathematical hypotheses. For the unsaturated flow, where $h < 0$, we assume the following:

(m_1) $\widetilde{C}^* : [h_r, 0) \to [S_r, S_s)$ is single-valued, positive, differentiable on $[h_r, 0)$, monotonically increasing ;

(m_2) $k : [h_r, 0) \to [K_r, K_s)$ is single-valued, positive, differentiable on $[h_r, 0)$, monotonically increasing and satisfies the property $k'(h_r) = 0$;

(m_3) $\widetilde{C} : [h_r, 0) \to (\widetilde{C}_0, \widetilde{C}_r]$ is single-valued, non-negative, differentiable on $[h_r, 0)$ monotonically decreasing and satisfies $\widetilde{C}'(h_r) = 0$;

In the saturated flow we have

(m_4) $\widetilde{C}^*(h) = S_s$, $k(h) = K_s$ and $\widetilde{C}(h) = 0$ for $h \geq 0$.

We denote

$$S_s = (\widetilde{C}^*)(0) > 0, \tag{2.1}$$

$$\widetilde{C}_0 = (\widetilde{C}^*)'(0) = \widetilde{C}(0) \geq 0, \tag{2.2}$$

$$K_s = k(0) > 0, \tag{2.3}$$

$$K_0' = \lim_{h \nearrow 0} k'(h),\ K_0' \in [0, \infty). \tag{2.4}$$

Therefore, the unsaturated flow is characterized either by $h < 0$ or $S_w \in [S_r, S_s)$ while the saturated one is indicated by $h \geq 0$ or $S_w = S_s$.

The positive values S_r, S_s and their corresponding conductivities K_r, K_s are soil characteristics and they are known for each type of soil apart. The properties $k'(h_r) = 0$ and $\widetilde{C}'(h_r) = 0$ were put into evidence by experiments (see [10]).

We notice that the functions $\widetilde{C}^*$ and k are continuous on $[h_r, \infty)$, and h_r is the maximum point for $\widetilde{C}$. Also $\widetilde{C}$ is continuous on $[h_r, \infty)$, except possibly at the point 0.

We stress the fact that these properties are verified by the empirical hydraulic models set up in the last decades (see e.g., [44]).

We emphasize that the main role is played by the increase rate of the functions $\widetilde{C}^*$ and k around 0, the significant contribution being given by the behaviour of the retention curve $\widetilde{C}^*$.

2.1. Strongly nonlinear saturated-unsaturated diffusive model

Let us assume $(m_1)-(m_4)$ and

$$\widetilde{C}_0 = 0$$

which is the main characteristic of this case. It follows then that $\widetilde{C}$ is continuous on $[h_r, \infty)$ and we can write $\widetilde{C}^* : [h_r, \infty) \to [S_r, S_s]$, as

$$\widetilde{C}^*(h) = \begin{cases} S_r + \int_{h_r}^{h} \widetilde{C}(\zeta)\mathrm{d}\zeta, \ h < 0, \\ S_s, \ h \geq 0. \end{cases} \tag{2.5}$$

Strongly nonlinear hydraulic conductivity. This situation corresponds to $K_0' \in \mathbf{R}_+ = (0, \infty)$.

We define a primitive of K by

$$K^*(h) = \begin{cases} K_r^* + \int_{h_r}^{h} k(\zeta)\mathrm{d}\zeta, \ h < 0, \\ K_s^* + K_s h, \ h \geq 0, \end{cases} \tag{2.6}$$

where $K^* : [h_r, \infty) \to [K_r^*, \infty)$ and

$$K_s^* = K^*(0) > 0. \tag{2.7}$$

The function K^* is differentiable, monotonically increasing on $[h_r, \infty)$ and with these notations Richards' equation (1.1) becomes

$$\frac{\partial(m(x)S_w)}{\partial t} - \Delta K^*(h) + \frac{\partial k(h)}{\partial x_3} = f \text{ in } Q. \tag{2.8}$$

By the initial condition (1.2) we obtain

$$S_w(x, 0) = S_{w0}, \ S_{w0} = \widetilde{C}^*(h_0).$$

We can also consider the initial condition

$$m(x)S_w(x, 0) = \theta_0(x) \text{ in } \Omega, \text{ where } \theta_0 = m(x)\widetilde{C}^*(h_0) \tag{2.9}$$

and corresponding replacements should be made in the boundary conditions (1.4)–(1.5).

Since it is more convenient to work with the variable S_w, we introduce from (2.5) the inverse of $\widetilde{C}^*$, $(\widetilde{C}^*)^{-1} : [S_r, S_s] \to [h_r, +\infty)$, by

$$(\widetilde{C}^*)^{-1}(S_w) = \begin{cases} (\widetilde{C}^*)^{-1}(S_w), \ S_w \in [S_r, S_s), \\ [0, +\infty), \ S_w = S_s, \end{cases} \tag{2.10}$$

which is multivalued at $S_w = \theta_s$ and continuous and monotonically increasing on $[S_r, S_s)$. Then, we replace it all over in (1.1)–(1.5).

Thus, instead of the conductivity written in function of pressure, we obtain the conductivity expressed in terms of water saturation

$$\widetilde{K} : [S_r, S_s] \to [K_r, K_s],\ \widetilde{K}(S_w) = (k \circ \widetilde{C}^*)^{-1}(S_w),\ S_w \in [S_r, S_s], \quad (2.11)$$

function that preserves some of the properties of k, i.e., it is positive, differentiable (except at S_s) and monotonically increasing, since for any $S_w \in [S_r, S_s)$ we have that

$$\widetilde{K}'(S_w) = k'((\widetilde{C}^*)^{-1}(S_w)) \cdot ((\widetilde{C}^*)^{-1})'(S_w) = \frac{k'((\widetilde{C}^*)^{-1}(S_w))}{\widetilde{C}((\widetilde{C}^*)^{-1}(S_w))} > 0. \quad (2.12)$$

We notice also that

$$\widetilde{K}'(S_r) = 0 \quad (2.13)$$

and

$$\lim_{S_w \nearrow S_s} \widetilde{K}'(S_w) = +\infty. \quad (2.14)$$

However, for $S_w \in [S_r, S_l]$ with $S_l < S_s$ the derivative of $\widetilde{K}$ is bounded, so that $\widetilde{K}$ follows to be Lipschitz on intervals strictly included in $[S_r, S_s)$

$$\left|\widetilde{K}(S_w) - \widetilde{K}(\overline{S_w})\right| \leq M_l \left|S_w - \overline{S_w}\right|,\ \forall S_w, \overline{S_w} \in [S_r, S_l],\ S_l < S_s, \quad (2.15)$$

where

$$M_l = \max_{S_w \in [S_r, S_l]} \frac{k'((\widetilde{C}^*)^{-1}(S_w))}{\widetilde{C}((\widetilde{C}^*)^{-1}(S_w))} < \infty. \quad (2.16)$$

Plugging (2.10) in (2.6) we get the function

$$\widetilde{\beta}^*(S_w) = \begin{cases} (K^* \circ (\widetilde{C}^*)^{-1})(S_w),\ S_w \in [S_r, S_s), \\ [K_s^*, +\infty),\ S_w = S_s \end{cases} \quad (2.17)$$

that is multivalued at $S_w = S_s$ but is continuous from the left at this point

$$\lim_{S_w \nearrow S_s} \widetilde{\beta}^*(S_w) = K_s^*. \quad (2.18)$$

For $S_w \in [S_r, S_s)$ the function $(\widetilde{C}^*)^{-1}$ is monotonically increasing, so that we can calculate $\widetilde{\beta}^*(S_w)$ by changing the variable in the integral (2.6) and denoting $\zeta = (\widetilde{C}^*)^{-1}(\xi)$. In this way we get

$$\widetilde{\beta}^*(S_w) = K_r^* + \int_{S_r}^{S_w} \beta(\xi)\mathrm{d}\xi,\ \text{for } S_w \in [S_r, S_s),$$

where

$$\widetilde{\beta}(S_w) = \frac{k((\widetilde{C}^*)^{-1}(S_w))}{\widetilde{C}((\widetilde{C}^*)^{-1}(S_w))}, \text{ for } S_w \in [S_r, S_s). \tag{2.19}$$

In this way we have rigorously recovered the definition of the *water diffusivity* function.

We notice that $\widetilde{\beta}$ has two important properties

$$\widetilde{\beta}(S_w) \geq \widetilde{\rho} = \widetilde{\beta}(S_r) = \frac{K_r}{\widetilde{C}_r} > 0, \ \forall S_w \in [S_r, S_s) \tag{2.20}$$

and

$$\lim_{S_w \nearrow Ss} \widetilde{\beta}(S_w) = +\infty. \tag{2.21}$$

Moreover, by the hypotheses made upon the functions $\widetilde{C}$ and k it follows that $\widetilde{\beta}$ is monotonically increasing, i.e.,

$$\widetilde{\beta}' = \frac{k'\widetilde{C} - k\widetilde{C}'}{\widetilde{C}^3} \geq 0, \text{ on } [S_r, S_s), \tag{2.22}$$

$$\widetilde{\beta}'(S_r) = 0. \tag{2.23}$$

Hence, $\widetilde{\beta}^*$ is twice differentiable and strictly monotonically increasing on $[S_r, S_s)$ and as a matter of fact we can write

$$\widetilde{\beta}^*(S_w) = \begin{cases} K_r^* + \int_{S_r}^{S_w} \widetilde{\beta}(\xi)\mathrm{d}\xi \text{ for } S_w \in [S_r, S_s), \\ [K_s^*, +\infty) \text{ for } S_w = S_s. \end{cases} \tag{2.24}$$

Moreover, by (2.20) and (2.24) we deduce that the function $\widetilde{\beta}^*$ satisfies the inequality

$$(\widetilde{\beta}^*(S_w) - \widetilde{\beta}^*(\overline{S_w}))(S_w - \overline{S_w}) \geq \rho(S_w - \overline{S_w})^2, \forall S_w, \overline{S_w} \in [S_r, S_s]. \tag{2.25}$$

In conclusion we can set

Model 1. Let us assume $(m_1) - (m_4)$, $\widetilde{C}_0 = 0$ and $K_0' \in \mathbf{R}_+$. Then, the diffusive model of the *strongly nonlinear saturated-unsaturated infiltration with a strongly nonlinear hydraulic conductivity* is given by

$$\frac{\partial(m(x)S_w)}{\partial t} - \Delta\widetilde{\beta}^*(S_w) + \frac{\partial\widetilde{K}(S_w)}{\partial x_3} = f \text{ in } Q, \tag{2.26}$$

$$m(x)S_w(x,0) = \theta_0(x) \text{ in } \Omega, \tag{2.27}$$

$$\text{boundary conditions in } S_w \text{ on } \Sigma, \tag{2.28}$$

where $\widetilde{\beta}^*$ is the multivalued function defined by (2.24), $\widetilde{\beta}$ is given by (2.19) and $\widetilde{K}$ is the single-valued function (2.11). Moreover, $\widetilde{\beta}^*$ is strongly monotone, $\widetilde{\beta}$ satisfies (2.20)–(2.23) and $\widetilde{K}$ has the properties (2.13)–(2.16).

The boundary conditions (1.4)–(1.5) become

$$S_w(x,t) = S_s \text{ on } \Sigma_u, \tag{2.29}$$

$$\left(\widetilde{K}(S_w)i_3 - \nabla\widetilde{\beta}^*(S_w)\right)\cdot\nu = f_\alpha \text{ on } \Sigma_\alpha. \tag{2.30}$$

The qualifier of *strongly nonlinear* is implied by the property of the function β which evolves highly nonlinear around the saturation point, S_s. This is justified by the fact that the typical representative for this behaviour (correlated with that of its primitive $\widetilde{\beta}^*$ which is finite at this point) is of the form

$$\widetilde{\beta}(S_w) = \frac{1}{(S_s - S_w)^{1-p}} \text{ for } 0 < p < 1.$$

We notice that this form of the diffusivity function reveals the character of *fast diffusion* of this process (see the review of diffusion-type processes in [3]).

However, $\widetilde{\beta}^*$ is multivalued and the sign equal (=) in (2.26) is not properly used. The appropriate symbol should be $\ni$. Also, we shall specify later the exact meaning of the solutions to (2.26)-(2.30). The fact that equation (2.26) is multivalued must not be surprising if one takes into account that it models a free boundary problem. This means that, at each time t, the domain Ω can be decomposed into two regions: the saturated one, $\{x;\ S_w(x,t) = S_s\}$ and the unsaturated one $\{x;\ S_w(x,t) < S_s\}$, separated by a free boundary. The extension of a nonlinear function arising in such a problem to a multivalued one is common in the theory of nonlinear differential equations with discontinuous coefficients as well as in that modelling free boundary processes.

Thus, equation (2.26) represents an extension of Richards' equation (written for the unsaturated infiltration) to the simultaneous saturated-unsaturated flow.

Weakly nonlinear hydraulic conductivity. A strongly nonlinear model, but with a weaker nonlinear behaviour of the conductivity may be obtained under conditions that lead to $\lim_{S_w \nearrow S_s} \widetilde{K}'(S_w) < \infty$. To reach such a situation we have to impose just from the beginning a stronger condition for k, namely that there exists $M > 0$, such that

$$k'(h) \le M\widetilde{C}(h),\ \forall h \in [h_r, 0], \tag{2.31}$$

which implies that

$$K_0' = 0,\ \lim_{h \nearrow 0} \frac{k'(h)}{\widetilde{C}(h)} = M. \tag{2.32}$$

In this way $\widetilde{K}$ turns out to be Lipschitz on $[S_r, S_s]$ with the constant M. We observe that the functions $\widetilde{\beta}$ and $\widetilde{K}$ remain monotonically increasing. This situation is put into evidence e.g., in the van Genuchten model (see [39]) for the model parameter m close to 1. This case can be resumed in

Model 2. Let us assume $(m_1) - (m_4)$, $\widetilde{C}_0 = 0$ and (2.31)–(2.32). Then, the diffusive model of *strongly nonlinear saturated-unsaturated infiltration with a weakly nonlinear hydraulic conductivity* is given by (2.26)–(2.28), where the functions $\widetilde{\beta}$ and $\widetilde{\beta}^*$ have the properties specified in Model 1 except for $\widetilde{K}$ which is given by (2.11), with

$$\lim_{S_w \nearrow S_s} \widetilde{K}'(S_w) = M < \infty.$$

2.2. Weakly nonlinear saturated-unsaturated diffusive model

For some hydraulic models the diffusivity is finite at $S_w = S_s$. We intend to reveal which properties of the functions $\widetilde{C}^*$ and k can provide such a value. Let us suppose that the retention curve increases from the left to its maximum value with a nonzero rate at the left of zero,

$$\widetilde{C}_0 > 0,$$

but very close to 0. In this case $\widetilde{C}^*$ is not differentiable at $h = 0$ and the function

$$\widetilde{C} : [h_r, \infty) \to [0, \widetilde{C}_r],\ \widetilde{C}(h) = \begin{cases} \dfrac{\mathrm{d}S_w}{\mathrm{d}h}(h),\ h < 0 \\ 0,\ h \geq 0 \end{cases} \tag{2.33}$$

is no longer continuous at $h = 0$, having the jump $\left|\widetilde{C}_0\right| = \lim\limits_{h \nearrow 0} \frac{\mathrm{d}S_w}{\mathrm{d}h}$.

The functions $\widetilde{K}$ and $\widetilde{\beta}^*$ and $\widetilde{\beta}$ will be defined in the same way as before, but in this case the value of $\widetilde{\beta}$ at $S_w = S_s$ exists and it is

$$\lim_{S_w \nearrow S_s} \widetilde{\beta}(S_w) = \frac{K_s}{\widetilde{C}_0} < \infty. \tag{2.34}$$

However, the function $\widetilde{\beta}^*(S_w)$ will be extended in a multivalued way, by $\widetilde{\beta}^*(S_w) = K_s^*$ at S_s.

Weakly nonlinear hydraulic conductivity. Assume that the derivative of k at $h = 0$, has a finite value, $K_0' < \infty$. Hence, $\widetilde{K}$ is Lipschitz with the constant

$$M = \max_{S_w \in [S_r, S_s]} \frac{k'((\widetilde{C}^*)^{-1}(S_w))}{\widetilde{C}((\widetilde{C}^*)^{-1}(S_w))} \leq \frac{K_0'}{\widetilde{C}_0}, \tag{2.35}$$

so that we can settle

Model 3. Let us assume $(m_1) - (m_4)$, $\widetilde{C}_0 > 0$ and $K_0' < \infty$. Then, the diffusive model of *weakly saturated-unsaturated infiltration with a weakly nonlinear hydraulic conductivity* is given by (2.26)-(2.28), where $\widetilde{\beta}^*$ is the multivalued function defined by (2.24), $\widetilde{\beta}$ is given by (2.19) and $\widetilde{K}$ is the single-valued function (2.11) with $\widetilde{K}'(S_w)$ finite on $[S_r, S_s]$. Moreover, $\widetilde{\beta}^*$ is strongly monotone, (2.25), $\widetilde{\beta}$ satisfies (2.20), (2.22)-(2.23) with

$$\lim_{S_w \nearrow S_s} \widetilde{\beta}(S_w) < +\infty \tag{2.36}$$

and K is Lipschitz on $[S_r, S_s]$, i.e., there exists $M > 0$ such that

$$\left|\widetilde{K}(S_w) - \widetilde{K}(\overline{S_w})\right| \leq M \left|S_w - \overline{S_w}\right|, \ \forall S_w, \overline{S_w} \in [S_r, S_s]. \tag{2.37}$$

It is obvious that this situation which is illustrated by nonsingular diffusivities including also power functions

$$\widetilde{\beta}(S_w) = S_w^p, \text{ with } p > 1,$$

is related to a *slow diffusion* and to the well-known *porous media equation* (see [3]).

We write the model in the dimensionless form, introducing for example

$$S_w^{\dim} = \frac{S_w - S_r}{S_s - S_r}, \ \widetilde{K}^{\dim}(S_w^{\dim}) = \frac{\widetilde{K}(S_w) - K_r}{K_s - K_r}, \ \widetilde{\beta}^{\dim}(S_w) = \frac{\widetilde{\beta}(S_w)}{\beta_d},$$

where β_d is a characteristic value for the diffusivity. Without entering into details we specify that the dimensionless model has the same form as (2.26)–(2.28). The dimensionless $S_{wr}^{\dim} = 0$ and $K_r = 0$ and for convenience, we shall extend $\widetilde{\beta}$ and $\widetilde{K}$ at the left of $S_{wr}^{\dim}$ by the constant values $\widetilde{\rho}$ and 0 (for all these details see [34]). For simplicity, further we shall no longer indicate dimensionless by the superscript $^{\dim}$.

3. Analysis of the porosity-degenerate model

In this part we shall approach Model 2 given by (2.26)–(2.27), (2.29)–(2.30) corresponding to the strongly nonlinear saturated-unsaturated case with a weakly nonlinear hydraulic conductivity. We shall study a limit case letting m to vanish on a subset Ω_0 strictly included in Ω, see Fig. 1. This characterizes the existence of possible solid intrusions in the soil and we shall call this model *porosity-degenerate*.

In fact we intend to treat a little more general mathematical problem, in which we shall consider that the function conductivity depends both on the space variables and the solution. Therefore the model reads

$$\frac{\partial(m(x)S_w)}{\partial t} - \Delta\widetilde{\beta}^*(S_w) + \frac{\partial\widetilde{K}(x,S_w)}{\partial x_3} \ni f \text{ in } Q, \tag{3.1}$$

$$m(x)S_w(x,0) = S_{w0}(x) \text{ in } \Omega, \tag{3.2}$$

$$S_w(x,t) = S_s \text{ on } \Sigma_u, \tag{3.3}$$

$$\left(\widetilde{K}(x,S_w)i_3 - \nabla\widetilde{\beta}^*(S_w)\right)\cdot\nu \ni f_\alpha \text{ on } \Sigma_\alpha. \tag{3.4}$$

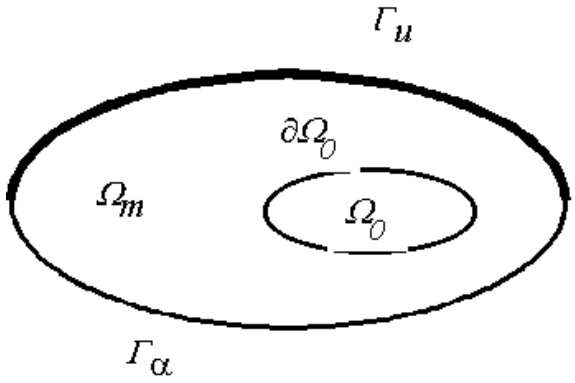


Fig. 1: The domain Ω.

At the points where m vanishes the equation degenerates. The function m is supposed to be essentially bounded, $m \in L^\infty(\Omega)$ with $0 \le m(x) \le 1$ a.e. $x \in \Omega$. However, we shall see that this assumption is not sufficient to get the

solution existence and a stronger hypothesis upon m is required. We specify once again the hypotheses made for the problem parameters, i.e.,

$$\widetilde{\beta}(r) \geq \widetilde{\rho} \text{ for } r < S_s,\ \widetilde{\beta}(r) = \widetilde{\rho} \text{ for } r \leq 0,\ \lim_{r \nearrow S_s} \widetilde{\beta}(r) = +\infty, \tag{3.5}$$

$$\widetilde{\beta}^*(r) = \begin{cases} \int_0^r \widetilde{\beta}(\xi)\mathrm{d}\xi,\ r < S_s \\ [K_s^*, +\infty),\ r = S_s, \end{cases} \tag{3.6}$$

$$\lim_{r \to -\infty} \widetilde{\beta}^*(r) = -\infty,\ \lim_{r \nearrow S_s} \widetilde{\beta}^*(r) = \widetilde{K}_s^* > 0, \tag{3.7}$$

$$(\widetilde{\beta}^*(r) - \widetilde{\beta}^*(\overline{r}))(r - \overline{r}) \geq \widetilde{\rho}(r - \overline{r})^2, \forall r, \overline{r} \in (-\infty, S_s]. \tag{3.8}$$

In what concerns $\widetilde{K}$ we assume that it has the form

$$\widetilde{K}(x, r) = \begin{cases} \widetilde{K}_0(x) \text{ on } \{x;\ m(x) = 0\} \\ \widetilde{K}_m(r) \text{ otherwise}, \end{cases} \tag{3.9}$$

$$\widetilde{K}(x, r) = 0 \text{ for } r \leq 0 \text{ and } \widetilde{K}(x, r) = \widetilde{K}_s \text{ for } r \geq S_s, \tag{3.10}$$

where $\widetilde{K}_s = \widetilde{K}(x, S_s) > 0$.

Moreover, we assume that $\widetilde{K}_0 \in H^1(\Omega_0)$ and $\widetilde{K}$ is Lipschitz with respect to r, uniformly with respect to x, i.e., there exists $M > 0$, such that

(i_K) $\left|\widetilde{K}(x, r) - \widetilde{K}(x, \overline{r})\right| \leq M\,|r - \overline{r}|\,,\ \forall r, \overline{r} \in \mathbf{R},\ \forall x \in \Omega.$

Finally we shall impose that

$$m \in C^1(\overline{\Omega}),\ 0 \leq m(x) \leq 1. \tag{3.11}$$

Functional framework. We perform a function replacement by denoting

$$w = S_w - S_s, \tag{3.12}$$

so that we are led to the system

$$\frac{\partial(m(x)w)}{\partial t} - \Delta\widetilde{\beta}^*(w + S_s) + \frac{\partial \widetilde{K}(x, w + S_s)}{\partial x_3} \ni f \text{ in } Q, \tag{3.13}$$

$$m(x)w(x, 0) = v_0(x) \text{ in } \Omega, \tag{3.14}$$

$$w(x, t) = 0 \text{ on } \Sigma_u, \tag{3.15}$$

$$\left(\widetilde{K}(x, w + S_s)i_3 - \nabla\widetilde{\beta}^*(w + S_s)\right) \cdot \nu \ni f_\alpha \text{ on } \Sigma_\alpha, \tag{3.16}$$

which we are going to study. Here $v_0(x) = S_{w0} - m(x)S_s$. We shall indicate the value of w at saturation by w_s (actually, by (3.12) it is equal to zero, but we shall keep the notation w_s in order to put into evidence the behaviour of the solution at this point).

We consider the spaces $L^2(\Omega)$ with the standard norm denoted $\|\cdot\|$,

$$V = \{w \in H^1(\Omega);\ w = 0 \text{ on } \Gamma_u\}, \tag{3.17}$$

with the norm

$$\|\psi\|_V = \left(\int_\Omega |\nabla\psi|^2 \,\mathrm{d}x\right)^{1/2}, \tag{3.18}$$

and its dual V' on which we introduce the scalar product by

$$(w, \overline{w})_{V'} = \langle w, \psi\rangle_{V',V},$$

where ψ is the solution to the boundary value problem

$$-\Delta\psi = \overline{w},\ \psi = 0 \text{ on } \Gamma_u,\ \nabla\psi\cdot\nu = 0 \text{ on } \Gamma_\alpha. \tag{3.19}$$

Let $f_\alpha \in L^2(0,T;L^2(\Gamma_\alpha))$. We define the functional $f_{\Gamma_\alpha} \in L^2(0,T;V')$ by

$$f_{\Gamma_\alpha}(t)(\psi) = -\int_{\Gamma_\alpha} f_\alpha(t)\psi \mathrm{d}\sigma \text{ for any } \psi \in V \tag{3.20}$$

and notice that

$$\|f_{\Gamma_\alpha}(t)\|_{V'} \le c_{tr}\,\|f_\alpha(t)\|_{L^2(\Gamma_\alpha)}$$

where c_{tr} is the constant provided by the trace theorem.

For the further mathematical developments it is more convenient to work with the multivalued function

$$\beta^*(r) = \widetilde{\beta}^*(r + S_s) - \widetilde{K}_s^*. \tag{3.21}$$

DEFINITION **3.1** *Let*

$$\begin{aligned} m &\in C^1(\overline{\Omega}),\ f \in L^2(0,T;V'),\ f_\alpha \in L^2(0,T;L^2(\Gamma_\alpha)), \\ v_0 &\in L^2(\Omega),\ \frac{v_0}{m} \in L^2(\Omega),\ \frac{v_0}{m} \le w_s,\ \textit{a.e. } x \in \Omega. \end{aligned} \tag{3.22}$$

We call w a solution to (3.13)-(3.16) *if*

$$\begin{aligned} w &\in L^2(0,T;V), \\ \zeta &\in L^2(0,T;V),\ \zeta \in \beta^*(w(x,t)) \textit{ a.e. on } Q, \\ mw &\in C([0,T];L^2(\Omega)) \cap W^{1,2}(0,T;V'), \end{aligned} \tag{3.23}$$

satisfies the equation

$$\left\langle \frac{\mathrm{d}(m(x)w)}{\mathrm{d}t}(t),\psi\right\rangle_{V',V}+\int_\Omega\left(\nabla\zeta(t)\cdot\nabla\psi-\widetilde{K}(x,w(t)+S_s)\frac{\partial\psi}{\partial x_3}\right)\mathrm{d}x= \\ =\langle f(t),\psi\rangle_{V',V}+\langle f_{\Gamma_\alpha}(t),\psi\rangle_{V',V},\ \textit{a.e. } t\in(0,T),\ \forall\psi\in V, \tag{3.24}$$

the initial condition $m(x)w(0)=v_0$ *and the property*

$$w\le w_s,\ \textit{a.e. } (x,t)\in Q. \tag{3.25}$$

Eq. (3.24) can be written also in the equivalent form

$$\begin{aligned}&\int_0^T\left\langle\frac{\mathrm{d}(m(x)w)}{\mathrm{d}t}(t),\phi(t)\right\rangle_{V',V}\mathrm{d}t \\ &+\int_Q\left(\nabla\zeta\cdot\nabla\phi-\widetilde{K}(x,w+S_s)\frac{\partial\phi}{\partial x_3}\right)\mathrm{d}x\mathrm{d}t \\ =&\int_0^T\langle f(t)+f_{\Gamma_\alpha}(t),\phi(t)\rangle_{V',V}\,\mathrm{d}t,\ \forall\phi\in L^2(0,T;V).\end{aligned} \tag{3.26}$$

Replacing S_w from (3.12) we get that S_w satisfies

$$\begin{aligned}S_w &\in L^2(0,T;H^1(\Omega)),\\ \widetilde{\zeta} &\in L^2(0,T;H^1(\Omega)),\ \widetilde{\zeta}\in\widetilde{\beta}^*(S_w(x,t))\text{ a.e. on }Q,\\ mS_w &\in C([0,T];L^2(\Omega))\cap W^{1,2}(0,T;V').\end{aligned}$$

We set

$$D(A)=\{\theta\in L^2(\Omega);\ \exists\eta\in V,\ \eta(x)\in\beta^*(\theta(x))\text{ a.e. }x\in\Omega\}$$

and we introduce the multivalued operator $A:D(A)\subset V'\to V'$ by

$$\langle A\theta,\psi\rangle_{V',V}=\int_\Omega\left(\nabla\eta\cdot\nabla\psi-\widetilde{K}(x,\theta+S_s)\frac{\partial\psi}{\partial x_3}\right)\mathrm{d}x,$$

for any $\psi\in V$, where $\eta\in\beta^*(\theta)$ a.e. $x\in\Omega$. Thus, we can write the problem

$$\begin{aligned}\frac{\mathrm{d}(m(x)w)}{\mathrm{d}t}+Aw &\ni f+f_{\Gamma_\alpha},\ \text{a.e. } t\in(0,T)\\ m(x)w(0) &= v_0.\end{aligned} \tag{3.27}$$

We consider now the multiplication operator

$$M:D(A)\to L^2(\Omega),\ Mw=mw, \tag{3.28}$$

whose inverse is multivalued and denoting

$$v(x,t) = m(x)w(x,t), \tag{3.29}$$

we can rewrite (3.27) in terms of v as

$$\begin{aligned} \frac{\mathrm{d}v}{\mathrm{d}t} + A_M v &\ni f + f_{\Gamma_\alpha}, \text{ a.e. } t \in (0,T) \\ v(0) &= v_0, \end{aligned} \tag{3.30}$$

where $A_M v = AM^{-1}v = A\left(\frac{v}{m}\right)$ for any $v \in D(A_M)$, where

$$D(A_M) = \left\{ v \in L^2(\Omega); \frac{v}{m} \in L^2(\Omega), \exists \eta \in V,\ \eta \in \beta^*\left(\frac{v}{m}\right) \text{ a.e. } x \in \Omega \right\}.$$

We see that $v \in D(A_M)$ implies $\frac{v}{m} \in D(A)$. Conversely, if $w = \frac{v}{m} \in D(A)$, then $v = mw \in D(A_M)$.

We still define $\widetilde{j} : \mathbf{R} \to (-\infty, +\infty]$ by

$$\widetilde{j}(r) = \begin{cases} \int_0^r \widetilde{\beta}^*(\xi)\mathrm{d}\xi,\ r \le S_s \\ +\infty,\ r > S_s, \end{cases}$$

where the left limit of $\widetilde{\beta}^*$ at S_s is specified in (3.7). This function is proper, convex, lower semicontinuous and

$$\partial\widetilde{j}(r) = \begin{cases} \widetilde{\beta}^*(r),\ r < S_s, \\ [\widetilde{K}_s^*, +\infty),\ r = S_s, \\ \varnothing,\ r > S_s. \end{cases} \tag{3.31}$$

(The proof is similar to that done for a slightly different function in [34], Sect. 5.3.)

3.1. Approximating problem

Since the operator A_M is multivalued, in order to prove the existence for (3.27) we introduce an approximating problem replacing m by

$$m_\varepsilon(x) = m(x) + \varepsilon, \text{ for } \varepsilon > 0$$

and $\widetilde{\beta}^*$ by the single-valued continuous function

$$\widetilde{\beta}^*_\varepsilon(r) = \begin{cases} \widetilde{\beta}^*(r),\ r < S_s - \varepsilon \\ \widetilde{\beta}^*(S_s - \varepsilon) + \frac{\widetilde{K}_s^* - \widetilde{\beta}^*(S_s - \varepsilon)}{\varepsilon}\left[r - (S_s - \varepsilon)\right],\ r \ge S_s - \varepsilon. \end{cases}$$

Then we define

$$\beta_\varepsilon^*(r) = \widetilde{\beta}_\varepsilon^*\left(r + S_s\right) - \widetilde{K}_s^* \tag{3.32}$$

and the single valued operator

$$A_\varepsilon : D(A_\varepsilon) \subset V' \to V',$$

$$\langle A_\varepsilon\theta, \psi\rangle_{V',V} = \int_\Omega \left(\nabla\beta_\varepsilon^*(\theta)\cdot\nabla\psi - \widetilde{K}(x,\theta+S_s)\frac{\partial\psi}{\partial x_3}\right)\mathrm{d}x,\ \forall\psi\in V,$$

with

$$D(A_\varepsilon) = \{\theta\in L^2(\Omega); \beta_\varepsilon^*\left(\theta\right)\in V\}.$$

We can write the approximating Cauchy problem (corresponding to (3.27))

$$\begin{aligned} \frac{\mathrm{d}(m_\varepsilon w_\varepsilon)}{\mathrm{d}t} + A_\varepsilon w_\varepsilon &= f + f_{\Gamma_\alpha},\ \text{a.e. } t\in(0,T), \\ m_\varepsilon w_\varepsilon(0) &= v_{0\varepsilon}, \end{aligned} \tag{3.33}$$

where

$$v_{0\varepsilon} = m_\varepsilon\frac{v_0}{m}. \tag{3.34}$$

DEFINITION **3.2** *Let $\varepsilon>0$ and*

$$\begin{aligned} m &\in C^1(\overline{\Omega}),\ f\in L^2(0,T;V'),\ f_\alpha\in L^2(0,T;L^2(\Gamma_\alpha)), \\ v_0 &\in L^2(\Omega),\ \frac{v_0}{m}\in L^2(\Omega),\ \frac{v_0}{m}\le w_s. \end{aligned}$$

A solution to (3.33) *is a function w_ε that satisfies*

$$\begin{aligned} w_\varepsilon &\in C([0,T];L^2(\Omega))\cap L^2(0,T;V)\cap W^{1,2}(0,T;V'), \\ \beta_\varepsilon^*\left(w_\varepsilon\right) &\in L^2(0,T;V), \end{aligned}$$

$$\begin{aligned} &\int_0^T \left\langle \frac{\mathrm{d}(m_\varepsilon w_\varepsilon)}{\mathrm{d}t}(t),\phi(t)\right\rangle_{V',V}\mathrm{d}t \\ &+\int_Q \left\{\nabla\beta_\varepsilon^*\left(w_\varepsilon\right)\cdot\nabla\phi - \widetilde{K}\left(x,w_\varepsilon+S_s\right)\frac{\partial\phi}{\partial x_3}\right\}\mathrm{d}x\,\mathrm{d}t \\ = &\int_0^T \langle f(t)+f_{\Gamma_\alpha}(t),\phi(t)\rangle_{V',V}\,\mathrm{d}t,\ \forall\phi\in L^2(0,T;V), \end{aligned} \tag{3.35}$$

and the initial condition $m_\varepsilon w_\varepsilon(0) = v_{0\varepsilon}$.

Then denoting

$$v_\varepsilon(x,t) = m_\varepsilon(x) w_\varepsilon(x,t), \tag{3.36}$$

we can write problem (3.33) in the equivalent form (corresponding to (3.30))

$$\begin{aligned} \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}t} + B_\varepsilon v_\varepsilon &= f, \text{ a.e. } t \in (0,T), \\ v_\varepsilon(0) &= v_{0\varepsilon}. \end{aligned} \tag{3.37}$$

The operator $B_\varepsilon : D(B_\varepsilon) \subset V' \to V'$ is single-valued, has the domain

$$D(B_\varepsilon) = \left\{ \theta \in L^2(\Omega);\ \beta_\varepsilon^* \left(\frac{\theta}{m_\varepsilon} \right) \in V \right\}$$

and is given by

$$\langle B_\varepsilon \theta, \psi \rangle_{V',V} = \int_\Omega \left(\nabla \beta_\varepsilon^* \left(\frac{\theta}{m_\varepsilon} \right) \cdot \nabla \psi - \widetilde{K} \left(x, \frac{\theta}{m_\varepsilon} + S_s \right) \frac{\partial \psi}{\partial x_3} \right) \mathrm{d}x,\ \forall \psi \in V.$$

Then (3.37) can be still written

$$\begin{aligned} &\int_0^T \left\langle \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \\ &+ \int_Q \left\{ \nabla \beta_\varepsilon^* \left(\frac{v_\varepsilon}{m_\varepsilon} \right) \cdot \nabla \phi - \widetilde{K} \left(x, \frac{v_\varepsilon}{m_\varepsilon} + S_s \right) \frac{\partial \phi}{\partial x_3} \right\} \mathrm{d}x \mathrm{d}t = \\ = &\int_0^T \langle f(t) + f_{\Gamma_\alpha}(t), \phi(t) \rangle_{V',V} \, \mathrm{d}t,\ \forall \phi \in L^2(0,T;V), \end{aligned} \tag{3.38}$$

which is in fact (3.35).

For a later use we define $\widetilde{j}_\varepsilon : \mathbf{R} \to \mathbf{R}$,

$$\widetilde{j}_\varepsilon(r) = \int_0^r \widetilde{\beta}_\varepsilon^*(\xi) \mathrm{d}\xi,$$

and notice that

$$\partial \widetilde{j}_\varepsilon(r) = \widetilde{\beta}_\varepsilon^*(r),\ \forall r \in \mathbf{R}. \tag{3.39}$$

First we shall prove that (3.37) has, for each $\varepsilon > 0$, a unique solution, v_ε in appropriate functional spaces.

3.2. Existence for the approximating problem

PROPOSITION **3.1** *Let*

$$\begin{aligned} m &\in C^1(\overline{\Omega}),\ 0 \le m \le 1, \\ f &\in L^2(0,T;V'),\ f_\alpha \in L^2(0,T;L^2(\Gamma_\alpha)), \\ v_0 &\in L^2(\Omega),\ \frac{v_0}{m} \in L^2(\Omega),\ \frac{v_0}{m} \le w_s \text{ a.e. on } \Omega. \end{aligned}$$

Then, the Cauchy problem (3.37) *has, for each* $\varepsilon > 0$, *a unique solution*

$$v_\varepsilon \in C([0,T];L^2(0,T)) \cap W^{1,2}(0,T;V') \cap L^2(0,T;V) \tag{3.40}$$

$$\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right) \in L^2(0,T;V), \tag{3.41}$$

$$\widetilde{j_\varepsilon}\left(\frac{v_\varepsilon}{m_\varepsilon}\right) \in L^\infty(0,T;L^1(\Omega)), \tag{3.42}$$

that satisfies the estimates

$$\begin{aligned} &\int_\Omega m_\varepsilon(x)\widetilde{j_\varepsilon}\left(\frac{v_\varepsilon}{m_\varepsilon}(x,t)+S_s\right)\mathrm{d}x + \int_0^t \left\|\frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau)\right\|_{V'}^2 \mathrm{d}\tau + \\ &+ \int_0^t \left\|\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right)\right\|_V^2 \mathrm{d}\tau \le \qquad (3.43) \\ \le\ &\beta_0\left(\int_0^T \|f(t)\|_{V'}^2\,\mathrm{d}t + \int_0^T \|f_\alpha(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t + 1\right), \end{aligned}$$

$$\left\|\sqrt{m_\varepsilon}\left(\frac{v_\varepsilon}{m_\varepsilon}(t)\right)\right\| \le c_0,\ \forall t \in [0,T], \tag{3.44}$$

$$\|v_\varepsilon(t)\| \le c_1,\ \forall t \in [0,T], \tag{3.45}$$

where β_0, c_0 *and* c_1 *do not depend on* ε.

Moreover, if v_ε *and* $\overline{v_\varepsilon}$ *are two solutions corresponding to the pairs of data* f, f_{Γ_α}, v_0 *and* $\overline{f}$, $\overline{f_{\Gamma_\alpha}}$, $\overline{v_0}$, *we have the estimate*

$$\begin{aligned} &\|v_\varepsilon(t) - \overline{v_\varepsilon}(t)\|_{V'}^2 + \int_0^t \|v_\varepsilon(\tau) - \overline{v_\varepsilon}(\tau)\|^2\,\mathrm{d}\tau \le \qquad (3.46) \\ \le\ &\alpha_0(\varepsilon)\left(\|v_0 - \overline{v_0}\|_{V'}^2 + \right. \\ &\left. + \int_0^T \|f(t) - \overline{f}(t)\|_{V'}^2\,\mathrm{d}t + \int_0^T \|f_\alpha(t) - \overline{f_\alpha}(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t\right). \end{aligned}$$

Proof. The proof is based on the quasi m-accretivity of the operator B_ε which is proved below. To show the quasi monotony we compute

$$\begin{aligned}&\left((\lambda I+B_\varepsilon)\theta-(\lambda I+B_\varepsilon)\overline{\theta},\theta-\overline{\theta}\right)_{V'}=\lambda\left\|\theta-\overline{\theta}\right\|_{V'}^2+\\&+\int_\Omega\nabla\left(\beta_\varepsilon^*\left(\frac{\theta}{m_\varepsilon}\right)-\beta_\varepsilon^*\left(\frac{\overline{\theta}}{m_\varepsilon}\right)\right)\cdot\nabla\psi\mathrm{d}x-\\&-\int_\Omega\left(\widetilde{K}(x,\frac{\theta}{m_\varepsilon}+S_s)-\widetilde{K}(x,\frac{\overline{\theta}}{m_\varepsilon}+S_s)\right)\frac{\partial\psi}{\partial x_3}\mathrm{d}x,\end{aligned}$$

where $-\Delta\psi=\theta-\overline{\theta}$, $\nabla\psi\cdot\nu=0$ on Γ_α and $\psi=0$ on Γ_u. Hence

$$\begin{aligned}&\left((\lambda I+B_\varepsilon)\theta-(\lambda I+B_\varepsilon)\overline{\theta},\theta-\overline{\theta}\right)_{V'}\geq\\\geq\;&\lambda\left\|\theta-\overline{\theta}\right\|_{V'}^2+\widetilde{\rho}\left\|\frac{\theta-\overline{\theta}}{\sqrt{m_\varepsilon}}\right\|^2-M\left\|\frac{\theta-\overline{\theta}}{m_\varepsilon}\right\|\left\|\theta-\overline{\theta}\right\|_{V'}\geq\\\geq\;&\left(\lambda-\frac{M^2}{2\widetilde{\rho}\varepsilon}\right)\left\|\theta-\overline{\theta}\right\|_{V'}^2+\frac{\widetilde{\rho}}{2}\left\|\frac{\theta-\overline{\theta}}{\sqrt{m_\varepsilon}}\right\|^2>0\end{aligned}$$

for $\lambda>\frac{M^2}{2\widetilde{\rho}\varepsilon}$. Here we used the fact that $\varepsilon\leq m_\varepsilon(x)\leq 1+\varepsilon$.

Next we have to prove that

$$R(I+B_\varepsilon)=V',$$

i.e., to show that the equation

$$v_\varepsilon+B_\varepsilon v_\varepsilon=g \tag{3.47}$$

has a solution $v_\varepsilon\in D(B_\varepsilon)$ for any $g\in V'$. Recall that ε is fixed.

If we denote $\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right)=\zeta\in V$, due to the fact that β_ε^* is continuous and monotonically increasing on $\mathbf{R}$ and $R(\beta_\varepsilon^*)=(-\infty,\infty)$ it follows that its inverse

$$G_\varepsilon(\zeta)=m_\varepsilon(\beta_\varepsilon^*)^{-1}(\zeta) \tag{3.48}$$

is continuous from V to $L^2(\Omega)$ because

$$\begin{aligned}&\left\|G_\varepsilon(\zeta)-G_\varepsilon(\overline{\zeta})\right\|= \qquad (3.49)\\=\;&\left\|m_\varepsilon\left((\beta_\varepsilon^*)^{-1}(\zeta)-(\beta_\varepsilon^*)^{-1}(\overline{\zeta})\right)\right\|\leq\\\leq\;&\frac{1+\varepsilon}{\widetilde{\rho}}\left\|\zeta-\overline{\zeta}\right\|\leq\frac{(1+\varepsilon)c_\Omega}{\widetilde{\rho}}\left\|\zeta-\overline{\zeta}\right\|_V,\ \forall\zeta,\overline{\zeta}\in V.\end{aligned}$$

Here we used (3.8) and Poincaré's inequality (with the constant c_Ω). So, (3.47) can be rewritten as

$$G_\varepsilon(\zeta)+B_0^\varepsilon\zeta=g \tag{3.50}$$

with $B_0^\varepsilon : V \to V'$ defined by

$$\langle B_0^\varepsilon \zeta, \psi \rangle_{V',V} = \int_\Omega \left(\nabla\zeta \cdot \nabla\psi - \widetilde{K}\left(x, \frac{G_\varepsilon(\zeta)}{m_\varepsilon} + S_s \right) \frac{\partial \psi}{\partial x_3} \right) \mathrm{d}x, \ \forall \psi \in V. \tag{3.51}$$

The operator $G_\varepsilon + B_0^\varepsilon$ is monotone, continuous and coercive for $\lambda > \frac{M^2}{2\widetilde{\rho}\varepsilon}$, hence it is surjective. Therefore (3.50) has a solution $\zeta \in V$, implying that (3.47) has a solution $v_\varepsilon \in D(B_\varepsilon)$.

a) Now we assume that $f \in W^{1,1}(0,T;V')$, $f_\alpha \in W^{1,1}(0,T;L^2(\Omega))$ and $\frac{v_0}{m} \in V$ which is equivalent to $v_{0\varepsilon} \in D(B_\varepsilon)$.

Therefore, the existence of a unique solution to (3.37)

$$v_\varepsilon \in W^{1,\infty}(0,T;V') \cap L^\infty(0,T;D(B_\varepsilon))$$

follows from the general theorems for evolution equations with m-accretive operators, hence $\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right) \in L^\infty(0,T;V)$. Since the inverse of β_ε^* is Lipschitz we deduce that $\frac{v_\varepsilon}{m_\varepsilon} \in L^\infty(0,T;V)$.

It follows that (3.33) has a solution

$$w_\varepsilon = \frac{v_\varepsilon}{m_\varepsilon}$$

in the same spaces.

To prove estimate (3.43) we test (3.37) at $\beta_\varepsilon^*(v_\varepsilon)$ and integrate over $(0,t)$. Taking into account (3.36) and (3.32) we have

$$\begin{aligned}
& \int_0^t \left\langle \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau), \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\rangle_{V',V} \mathrm{d}\tau + \int_0^t \left\| \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\|_V^2 \mathrm{d}\tau \\
\leq\ & \int_0^t \left\| \widetilde{K}\left(\cdot, \frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\| \left\| \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\|_V \mathrm{d}\tau \\
& + \int_0^t \|f(\tau)\|_{V'} \left\| \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\|_V \mathrm{d}\tau + \int_0^t \|f_{\Gamma_\alpha}(\tau)\|_{V'} \left\| \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right) \right\|_V \mathrm{d}\tau \\
\leq\ & \frac{1}{2} \int_0^t \left\| \beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)(\tau)\right) \right\|_V^2 \mathrm{d}\tau + C_0,
\end{aligned}$$

where we have used the boundedness of $\widetilde{K}$ and

$$C_0 = \frac{3}{2}\left\{ \widetilde{K}_s^2 T \mathrm{meas}(\Omega) + \int_0^T \|f(\tau)\|_{V'}^2 \mathrm{d}\tau + c_{tr}^2 \int_0^T \|f_\alpha(\tau)\|_{L^2(\Gamma_\alpha)}^2 \mathrm{d}\tau \right\}.$$

Next, we take into account that

$$\begin{aligned}
&\int_0^t \left\langle \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau), \beta_\varepsilon^* \left(\frac{v_\varepsilon}{m_\varepsilon}(\tau) \right) \right\rangle_{V',V} \mathrm{d}\tau \\
= &\int_0^t \left\langle \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau), \widetilde{\beta}_\varepsilon^* \left(\frac{v_\varepsilon}{m_\varepsilon}(\tau) + S_s \right) - \widetilde{K}_s^* \right\rangle_{V',V} \mathrm{d}\tau \\
= &\int_\Omega m_\varepsilon(x) \widetilde{j}_\varepsilon \left(\frac{v_\varepsilon(x,t)}{m_\varepsilon} + S_s \right) \mathrm{d}x - \int_\Omega m_\varepsilon(x) \widetilde{j}_\varepsilon \left(\frac{v_0}{m}(x) + S_s \right) \mathrm{d}x \\
&- \int_\Omega \widetilde{K}_s^* v_\varepsilon(x,t) \mathrm{d}x + \int_\Omega \widetilde{K}_s^* v_{0\varepsilon} \mathrm{d}x
\end{aligned}$$

and obtain that

$$\begin{aligned}
&\int_\Omega m_\varepsilon(x) \widetilde{j}_\varepsilon \left(\frac{v_\varepsilon(x,t)}{m_\varepsilon} + S_s \right) \mathrm{d}x + \frac{1}{2} \int_0^t \left\| \beta_\varepsilon^* \left(\frac{v_\varepsilon}{m_\varepsilon}(\tau) \right) \right\|_V^2 \mathrm{d}\tau \leq \\
\leq &\int_\Omega m_\varepsilon(x) \widetilde{j}_\varepsilon \left(\frac{v_0}{m}(x) + S_s \right) \mathrm{d}x + \int_\Omega \widetilde{K}_s^* v_\varepsilon(t) \mathrm{d}x + C_1, \qquad (3.52)
\end{aligned}$$

where

$$C_1 = \frac{1}{2} \widetilde{K}_s^{*2} \mathrm{meas}(\Omega) + \frac{1}{2} \left\| \frac{v_0}{m} \right\|^2 + C_0.$$

Since

$$\widetilde{j}_\varepsilon(r) \geq \frac{\widetilde{\rho}}{2} r^2, \ \forall r \in \mathbf{R},$$

we have

$$\begin{aligned}
&\int_\Omega m_\varepsilon(x) \widetilde{j}_\varepsilon \left(\frac{v_\varepsilon(x,t)}{m_\varepsilon} + S_s \right) \mathrm{d}x \geq \\
\geq &\frac{\widetilde{\rho}}{2} \int_\Omega m_\varepsilon(x) \left(\frac{v_\varepsilon(x,t)}{m_\varepsilon} + S_s \right)^2 \mathrm{d}x \geq \frac{\widetilde{\rho}}{2} \int_\Omega m_\varepsilon \left\{ \frac{1}{2} \left(\frac{v_\varepsilon(x,t)}{m_\varepsilon} \right)^2 - S_s^2 \right\} \mathrm{d}x.
\end{aligned}$$

On the other hand we recall that $\frac{v_0}{m} \leq w_s = 0$ and notice that

$$\begin{aligned}
&\widetilde{j}_\varepsilon \left(\frac{v_{0\varepsilon}}{m_\varepsilon} + S_s \right) = \int_0^{\frac{v_0}{m} + S_s} \widetilde{\beta}_\varepsilon^*(r) \mathrm{d}r \leq \int_0^{S_s} \widetilde{\beta}_\varepsilon^*(r) \mathrm{d}r = \\
= &\lim_{\delta \nearrow 0} \int_0^{S_s - \delta} \widetilde{\beta}_\varepsilon^*(r) \mathrm{d}r = \lim_{\delta \nearrow 0} \int_0^{S_s - \delta} \widetilde{\beta}^*(r) \mathrm{d}r \leq \widetilde{K}_s^* S_s.
\end{aligned}$$

Thus we obtain by (3.52) that

$$\frac{\widetilde{\rho}}{4}\int_\Omega m_\varepsilon(x)\left(\frac{v_\varepsilon(x,t)}{m_\varepsilon}\right)^2\mathrm{d}x+\int_0^t\left\|\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right)\right\|_V^2\mathrm{d}\tau\leq \tag{3.53}$$

$$\leq\ 2\widetilde{K}_s^*S_s\mathrm{meas}(\Omega)+\int_\Omega\widetilde{K}_s^*m_\varepsilon\left(\frac{v_\varepsilon}{m_\varepsilon}(t)\right)\mathrm{d}x+C_1+\frac{\widetilde{\rho}}{2}S_s^2\int_\Omega m_\varepsilon(x)\mathrm{d}x\leq$$

$$\leq\ C_2+\frac{\widetilde{\rho}}{8}\int_\Omega m_\varepsilon(x)\left(\frac{v_\varepsilon(x,t)}{m_\varepsilon}\right)^2\mathrm{d}x+\frac{4}{\widetilde{\rho}}\widetilde{K}_s^{*2}\mathrm{meas}(\Omega).$$

We have used several times that $m_\varepsilon\leq 1+\varepsilon\leq 2$. We can conclude that

$$\left\|\sqrt{m_\varepsilon}\frac{v_\varepsilon}{m_\varepsilon}(t)\right\|\leq c_0,\ \forall t\in[0,T]. \tag{3.54}$$

Next, from the relation

$$v_\varepsilon(t)=\sqrt{m_\varepsilon}\frac{v_\varepsilon}{m_\varepsilon}(t)\sqrt{m_\varepsilon} \tag{3.55}$$

we get that

$$\|v_\varepsilon(t)\|^2=\int_\Omega\left(\sqrt{m_\varepsilon(x)}\frac{v_\varepsilon(t)}{m_\varepsilon}\right)^2m_\varepsilon(x)\mathrm{d}x\leq 2\left\|\sqrt{m_\varepsilon}\frac{v_\varepsilon}{m_\varepsilon}(t)\right\|^2$$

and therefore

$$\|v_\varepsilon(t)\|\leq c_1,\ \forall t\in[0,T] \tag{3.56}$$

where c_0, c_1, C_0, C_1, C_2 are independent of ε. Replacing this in (3.52) we deduce

$$\int_\Omega m_\varepsilon(x)\widetilde{j}_\varepsilon\left(\frac{v_\varepsilon(x,t)}{m_\varepsilon}+S_s\right)\mathrm{d}x+\int_0^t\left\|\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right)\right\|_V^2\mathrm{d}\tau\leq \tag{3.57}$$

$$\leq\ C_2\left(\int_0^T\|f(t)\|_{V'}^2\,\mathrm{d}t+\int_0^T\|f_\alpha(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t+1\right). \tag{3.58}$$

Then we multiply (3.37) scalarly in V' by $\frac{\mathrm{d}v_\varepsilon}{\mathrm{d}t}(t)$, integrate over $(0,t)$ and proceeding as before we get

$$\int_\Omega m_\varepsilon(x)\widetilde{j}_\varepsilon\left(\frac{v_\varepsilon(x,t)}{m_\varepsilon}+S_s\right)\mathrm{d}x+\int_0^t\left\|\frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau)\right\|_{V'}^2\mathrm{d}\tau\leq \tag{3.59}$$

$$\leq\ C_2\left(\int_0^T\|f(t)\|_{V'}^2\,\mathrm{d}t+\int_0^T\|f_\alpha(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t+1\right).$$

Adding this relation with (3.58) we obtain

$$\begin{aligned}
&\int_\Omega m_\varepsilon(x)\widetilde{j_\varepsilon}\left(\frac{v_\varepsilon}{m_\varepsilon}(x,t)+S_s\right)\mathrm{d}x+\int_0^t\left\|\frac{\mathrm{d}v_\varepsilon}{\mathrm{d}\tau}(\tau)\right\|_{V'}^2\mathrm{d}\tau+ \qquad (3.60)\\
&+\int_0^t\left\|\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}(\tau)\right)\right\|_V^2\mathrm{d}\tau\le\\
\le\ &\beta_0\left(\int_0^T\|f(t)\|_{V'}^2\,\mathrm{d}t+\int_0^T\|f_\alpha(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t+1\right),
\end{aligned}$$

with β_0 independent of ε.

To show the estimate (3.46) we write two equations (3.37) corresponding to different pairs of data, subtract them, multiply the difference scalarly in V' by $v_\varepsilon-\overline{v_\varepsilon}$ and integrate over $(0,t)$. We get

$$\begin{aligned}
&\frac12\|v_\varepsilon(t)-\overline{v_\varepsilon}(t)\|_{V'}^2+\frac{\widetilde{\rho}}{2}\int_0^t\int_\Omega\frac{1}{m_\varepsilon}(v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau))^2\mathrm{d}\tau\mathrm{d}x\le\\
\le\ &\frac12\|v_0-\overline{v_0}\|_{V'}^2+\frac{M^2}{2\widetilde{\rho}\varepsilon}\int_0^t\|v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau)\|_{V'}^2\,\mathrm{d}\tau+\\
&+\int_0^t\left\|f(\tau)-\overline{f}(\tau)\right\|_{V'}^2\|v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau)\|_{V'}\,\mathrm{d}\tau+\\
&+c_{tr}^2\int_0^t\left\|f_\alpha(\tau)-\overline{f_\alpha}(\tau)\right\|_{L^2(\Gamma_\alpha)}^2\|v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau)\|_{V'}\,\mathrm{d}\tau
\end{aligned}$$

and moreover

$$\begin{aligned}
&\|v_\varepsilon(t)-\overline{v_\varepsilon}(t)\|_{V'}^2+\widetilde{\rho}\int_0^t\int_\Omega\frac{(v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau))^2}{m_\varepsilon}\mathrm{d}\tau\mathrm{d}x\le\\
\le\ &\|v_0-\overline{v_0}\|_{V'}^2+\left(\frac{M^2}{\widetilde{\rho}\varepsilon}+2\right)\int_0^t\|v_\varepsilon(\tau)-\overline{v_\varepsilon}(\tau)\|_{V'}^2\,\mathrm{d}\tau+\\
&+\int_0^T\left\|f(\tau)-\overline{f}(\tau)\right\|_{V'}^2\mathrm{d}\tau+c_{tr}^2\int_0^T\left\|f_\alpha(\tau)-\overline{f_\alpha}(\tau)\right\|_{L^2(\Gamma_\alpha)}^2\mathrm{d}\tau.
\end{aligned}$$

We obtain the estimate (3.46), via Gronwall lemma with α_0 depending on ε.

b) Now, we assume that $f\in L^2(0,T;V')$ and $\frac{v_0}{m}\in L^2(\Omega)$, $\frac{v_0}{m}\le w_s$.

Due to some obvious densities we can take $\{f_n\}_{n\ge1}\subset W^{1,1}(0,T;V')$, $\{f_\alpha^n\}_{n\ge1}\subset W^{1,1}(0,T;L^2(\Gamma_\alpha))$ and $\{v_0^n\}_{n\ge1}\subset D(B_\varepsilon)=V$, such that

$$\begin{aligned}
f_n &\to f \text{ strongly in } L^2(0,T;V'), \qquad (3.61)\\
f_\alpha^n &\to f_\alpha \text{ strongly in } L^2(0,T;L^2(\Gamma_\alpha))\\
v_0^n &\to v_0 \text{ strongly in } L^2(\Omega).
\end{aligned}$$

Then, for each $\varepsilon > 0$, the problem

$$\begin{aligned} \frac{\mathrm{d}v_\varepsilon^n}{\mathrm{d}t} + B_\varepsilon v_\varepsilon^n &= f_n + f_{\Gamma_\alpha}^n, \text{ a.e. } t \in (0,T), \\ v_\varepsilon^n(0) &= v_{0\varepsilon}^n \end{aligned} \tag{3.62}$$

has a unique solution v_ε^n according to a), satisfying the estimate (3.60) with the right-hand side independent of n, namely,

$$\begin{aligned} &\int_\Omega m_\varepsilon(x) j_\varepsilon\left(\frac{v_\varepsilon^n}{m_\varepsilon}(t) + S_s\right)\mathrm{d}x + \int_0^t \left\|\frac{\mathrm{d}v_\varepsilon^n}{\mathrm{d}\tau}(\tau)\right\|_{V'}^2 \mathrm{d}\tau + \\ &+ \int_0^t \left\|\beta_\varepsilon^*\left(\frac{v_\varepsilon^n}{m_\varepsilon}(\tau)\right)\right\|_V^2 \mathrm{d}\tau \le \\ \le\ &\beta_0\left(\int_0^T \|f_n(t)\|_{V'}^2\,\mathrm{d}t + \int_0^T \|f_\alpha^n(t)\|_{L^2(\Gamma_\alpha)}^2\,\mathrm{d}t + 1\right). \end{aligned} \tag{3.63}$$

We stress that ε is fixed and the second term in the previous relation is uniformly bounded due to (3.61). By this estimate we deduce that $\left\{\beta_\varepsilon^*\left(\frac{v_\varepsilon^n}{m_\varepsilon}\right)\right\}_n$ is in a bounded subset of $L^2(0,T;V)$ and $\left\{\frac{\mathrm{d}v_\varepsilon^n}{\mathrm{d}t}\right\}_n$ is in a bounded subset of $L^2(0,T;V')$, so we can select a subsequence such that

$$\beta_\varepsilon^*\left(\frac{v_\varepsilon^n}{m_\varepsilon}\right) \to \zeta_\varepsilon \text{ weakly in } L^2(0,T;V) \text{ as } n \to \infty,$$

and

$$\frac{\mathrm{d}v_\varepsilon^n}{\mathrm{d}t} \to \frac{\mathrm{d}v_\varepsilon}{\mathrm{d}t} \text{ weakly in } L^2(0,T;V') \text{ as } n \to \infty.$$

We get immediately that

$$\frac{v_\varepsilon^n}{m_\varepsilon} \to w_\varepsilon \text{ weakly in } L^2(0,T;V) \text{ as } n \to \infty.$$

But $m_\varepsilon \in C^1(\overline{\Omega})$ and so the sequence $\{v_\varepsilon\}_n = \left\{m_\varepsilon \frac{v_\varepsilon^n}{m_\varepsilon}\right\}_n$ is bounded in $L^2(0,T;V)$, whence

$$v_\varepsilon^n \to v_\varepsilon \text{ weakly in } L^2(0,T;V) \text{ as } n \to \infty.$$

Since V is compact in $L^2(\Omega)$ it follows by Lions-Aubin's theorem that

$$v_\varepsilon^n \to v_\varepsilon \text{ strongly in } L^2(0,T;L^2(\Omega)) \text{ as } n \to \infty. \tag{3.64}$$

By (3.37) we have that $\{B_\varepsilon v_\varepsilon^n\}_n$ is bounded in $L^2(0,T;V')$ so that

$$B_\varepsilon v_\varepsilon^n \to \chi \text{ weakly in } L^2(0,T;V') \text{ as } n \to \infty. \tag{3.65}$$

But B_ε is quasi m-accretive so its realization on $L^2(0,T;V')$ is quasi m-accretive too, hence it is demiclosed and by (3.64) and (3.65) we get that $\chi = Bv_\varepsilon$ a.e. on Q.

Now we can pass to the limit in (3.62) as $n \to \infty$ and get (3.37), proving thus that this problem has the solution $v_\varepsilon \in C([0,T], L^2(\Omega)) \cap W^{1,2}(0,T;V') \cap L^2(0,T;V)$.

Finally, passing to the limit in (3.63), as $n \to \infty$, and using the lower semicontinuity property we get (3.43) as claimed. Estimates (3.44)–(3.45) have been proved in (3.54)–(3.55).

The uniqueness of the approximating solution follows by (3.46). ∎

3.3. Existence for the original problem

As we specified before the domains

$$\Omega_m = \{x \in \Omega;\ m(x) > 0\} \text{ and } \Omega_0 = \text{int}\{x \in \Omega;\ m(x) = 0\}$$

have the common C^1-boundary, $\partial\Omega_0$, see again Fig. 1. Here, the notation "int" represents the interior of the subset.

THEOREM **3.1** *Let*

$$\begin{aligned} m &\in C^1(\overline{\Omega}),\ 0 \le m \le 1,\ f \in L^2(0,T;V'),\ f_\alpha \in L^2(0,T;L^2(\Gamma_\alpha)), \\ v_0 &\in L^2(\Omega),\ \frac{v_0}{m} \in L^2(\Omega),\ \frac{v_0}{m} \le w_s \text{ a.e. on } \Omega. \end{aligned}$$

Then, the Cauchy problem (3.27) *has a solution*

$$w \in L^2(0,T;V), \tag{3.66}$$

such that

$$\zeta \in L^2(0,T;V),\ \zeta \in \beta^*(w(x,t)) \text{ a.e. on } Q, \tag{3.67}$$

$$mw \in C([0,T];L^2(\Omega)) \cap W^{1,2}(0,T;V'), \tag{3.68}$$

$$w \le w_s \text{ a.e. } (x,t) \in Q. \tag{3.69}$$

Proof. By the hypotheses it follows that the approximating problem (3.37) (and consequently (3.33)) has, for each ε, a unique solution according to

Proposition 3.1, including the estimates (3.43)–(3.45). These do not depend on ε and imply that we can select a subsequence such that

$$\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right) \to \zeta \text{ weakly in } L^2(0,T;V), \tag{3.70}$$

$$\widetilde{\beta}_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}+S_s\right) \to \zeta+\widetilde{K}_s^* \text{ weakly in } L^2(0,T;H^1(\Omega)), \tag{3.71}$$

$$\frac{\mathrm{d}v_\varepsilon}{\mathrm{d}t} \to \mu \text{ weakly in } L^2(0,T;V'), \tag{3.72}$$

$$w_\varepsilon=\frac{v_\varepsilon}{m_\varepsilon} \to w \text{ weakly in } L^2(0,T;V). \tag{3.73}$$

We also get that the trace of $\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right)$ on Σ_u is well defined and since $\beta_\varepsilon^*\left(\frac{v_\varepsilon}{m_\varepsilon}\right) \in L^2(0,T;V)$ it follows that $\zeta=0$ on Σ_u. Now

$$v_\varepsilon=m_\varepsilon\frac{v_\varepsilon}{m_\varepsilon} \tag{3.74}$$

and since $m_\varepsilon \to m$ uniformly on Ω and $m \in C(\overline{\Omega})$ it follows that

$$v_\varepsilon \to v \text{ weakly in } L^2(0,T;L^2(\Omega)). \tag{3.75}$$

By (3.73) and (3.75) we get

$$v=mw \tag{3.76}$$

and obviously

$$v=0, \text{ a.e. on } Q_0=\Omega_0\times(0,T). \tag{3.77}$$

Using (3.73), (3.74) and (3.75) we still obtain that

$$\sqrt{m_\varepsilon}\frac{v_\varepsilon}{m_\varepsilon} \to \sqrt{m}w \text{ weak-star in } L^\infty(0,T;L^2(\Omega)),$$

$$v_\varepsilon=\sqrt{m_\varepsilon}\frac{v_\varepsilon}{m_\varepsilon}\sqrt{m_\varepsilon} \to v \text{ weak-star in } L^\infty(0,T;L^2(\Omega)).$$

Again by (3.74) and $m \in C^1(\overline{\Omega})$ we deduce that

$$\|v_\varepsilon\|_{L^2(0,T;V)} \leq \text{ constant independent of } \varepsilon. \tag{3.78}$$

By Lions-Aubin compactness theorem we conclude then that $\{v_\varepsilon\}_\varepsilon$ is compact in $L^2(0,T;L^2(\Omega))$, i.e.,

$$v_\varepsilon \to v \text{ strongly in } L^2(0,T;L^2(\Omega)) \text{ as } \varepsilon \to 0, \tag{3.79}$$

and $\mu = \frac{\mathrm{d}v}{\mathrm{d}t}$. Also, by Ascoli-Arzelà theorem we can prove that $v_\varepsilon(t) \to v(t)$ strongly in V' (using (3.72) and (3.78)). Using (3.76) we can deduce by letting $\varepsilon \to 0$ in the second equation in (3.37) that

$$mw(0) = v_0. \tag{3.80}$$

We set now

$$\Omega_\delta = \{x \in \Omega;\ m(x) > \delta\} \text{ for arbitrary } \delta > 0,$$

$$Q_\delta = \Omega_\delta \times (0,T),\ Q_m = \Omega_m \times (0,T),$$

and notice that Ω_δ and Ω_m are open because $m \in C^1(\overline{\Omega})$. We have

$$\frac{1}{m_\varepsilon} = \frac{1}{m+\varepsilon} < \frac{1}{m} < \frac{1}{\delta} \text{ on } \Omega_\delta$$

and by (3.79)

$$w_\varepsilon = \frac{1}{m_\varepsilon} v_\varepsilon \to \frac{v}{m} = w \text{ strongly in } L^2(0,T;L^2(\Omega_\delta)),\ \forall \delta > 0.$$

Recall that $\beta_\varepsilon^*(r) = \widetilde{\beta}_\varepsilon^*(r + S_s) - \widetilde{K}_s^*$.

Let us fix $(x,t) \in Q_\delta$. Using the same argument like in the proof of Theorem 3.1, in Sect. 5.3 in [34], we obtain that

$$\widetilde{\beta}_\varepsilon^*(w_\varepsilon + S_s) \to \widetilde{\zeta} \in \widetilde{\beta}^*(w + S_s) \text{ weakly in } L^2(0,T;H^1(\Omega_\delta)).$$

By (3.32) and (3.71) we get that

$$\beta_\varepsilon^*(w_\varepsilon + S_s) \to \widetilde{\beta}^*(w + S_s) - \widetilde{K}_s^* \text{ weakly in } L^2(0,T;H^1(\Omega_\delta)).$$

Since δ is arbitrary we obtain

$$\zeta(x,t) \in \widetilde{\beta}^*(w(x,t) + S_s) - \widetilde{K}_s^* \text{ a.e. } (x,t) \in Q_m = \bigcup_{\delta > 0} Q_\delta. \tag{3.81}$$

Proving that the subset

$$Q_m^+ = \{(x,t) \in Q_m;\ w(x,t) > w_s\}$$

has a zero measure, we deduce similarly to the proof of Corollary 3.3 in Sect. 5.3 in [34], that $w \le w_s$ a.e. $(x,t) \in Q_m$.

Finally, since $\left\{\widetilde{K}(x, w_\varepsilon + S_s)\right\}_\varepsilon$ is bounded in $L^2(Q)$, we have

$$\widetilde{K}(x, w_\varepsilon + S_s) \to \kappa \text{ weakly in } L^2(0,T;L^2(\Omega)), \tag{3.82}$$

and we assert that

$$\kappa(x,t) = \widetilde{K}(x, w(x,t)), \text{ a.e. } (x,t) \in Q.$$

Indeed, $\left\{\widetilde{K}_m(w_\varepsilon + S_s)\right\}_\varepsilon$ is weakly convergent to κ, on Q_m, too. On the other hand, it is strongly convergent to $\widetilde{K}_m(w+S_s)$ on each Q_δ, because $\widetilde{K}_m$ is Lipschitz. By the uniqueness of the limit the restriction of the weak limit to Q_δ should coincide with $\widetilde{K}_m(w + S_s)$. This implies that

$$\kappa = \widetilde{K}(x, w + S_s), \text{ a.e. on } Q_m. \tag{3.83}$$

On the subset Q_0 the function $\widetilde{K}$ does not depend on w, so the limit is equal to $\widetilde{K_0}(x)$.

Now we can pass to limit as $\varepsilon \to 0$ in (3.38) and obtain

$$\begin{aligned}&\int_0^T \left\langle \frac{\mathrm{d}(mw)}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \int_Q \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}(x, w + S_s)\frac{\partial\phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t = \\ &= \int_0^T \langle f(t) + f_{\Gamma_\alpha}(t), \phi(t)\rangle_{V',V}\, \mathrm{d}t, \ \forall \phi \in L^2(0,T;V),\end{aligned} \tag{3.84}$$

where ζ is given by (3.70).

In (3.84) taking $\phi \in L^2(0,T;H_0^1(\Omega_m))$ we still deduce that w is the solution to (3.27) on Q_m too,

$$\begin{aligned}&\int_0^T \left\langle \frac{\mathrm{d}(mw)}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \int_{Q_m} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}_m(w + S_s)\frac{\partial\phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t = \\ &= \textstyle\int_0^T \langle f(t) + f_{\Gamma_\alpha}(t), \phi(t)\rangle_{V',V}\, \mathrm{d}t, \ \forall \phi \in L^2(0,T;H_0^1(\Omega_m)),\end{aligned} \tag{3.85}$$

where $\zeta(x,t) \in \beta^*(w(x,t))$ a.e. on Q_m.

Taking now $\phi \in L^2(0,T;H_0^1(\Omega_0))$, we obtain the weak form of the equation on this subset

$$\int_{Q_0} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}_0(x)\frac{\partial\phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t = 0, \ \forall \phi \in L^2(0,T;H_0^1(\Omega_0)), \tag{3.86}$$

where ζ is given by (3.70).

On the other hand, (3.84) corresponds to the problem

$$\begin{aligned} \frac{\partial(mw)}{\partial t} - \Delta\zeta + \frac{\partial \widetilde{K}(x, w+S_s)}{\partial x_3} &= f \text{ in } Q, \\ \zeta &= 0 \text{ on } \Sigma_u, \\ (\widetilde{K}(x, w+S_s)i_3 - \nabla\zeta)\cdot\nu &= f_\alpha \text{ on } \Sigma_\alpha, \end{aligned} \tag{3.87}$$

and (3.85)–(3.86) to the problem

$$\begin{aligned} \frac{\partial(mw)}{\partial t} - \Delta\zeta + \frac{\partial \widetilde{K}_m(w+S_s)}{\partial x_3} &= f \text{ in } Q_m, \\ -\Delta\zeta + \frac{\partial \widetilde{K}_0(x)}{\partial x_3} &= f \text{ in } Q_0, \\ \zeta &= 0 \text{ on } \Sigma_u, \\ (\widetilde{K}_m(w+S_s)i_3 - \nabla\zeta)\cdot\nu &= f_\alpha \text{ on } \Sigma_\alpha. \end{aligned} \tag{3.88}$$

We recall that the common boundary of the domains Ω_m and Ω_0 is regular due to the fact that $m \in C^1(\overline{\Omega})$. Since $\zeta \in L^2(0,T;V)$, we deduce that the trace of $\zeta(t) \in \beta^*(w(t))$ belongs to V a.e. t, so it is continuous across the boundary $\partial\Omega_0$ (more exactly along lines $\mathcal{L}$ that cross the boundary), a.e. $t \in (0,T)$. Thus if we take $x_0 \in \partial\Omega_0$ and denote

$$\zeta^+(t) = \lim_{x\to x_0,\ x\in\mathcal{L}\cap\Omega_m} \zeta(t),$$

then we have

$$\zeta^+(t) = \lim_{x\to x_0,\ x\in\mathcal{L}\cap\Omega_0} \zeta(t) \text{ a.e. } t \in (0,T).$$

We take into account that $\zeta^+ \in \beta^*(w(t))$ a.e. on Q_m, hence ζ turns out to be the solution to the elliptic problem

$$\begin{aligned} -\Delta\zeta(t) &= f(t) + f_{\Gamma_\alpha}(t) \text{ in } \Omega_0 \\ \zeta(t) &= \zeta^+(t) \in \beta^*(w(t)) \text{ on } \partial\Omega_0, \text{ a.e. } t \in (0,T) \end{aligned} \tag{3.89}$$

for a.e. t fixed in $(0,T)$, and w is the solution to (3.85) (equivalently to (3.24)) in Q_m.

Then, we define the function

$$w^*(x,t) = \begin{cases} w(x,t), \text{ if } (x,t) \in Q_m \\ (\beta^*)^{-1}(\zeta(x,t)), \text{ if } (x,t) \in Q_0 = \Omega_0 \times (0,T), \end{cases} \tag{3.90}$$

where ζ is the solution to (3.89) and show that it is the solution to (3.27) in the sense of Definition 3.1. Indeed, $\zeta(x,t) \in \beta^*(w^*(x,t))$ and $\zeta \in L^2(0,T;V)$,

so it follows that $w^* \in D(A)$, implying that $w^* \leq w_s$ a.e. on Q. Then, mw^* belongs to the spaces specified in (3.23) (we take into account that $mw^* = 0$ on Q_0). Finally, we have to check that w^* satisfies the equation (3.26) and this follows by a straightforward computation using (3.84)–(3.86). Indeed, if we replace w^* in (3.26) we obtain

$$\begin{aligned}
&\int_0^T \left\langle \frac{\mathrm{d}(mw^*)}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \\
&+ \int_0^T \int_{\Omega_m} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}(x, w + S_s) \frac{\partial \phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t + \\
&+ \int_0^T \int_{\Omega_0} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}(x, w^*) \frac{\partial \phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t = \\
= \quad &\int_0^T \left\langle \frac{\mathrm{d}(mw)}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \\
&+ \int_Q \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}(x, w + S_s) \frac{\partial \phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t = \\
= \quad &\int_0^T \langle f(t) + f_{\Gamma_\alpha}, \phi(t) \rangle_{V',V} \, \mathrm{d}t, \ \forall \phi \in L^2(0,T;V).
\end{aligned}$$

We took into account the expressions assigned to w^* and $\widetilde{K}(x, w + S_s)$ on each subset, (3.81) and (3.84). □

COROLLARY **3.1** *Under the assumptions of Theorem* 3.1 *the solution to (3.27) is unique if in addition*

$$\widetilde{\rho} > c_\Omega M. \tag{3.91}$$

Proof. Let us denote by w_1^* and w_2^* two solutions to (3.27) corresponding to the same data. We multiply the difference of equations (3.27) written for w_1^* and w_2^* by $(w_1^* - w_2^*)$ scalarly in V', integrate on $(0,T)$ and use the Lipschitz property of $\widetilde{K}$. We get

$$\begin{aligned}
&\|m(w_1^*(\tau) - w_2^*(\tau))\|_{V'}^2 + \widetilde{\rho} \int_0^T \|w_1^*(\tau) - w_2^*(\tau)\|^2 \, \mathrm{d}\tau \leq \quad (3.92) \\
\leq \quad &\frac{M^2}{\widetilde{\rho}} \int_0^T \|w_1^*(\tau) - w_2^*(\tau)\| \, \|w_1^*(\tau) - w_2^*(\tau)\|_{V'} \, \mathrm{d}\tau \leq \\
\leq \quad &\frac{M^2}{\widetilde{\rho}} c_\Omega^2 \int_0^T \|w_1^*(\tau) - w_2^*(\tau)\|^2 \, \mathrm{d}\tau
\end{aligned}$$

where c_Ω is the constant in Poincaré's inequality. Here we took into account that for $w \in L^2(\Omega)$ we have $\|w\|_{V'} \leq c_\Omega \|w\|$.

It follows by (3.91) that $mw_1^* = mw_2^*$ a.e. on Q and $w_1^* = w_2^*$ a.e. on Q_m where $m(x) > 0$. Now we subtract the equations (3.88) corresponding to w_1^* and w_2^* and get

$$\begin{aligned} -\Delta(\zeta_1 - \zeta_2) &= 0 \text{ in } Q, \\ \zeta_1 - \zeta_2 &= 0 \text{ on } \Sigma_u, \\ -\nabla(\zeta_1 - \zeta_2) \cdot \nu &= 0 \text{ on } \Sigma_\alpha, \end{aligned}$$

where $\zeta_1 \in \beta^*(w_1^*)$, $\zeta_2 \in \beta^*(w_2^*)$ a.e. on Q. Hence $\zeta_1 = \zeta_2$ and since $(\beta^*)^{-1}$ is single valued then $w_1^* = w_2^*$ a.e. on Q. □

Remark 3.1 We observe that in the degenerate case the uniqueness of the solution can be obtained only if the transport is dominated in a sense (see (3.91)) by the diffusivity. In particular, this is true when $\widetilde{K} = 0$, i.e., when we deal with a horizontal infiltration, also called sorption.

Remark 3.2 By the proof of the solution existence we also ascertain a consequence that can be inferred at an intuitive level, i.e., the boundary value problem is separated into two problems corresponding to the domains Q_m and Q_0, connected by the flux continuity.

Indeed, if we test the first two equations in (3.88) at $\phi \in L^2(0,T;V)$ and integrate the sum over $(0,T)$ we obtain

$$\begin{aligned} &\int_0^T \left\langle \frac{\mathrm{d}(mw)}{\mathrm{d}t}(t), \phi(t) \right\rangle_{V',V} \mathrm{d}t + \\ &+ \int_0^T \int_{\Omega_m} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}_m(w + S_s) \frac{\partial \phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t - \\ &- \int_0^T \int_{\partial\Omega_m} \left(\widetilde{K}_m(w + S_s) i_3 - \nabla\zeta \right) \cdot \nu^+ \phi \, \mathrm{d}\sigma\mathrm{d}t + \\ &+ \int_0^T \int_{\Omega_0} \left(\nabla\zeta \cdot \nabla\phi - \widetilde{K}_0(x) \frac{\partial \phi}{\partial x_3} \right) \mathrm{d}x\mathrm{d}t - \\ &- \int_0^T \int_{\partial\Omega_0} \left(\widetilde{K}_0(x) i_3 - \nabla\zeta \right) \cdot \nu^- \phi \, \mathrm{d}\sigma\mathrm{d}t = \\ = &\int_0^T \int_\Omega \langle f(t) + f_{\Gamma_\alpha}(t), \phi(t) \rangle_{V',V} \, \mathrm{d}x\mathrm{d}t, \end{aligned}$$

for any $\phi \in L^2(0,T;V)$, where ν^+ is the outer normal to $\partial\Omega_m$, ν^- is the outer normal to $\partial\Omega_0$ and $\zeta \in \beta^*(w)$ a.e. on Q_m. Taking into account (3.84)

we obtain the flux continuity on the common boundary $\partial\Omega_0 \times (0,T)$

$$\left(\widetilde{K}_m(w+S_s)i_3 - \nabla\zeta\right)\cdot\nu^+ = \left(\widetilde{K}_0(x)i_3 - \nabla\zeta\right)\cdot\nu^+ \text{ on } \partial\Omega_0\times(0,T). \quad (3.93)$$

The previous integrals on $\partial\Omega_m$ and $\partial\Omega_0$ are considered in the sense of distributions, e.g., as the value of $\left(\widetilde{K}(x,w+S_s)i_3 - \nabla\zeta\right)\cdot\nu$ at ϕ. By the trace theorem we see that, generally, the flux $\left(\widetilde{K}(x,w+S_s)i_3 - \nabla\zeta\right)\cdot\nu$ is well defined as an element of the space $L^2(0,T;H^{-1/2}(\partial\Omega_0))$.

References

[1] H.W. Alt, S. Luckhaus, *Quasi-linear elliptic-parabolic differential equations*. Math. Z., **183** (1983), 311–341.

[2] H.W. Alt, S. Luckhaus, A. Visintin, *On nonstationary flow through porous media*. Ann. Mat. Pura Appl., **136** (1984), 303–316.

[3] D.G. Aronson, *The porous medium equation*. In: A. Fasano, M. Primicerio (Eds.), *Some Problems in Nonlinear Diffusion*, Lecture Notes in Mathematics, **1224**, Springer, Berlin 1986.

[4] C. Baiocchi, *Su un problema di frontiera libera conesso a questioni di idraulica*. Ann. Mat. Pura Appl., **92** (1972), 107–127.

[5] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International Publishing, Leyden 1976.

[6] V. Barbu, G. Marinoschi, *Existence for a time dependent rainfall infiltration model with a blowing up diffusivity*. Nonlinear Analysis Real World Applications, **5** (2004), *2*, 231–245.

[7] J. Bear, *Hydraulics of Grounwater*. McGraw-Hill, Inc., New York 1979.

[8] P. Benilan, S.N. Krushkov, *Quasilinear first-order equations with continuous nonlinearities*. Russian Acad. Sci. Dokl. Math., **50** (1995), *3*, 391–396.

[9] I. Borsi, A. Farina, A. Fasano, *On the infiltration of rain water through the soil with runoff of the excess water*. Nonlinear Analysis Real World Applications, **5** (2004), 763–800.

[10] P. Broadbridge, I. White, *Constant Rate Rainfall Infiltration, A versatile nonlinear model. 1. Analytic solution.* Water Resources Research, **24** (1988), *1,* 145–154.

[11] P. Broadbrigde, J.H. Knight, C. Rogers, *Constant rate rainfall in a bounded profile: Solutions of a nonlinear model.* Soil Sci. Soc. Am. J., **52** (1988), 1526–1533.

[12] J. R. Cannon, R.B. Guenther, F. A. Mohamed, *On the rainfall infiltration through a soil medium.* SIAM J. Appl. Math., **49** (1989), 720–729.

[13] J. Carillo, *Unicité des solutions du type Krushkov pour des problemes elliptique avec des termes de transport non linéaires.* C. R. Acad. Sci. Paris, Serie I, **303** (1986), 189–192.

[14] J. Carillo, *On the uniqueness of the solution of the evolution dam problem.* Nonlinear Analysis, **22** (1994), 573–607.

[15] J. Carillo, *Entropy solutions for nonlinear degenerate problems.* Arch. Rational Mech. Anal., **147** (1999), 269–361.

[16] J. Carillo, P. Wittbold, *Uniqueness of renormalized solutions of degenerate elliptic-parabolic problems.* J. Differential Equations, **156** (1999), 93–121.

[17] E. Chasseigne, J.L. Vázquez, *Theory of extended solutions for fast diffusion equations in optimal classes of data. Radiation from singularities.* Archive Rat. Mech. Anal. **164** (2002), 133–187.

[18] J.R. Esteban, A. Rodriguez, J.L. Vázquez, *A nonlinear heat equation with singular diffusivity.* Comm. Partial Differential Equations, **13** (1988), 985–1039.

[19] A. Fasano, M. Primicerio, *Free boundary problems for nonlinear parabolic equations with nonlinear free bounadry conditions.* J. Math. Anal. Appl., **72** (1979), 247–273.

[20] A. Fasano, M. Primicerio, *Liquid flow in partially saturated porous media.* J. Inst. Math. Appl., **23** (1979), 503–517.

[21] A. Favini, M. Fuhrman, *Approximation results for semigroups generated by multivalued linear operators and applications.* Differential and Integral Equations, **11** (1998), *5*, 781–805.

[22] A. Favini, A. Yagi, *Degenerate differential equations in Banach spaces.* Marcel Dekker, Inc., New York, 1999.

[23] A. Favini, A. Yagi, *Quasilinear degenerate evolution equations in Banach spaces*, J. Evol. Eq. 4 (2004), 421–449.

[24] A. Favini, G. Marinoschi, *Existence for a degenerate diffusion problem with a nonlinear operator*. J. Evol. Equ., **7** (2007) 743–764.

[25] J. Filo, S. Luckhaus, *Modelling surface runoff and infiltration of rain by an elliptic-parabolic equation coupled with a first-order equation on the boundary*. Arch. Rational Mech. Anal., **146** (1999), 157–182.

[26] R. Gianni, *A filtration problem with ponding*. Boll. Un. Mat. Ital., **5** B (1991), 875–891.

[27] G. Gilardi, *A new approach to evolution free boundary problems*. Comm. Partial Differential Equations, **4** (1979), 1099–1122.

[28] S.N. Krushkov, *Generalized solutions of the Cauchy problem in the large for first-order nonlinear equations*. Soviet Math. Dokl., **10** (1969), 785–788.

[29] G. Marinoschi, *Nonlinear infiltration with a singular diffusion coefficient*. Differential Integral Equations, **16** (2003), *9*, 1093–1110.

[30] G. Marinoschi, *On a nonlinear boundary value problem related to infiltration in unsaturated media*. In: M. Suliciu (Ed.), *New Trends in Continuum Mechanics*, Proceedings of International Conference, Constanta, September 2003, pp. 175–184, Theta Publishing, Bucharest, 2005.

[31] G. Marinoschi, *On some problems concerning the nonlinear infiltration in unsaturated media*. Proceedings of the International Conference NPDE 2003, Alushta, Ukraine, September 2003.

[32] G. Marinoschi, *A free boundary problem describing the saturated unsaturated flow in a porous medium*. Abstr. Appl. Anal., **9** (2004), 729–755.

[33] G. Marinoschi, *A free boundary problem describing the saturated unsaturated flow in a porous medium. II. Existence of the free boundary in the 3-D case*. Abstr. Appl. Anal., **8** (2005), 813–854.

[34] G. Marinoschi, *Functional Approach to Nonlinear Models of Water Flow in Soils*. Mathematical Modelling: Theory and Applications, volume **21**, Springer 2006.

[35] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*. J. Differential Equations, **131** (1996), 20–38.

[36] F. Otto, *L^1-contraction and uniqueness for unstationary saturated-unsaturated porous media flow.* Adv. Math. Sci. Appl., **7** (1997), 537–553.

[37] A. Torelli, *Su un problema a frontiera libera di evoluzione.* Boll. Un. Mat. Ital., **11** (1975), 559–570.

[38] C.J. van Duyn, J.B. McLeod, *Nonstationary filtration in partially saturated porous media.* Arch. Rational Mech. Anal., **78** (1982), 173–198.

[39] M.T. van Genuchten, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils.* Soil Sci. Soc. Am. J., **44** (1980), 892–898.

[40] J.L. Vázquez, *Nonexistence of solutions for nonlinear heat equations of fast-diffusion type.* J. Math. Pures Appl., **71** (1992), 503–526.

[41] J.L. Vázquez, *Darcy's law and the theory of shrinking solutions of fast diffusion equations.* SIAM J. Math. Anal., **35** (2003), *4*, 1005–1028.

[42] J.L. Vázquez, *Symmetrization and mass comparison for degenerate nonlinear parabolic and related elliptic equations.* Advanced Nonlinear Studies **5** (2005), 87–131.

[43] J.L. Vázquez, *Smoothing and Decay Estimates for Nonlinear Diffusion Equations. Equations of Porous Medium Type.* Series Oxford Lecture Series in Mathematics and its Applications no. 33, Oxford University Press 2006.

[44] A. W. Warrick, P. Broadbridge, *Sorptivity and macroscopic capillary length relationships.* Water Resources Research, **28** (1992), 2, 427–431.