SEMA SIMAI Springer Series 3

Carlos Parés
Carlos Vázquez
Frédéric Coquel
Editors

Advances in Numerical Simulation in Physics and Engineering

Lecture Notes of the XV 'Jacques-Louis Lions' Spanish-French School

SEMA SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE
Springer

SEMA SIMAI Springer Series

Volume 3

The SEMA SIMAI Springer Series is a joint series aiming to publish advanced textbooks, research-level monographs and collected works that focus on applications of mathematics to social and industrial problems, including biology, medicine, engineering, environment and finance. Mathematical and numerical modeling is playing a crucial role in the solution of the complex and interrelated problems faced nowadays not only by researchers operating in the field of basic sciences, but also in more directly applied and industrial sectors. This series is meant to host selected contributions focusing on the relevance of mathematics in real life applications and to provide useful reference material to students, academic and industrial researchers at an international level. Interdisciplinary contributions, showing a fruitful collaboration of mathematicians with researchers of other fields to address complex applications, are welcomed in this series

For further volumes:
http://www.springer.com/series/10532

Carlos Parés • Carlos Vázquez •
Frédéric Coquel
Editors

Advances in Numerical Simulation in Physics and Engineering

Lecture Notes of the XV 'Jacques-Louis Lions' Spanish-French School

Springer

Editors
Carlos Parés
Facultad de Ciencias
Universidad de Málaga
Málaga
Spain

Carlos Vázquez
Facultad de Informática
Universidade da Coruña
A Coruña
Spain

Frédéric Coquel
Centre de Mathématiques Appliquées
CNRS and Ecole Polytechnique
Paris
France

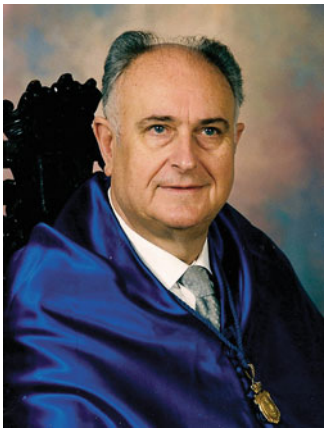
ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISBN 978-3-319-02838-5 ISBN 978-3-319-02839-2 (eBook)
DOI 10.1007/978-3-319-02839-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014943804

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to the memory of
Prof. Antonio Valle Sánchez (1930–2012)



Antonio Valle Sánchez, 1930–2012

Preface

This two-part book contains the lecture notes of the XV Spanish–French School on Numerical Simulation in Physics and Engineering that took place in Torremolinos (Málaga, Spain) in September 2012: Part I corresponds to the four courses and Part II to the invited talks. This series of Schools is organized every 2 years since 1984 and it intends to bring together professionals, researchers, and students interested in numerical methods. The previous editions were held in Santiago de Compostela (1984), Benalmádena, Málaga (1986), Madrid (1988), Santiago de Compostela (1990), Benicassim, Castellón (1992), Sevilla (1994), Oviedo (1996), Córdoba (1998), Laredo, Cantabria (2000), Jaca, Huesca (2002), Cádiz (2004), Castro Urdiales, Cantabria (2006), Valladolid (2008), and La Coruña (2010). Next edition will take place in Pamplona, Navarra in September 2014. Since its foundation in 1991, the Sociedad Española de Matemática Aplicada (SEMA) is actively involved in the organization of these Schools. The Spanish–French Schools and the Congresos de Ecuaciones Diferenciales y Aplicaciones/Congresos de Matemática Aplicada constitute the two main series of scientific meetings sponsored by the Society. In 2004, it was decided to honour the French mathematician Jacques-Louis Lions by adding his name to the Schools. Since 2008, the Société de Mathématiques Appliquées et Industrielles (SMAI) co-organizes these meetings. The main goals of the Schools are the following:

- To initiate people interested in Applied Mathematics in research topics, in particular in mathematical modeling and numerical simulation related to the research areas developed in France and Spain.
- To be a meeting point for researchers, teachers, industrial technicians, and students from both countries.
- To show current applications of numerical simulation in industry, in particular in French and Spanish companies.

The School is mainly destined to young holders of engineering or science degrees who want to start working in numerical simulation, either in research or in the field

of industrial applications. The School is also oriented to technicians working in industry who are interested in the use of numerical techniques in some problems similar to those handled by them, or who want to know the research lines developed in French and Spanish universities and scientific organisms. The School is also of interest for other university people, as it permits the exchange of experience and knowledge concerning the research developed in different laboratories.

Each edition is organized around several main courses and conferences delivered by renowned French and Spanish scientists. On this last occasion there were four 6-h courses, whose lecturers were Begoña Calvo and Estefanía Peña, Enrique Domingo Fernández Nieto, Emmanuel Gobet, and Philippe LeFloch, together with five 1-h talks given by Carlos Castro, Emanuele Schiavi, Michel Langlais, Fabien Mangeant, and Denis Talay. The XV School was a very special one for the Spanish and French communities of Numerical Analysis: it was a tribute to Prof. Antonio Valle Sánchez who sadly passed away on June 24, 2012. Professor Valle was the first Spanish student of Prof. Jacques-Louis Lions and he was among the main promoters of the research in modern Applied Mathematics in Spain. He was the founder of a very large Spanish community of Numerical Methods in Partial Differential Equations that grew up from the three universities in which he was a Professor: Santiago de Compostela, Sevilla, and Málaga. He also promoted the scientific collaboration with French researchers in this field. Professor Antonio Valle was the first President of the Spanish Society for Applied Mathematics (SEMA) and one of the promoters of the Spanish–French Schools on Numerical Simulation in Physics and Engineering. In particular, Prof. Valle was the President of the Organizing Committee of the second edition of these schools, which took place in Benalmádena (Málaga). The fifteenth edition of the Schools took place again in Málaga, the city where Prof. Valle was born and where he concluded his career. It was therefore natural to pay a special tribute to him and his invaluable work in promoting the development of research groups whose activity is strongly related to the subject of the Schools in collaboration with first level French researchers. A special session in his honor was included in the program, in which, besides the academic authorities and the President of SEMA, Professors Alfredo Bermúdez de Castro (University of Santiago de Compostela) and Michel Bernadou (Pôle Universitaire Léonard de Vinci) spoke on behalf of the many Spanish disciples of Prof. Valle and their French collaborators, respectively. It is worth mentioning that Prof. Michel Bernadou has been involved in the organization of the Schools from the beginning.

The Editors warmly thank all the speakers and participants for their contribution to the success of the School. In particular, we would like to acknowledge the efforts of all the lecturers and speakers who have contributed to this volume. We are also grateful to the Organizing and the Scientific Committees for their efforts in the preparation of the School. We extend our thanks and gratitude to all sponsors and supporting institutions for their valuable contribution: SEMA, SMAI, Universidad

de Málaga, the French Embassy in Spain, and the Spanish Ministry of Economy and Competitiveness that awarded the grant MTM2011-14775-E.

Málaga, Spain — Carlos Parés
A Coruña, Spain — Carlos Vázquez
Paris, France — Frédéric Coquel
January 2014

Contents

Part I
Courses

Fundamental Aspects in Modelling the Constitutive Behaviour of Fibered Soft Tissues

Begoña Calvo and Estefanía Peña

Abstract Fibered soft tissues like ligament, tendons, cartilage or those composing the cardiovascular system among others are characterized by a complex behaviour derived from their specific internal composition and architecture that has to be considered when trying to simulate their response under physiological or pathological external loads, their interaction with external implants or during and after surgery. The evaluation of the acting stresses and strains on these tissues is essential in predicting possible failure (i.e., aneurisms, atherosclerotic plaques, ligaments rupture) or the evolution of their microstructure under changing mechanical environment (i.e. cardiac aging, atherosclerosis, ligament remodeling). As structural materials, fibered soft tissues undergo large deformations even under physiological loads and are almost incompressible and highly anisotropic, mainly due to the directional distribution of the different composing families of collagen fibers. In addition, they are non-linearly elastic under slowly-acting loads, viscoelastic, due both to the moving internal fluid in some tissues (i.e. cartilage) or to the inherent viscoelasticity of the solid matrix. They are also subjected to non-negligible initial stresses due to the growth and remodeling processes that act along their whole live. Finally, they are susceptible to suffer damage induced by the rupture of the fibers or tearing of the surrounding matrix. All these aspects should be considered for a full description of the constitutive behaviour of these materials, requiring an appropriate mathematical formulation and finite element implementation to get efficient simulations useful for a better understanding of their phsyiological function, the effect of pathologies or surgery as well as for surgery planning and design of implants among many other usual applications. In this work, formulations of all the different phenomena commented above in fibered soft tissues are presented. The effect of each of these

B. Calvo • E. Peña (✉)
Aragón Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain

Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain
e-mail: bcalvo@unizar.es; fany@unizar.es

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_1,

aspects is analyzed in simplified examples to demonstrate the applicability of the models. Finally, different applications of clinical interest are discussed.

1 Introduction

Biological soft tissues are subjected to large deformations with negligible volume change and show a highly non-linear anisotropic mechanical response due to their internal structure. The extra-cellular matrix is composed of a network of collagen fibrils and elastin fibers embedded in a viscous and quasi-isotropic ground substance [20]. Typical examples of fibered soft biological tissues are blood vessels, tendons, ligaments, cornea and cartilage.

The purely elastic response of soft tissues is often modelled within the framework of continuum mechanics by means of the definition of a strain energy function expressed in terms of kinematic invariants, first developed by Spencer [66]. This approach was further tuned and applied to finite element simulations of soft collagenous biological tissues (see for example Weiss et al. [68], Peña et al. [47] for ligaments, Holzapfel et al. [30] for arteries and Alastrué et al. [2] for cornea).

Fibered soft tissues are also exposed to a complex distribution of "in vivo" initial strains. This state is a consequence of the continuous growth, remodeling, damage and viscoplastic strains that suffer these living materials along their whole life. Due to the non-linear behaviour of this kind of materials and the non-uniform distribution of the residual stresses, a wrong inclusion of the initial strain state in computational models of soft tissues can lead to large errors (usually an important underestimation of the stress level) [48].

Many fibered soft tissues exhibit simultaneously elastic and viscous material behavior. The rate-dependent material behavior of this kind of materials has been well-documented and quantified in the literature. For example, works on ligaments [58], tendons [36], blood vessels [31, 54], cornea [55] and articular cartilage [26]. Furthermore, non-physiological loads drive soft tissue to damage that may induce a strong reduction of the stiffness. Damage may arise from two possible mechanisms: tear or plastic deformation of the fibers, or biochemical degradation of the extracellular matrix from protease release associated with the observed cellular necrosis [57].

It is important to note that accuracy of the biomechanical models strongly depends on a precise geometrical reconstruction and on an accurate mathematical description of the behavior of the biological tissues involved, and their interactions with the surrounding environment, see for example for a knee joint in Fig. 1. The acquisition of an accurate geometry is a fundamental requirement for the construction of three-dimensional finite element (FE) models. Both magnetic resonance imaging (MRI) and computerized tomography (CT) are used to acquire joint geometry. MRI provides detailed images of soft tissues in diarthrodial joints while CT provides excellent images of the bones. Once the geometrical model has been reconstructed from the 3D image dataset, it is necessary to generate the FE

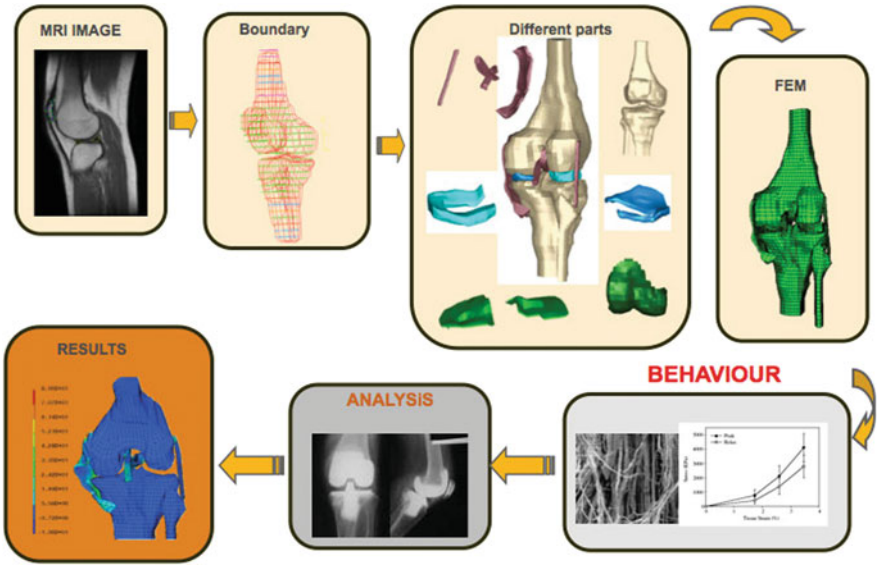


Fig. 1 Steps for the numerical resolution of biomechanics problem using finite element method

mesh. Constitutive equations are used to describe the mechanical behavior of ideal materials through the establishment of the dependence of the stress on different variables, such as the deformation gradient, rate of deformation or temperature. Another important aspect in tissue modelling is the evaluation of the constitutive parameters. Material coefficients may be based on subject-specific measurements or on population averages. In the former, uncertainty is related to inherent errors in experimental measurements and their extension to other individuals. The comparison of model predictions to experimental measurements, or clinical evidences constitutes the validation process. There is no way to completely validate a model. Therefore, one must pose specific hypotheses about model predictions along with tolerable errors. Validation is the most challenging aspect of the FE modelling of tissue mechanics, as it requires accurate experimental measurements of quantities that are difficult to obtain.

Taken all this into account, this chapter is focused on the development of constitutive models for soft fibred tissues and organized as follows. In Sect. 2 the constitutive equations of anisotropic hyperelastic materials are reviewed. In Sect. 3, we present the weak form and linearized weak form of the continuum problem. Section 4 addresses the different methodologies used to enforce initial stresses. Section 5 considered an anisotropic visco-hyperelastic model and Sect. 5.2 an anisotropic damage model for biological soft tissue. The application of these methodologies to some examples is discussed in Sect. 6. Finally, Sect. 7 includes some concluding remarks.

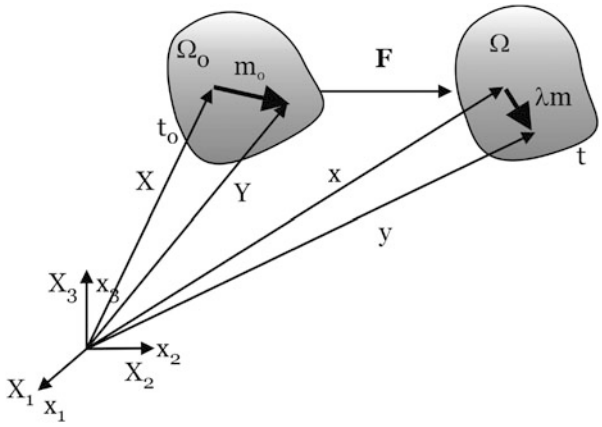


Fig. 2 Kinematics of a unit vector field $\mathbf{m}_0(\mathbf{X})$

2 Hyperelastic Behavior

This section deals with the formulation of standard finite strain material models for soft biological tissues. To clarify the framework it is necessary to summarize the formulation of finite strain hyperelasticity in terms of invariants with uncoupled volumetric/deviatoric responses, first suggested in Flory [19], generalized in Simo et al. [65] and employed for anisotropic soft biological tissues in Weiss et al.[68], Holzapfel et al. [30] among others.

Let $\Omega_0 \subset \mathbb{E}^3$ be a reference or rather material configuration of a body of interest. The notation $\boldsymbol{\varphi} : \Omega_0 \times \mathcal{T} \rightarrow \Omega_t$ represents the one to one mapping, continuously differentiable, transforming a material point $\mathbf{X} \in \Omega_0$ to a position $\mathbf{x} = \boldsymbol{\varphi}(\mathbf{X}, t) \in \Omega_t \subset \mathbb{E}^3$, where Ω_t represents the deformed configuration at time $t \in \mathcal{T} \subset \mathbb{R}$. The mapping $\boldsymbol{\varphi}$ represents a motion of the body that establishes the trajectory of a given point when moving from its reference position $\mathbf{X}$ to $\mathbf{x}$. The two-point deformation gradient tensor is defined as $\mathbf{F}(\mathbf{X}, t) := \nabla_{\mathbf{X}} \boldsymbol{\varphi}(\mathbf{X}, t)$, with $J(\mathbf{X}) = \det(\mathbf{F}) > 0$ the local volume variation.

The direction of a fiber at a point $\mathbf{X} \in \Omega_0$ is defined by a unit vector field $\mathbf{m}_0(\mathbf{X})$, $|\mathbf{m}_0| = 1$ (see Fig. 2). It is usually assumed that, under deformation, the fiber moves with the material points of the continuum body, that is, it follows an affine deformation. Therefore, the stretch λ of the fiber defined as the ratio between its lengths at the deformed and reference configurations can be expressed as

$$\lambda \mathbf{m}(\mathbf{x}, t) = \mathbf{F}(\mathbf{X}, t)\mathbf{m}_0(\mathbf{X}), \tag{1}$$

where $\mathbf{m}$ is the unit vector of the fiber in the deformed configuration and

$$\lambda^2 = \mathbf{m}_0 \cdot \mathbf{F}^T\mathbf{F} \cdot \mathbf{m}_0 = \mathbf{m}_0 \cdot \mathbf{C}\mathbf{m}_0 \tag{2}$$

stands for the stretch along the fiber direction at point $\mathbf{X}$. In (2) $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ is the standard deformation gradient and the corresponding right Cauchy–Green strain measure. The introduced kinematics for one family of fibers can be applied to a second fiber family in an analogous manner. We shall denote a second preferred fiber orientation by the unit vector field $\mathbf{n}_0(\mathbf{X})$.

It is sometimes useful to consider the multiplicative decomposition of $\mathbf{F}$

$$\mathbf{F} := J^{1/3}\mathbf{I}\cdot\bar{\mathbf{F}}. \tag{3}$$

Hence, deformation is split into a dilatational part, $J^{1/3}\mathbf{I}$, where $\mathbf{I}$ represents the second-order identity tensor, and an isochoric contribution, $\bar{\mathbf{F}}$, so that $\det(\bar{\mathbf{F}}) = 1$ [19]. With these quantities at hand, the isochoric counterparts of the right Cauchy–Green deformation tensors associated with $\bar{\mathbf{F}}$ are defined as $\bar{\mathbf{C}} := \bar{\mathbf{F}}^{\mathrm{T}}\cdot\bar{\mathbf{F}} = J^{-2/3}\mathbf{C}$.

The free energy function (SEF) is given by a scalar-valued function Ψ defined per unit reference volume in the reference configuration and for isothermal processes. Flory [19] postulated the additive decoupled representation of this SEF in volumetric and isochoric parts. To differentiate between the isotropic and the anisotropic parts, the free energy density function can be split up again as

$$\Psi = \Psi_{\mathrm{vol}} + \bar{\Psi}_{\mathrm{iso}} + \bar{\Psi}_{\mathrm{ani}}, \tag{4}$$

where Ψ_{vol} describes the free energy associated to changes of volume, $\bar{\Psi}_{\mathrm{iso}}$ is the isochoric isotropic contribution of the free energy (usually associated to the ground matrix) and $\bar{\Psi}_{\mathrm{ani}}$ takes into account the isochoric anisotropic contribution (associated to the fibers) [66].

This strain-energy density function must satisfy the principle material frame invariance $\Psi(\mathbf{C},\mathbf{M},\mathbf{N}) = \Psi(\mathbf{Q}\cdot\mathbf{C},\mathbf{Q}\cdot\mathbf{M},\mathbf{Q}\cdot\mathbf{N})$ for all $[\mathbf{C},\mathbf{Q}] \in [\mathbb{S}^3_+ \times \mathbb{Q}^3_+]$. Because of the directional dependence on the deformation, we require that the function Ψ explicitly depends on both the right Cauchy–Green tensor $\mathbf{C}$ and the fibers directions in the reference configuration ($\mathbf{m}_0$ and $\mathbf{n}_0$ in the case of two fiber families). Since the sign of $\mathbf{m}_0$ and $\mathbf{n}_0$ is not significant, Ψ must be an even function of $\mathbf{m}_0$ and $\mathbf{n}_0$ and so it may be expressed by $\Psi = \Psi(\mathbf{C},\mathbf{M},\mathbf{N})$ where $\mathbf{M} = \mathbf{m}_0 \otimes \mathbf{m}_0$ and $\mathbf{N} = \mathbf{n}_0 \otimes \mathbf{n}_0$ are structural tensors [66]. In terms of the strain invariants [66], Ψ can be written as

$$\Psi = \Psi_{\mathrm{vol}}(J) + \bar{\Psi}_{\mathrm{iso}}(\bar{I}_1,\bar{I}_2) + \bar{\Psi}_{\mathrm{ani}}(\bar{I}_4,\bar{I}_5,\bar{I}_6,\bar{I}_7,\bar{I}_8,\bar{I}_9), \tag{5}$$

with $\bar{I}_1$ and $\bar{I}_2$ the first two modified strain invariants of the symmetric modified Cauchy–Green tensor $\bar{\mathbf{C}}$ (note that $I_3 = J^2$). Finally, the anisotropic invariants $\bar{I}_4,\ldots,\bar{I}_9$ characterize the constitutive response of the fibers [66]:

$$\begin{aligned}
&\bar{I}_4 = \bar{\mathbf{C}} : \mathbf{M} = \bar{\lambda}_m^2, \quad \bar{I}_5 = \bar{\mathbf{C}}^2 : \mathbf{M} \\
&\bar{I}_6 = \bar{\mathbf{C}} : \mathbf{N} = \bar{\lambda}_n^2, \quad \bar{I}_7 = \bar{\mathbf{C}}^2 : \mathbf{N} \\
&\bar{I}_8 = [\mathbf{m}_0\cdot\mathbf{n}_0]\mathbf{m}_0\cdot\bar{\mathbf{C}}\mathbf{n}_0 \quad \bar{I}_9 = [\mathbf{m}_0\cdot\mathbf{n}_0]^2.
\end{aligned} \tag{6}$$

Remark. While the invariants $\bar{I}_4$ and $\bar{I}_6$ have a clear physical meaning, the square of the stretch λ in the fibers directions, the influence of $\bar{I}_5$, $\bar{I}_7$ and $\bar{I}_8$ is difficult to evaluate due to the high correlation between them [28]. For this reason and the lack of sufficient experimental data it is usual not to include these invariants in the definition of Ψ for soft biological tissues. Finally, $\bar{I}_9$ does not depend on the deformation, so it has not included in (5).

The second Piola–Kirchhoff stress tensor is obtained by derivation of (4) with respect to the right Cauchy–Green tensor [40]. Thus, the stress tensor consists of a purely volumetric and a purely isochoric contribution, i.e. $\mathbf{S}_{\text{vol}}$ and $\mathbf{S}_{\text{ich}}$, so the total stress is

$$\begin{aligned}\mathbf{S} &= \mathbf{S}_{\text{vol}} + \bar{\mathbf{S}} = 2\frac{\partial \Psi_{\text{vol}}(J)}{\partial \mathbf{C}} + 2\frac{\partial \bar{\Psi}(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N})}{\partial \mathbf{C}} \\ &= \left[\frac{\partial \Psi_{\text{vol}}(J)}{\partial J}\frac{\partial J}{\partial \mathbf{C}} + \frac{\partial \bar{\Psi}(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N})}{\partial \bar{\mathbf{C}}}\frac{\partial \bar{\mathbf{C}}}{\partial \mathbf{C}}\right] \\ &= J\, p\mathbf{C}^{-1} + \sum_{j=1,2,4,6} \mathsf{P} : 2\frac{\partial \bar{\Psi}}{\partial \bar{I}_j}\frac{\partial \bar{I}_j}{\partial \bar{\mathbf{C}}} = J\, p\mathbf{C}^{-1} + \mathsf{P} : \tilde{\mathbf{S}}, \end{aligned} \tag{7}$$

where the second Piola–Kirchhoff stress $\mathbf{S}$ consists of a purely volumetric contribution and a purely isochoric one. Moreover, one obtains the following noticeable relations $\partial_{\mathbf{C}} J = \frac{1}{2} J \mathbf{C}^{-1}$ and $\mathsf{P} = \partial_{\mathbf{C}} \bar{\mathbf{C}} = J^{-2/3}[\mathsf{I} - \frac{1}{3}\mathbf{C} \otimes \mathbf{C}^{-1}]$. P is the fourth-order projection tensor and I denotes the fourth-order unit tensor, which, in index notation, has the form $\mathsf{I}_{IJKL} = \frac{1}{2}[\delta_{IK}\delta_{JL} + \delta_{IL}\delta_{JK}]$. The projection tensor P furnishes the physically correct deviatoric operator in the Lagrangian description, i.e. $DEV[\cdot] = (\cdot) - 1/3(\bar{\mathbf{C}} : (\cdot))\bar{\mathbf{C}}^{-1}$.

Note that it is possible to obtain the Cauchy stress tensor by applying the push-forward operation to (7) $\boldsymbol{\sigma} = J^{-1}\boldsymbol{\chi}_*(\mathbf{S})$ [40]. Hence:

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_{\text{vol}} + \bar{\boldsymbol{\sigma}} = p\mathbf{1} + \frac{1}{J} dev\left[\bar{\mathbf{F}}\tilde{\mathbf{S}}\bar{\mathbf{F}}^{\mathbf{T}}\right] = p\mathbf{1} + \frac{1}{J} dev[\tilde{\boldsymbol{\sigma}}] = p\mathbf{1} + \mathsf{p} : \tilde{\boldsymbol{\sigma}}, \tag{8}$$

where we have introduced the projection tensor $\mathsf{p} = J^{-1}[\mathsf{I} - \frac{1}{3}\mathbf{1} \otimes \mathbf{1}]$ which furnishes the physically correct deviatoric operator in the Eulerian description, i.e. $dev[\cdot] = (\cdot) - \frac{1}{3} tr[\cdot]\mathbf{1}$.

Based on the kinematic decomposition of the deformation gradient tensor, the tangent operator, also known as the elasticity tensor when dealing with elastic constitutive laws, is defined in the reference configuration as

$$\begin{aligned}\mathsf{C} &= 2\frac{\partial \mathbf{S}(\mathbf{C}, \mathbf{M}, \mathbf{N})}{\partial \mathbf{C}} = \mathsf{C}_{\text{vol}} + \bar{\mathsf{C}} = 2\frac{\partial \mathbf{S}_{\text{vol}}}{\partial \mathbf{C}} + 2\frac{\partial \bar{\mathbf{S}}}{\partial \mathbf{C}} \\ &= 4\left[\frac{\partial^2 \Psi_{\text{vol}}(J)}{\partial \mathbf{C} \otimes \partial \mathbf{C}} + \frac{\partial^2 \bar{\Psi}(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N})}{\partial \mathbf{C} \otimes \partial \mathbf{C}}\right], \end{aligned} \tag{9}$$

where

$$\mathbf{C}_{\text{vol}} = 2\mathbf{C}^{-1} \otimes \left(p\frac{\partial J}{\partial \mathbf{C}} + J\frac{\partial p}{\partial \mathbf{C}} + 2Jp\frac{\partial \mathbf{C}^{-1}}{\partial \mathbf{C}} \right) = J\tilde{p}\mathbf{C}^{-1} \otimes \mathbf{C}^{-1} - 2J\mathsf{I}_{C^{-1}}, \tag{10}$$

with $(\mathsf{I}_{C^{-1}})_{IJKL} = -(\mathbf{C}^{-1} \odot \mathbf{C}^{-1})_{IJKL} = -\frac{1}{2}(C^{-1}_{IK}C^{-1}_{JL} + C^{-1}_{IL}C^{-1}_{JK})$, and $\tilde{p} = p + J\frac{dp}{dJ}$.

The term $\bar{\mathbf{C}}$ corresponding to the deviatoric part is given by:

$$\begin{aligned}\bar{\mathbf{C}} = &-\frac{4}{3}J^{-\frac{4}{3}}\left(\frac{\partial \bar{\Psi}}{\partial \bar{\mathbf{C}}} \otimes \bar{\mathbf{C}}^{-1} + \bar{\mathbf{C}}^{-1} \otimes \frac{\partial \bar{\Psi}}{\partial \bar{\mathbf{C}}}\right) \\ &+ \frac{4}{3}J^{-\frac{4}{3}}\left(\frac{\partial \bar{\Psi}}{\partial \bar{\mathbf{C}}} : \bar{\mathbf{C}}\right)\left(\mathbb{I}_{\bar{C}^{-1}} - \frac{1}{3}\bar{\mathbf{C}}^{-1} \otimes \bar{\mathbf{C}}^{-1}\right) + J^{-\frac{4}{3}}\bar{\mathbf{C}}_{\bar{w}},\end{aligned} \tag{11}$$

where term $\bar{\mathbf{C}}_{\bar{w}}$ is defined as:

$$\begin{aligned}\bar{\mathbf{C}}_{\bar{w}} = &\, 4\frac{\partial^2 \bar{\Psi}}{\partial \bar{\mathbf{C}} \partial \bar{\mathbf{C}}} - \frac{4}{3}\left[\left(\frac{\partial^2 \bar{\Psi}}{\partial \bar{\mathbf{C}} \partial \bar{\mathbf{C}}} : \bar{\mathbf{C}}\right) \otimes \bar{\mathbf{C}}^{-1} + \bar{\mathbf{C}}^{-1} \otimes \left(\frac{\partial^2 \bar{\Psi}}{\partial \bar{\mathbf{C}} \partial \bar{\mathbf{C}}} : \bar{\mathbf{C}}\right)\right] \\ &+ \frac{4}{9}\left(\bar{\mathbf{C}} : \frac{\partial^2 \bar{\Psi}}{\partial \bar{\mathbf{C}} \partial \bar{\mathbf{C}}} : \bar{\mathbf{C}}\right)\bar{\mathbf{C}}^{-1} \otimes \bar{\mathbf{C}}^{-1}.\end{aligned} \tag{12}$$

Note that its spatial counterpart of (9) is obtained from the application of the push-forward operation to (9) $\mathbf{c} = J^{-1}\boldsymbol{\chi}_*(\mathbf{C})$ [10]. Hence

$$\mathbf{c} = \mathbf{c}_{\text{vol}} + \bar{\mathbf{c}}, \tag{13}$$

where

$$\mathbf{c}_{\text{vol}} = (\tilde{p}\mathbf{1} \otimes \mathbf{1} - 2p\mathsf{I}). \tag{14}$$

The deviatoric term, $\bar{\mathbf{c}}$, can be obtained using the expression

$$\bar{\mathbf{c}} = \frac{2}{3}tr(\tilde{\boldsymbol{\sigma}})_{\mathsf{P}} - \frac{2}{3}(\mathbf{1} \otimes dev(\bar{\boldsymbol{\sigma}}) + dev(\bar{\boldsymbol{\sigma}}) \otimes \mathbf{1}) + \bar{\mathbf{c}}_{\bar{w}}, \tag{15}$$

where $\bar{\mathbf{c}}_{\bar{w}}$ in (15) is the weighted push forward of $\bar{\mathbf{C}}_{\bar{w}}$

$$\bar{\mathbf{c}}_{\bar{w}} = {}_{\mathsf{P}} : \bar{\mathbf{c}} : {}_{\mathsf{P}}. \tag{16}$$

For a more detailed derivation of the material and spatial elasticity tensors for fully incompressible or compressible fibered hyperelastic materials and their explicit expressions, see i.e. [30] or [50].

2.1 Phenomenological Soft Tissue Models

Following this approach, different strain energy functions have been proposed in order to take into account both the isotropy related to the solid matrix and the anisotropy introduced by the collagen fibers.

The most used transverse model for ligaments is the one proposed by Weiss [68]. The strain energy function for quasi-incompressible material was divided into an isotropic part (F_1) that corresponds to a Neo–Hookean model and other depending on the collagen fibers (F_λ).

$$\bar{\Psi} = c_1[\bar{I}_1 - 3] + F_\lambda(\lambda), \tag{17}$$

where c_1 is the Neo–Hookean constant and D the inverse of the bulk modulus $k = \frac{1}{D}$. Following physical observations in human ligaments, they assumed that collagen fibers do not support compressive loads. Secondly, the stress-strain relation curves for ligaments have two well-defined parts: an initial curve with increasing stiffness (toe region) and a second part with stiffness almost constant (linear region) [67]. The derivatives of the term of the free-energy function related to the fibers were initially proposed by Weiss et al. [68] as:

$$\begin{aligned} \lambda\bar{\Psi}_\lambda &= 0 \qquad \lambda < 1 \\ \lambda\bar{\Psi}_\lambda &= c_3[\exp(c_4[\lambda - 1]) - 1] \qquad \lambda < \lambda^* \\ \lambda\bar{\Psi}_\lambda &= c_5\lambda + c_6 \qquad \lambda > \lambda^*, \end{aligned} \tag{18}$$

where $\bar{\Psi}_\lambda = \frac{\partial F_\lambda}{\partial \lambda}$, λ^* is the stretch at which collagen fibers start to be straightened, changing $\bar{\Psi}_\lambda$ from exponential to linear, c_3 scales the exponential stress, c_4 is related to the rate of collagen uncrimping and c_5 is the elastic modulus of the straightened collagen fibers.

Pioletti et al. [56] proposed an isotropic SEF for ligaments that was later modified in order to consider the anisotropy of the soft tissues by Natali et al. [42]. It consists in an exponential function where $c_1 > 0$ and $c_2 > 0$ are stress-like parameters, and $c_3 > 0$ is dimensionless

$$\bar{\Psi}(\mathbf{C}, \mathbf{M}) = c_1[\bar{I}_1 - 3] + \frac{c_2}{c_3}[\exp(c_3[\bar{I}_4 - 1]) - c_3[\bar{I}_4 - 1] - 1]. \tag{19}$$

The Weiss's strain-energy functions (SEF) (17) was modified by Calvo et al. [14] in order to obtain an analytical expression for the strain energy function as shown in Eq. (20) to model ligaments and passive behaviour of muscular tissue

$$\begin{aligned}
\bar{\Psi} &= c_1[\bar{I}_1 - 3] + \bar{\Psi}_f \\
\bar{\Psi}_f &= 0 \qquad \bar{I}_4 < \bar{I}_{4_0} \\
\bar{\Psi}_f &= \frac{c_3}{c_4}[\exp(c_4[\bar{I}_4 - \bar{I}_{4_0}]) - c_4[\bar{I}_4 - \bar{I}_{4_0}] - 1] \qquad \bar{I}_4 > \bar{I}_{4_0} \text{ and } \bar{I}_4 < \bar{I}_{4_{ref}} \\
\bar{\Psi}_f &= 2c_5\sqrt{\bar{I}_4} + c_6 \ln(\bar{I}_4) + c_7 \qquad \bar{I}_4 > \bar{I}_{4_{ref}},
\end{aligned} \tag{20}$$

where $\bar{I}_{4_{ref}}$ characterizes the stretch at which collagen fibers start to be straightened, $c_1 > 0$, $c_3 > 0$, $c_5 > 0$, and $c_6 > 0$ are stress-like parameters, $c_4 > 0$ is dimensionless and c_7 is a strain parameter. Moreover, it has been assumed that the strain energy corresponding to the anisotropic terms only contributes to the global mechanical response of the tissue when stretched, that is $\bar{I}_4 > \bar{I}_{4_0}$. Note that c_5, c_6 and $c_7 > 0$ are dependent parameters that enforce strain, stress and stress derivative's continuity respectively.

To model cardiac tissue, Humphrey and Yin [35] proposed an exponential function

$$\bar{\Psi}(\bar{\mathbf{C}}, \lambda) = c[\exp(b[\bar{I}_1 - 3]) - 1] + d[\exp(a[\bar{\lambda} - 1]^2) - 1], \tag{21}$$

where c and d are stress-like and b and a dimensionless parameters.

Another transversely isotropic function for modeling the cardiac tissue was proposed by Lin and Yin [39]

$$\bar{\Psi}(\bar{\mathbf{C}}, \mathbf{M}) = c_1[\exp(Q) - 1], \tag{22}$$

with

$$Q = c_2[\bar{I}_1 - 3]^2 + c_3[\bar{I}_1 - 3][\bar{I}_4 - 1] + c_4[\bar{I}_4 - 1]^2, \tag{23}$$

where the material constants $c_2 > 0$, $c_4 > 0$, and c_3 are dimensionless, whereas $c_1 > 0$ is a stress-like parameter. Furthermore, note that the convexity of this function is not guaranteed for all the combinations of the constants in (23).

The most used SEF specifically designed for the arterial tissue that included two directions of anisotropy was proposed by Holzapfel et al. [30]

$$\bar{\Psi}(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N}) = \mu[\bar{I}_1 - 3] + \frac{k_1}{2k_2}[\exp(k_2[\bar{I}_4 - 1]^2) - 1] + \frac{k_3}{2k_4}[\exp(k_4[\bar{I}_6 - 1]^2) - 1], \tag{24}$$

where the parameters μ, k_1 and k_3 are stress-like, whereas k_2 and k_4 are dimensionless.

In order to include certain amount of fiber dispersion around the anisotropy directions characterized by the invariants $\bar{I}_4$ and $\bar{I}_6$, the SEF (24) model was modified in Holzapfel et al. [32]

$$\Psi(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N}) = \mu[\bar{I}_1 - 3] + \frac{k_1}{k_2}[\exp(k_2[[1-\rho][\bar{I}_1 - 3]^2 + \rho[\bar{I}_4 - 1]^2]) - 1] + \frac{k_3}{k_4}[\exp(k_4[[1-\rho][\bar{I}_1 - 3]^2 + \rho[\bar{I}_6 - 1]^2]) - 1]. \quad (25)$$

In this equation $\mu > 0$, $k_1 > 0$ and $k_3 > 0$ are stress-like parameters and $k_2 > 0$ and $k_4 > 0$ are dimensionless. This expression includes the exponential dependence on the factor $[\bar{I}_1 - 3]$ as well as the dimensionless weighting parameter $\rho \in [0, 1]$ in order to improve the ability of the original model (24). By means of this parameter it is possible to regulate the degree of anisotropy in such way that (24) is recovered when $\rho = 1$ whereas the isotropic exponential model is obtained when $\rho = 0$.

Regarding to the volumetric contribution to the SEF, the high water content of soft biological tissues has frequently leaded to the assumption of expressions with the form

$$\Psi_{\text{vol}} = k\,\mathcal{P}(J), \quad (26)$$

where $\mathcal{P}(J)$ is a strictly convex function satisfying $\mathcal{P}(1) = 0$.

Thus, compressibility is enforced as a function of the constant parameter k, and (26) renders a fully incompressible behaviour when $k \to \infty$. A wide variety of particular forms for the volumetric contribution to the strain energy function are found in the literature. For example, the functions

$$\mathcal{P}(J) = [J - 1]^2 \quad (27)$$

or

$$\mathcal{P}(J) = \ln(J) \quad (28)$$

and combinations of them are widely used in modelling the vascular tissue behaviour. As another example, similarly to (26), Ogden [43] proposed the relation

$$\Psi_{\text{vol}} = \beta^{-2}\left[\beta \ln(J) + J^{-\beta} + 1\right], \quad (29)$$

with $\beta > 0$.

2.2 *Micro-structurally Soft Tissue Models*

The models proposed for soft tissues could be classified into two groups. The first comprises the macroscopic models previously presented, in which a SEF is obtained disregarding the nature of the micro-structural components of the tissue. Second, a group of micro-structurally based models are presented in this section, in which the

macroscopic mechanical properties are obtained by assuming a constitutive relation for the microscopic components along each direction, whereas the macroscopic behaviour is obtained by integration of the contributions in all directions of space.

Gasser et al. [24] proposed the SEF

$$\bar{\Psi}(\mathbf{C},\mathbf{M},\mathbf{N}) = \mu[\bar{I}_1 - 3] + \frac{k_1}{2k_2}[\exp(k_2[\kappa \bar{I}_1 + [1-3\kappa]\bar{I}_4 - 1]^2) - 1] \tag{30}$$

$$+ \frac{k_3}{2\,k_4}[\exp(k_4[\kappa \bar{I}_1 + [1-3\kappa]\bar{I}_6 - 1]^2) - 1], \tag{31}$$

where $\kappa \in [0, 1/3]$ is a measure of the dispersion of the fibers around the preferred orientations This parameter is a result of considering the fibers oriented following the von Mises orientation density function. Thus, $\kappa = 1/3$ means isotropy and $\kappa = 0$ no fiber dispersion.

Histological studies performed in a number of soft tissues [18, 33] have shown that elastic fibers appear to be wavy and distributed about preferential directions [38]. Thus, as the load is applied, more and more fibers start to bear load. However, the degree of straightening of each fiber will also depend upon its orientation relative to the loading and the interstitial matter which might avoid its complete straightening. A model that consider the wavy nature of elastic fibers was proposed by Rodríguez et al. [60, 61]. Each bundle of fibers is assumed to behave following the worm-like eight-chain model proposed by Arruda and Boyce [8]

$$n\bar{\Psi}_{\mathrm{f}}(\bar{\lambda}) = \begin{cases} 0 & \text{if } \bar{\lambda} < 1 \\ B\left[2\,\dfrac{\bar{r}^2}{L^2} + \dfrac{1}{1-\bar{r}/L} - \dfrac{\bar{r}}{L}\right. & \\ \left. -\dfrac{\ln(\bar{\lambda}^{4\,r_0^2})}{4\,r_0\,L}\left[4\,\dfrac{r_0}{L} + \dfrac{1}{[1-r_0/L]^2} - 1\right] - \Psi_{\mathrm{r}}\right] & \text{if } \bar{\lambda} \geq 1, \end{cases} \tag{32}$$

with $B = \frac{1}{4}\,n\,K\,\Theta\,r_0/A$ a stress-like material parameter, L the maximum fiber length, r_0 the fiber length in the undeformed configuration, $\bar{r} = \bar{\lambda}\,r_0 < L$ the actual fiber length, $\bar{\lambda}$ the actual isochoric fiber stretch, and

$$\Psi_{\mathrm{r}} = 2\,\frac{r_0^2}{L^2} + \frac{1}{1-r_0/L} - \frac{r_0}{L}, \tag{33}$$

being a repository constant accounting for a zero strain energy at $\bar{\lambda} = 1$. This model considers the maximum fiber length, L, as a Beta random variable, and assumes the same average orientation for all fibers within the bundle as well as that fibers do not bear compressive loads. Hence, the strain energy density function for a bundle of fibers is given by

$$\bar{\Psi}_{\text{bun}}(\bar{\lambda},\, \bar{\lambda}_{t^*}) \;=\; \begin{cases} 0, & \bar{\lambda} < 1, \\ \displaystyle\int_1^{\bar{\lambda}} \int_{a(r_0 \bar{\lambda}_{t^*})}^{\iota} \bar{\Psi}'_{\text{f}}\,(\xi,\, x)\; \ell_L(x)\, \mathrm{d}x\, \mathrm{d}\xi, & \bar{\lambda} \geq 1, \end{cases} \tag{34}$$

where $a(r_0 \bar{\lambda}_{t^*})$ is a monotonically increasing function that determines the minimal fiber length within the bundle for which failure has not yet occurred,[1] $\Psi'_{\text{f}} = n\, \partial \Psi_{\text{f}} / \partial \bar{\lambda}$, and $\ell_L(x)$ is a Beta probability density function with parameters γ and η

$$\ell_L(x) \;=\; \frac{1}{\iota - r_0}\, \frac{\Gamma(\eta + \gamma)}{\Gamma(\eta)\, \Gamma(\gamma)} \left[\frac{x - r_0}{\iota - r_0} \right]^{\gamma - 1} \left[1 - \frac{x - r_0}{\iota - r_0} \right]^{\eta - 1}, \; x \in [r_0, \iota]. \tag{35}$$

The parameter $\bar{\lambda}_{t^*}$ in (34) corresponds to the maximum isochoric fiber stretch attained by the bundle over the past history up to time $t \in \mathcal{T}_+$. Therefore, the damage of the fiber bundle increases whenever $\bar{\lambda}_t - \bar{\lambda}_{t^*} \geq 0$ and, therefore, it is strain driven. On the other hand, function $a(r_0 \bar{\lambda}_{t^*})$ determines the minimum fiber length within the bundle for which failure has not yet occurred, and is given by

$$a(r_0 \bar{\lambda}_{t^*}) \;=\; \exp\left(\left[\frac{r_0 \bar{\lambda}_{t^*}}{\delta} \right]^{\varpi} \right) r_0 \bar{\lambda}_{t^*}, \tag{36}$$

where ϖ and δ are dimensionless model parameters. Note that with this form of $a(r_0 \bar{\lambda}_{t^*})$, the bundle will degrade faster as the deformation gets larger (i.e., longer fiber will fail at a smaller fraction of their maximum length).

With these considerations at hand, fiber damage is quantified as

$$\begin{aligned} D_{\text{f}} &= \frac{1}{\iota - r_0}\, \frac{\Gamma\,(\eta + \gamma)}{\Gamma\,(\eta) + \Gamma\,(\gamma)} \int_{r_0}^{a(r_0\, \bar{\lambda}_t)} \left[\frac{x - r_0}{\iota - r_0} \right]^{\gamma - 1} \left[1 - \frac{x - r_0}{\iota - r_0} \right]^{\eta - 1} \mathrm{d}x \\ &= \text{Beta}\left(\frac{a(r_0\, \bar{\lambda}_t)}{\iota - r_0},\, \gamma,\, \eta \right). \end{aligned} \tag{37}$$

Anisotropy can straightforwardly introduced in micro-structurally based models by considering an orientation density function, ρ, weighting the contribution of the fibers in space

$$\bar{\Psi} \;=\; \frac{3}{4\, \pi} \int_{\mathbb{U}^2} \rho\, \bar{\Psi}\, \mathrm{d}A. \tag{38}$$

First contributions considering this approach are due to Lanir [37], who proposed a structural model for planar tissues assuming that fibers are arranged

[1] Notice that x is a dummy variable used for integration purposes.

in three-dimensional but almost planar wavy array. Thus, collagen fibers were restricted to a plane in which they were oriented following a Gaussian distribution around a mean preferred direction. The same assumption was adopted for the elastin fibers, which were oriented following a different distribution.

More recently, Alastrué et al. [6] proposed a hyperelastic microsphere-based model with statistically distributed fibers. In that model, it is assumed the existence of a uniaxial orientation distribution function $\rho(\mathbf{r};\mathbf{a}) = \rho(-\mathbf{r};\mathbf{a})$ for $\mathbf{r} \in \mathbb{U}^2$ a referential unit vector and $\mathbb{U}^2$ the unit sphere surface. The macroscopic strain energy density corresponding to one family of fibers associated with the so-called preferred direction $\mathbf{a}$ and with n fibers per unit volume is then defined as

$$\bar{\Psi}_{\mathrm{f}} = \sum_{i=1}^{n} \rho(\mathbf{r}^i;\mathbf{a})\, \bar{\Psi}_{\mathrm{f}}^i, \tag{39}$$

where $\mathbf{r}^i$ are referential unit vectors associated with the direction of the i-th fiber, and Ψ_{f}^i is the fiber's strain energy according to the deformation in the direction of $\mathbf{r}^i$. When expanding this expression in order to account for N preferred orientations $\mathbf{a}_I$ related to different families of fibers one obtains

$$\bar{\Psi}_{\mathrm{ani}} = \sum_{I=1}^{N} \Psi_{\mathrm{f}}^I = \sum_{I=1}^{N} \langle n\, \rho_I\, \Psi_{\mathrm{f}}(\bar{\lambda}) \rangle = \frac{1}{4\,\pi} \int_{\mathbb{U}^2} n\, \rho_I\, \Psi_{\mathrm{f}}\, \mathrm{d}A. \tag{40}$$

Apart from the symmetry condition $\rho(\mathbf{r};\mathbf{a}) = \rho(-\mathbf{r};\mathbf{a})$ it was considered that fibers are rotationally symmetrically distributed with respect to the preferred mean orientation $\mathbf{a}$—in other words, $\rho(\mathbf{Q}\cdot\mathbf{r};\mathbf{a}) = \rho(\mathbf{r};\mathbf{a})\ \forall\ \mathbf{Q} \in \mathbb{Q}_+^3$ with rotation axis $\mathbf{a}$. As a consequence of the uniaxial distribution assumed for the one family of fibers considered, ρ can be defined as a function of the so-called mismatch angle $\omega = \arccos(\mathbf{a}\cdot\mathbf{r})$.

In Alastrué et al. [6], it was adopted the frequently applied π-periodic von Mises orientation distribution function (ODF)

$$\rho(\theta) = 4\sqrt{\frac{b}{2\pi}}\, \frac{\exp\left(b\,[\cos(2\,\theta)+1]\right)}{\mathrm{erfi}(\sqrt{2\,b})}, \tag{41}$$

where the positive concentration parameter b constitutes a measure of the degree of anisotropy. Moreover, $\mathrm{erfi}(x) = -i\,\mathrm{erf}(x)$ denotes the imaginary error function with $\mathrm{erf}(x)$ given by

$$\mathrm{erf}(x) = \sqrt{\frac{2}{\pi}} \int_0^x \exp(-\xi^2)\, \mathrm{d}\xi. \tag{42}$$

Recently, the ODF Bingham [9] was proposed by Alastrué et al. [5] to account for the dispersion of the collagen fibrils with respect to their preferential orientation. That function is expressed as

$$\rho(\boldsymbol{r};\,\boldsymbol{A})\,\frac{\mathrm{d}A}{4\,\pi}\,=\,[K(\boldsymbol{A})]^{-1}\,\exp\left(\boldsymbol{r}^{\mathrm{t}}\cdot\boldsymbol{A}\cdot\boldsymbol{r}\right)\frac{\mathrm{d}A}{4\,\pi},\tag{43}$$

where $\boldsymbol{A}$ is a symmetric 3×3 matrix, $\mathrm{d}A$ is the Lebesgue invariant measure on the unit sphere, $\boldsymbol{r}\,\in\,\mathbb{U}^2$ and $K(\boldsymbol{A})$ is a normalizing constant. As its main features, it is worth noting that this distribution always exhibits antipodal symmetry, but not rotational symmetry for the general case.

Applying straightforward transformations, Eq. (43) can be rewritten as

$$\rho(\boldsymbol{r};\,\boldsymbol{Z},\,\boldsymbol{Q})\,\frac{\mathrm{d}A}{4\,\pi}\,=\,[F_{000}(\boldsymbol{Z})]^{-1}\,\mathrm{etr}\left(\boldsymbol{Z}\cdot\boldsymbol{Q}^{\mathrm{t}}\cdot\boldsymbol{r}\cdot\boldsymbol{r}^{\mathrm{t}}\cdot\boldsymbol{Q}\right)\frac{\mathrm{d}A}{4\,\pi},\tag{44}$$

where $\mathrm{etr}\,(\bullet)\,\equiv\,\exp\left(\mathrm{tr}\,(\bullet)\right)$, $\boldsymbol{Z}$ is a diagonal matrix with eigenvalues $\kappa_{1,2,3}$, $\boldsymbol{Q}\,\in\,\mathbb{Q}^3$ such that $\boldsymbol{A}\,=\,\boldsymbol{Q}\cdot\boldsymbol{Z}\cdot\boldsymbol{Q}^{\mathrm{T}}$ and $F_{000}(\boldsymbol{Z})$ may be written as

$$F_{000}(\boldsymbol{Z})\,=\,[4\,\pi]^{-1}\int_{\mathbb{U}^2}\mathrm{etr}\left(\boldsymbol{Z}\cdot\boldsymbol{r}\cdot\boldsymbol{r}^{\mathrm{t}}\right)\mathrm{d}A\,=\,{}_1F_1(\frac{1}{2};\,\frac{3}{2};\,\boldsymbol{Z}),\tag{45}$$

with ${}_1F_1$ a confluent hypergeometric function of matrix argument as defined by Herz [27].

Thus, the probability concentration is controlled by the eigenvalues of $\boldsymbol{Z}$, which might be interpreted as concentration parameters. Specifically, the difference between pairs of $\kappa_{1,2,3}$—i.e., $[\kappa_1-\kappa_2]$, $[\kappa_1-\kappa_3]$ and $[\kappa_2-\kappa_3]$—determines the shape of the distribution over the surface of the unit sphere. Therefore, the value of one of these three parameters may be fixed to a constant value without reducing the versatility of (44). In fact, setting two of the parameters equal to zero the Von Mises ODF is obtained and when two parameters come close up, a rotational symmetry is achieved.

3 Weak Form and Linearized Weak Form of the Continuum Problem in Spatial Description

The principle of virtual work can be defined by the current or the reference configurations. The spatial version is written as:

$$\delta W(\mathbf{u},\delta\mathbf{u})=\delta W_{int}(\mathbf{u},\delta\mathbf{u})+\delta W_{ext}(\mathbf{u},\delta\mathbf{u}),\tag{46}$$

where

$$\delta W_{int}(\mathbf{u},\delta\mathbf{u})=\int_{\Omega}\boldsymbol{\sigma}:\delta\mathbf{e}\mathrm{d}v,\tag{47}$$

with $\mathbf{e} = \frac{1}{2}(\mathbf{I} - \mathbf{F}^{-T}\mathbf{F}^{-1})$ the Euler-Almansi strain tensor, and

$$\delta W_{ext}(\mathbf{u}, \delta\mathbf{u}) = \int_{\Omega} \rho(\mathbf{b} - \ddot{\mathbf{u}}).\delta\mathbf{u}\mathrm{d}v + \int_{\partial\Omega_\sigma} \bar{\mathbf{t}}.\delta\mathbf{u}\mathrm{d}s. \tag{48}$$

We shall consider a purely static problem, so that $\ddot{\mathbf{u}} = 0$. In addition, we assume that the load (the body force $\mathbf{b}$ and the surface traction $\bar{\mathbf{t}}$) are 'dead' (independent of the deformation), so that the linearization of the external virtual work vanishes, i.e. $D_{\triangle\mathbf{u}}\delta W_{ext}(\mathbf{u}, \delta\mathbf{u}) = \mathbf{0}$. Hence, the linearization of the variational equation (46) only affects the internal virtual work δW_{int}, which we shall consider below. The idea is first to pull-back the spatial quantities to the reference configuration (internal virtual work in the material description), then to linearize and to push-forward again. Starting with the equivalence pull-back

$$\delta W_{int}(\mathbf{u}, \delta\mathbf{u}) = \int_{\Omega} \sigma(\mathbf{u}) : \delta\mathbf{e}(\mathbf{u})\mathrm{d}v = \int_{\Omega_0} \mathbf{S}(\mathbf{E}(\mathbf{u})) : \delta\mathbf{E}(\mathbf{u})\mathrm{d}V, \tag{49}$$

with $\mathbf{E} = \frac{1}{2}(\mathbf{F}^T\mathbf{F} - \mathbf{I})$ the Green–Lagrange strain tensor, we consider now the linearization of the internal virtual work in the material description

$$D_{\triangle\mathbf{u}}\delta W_{int}(\mathbf{u}, \delta\mathbf{u}) = \int_{\Omega_0} [\delta\mathbf{E}(\mathbf{u}) : D_{\triangle\mathbf{u}}\mathbf{S}(\mathbf{E}(\mathbf{u})) + \mathbf{S}(\mathbf{E}(\mathbf{u})) : D_{\triangle\mathbf{u}}\delta\mathbf{E}(\mathbf{u})]\mathrm{d}V. \tag{50}$$

The first term corresponds to the material stiffness matrix and the second to the geometric part of the stiffness matrix. We can write

$$D_{\triangle\mathbf{u}}\delta W_{int}(\mathbf{u}, \delta\mathbf{u}) = \int_{\Omega_0} [\delta\mathbf{E}(\mathbf{u}) : \mathsf{C}(\mathbf{u}) : D_{\triangle\mathbf{u}}\mathbf{E}(\mathbf{u}) + \mathbf{S}(\mathbf{E}(\mathbf{u})) : D_{\triangle\mathbf{u}}\delta\mathbf{E}(\mathbf{u})]\mathrm{d}V. \tag{51}$$

Considering the push-forward operations already derived, from (51) and taking into account the relation $dv = JdV$, the linearized virtual work in the spatial description may be written as [32]

$$\begin{aligned} D_{\triangle\mathbf{u}}\delta W_{int}(\mathbf{u}, \delta\mathbf{u}) &= \int_{\Omega} \left(\frac{\partial\delta u_a}{\delta x_b}\mathsf{c}_{abcd}\frac{\partial\triangle u_c}{\delta x_d} + \frac{\partial\delta u_a}{\delta x_b}\frac{\partial\triangle u_a}{\delta x_d}\sigma_{bd}\right)\mathrm{d}v \\ &= \int_{\Omega} \frac{\partial\delta u_a}{\delta x_b}(\mathsf{c}_{abcd} + \delta_{ac}\sigma_{bd})\frac{\partial\triangle u_c}{\delta x_d}\mathrm{d}v, \end{aligned} \tag{52}$$

where $(\mathsf{c} + \boldsymbol{\delta} \otimes \boldsymbol{\sigma})$ represents the effective elasticity tensor that includes the material and geometric parts of the consistent tangent stiffness matrix.

4 Residual Stresses

Biological soft tissues are usually exposed to a complex distribution of "in vivo" initial strains or residual stresses. This state is a consequence of the continuous growth, remodeling, damage and viscoelastic strains that suffer these living materials along their whole life. Usually, the distinction between residual and initial strains or stresses refers to their origin. In the context of living tissues, residual strains are commonly considered a consequence of different phenomena such as growing or adaptation. On the contrary, initial stresses have a more general sense, and can be considered as any autobalanced stress distribution defined in the reference configuration, independently of its origin. Their most important aim is to homogenize the stress distribution at different stages of tissue deformation. For example, in arteries, their effect is to decrease the circumferential stresses at the inner wall and to reduce the stress gradient through the arterial thickness [15]. It has been assumed by different authors that the physiological state of a healthy artery requires constant circumferential stress in each layer. This situation is only possible by the presence of initial stresses [59]. In ligaments of diarthrodial joints, initial stretches provide joint stability even in a relatively unloaded joint configuration [23]. Typical residual strains are approximately 3–5 % in ligaments and 20 % for arteries. Initial strains can be relieved by selective cutting of the living tissue and removal of its internal constraints. Due to the non-linear behavior of this kind of materials and the non-uniform distribution of the residual stresses, a wrong inclusion of the initial strain state in computational models of soft tissues can lead to large errors (usually an important underestimation of the stress level).

The most common procedure to incorporate residual stresses in computational models consists on the numerical simulation of the reverse of the stress free configuration. In order to apply this methodology, a suitable discretisation of the assumed stress-free geometry is considered as the starting configuration for the analysis. This strategy has some drawbacks. First, it becomes an extremely complicated task to determine the displacement field required to get the closed geometry from the stress geometry when dealing with initial geometries. Another key disadvantage inherent to this methodology is related to the simulation of patient-specific geometries for clinical purposes. Since it requires starting from the stress free configuration, it cannot be applied without obtaining the open geometry, which implies the destruction of the soft tissue.

An alternative method which overcomes some of the disadvantages of the previous method is the Finite Element implementation of the deformation gradient tensor decomposition strategy represented in Fig. 3 [1,4,48]. There, $\mathcal{B}_{\text{rs}}$ corresponds to the closed starting configuration, namely the configuration corresponding to the geometry obtained from medical imaging techniques, so it constitutes the initial configuration on which external loads represented by the deformation $\boldsymbol{\varphi} : \mathcal{B}_{\text{rs}} \times \mathcal{T} \rightarrow \mathbb{R}^3$ can be applied.

The two-point tensor $\boldsymbol{F}_{\text{oa}}$, that maps elements of the tangent space $T\mathcal{B}_0$ associated to the open configuration $\mathcal{B}_0$ onto $T\mathcal{B}_{\text{rs}}$, is assumed to account for an

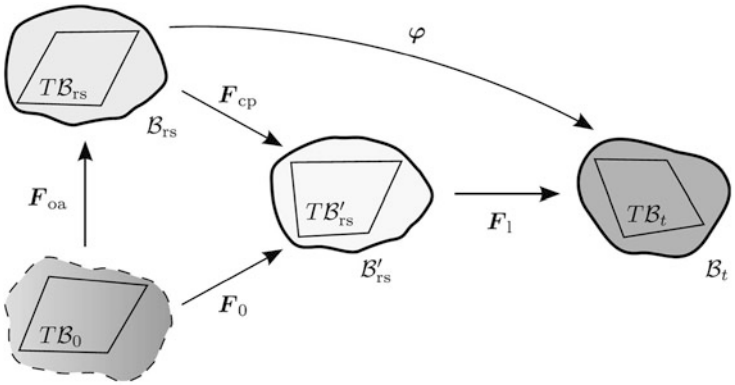


Fig. 3 Graphical illustration of the transformation operations and configurations related to the overall deformation and the incorporation of residual strains. Adapted from [1,4]

arbitrary residual strains field coming, e.g., from the opening angle experiment. $T\mathcal{B}_0$ is associated to an, in general, unknown, say virtual, open configuration. Due to this reason, the application of $\boldsymbol{F}_{\mathrm{oa}}$ on $T\mathcal{B}_0$, provides a non-equilibrated residual stress configuration, namely $\mathcal{B}_{\mathrm{rs}}$, that constitutes an intermediate step in obtaining a compatible residually stressed configuration $\mathcal{B}'_{\mathrm{rs}}$, reached by means of the application of the elastic deformation[2] represented by $\boldsymbol{F}_{\mathrm{cp}}$. Thus, $\mathcal{B}'_{\mathrm{rs}}$ includes the information from $\boldsymbol{F}_0 = \boldsymbol{F}_{\mathrm{cp}} \cdot \boldsymbol{F}_{\mathrm{oa}}$.

Note that, a good choice of $\boldsymbol{F}_{\mathrm{oa}}$ will cause a very small geometrical modification of $\mathcal{B}_{\mathrm{rs}}$, so that the total residual strain field $\boldsymbol{F}_0$ due to residual strains will mainly correspond to the estimated $\boldsymbol{F}_{\mathrm{oa}}$. In other words, it is convenient that the residual stress distribution present in $\mathcal{B}'_{\mathrm{rs}}$ to be, as much as possible, the consequence of $\boldsymbol{F}_{\mathrm{oa}}$, so that $\boldsymbol{F}_{\mathrm{cp}} \simeq \boldsymbol{I}$ has to be satisfied in order to obtain accurate results. Finally, $\boldsymbol{F}_1$, representing the deformation gradient tensor associated to load, is applied so that the residually-stressed loaded configuration $\mathcal{B}_t$ is obtained.

To introduce initial strains into the finite element formulation, it is necessary to specify $\boldsymbol{F}_{\mathrm{oa}}$ pointwise within the finite element mesh. An equilibrium step is firstly applied with zero forces with the constitutive behaviour defined by $\Psi_{\Omega_{sf}}$ in order to obtain a balanced, although not fully compatible configuration. A second load step will result in the deformation gradient $\mathbf{F}$ that balances the externally applied forces [3,50].

$\boldsymbol{F}_{\mathrm{oa}}$ is difficult to determine from experiments. In the case of ligaments and tendons, Gardiner et al. [23] proposed a relatively easy form to measure length variations along the fiber direction at different points, λ_{oa}. They assumed that $\mathbf{F}_{\mathrm{oa}}$ corresponds to an axial stretch λ_{oa} along the fiber direction $\mathbf{m}_0$ in the reference

[2]Note that, thinking about the numerical implementation of this procedure, the elastic strain tensor $\boldsymbol{F}_{\mathrm{cp}}$ corresponds to the strain field associated to the displacement needed to make $\mathcal{B}_{\mathrm{rs}}$ to satisfy the equilibrium equations. Thus, it constitutes an output of the Finite Element Method.

Fig. 4 Arterial ring before carrying out the cut in the radial direction and after releasing of the residual stresses

state $\mathcal{B}_0$ that in ligaments closely follows the direction of maximal length. The concomitant contraction in the perpendicular plane is dictated by incompressibility, usual assumption in biological soft tissues.

In a coordinate system (*) where the fiber direction $\mathbf{m}_0$ is aligned with the X_1 axis, $\mathbf{F}_{\text{oa}}$ can be written as:

$$\left[\boldsymbol{F}^{*}_{\text{oa}}\right] = \begin{bmatrix} \lambda_{\text{oa}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\lambda_{\text{oa}}}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\lambda_{\text{oa}}}} \end{bmatrix} \tag{53}$$

and transformed to the global systems:

$$\boldsymbol{F}_{\text{oa}} = \mathbf{R}\boldsymbol{F}^{*}_{\text{oa}} \tag{54}$$

with $\mathbf{R}$ the rotation tensor from this local system to the global one [50].

A number of feasible alternatives exist to determine $\boldsymbol{F}_{\text{oa}}$ for blood vessels. Residual stresses present in blood vessels, their functional role is to homogenize the circumferential stress distribution, optimizing the mechanical performance of the vessel wall under physiological loads [20]. The usual procedure to measure longitudinal strain is to compare the "in vivo" and the "in vitro" vessel lengths [16]. On the other hand, circumferential strain field is usually quantified by performing the so-called opening angle experiment (see Fig. 4), which consists on performing a radial cut on a vessel ring. Then, the ring springs open releasing circumferential residual stresses which can be estimated as a function of the ring opening angle [21].

Most large and middle size blood vessels show a conduit-type geometry, which has frequently leaded to consider them as purely cylindrical. This assumption has been widely used in the analytical modelling of residual stresses. Considering a cylindrical geometry, the solution of the problem consisting on closing an open thick-walled cylindrical geometry subjected to pure bending constitutes one of the most widely used approximations in order to analytically determine the residually strained configuration. A schematic representation of the named problem is given in Fig. 5. There, $\mathcal{B}_{\text{op}}$ constitutes the stress-free open configuration that transforms to the close configuration $\mathcal{B}_{\text{cl}}$ by means of the application of a suitable deformation

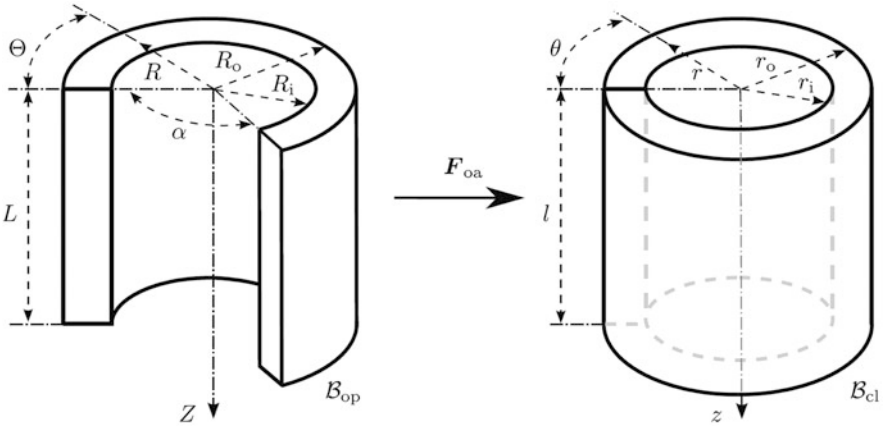


Fig. 5 Arterial ring in the open ($\mathcal{B}_{op}$) and closed ($\mathcal{B}_{cl}$) configurations. Adapted from [4]

mapping, $\boldsymbol{F}_{oa}$. A cylindrical coordinate system $\{\boldsymbol{E}_I, I = R, \Theta, Z\}$ is defined in $\mathcal{B}_{op}$, where the geometrical constraints

$$R_i \leq R \leq R_o, \qquad 0 \leq \Theta \leq (2\pi - \alpha), \qquad 0 \leq Z \leq L, \quad (55)$$

must be satisfied, with R_i and R_o denoting the inner and outer radii, respectively, α is the opening angle, and L the length of the open cylinder (see the $\mathcal{B}_{op}$ configuration in Fig. 5).

The cylindrical coordinate system $\{\boldsymbol{e}_i, i = r, \theta, z\}$ belonging to $\mathcal{B}_{cl}$ is analogously defined, so the new geometrical constraints can be written as

$$r_i \leq r \leq r_o, \qquad 0 \leq \theta \leq 2\pi, \qquad 0 \leq z \leq l, \quad (56)$$

where r_i, r_o and l denote the inner and outer radii and length of the arterial ring in $\mathcal{B}_{cl}$, respectively.

At this point, imposition of the incompressibility constraint, namely $\det(\boldsymbol{F}_{oa}) = 1$, allows to express the values of the cylindrical coordinates in the $\mathcal{B}_{cl}$ configuration as

$$r = \sqrt{\frac{R^2 - R_i^2}{\kappa \lambda_z} + r_i^2}, \qquad \theta = \kappa \Theta, \qquad z = \lambda_z Z, \quad (57)$$

where λ_z, assumed constant, is the axial stretch, and the parameter $\kappa = \frac{2\pi}{(2\pi - \alpha)}$ is a measure of the opening angle of the ring [16]. The principal stretches in radial and circumferential direction can be expressed as

$$\lambda_r(R) = \frac{\partial r}{\partial R} = \frac{R}{r \kappa \lambda_z}, \qquad \lambda_\theta(R) = \frac{r}{R} \frac{\partial \theta}{\partial \Theta} = \frac{\kappa r}{R}. \quad (58)$$

Finally, the two-point tensor $\boldsymbol{F}_{\text{oa}}$ can be written as a function of principal stretches and the unit vectors of the basis declared in the open $\mathcal{B}_{\text{op}}$ configuration and in the closed $\mathcal{B}_{\text{cl}}$ configuration as

$$\boldsymbol{F}_{\text{oa}} = \lambda_r \, \boldsymbol{e}_r \otimes \boldsymbol{E}_R + \lambda_\theta \, \boldsymbol{e}_\theta \otimes \boldsymbol{E}_\Theta + \lambda_z \, \boldsymbol{e}_z \otimes \boldsymbol{E}_Z. \tag{59}$$

5 Inelastic Effects

Many fibred soft tissues exhibit simultaneously elastic and viscous material behavior. The rate-dependent material behavior of this kind of materials has been well-documented and quantified in the literature. This includes works on ligaments [69], blood vessels [34], cornea [55] and articular cartilage [26]. This behavior can arise from the fluid flow inside the tissue, from the inherent viscoelasticity of the solid phase, or from viscous interactions between the tissue phases. Furthermore, non-physiological loads drive soft tissue to damage that may induce a strong reduction of the stiffness. Damage may arise from two possible mechanisms: tear or plastic deformation of the fibres, or biochemical degradation of the extracellular matrix from protease release associated with the observed cellular necrosis.

5.1 Time-Dependent Response

In order to describe the viscoelastic response, the finite-strain anisotropic viscoelastic constitutive model proposed by Peña et al. [49] is here considered. The concept of internal variable [62] is also applied, postulating the existence of an uncoupled free energy density function Ψ in the form

$$\Psi(\mathbf{C}, \mathbf{M}, \mathbf{N}, \mathbf{Q}_{ik}) = \Psi^0_{\text{vol}}(J) + \bar{\Psi}^0 - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=m,f_1,f_2} [\bar{\mathbf{C}} : \mathbf{Q}_{ik}], \tag{60}$$

where Ψ^0_{vol} and $\bar{\Psi}^0$ are the effective volumetric and is isochoric elastic responses [66], $J > 0$ the local volume variation or jacobian [40], and $\mathbf{C}$ and $\bar{\mathbf{C}}$ are the right and left Cauchy–Green deformation tensors and their isochoric counterparts [66]. The internal variables $\mathbf{Q}_{ik}$ may be interpreted as non-equilibrium stresses, in the sense of non-equilibrium thermodynamics that remain unaltered under superposed spatial rigid body motions. $\mathbf{Q}_{im}$ are the isotropic contributions due to the matrix material, and $\mathbf{Q}_{if}$ is the anisotropic contribution due to the family of fibres [49].

Standard arguments based on the Clausius–Duhem inequality lead to the representation

$$\mathbf{S} = 2\frac{\partial \Psi(\mathbf{C}, \mathbf{M}, \mathbf{N}, \mathbf{Q}_{ik})}{\partial \mathbf{C}} = \mathbf{S}_{\text{vol}} + \bar{\mathbf{S}}^0 - J^{-\frac{2}{3}} \sum_{i=1}^{n} \sum_{k=m,f_1,f_2} DEV[\mathbf{Q}_{ik}]. \tag{61}$$

The non-equilibrium second Piola–Kirchhof stress tensor in (61), $\mathbf{Q}_{ik}$, are then assumed to be governed by a set of rate equations [49]

$$\dot{\mathbf{Q}}_{ik} + \frac{1}{\tau_{ik}}\mathbf{Q}_{ik} = \frac{\gamma_{ik}}{\tau_{ik}} DEV[2\frac{\partial \bar{\Psi}_k^0(\bar{\mathbf{C}}, \mathbf{M}, \mathbf{N})}{\partial \bar{\mathbf{C}}}], \qquad \lim_{t\to\infty} \mathbf{Q}_{ik} = \mathbf{0}, \tag{62}$$

where $\bar{\Psi}_k^0$ is the isochoric strain energy associated to the k contribution and $\gamma_{ik} \in [0, 1]$ are dimensionless free energy factors associated with the relaxation times $\tau_{ik} > 0$.

The evolution equation (62) explicitly leads to the following convolution representation

$$\mathbf{Q}_{ik}(t) = \int_{-\infty}^{t} \frac{\gamma_{ik}}{\tau_{ik}} \exp\left(\frac{-[t-s]}{\tau_{ik}}\right) DEV[2\frac{\partial \bar{\Psi}_k^0}{\partial \bar{\mathbf{C}}}]\mathrm{d}s. \tag{63}$$

The evolution equations obtained are linear, discarding therefore the stretch-dependence of the relaxation rate which occurs in some kind of tissues [22, 69]. For those cases, we use a modified set of evolution equations that consider a set of strain-dependent reduced relaxation and time functions γ_{ik} and τ_{ik}, respectively, following the relationships below [54]

$$\gamma_{im}(\bar{I}_1) = \gamma_{im}^a e^{\gamma_{im}^b(\bar{I}_1-3)}, \quad \gamma_{if_1}(\bar{I}_4) = \gamma_{if_1}^a e^{\gamma_{if_1}^b(\bar{I}_4-1)}, \quad \gamma_{if_2}(\bar{I}_6) = \gamma_{if_2}^a e^{\gamma_{if_2}^b(\bar{I}_6-1)}, \tag{64}$$

$$\tau_{mi}(\bar{I}_1) = \tau_{im}^a e^{\tau_{im}^b(\bar{I}_1-3)}, \quad \tau_{if_1}(\bar{I}_4) = \tau_{if_1}^a e^{\tau_{if_1}^b(\bar{I}_4-1)}, \quad \gamma_{if_2}(\bar{I}_6) = \gamma_{if_2}^a e^{\gamma_{if_2}^b(\bar{I}_6-1)}, \tag{65}$$

where γ_{ik}^a, γ_{ik}^b and τ_{ik}^b are dimensionless parameters and τ_{ik}^a has a time dimension.

The corresponding evolution equations result as

$$\begin{aligned}
\mathbf{Q}_{im}(t) &= \int_{-\infty}^{t} \frac{\gamma_{im}^a e^{\gamma_{im}^b(\bar{I}_1-3)}}{\tau_{im}^a e^{\tau_{im}^b(\bar{I}_1-3)}} \exp[\frac{-(t-s)}{\tau_{im}^a e^{\tau_{im}^b(\bar{I}_1-3)}}]\ DEV[2\frac{\partial \bar{\Psi}_m^0}{\partial \bar{\mathbf{C}}}]\mathrm{d}s, \\
\mathbf{Q}_{if_1}(t) &= \int_{-\infty}^{t} \frac{\gamma_{if_1}^a e^{\gamma_{if_1}^b(\bar{I}_4-1)}}{\tau_{if_1}^a e^{\tau_{if_1}^b(\bar{I}_4-1)}} \exp[\frac{-(t-s)}{\tau_{if_1}^a e^{\tau_{if_1}^b(\bar{I}_4-1)}}]\ DEV[2\frac{\partial \bar{\Psi}_{f_1}^0}{\partial \bar{\mathbf{C}}}]\mathrm{d}s, \\
\mathbf{Q}_{if_2}(t) &= \int_{-\infty}^{t} \frac{\gamma_{if_2}^a e^{\gamma_{if_2}^b(\bar{I}_6-1)}}{\tau_{if_2}^a e^{\tau_{if_2}^b(\bar{I}_6-1)}} \exp[\frac{-(t-s)}{\tau_{if_2}^a e^{\tau_{if_2}^b(\bar{I}_6-1)}}]\ DEV[2\frac{\partial \bar{\Psi}_{f_2}^0}{\partial \bar{\mathbf{C}}}]\mathrm{d}s.
\end{aligned} \tag{66}$$

The basic idea in the numerical integration of the constitutive equations is to evaluate the convolution integral in (63) through a recursive relation. The key idea is

to transform the convolution representation discussed in the preceding section into a two-step recursive formula involving internal variables stored at the quadrature points of a finite-element mesh [63].

First at all, we introduce the following internal algorithmic history variables

$$\mathbf{H}^{(ik)} = \int_{-\infty}^{t} \exp[\frac{-(t-s)}{\tau_{ik}}]\frac{d}{ds}\left\{ DEV[2\frac{\partial \bar{\Psi}_k^0(\bar{\mathbf{C}},\mathbf{M},\mathbf{N})}{\partial \bar{\mathbf{C}}}(s)]\right\} \mathrm{d}s. \tag{67}$$

Let $[t_0, T] \subset \mathbb{R}$, with $t_0 < T$, be the time interval of interest. Without loss of generality, we take $t_0 = -\infty$. Further, let $[t_0, T] = \bigcup_{n\in\mathbb{I}}[t_n, t_{n+1}]$, be a partition of the interval $[t_0, T]$ with $\mathbb{I}$ an appropriate subset of the natural numbers and $\Delta t_n = t_{n+1} - t_n$ the associated time increment. From an algorithmic standpoint, the problem is defined in the usual strain-driven format and we assume that at certain times t_n and t_{n+1} all relevant kinematic quantities are known.

Using the semigroup property of the exponential function, the property of additivity of the integral over the interval of integration and the midpoint rule to approximate the integral over $[t_n, t_{n+1}]$ we can arrive to the update formula [63]

$$\mathbf{H}_{n+1}^{(ik)} = \exp[\frac{-\Delta t_n}{\tau_{ik}}]\mathbf{H}_n^{(ik)} + \exp[\frac{-\Delta t_n}{2\tau_{ij}}](\bar{\mathbf{S}}_{k_{n+1}}^0 - \bar{\mathbf{S}}_{k_n}^0), \tag{68}$$

where $\bar{\mathbf{S}}_{k_{n+1}}^0 = DEV[2\frac{\partial \bar{\Psi}_k^0(\bar{\mathbf{C}}_{n+1},\mathbf{M},\mathbf{N})}{\partial \bar{\mathbf{C}}}(s)]$ is the term of the initial stress response corresponding to I_k, i.e., $\bar{\mathbf{S}}_{1_{n+1}}^0$ is due to the matrix material and $\bar{\mathbf{S}}_{4_{n+1}}^0$ and $\bar{\mathbf{S}}_{6_{n+1}}^0$ are due to the fibers.

Following the convolution representation (63), the algorithmic approximation for the second Piola–Kirchhoff stress takes the form

$$\begin{aligned}\mathbf{S}_{n+1} = {} & J_{n+1}p_{n+1}\mathbf{C}_{n+1}^{-1} + J_{n+1}^{-\frac{2}{3}}\sum_{k=m,f_1,f_2}[(1-\sum_{i=1}^{n}\gamma_{ik})\bar{\mathbf{S}}_{k_{n+1}}^0]) \\ & + J_{n+1}^{-\frac{2}{3}}\sum_{k=m,f_1,f_2}\sum_{i=1}^{n}[\gamma_{ik}\{DEV[\mathbf{H}_{n+1}^{(ik)}]\}].\end{aligned} \tag{69}$$

Also, we can calculate the Cauchy stress tensor as:

$$\begin{aligned}\boldsymbol{\sigma}_{n+1} = {} & p_{n+1}\mathbf{1} + \frac{1}{J_{n+1}}\sum_{k=m,f_1,f_2}[(1-\sum_{i=1}^{n}\gamma_{ik})dev\left\{\bar{\mathbf{F}}_{n+1}[2\frac{\partial \bar{\Psi}_k^0(\bar{\mathbf{C}}_{n+1},\mathbf{M},\mathbf{N})}{\partial \bar{\mathbf{C}}}]\bar{\mathbf{F}}_{n+1}^T\right\}]) + \\ & + \frac{1}{J_{n+1}}\sum_{k=m,f_1,f_2}\sum_{i=1}^{n}[\gamma_{ik}\left\{dev[\bar{\mathbf{F}}_{n+1}[\mathbf{H}_{n+1}^{(ik)}]\bar{\mathbf{F}}_{n+1}^T]\right\}].\end{aligned} \tag{70}$$

The tangent modulus plays a crucial role in the numerical solution of the boundary value problem by Newton-type iterative methods [10]. The use of

consistently linearized moduli is essential to preserve the quadratic rate of the asymptotic convergence that characterizes full Newton's method [63]. In order to obtain an easier recursive update procedure, we rewrite the update formula (68) as

$$\tilde{\mathbf{H}}_n^{(ik)} = \exp[\frac{-\Delta t_n}{\tau_{ik}}]\mathbf{H}_n^{(ik)} - \exp[\frac{-\Delta t_n}{2\tau_{ik}}]\bar{\mathbf{S}}_{k_n}^0 \tag{71}$$

$$\mathbf{H}_{n+1}^{(ik)} = \tilde{\mathbf{H}}_n^{(ik)} + \exp[\frac{-\Delta t_n}{2\tau_{ik}}]\bar{\mathbf{S}}_{k_{n+1}}^0. \tag{72}$$

With this notation

$$\mathbf{S}_{n+1} = J_{n+1}p_{n+1}\mathbf{C}_{n+1}^{-1} + J_{n+1}^{-\frac{2}{3}} \sum_{k=m,f_1,f_2} [(1-\gamma_k+\nu_k)\bar{\mathbf{S}}_{k_{n+1}}^0 + \sum_{i=1}^{n} \gamma_{ik}\{DEV[\tilde{\mathbf{H}}_n^{(ik)}]\}], \tag{73}$$

$$\boldsymbol{\sigma}_{n+1} = p_{n+1}\mathbf{1} + \sum_{k=m,f_1,f_2} [(1-\gamma_k+\nu_k)dev[\boldsymbol{\sigma}_{k_{n+1}}^0] + \frac{1}{J_{n+1}} \sum_{i=1}^{n} \gamma_{ik}\{dev[\tilde{\mathbf{h}}_n^{(ik)}]\}], \tag{74}$$

where $\gamma_k = \sum_{i=1}^{n} \gamma_{ik}$ and $\nu_k = \sum_{i=1}^{n} \gamma_{ik} \exp[\frac{-\Delta t_n}{2\tau_{ik}}]$. Note that $\tilde{\mathbf{H}}_n^{(ik)}$ is a constant at time t_{n+1} in the linearization process.

Using (9) and (73) we obtain

$$\begin{aligned}\mathbf{C}_{n+1} =& \mathbf{C}_{\text{vol } n+1}^0 + \sum_{k=m,f_1,f_2} [(1-\gamma_k+\nu_k)\bar{\mathbf{C}}_{k_{n+1}}^0 + \\ &-\frac{2}{3}J_{n+1}^{-\frac{4}{3}} \sum_{i=1}^{n} \gamma_{ik}\{DEV[\tilde{\mathbf{H}}_n^{(ik)}] \otimes \bar{\mathbf{C}}_{n+1}^{-1} + \bar{\mathbf{C}}_{n+1}^{-1} \otimes DEV[\tilde{\mathbf{H}}_n^{(ik)}] - \\ &-(\tilde{\mathbf{H}}_n^{(ik)} : \bar{\mathbf{C}})(\mathbf{I}_{\mathbf{C}_{n+1}}^{-1} - \frac{1}{3}\bar{\mathbf{C}}_{n+1}^{-1} \otimes \bar{\mathbf{C}}_{n+1}^{-1})\}]\end{aligned} \tag{75}$$

and the spatial tangent modulus defined in (13) takes the form

$$\begin{aligned}\mathbf{c}_{n+1} =& \mathbf{c}_{\text{vol } n+1}^0 + \sum_{k=m,f_1,f_2} [(1-\gamma_k+\nu_k)\bar{\mathbf{c}}_{k_{n+1}}^0 + \\ &-\frac{2}{3J_{n+1}} \sum_{i=1}^{n} \gamma_{ik}\{dev[\tilde{\mathbf{h}}_n^{(ik)}] \otimes \mathbf{1}_{n+1} + \mathbf{1}_{n+1} \otimes dev[\tilde{\mathbf{h}}_n^{(ik)}] - \\ &-tr[\tilde{\mathbf{h}}_n^{(ik)}](\mathbf{I} - \frac{1}{3}\mathbf{1} \otimes \mathbf{1})\}]\end{aligned} \tag{76}$$

where $\tilde{\mathbf{h}}_n^{(ik)} = \bar{\mathbf{F}}_{n+1}\tilde{\mathbf{H}}_n^{(ik)}\bar{\mathbf{F}}_{n+1}^T$.

Note that in the recursive update procedure presented herein it is necessary to know at time t_{n+1} the variables $\mathbf{H}_n$ and $\bar{\mathbf{S}}^0_n$, so, the integration algorithm for the constitutive equations does not oblige to a constant time increment during the simulation. Frequently, biomechanical problems include highly large deformations and consequently it is convenient to use a variable time increment approach. In the case of Δt_n constant, it is not necessary to store $\bar{\mathbf{S}}^{0(j)}_n$ being only needed to compute and store $\tilde{\mathbf{H}}_n$ and the computational cost of the recursive update procedure is lower.

For the reader's convenience, we have summarized the overall implementation of the developed algorithm in the following scheme.

---

1. Database at each Gaussian point
   $\{\bar{\mathbf{S}}_{k_n},\ \{\mathbf{H}^{(ik)}_n\}\ i = 1\ldots$ internal variables and $k = m, f_1, f_2$
2. Compute the initial elastic stress Cauchy tensor
   $dev[\boldsymbol{\sigma}^0_{k_{n+1}}] = \frac{1}{J_{n+1}} dev\left\{\bar{\mathbf{F}}_{n+1}[2\frac{\partial\bar{\Psi}^0_k(\bar{\mathbf{C}}_{n+1},\mathbf{M},\mathbf{N})}{\partial\bar{\mathbf{C}}}]\bar{\mathbf{F}}^T_{n+1}\right\}$
3. Update algorithmic internal variables
   $\bar{\mathbf{S}}^0_{k_{n+1}} = \bar{\mathbf{F}}_{n+1}(J_{n+1} dev[\boldsymbol{\sigma}^0_{k_{n+1}}])\bar{\mathbf{F}}^T_{n+1}$
   $\tilde{\mathbf{H}}^{(ik)}_n = \exp[\frac{-\Delta t_n}{\tau_{ik}}]\mathbf{H}^{(ik)}_n - \exp[\frac{-\Delta t_n}{2\tau_{ij}}]\bar{\mathbf{S}}^0_{k_n}$
   $\mathbf{H}^{(ik)}_{n+1} = \tilde{\mathbf{H}}^{(ik)}_n + \exp[\frac{-\Delta t_n}{2\tau_{ik}}]\bar{\mathbf{S}}^0_{k_{n+1}}$
4. Compute the Cauchy stress tensor
   $p_{n+1} = \frac{d\Psi_{\text{vol}}(J_{n+1})}{dJ}|_{n+1}$
   $\tilde{\mathbf{h}}^{(k)}_n = \sum_{i=1}^n \gamma_{ik} dev[\bar{\mathbf{F}}_{n+1}\tilde{\mathbf{H}}^{(ik)}_n\bar{\mathbf{F}}^T_{n+1}]$
   $\tilde{h}^{(k)}_n = \sum_{i=1}^n \gamma_{ik} tr[\bar{\mathbf{F}}_{n+1}\tilde{\mathbf{H}}^{(ik)}_n\bar{\mathbf{F}}^T_{n+1}]$
   $\boldsymbol{\sigma}_{n+1} = p_{n+1}\mathbf{1} + \sum_{j=1,j\neq3}^5[(1-\gamma_k+\nu_k)dev[\boldsymbol{\sigma}^0_{k_{n+1}}] + \frac{1}{J_{n+1}}\tilde{\mathbf{h}}^{(k)}_n]$
5. Compute initial elastic modulus
   $\mathbf{c}^0_{\text{vol}\,n+1}$ and $\bar{\mathbf{c}}^0_{k_{n+1}}$
6. Introduce viscoelastic effects
   $\mathbf{c}_{iso\,n+1} = \sum_{j=1,j\neq3}^5[(1-\gamma_k+\nu_k)\bar{\mathbf{c}}^0_{k_{n+1}} -$
   $-\frac{2}{3J_{n+1}}[\tilde{\mathbf{h}}^{(k)}_n\otimes\mathbf{1} + \mathbf{1}\otimes\tilde{\mathbf{h}}^{(k)}_n - \tilde{h}^{(k)}_n(\mathbf{I} - \frac{1}{3}\mathbf{1}\otimes\mathbf{1})]\}]$
7. Compute elastic modulus
   $\mathbf{c}_{n+1} = \mathbf{c}^0_{\text{vol}\,n+1} + \bar{\mathbf{c}}_{n+1}$

5.2 *Softening and Damage Effects*

In continuum damage mechanics, the free energy for the fibers is assumed to be of the form

$$\Psi(\mathbf{C},\mathbf{M},\mathbf{N},D_k) = \Psi^0_{\text{vol}}(J) + \sum_{k=m,f_1,f_2}\bar{\Psi} = \Psi^0_{\text{vol}}(J) + \sum_{k=m,f_1,f_2}[1-D_k]\bar{\Psi}^0 \tag{77}$$

where $(1 - D_k)$ are known as the reduction factors [53, 62], being the internal variables $D_k \in [0, 1]$ normalized scalars referred to as the damage variables, for the matrix, D_m, and the two families of fibres, D_{f4} and D_{f6} respectively [12].

Using standard arguments based on the Clausius–Duhem inequality [10]

$$\mathcal{D}_{int} = -\dot{\Psi} + \frac{1}{2}\mathbf{S} : \dot{\mathbf{C}} \geq 0 \tag{78}$$

and Eqs. (77)–(78) gives [12]:

$$\mathcal{D}_{int} = \left[\mathbf{S} - J \frac{d\,\Psi^0_{\text{vol}}(J)}{dJ} \mathbf{C}^{-1} - 2J^{\frac{-2}{3}} \sum_{k=m,f_1,f_2} \mathbf{P} : [1 - D_k] \frac{\partial \bar{\Psi}^0_{(\text{k})}}{\partial \bar{\mathbf{C}}} \right] : \frac{\dot{\bar{\mathbf{C}}}}{2}$$
$$+ \sum_{k=m,f_1,f_2} \frac{\partial \bar{\Psi}_{(\text{k})}}{\partial D_k} \dot{D}_k \geq 0, \tag{79}$$

where $\bar{\Psi}^0_{(\text{k})}(k = m, f_1, f_2)$ being the contributions of the matrix and the two families of fibres respectively. Equation (79) leads to the representation

$$\mathbf{S} = 2\frac{\partial \Psi(\mathbf{C}, \mathbf{M})}{\partial \mathbf{C}} = \mathbf{S}_{\text{vol}} + \sum_{k=m,f_1,f_2} [1 - D_k]\bar{\mathbf{S}}^0, \tag{80}$$

where $\mathbf{S}_{\text{vol}}$ and $\mathbf{S}^0$ denote a purely volumetric and a purely isochoric effective contribution of the stress tensor of the undamaged material (7), whereas the principle of positive dissipation leads to

$$\mathcal{D}_{int} = \sum_{k=m,f_1,f_2} f_k \dot{D}_k \geq 0, \tag{81}$$

with f_k conjugate state functions of the internal variables D_k defined as

$$f_m = -\frac{\partial \bar{\Psi}_{(\text{m})}}{\partial D_m} = \bar{\Psi}^0_{(\text{m})}(\bar{\mathbf{C}}) \geq 0,$$
$$f_{f_1} = -\frac{\partial \bar{\Psi}_{(\text{f1})}}{\partial D_{f_1}} = \bar{\Psi}^0_{(\text{f1})}(\bar{\mathbf{C}}, \mathbf{M}) \geq 0,$$
$$f_{f_2} = -\frac{\partial \bar{\Psi}_{(\text{f2})}}{\partial D_{f_2}} = \bar{\Psi}^0_{(\text{f2})}(\bar{\mathbf{C}}, \mathbf{N}) \geq 0. \tag{82}$$

In order to complete the constitutive model we have to determine the evolution equation for the internal damage variables D_k. Firstly, a Mullins-type discontinuous damage evolution is assumed where the damage accumulation occurs only within the first cycle of a strain-controlled loading process. Further strain cycles below the

maximum effective strain energy reached will not contribute to this type of damage. Secondly we take into account independently of the mechanism above, a continuous damage accumulation within the whole strain history of the deformation process which is also governed by the local effective strain energy. The total damage is then described by the constitutive expression

$$D_k \doteq D_k^{\alpha}(\alpha) + D_k^{\beta}(\beta), \tag{83}$$

where $D_k^{\alpha} : \mathbb{R}_+ \to \mathbb{R}_+$ and $D_k^{\beta} : \mathbb{R}_+ \to \mathbb{R}_+$ are monotonically increasing smooth functions with the following properties $D_k^{\alpha}(0) = 0$, $D_k^{\beta}(0) = 0$ and $D_k^{\alpha}(\alpha) + D_k^{\beta}(\beta) \in [0, 1]\ \forall \alpha, \beta$. They can be considered as shape functions which relate the damage variables D_k to the new variables α and β which describe the discontinuous and the continuous damage, respectively. These new variables are related to the evolution of the damage driving forces f_k as follows.

The *discontinuous damage* (Mullins-type) is assumed to be governed by the variable

$$\alpha_k(t) \doteq \max_{s \in (-\infty, t)} \sqrt{2\bar{\Psi}^0_{(\mathrm{k})}(\bar{\mathbf{C}}(s))}. \tag{84}$$

Thus $\alpha(t)$ is simply the maximum thermodynamic force or effective strain energy which has been achieved within the loading history interval $[0, t]$. We define a damage criterion in the strain space by the condition that, at any time t of the loading process, the following expression is fulfilled [62]

$$\Phi_k(\mathbf{C}(t), \Xi_{k_t}) = \sqrt{2\bar{\Psi}^0_{(\mathrm{k})}(\bar{\mathbf{C}}(t))} - \alpha_k(t) = \Xi_k - \alpha_k(t) \leq 0. \tag{85}$$

The equation $\Phi_k(\mathbf{C}(t), \Xi_{k_t}) = 0$ defines a damage surface in the strain space. Finally, the evolution of the damage parameters D_k is characterized by an irreversible equation of evolution such as [12]

$$\frac{dD_k^{\alpha}}{dt} = \begin{cases} \bar{h}_k(\Xi_k, \alpha_k)\dot{\Xi}_k & \text{if} \quad \Phi_k = 0 \quad \text{and} \quad \mathbf{N}_k : \dot{\mathbf{C}} > 0. \\ 0 & \text{otherwise} \end{cases} \tag{86}$$

This underlines the discontinuous character of this damage effect. There is no damage accumulation if the thermodynamic force f_k lies inside an undamaged domain $\mathbb{D}_{\alpha} := \{\Xi_k \in \mathbb{R}_+ | \Xi_k - \alpha_k(t) \leq 0\}$. Here, $\mathbf{N}_k := \frac{\partial \phi_k}{\partial \mathbf{C}}$ is the normal to the damage surface in the strain space, Ξ_k are defined at the current time s and $\bar{h}_k(\Xi_k, \alpha_k)$ are given functions that characterize the damage evolution in the material.

Continuous damage is assumed to be governed by the arclength of the respective driving damage force or effective strain energy.

$$\beta_k(t) \doteq \int_0^t |\dot{f}_k(s)| \mathrm{d}s. \tag{87}$$

Thus we have the simple evolution equation

$$\dot{\beta}_k = |\dot{f}_k| = \mathrm{sign}(\dot{\bar{\Psi}}^{\,0}_{(\mathrm{k})}),$$

with the initial condition $\beta_k(0) = 0$. Therefore β_k monotonically increase within the deformation process.

The iterative Newton procedure to solve a nonlinear finite element problem requires the determination of the consistent tangent material operator. This can be derived analytically for the given material equation (9). The symmetric algorithmic material tensor is expressed as [62]

$$\mathbf{C} = \mathbf{C}^0_{\mathrm{vol}} + \sum_{k=m,f_1,f_2} \left[[1 - D_k] \bar{\mathbf{C}}^0_{(k)} - \bar{g}'_{(k)} \bar{\mathbf{S}}^0_{(\mathrm{k})} \otimes \bar{\mathbf{S}}^0_{(\mathrm{k})} \right], \tag{88}$$

with the continuous tangent factor $\bar{g}'_{(k)}$ defined as

$$\bar{g}'_{(k)} = \begin{cases} \dot{D}^\alpha_k(\alpha) + \dot{D}^\beta_k(\beta)\mathrm{sign}(\dot{f}_k) & \text{if} \quad \Phi_k = 0 \quad \text{and} \quad \mathbf{N}_k : \dot{\mathbf{C}} > 0. \\ \dot{D}^\beta_k(\beta)\mathrm{sign}(\dot{f}_k) & \text{otherwise} \end{cases} \tag{89}$$

Typical evolution equations for the discontinuous damage variables, D^α_k, proposed in the literature for fibred materials such as soft biological tissues have been used. They correspond to the following expressions [12, 14, 61]

$$D^\alpha_k(\Xi_{k_t}) \doteq \begin{cases} 0 & \text{if} \quad \Xi_{k_t} < \Xi^0_{min_k} \\ 1 - \frac{1-\exp(\mu_k[\Xi^k_t - \Xi_{max_k}])}{1-\exp(\mu_k[\Xi_{min_k} - \Xi_{max_k}])} & \text{if} \quad \Xi^0_{min_k} \le \Xi_{k_t} \le \Xi^0_{max_k} \\ 1 & \text{if} \quad \Xi_{k_t} > \Xi^0_{max_k} \end{cases}, \tag{90}$$

$$D^\alpha_k(\Xi_{k_t}) \doteq \begin{cases} 0 & \text{if} \quad \Xi_{k_t} < \Xi^0_{min_k} \\ \Lambda^2_k[1 - \eta_k[\Lambda^2_k - 1]] & \text{if} \quad \Xi^0_{min_k} \le \Xi_{k_t} \le \Xi^0_{max_k} \\ 1 & \text{if} \quad \Xi_{k_t} > \Xi^0_{max_k} \end{cases}, \tag{91}$$

$$D^\alpha_k(\Xi_{k_t}) \doteq \frac{1}{2}\left[1 + \frac{2\xi_k \Xi_{k_t} \exp(2\xi_k[[2\Xi_{k_t}/\rho_k] - 1]) - 1}{2\xi_k \Xi_{k_t} \exp(2\xi_k[[2\Xi_{k_t}/\rho_k] - 1]) + 1}\right], \tag{92}$$

with $0 \le \Lambda = \frac{\Xi_{k_t} - \Xi^0_{min_k}}{\Xi^0_{max_k} - \Xi^0_{min_k}} \le 1$ a dimensionless variable and $\Xi^0_{min_k}$ the variables (84) associated to the strain energies at initial damage for matrix and fibres respectively,

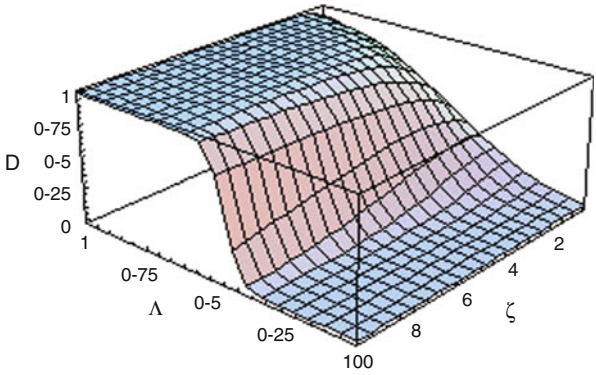


Fig. 6 Damage evolution for the modified sigmoidal function (93) with $0 \leq \Lambda \leq 1$ and $\zeta_k \geq 0$. *Source*: [53]

$\Xi^0_{max_k}$ the variables (84) associated to the strain energy at total damage for matrix and fibres, and $\eta_k \in [-1.0, 1.0]$, $\mu_k \geq 0$, $\xi_k \geq 0$ and $\rho_k \geq 0$ are model parameters.

Some remarks are needed regarding the previous evolution equations for the discontinuous damage variables, D^α_k. When $\Xi_{k_t} = \Xi^0_{max_k}$, Eq. (90) has not first continuous derivative, so some numerical problems could appear. Equation (91) is a non-monotonically increasing function for the material parameter η_k outside the interval $[-1.0, 1.0]$. This implies that the quality of the fitting of experimental data may be low when constants are restricted by stability considerations. In addition, in Eq. (92), we can not control damage initiation since the parameters $\Xi^0_{min_k}$ and $\Xi^0_{max_k}$ are not considered. With all this in the mind, Peña [44] proposed the new evolution equation

$$D^\alpha_k(\Xi_{k_t}) \doteq \begin{cases} 0 & \text{if} \quad \Xi_{k_t} < \Xi^0_{min_k} \\ \frac{1}{2}\left[1 + \frac{2\zeta_k \Lambda_k \exp(2\zeta_k[2\Lambda_k - 1]) - 1}{2\zeta_k \Lambda_k \exp(2\zeta_k[2\Lambda_k - 1]) + 1}\right] & \text{if} \quad \Xi^0_{min_k} \leq \Xi_{k_t} \leq \Xi^0_{max_k} \\ 1 & \text{if} \quad \Xi_{k_t} > \Xi^0_{max_k} \end{cases} \tag{93}$$

that is convex for $\zeta_k \geq 0$, Fig. 6.

It can be shown that depending on the parameter ζ_k when $\Lambda_k = 1$ the value of the damage (93) is not always equal to 1, so the total damage function can not be reached. In order to solve these concerns arising from the previously damage equations, a simple sigmoid function is proposed. A sigmoid curve is produced by a mathematical function having a classical "S" shape

$$D_k(\Xi_{k_t}) = \frac{1}{1 + \exp(-\alpha_k[\Xi_{k_t} - \gamma_k])}. \tag{94}$$

The parameter α_k controls the slope and γ_k defines the value Ξ_{k_t} such that $D_k(\Xi_{k_t}) = 0.5$.

Finally, the continuous damage D^β_k is assumed to have the form proposed by Peña [53]

$$D_k^\beta = d_{k\infty}^\beta \left[1 - \exp\left(-\frac{\beta}{\gamma_k^\beta} \right) \right]. \tag{95}$$

Note that the parameters $d_{k\infty}^\beta$ describe the maximum possible continuous damage. Thus we have the constraint $d_{k\infty}^\beta \in [0, 1]$. We refer to γ_k^β as the damage saturation parameters.

This completes the constitutive formulation of anisotropic finite strain elasticity with damage-caused energy-based softening effects. This results in a symmetric algorithmic tangent modulus, essential for the solution of the implicit finite element equations. If the material state is known at a time t_n and the deformation is known at a time $t_{n+1} = t_n + \Delta t$, we summarize the computational algorithm in the following scheme.

---

1. Database at each Gaussian point $D_{\alpha_{k_n}}$, $\Xi_{k_{t_n}}$, $D_{\beta_{k_n}}$, $\bar{\Psi}^0_{k_n}$
2. Compute the initial elastic stress tensors ($\bar{\boldsymbol{\sigma}}^0_{k_{n+1}}$) and the initial elastic modulus ($\mathsf{c}^0_{\mathrm{vol}_{n+1}}$ and $\bar{\mathsf{c}}^0_{k_{n+1}}$)
3. Compute the current equivalent measures
   $\Xi_{k_{n+1}} = \sqrt{2\bar{\Psi}^0_{k_{n+1}}}$
4. Check the discontinuous damage criterion
   $\Phi_{k_{n+1}}(\bar{\mathbf{C}}_{n+1}, \Xi_{k_{t_n}}) = \sqrt{2\bar{\Psi}^0_{k_{n+1}}} - \Xi_{k_{t_n}} = \Xi_{k_{n+1}} - \Xi_{k_{t_n}} > 0$
   YES: update discontinuous damage internal variables
   $D^\alpha_{k_{n+1}}(\Xi_{k_{t_{n+1}}}) \doteq \frac{1}{2}\left[1 + \frac{2\zeta_k \Lambda_{k_{n+1}} \exp(2\zeta_k[2\Lambda_{k_{n+1}} - 1]) - 1}{2\zeta_k \Lambda_{k_{n+1}} \exp(2\zeta_k[2\Lambda_{k_{n+1}} - 1]) + 1}\right]$
   $\bar{\mathbf{S}}^\alpha_{k_{n+1}} = D'_{\alpha_{k_{n+1}}} \bar{\mathbf{S}}^0_{k_{n+1}} \otimes \bar{\mathbf{S}}^0_{k_{n+1}}$
   $\Xi_{k_{t_{n+1}}} = \Xi_{k_{n+1}}$
   NO: no additional damage $D_{\alpha k_{n+1}} = D_{\alpha k_n}$ and $\bar{\mathbf{S}}^\alpha_{k_{n+1}} = 0$.
5. Update continuous damage internal variables
   $\beta_{k_{n+1}} = \beta_{k_n} + |\bar{\Psi}^0_{k_{n+1}} - \bar{\Psi}^0_{k_n}|$.
   $D_{\beta_{k_{n+1}}} = d_\infty^{\beta_k}\left[1 - \exp\left(-\frac{\beta_{k_{n+1}}}{\zeta_k}\right)\right]$
   $\bar{\mathbf{S}}^\beta_{k_{n+1}} = D'_{\beta_{k_{n+1}}} \operatorname{sign}(\dot{f}_{k_{n+1}}) \bar{\mathbf{S}}^0_{k_{n+1}} \otimes \bar{\mathbf{S}}^0_{k_{n+1}}$
6. Compute the Cauchy stress tensor
   $p_{n+1} = \frac{d\Psi_{\mathrm{vol}}(J_{n+1})}{dJ}|_{n+1}$
   $\boldsymbol{\sigma}_{n+1} = p_{n+1}\mathbf{1} + \sum_{k=m,f_1,f_2}[1 - D_{k_{n+1}}]\mathrm{dev}(\bar{\boldsymbol{\sigma}}^0_{k_{n+1}})$
7. Compute the extra term of the elastic modulus
   $\bar{\mathbf{S}}_{k_{n+1}} = \bar{\mathbf{S}}^\alpha_{k_{n+1}} + \bar{\mathbf{S}}^\beta_{k_{n+1}}$
8. Compute the elastic modulus
   $\mathbf{c}_{n+1} = \mathbf{c}^0_{\mathrm{vol}n+1} + \sum_{k=m,f_1,f_2}[1 - D_{k_{n+1}}]\bar{\mathbf{c}}^0_{k_{n+1}} - \bar{\mathbf{s}}_{k_{n+1}}$ with $\bar{\mathbf{s}}_{k_{n+1}} = J^{-1}\boldsymbol{\chi}_*(\bar{\mathbf{S}}_{k_{n+1}})$

Strain-softening and loss of strong ellipticity phenomena associated with damage mechanism impose numerical difficulties in finite element computations. Following [45], a viscous damage mechanism is presented in this section to regularize

the localization problems. Rate equations governing visco-damage behavior are obtained from their rate-dependent counterparts (93), by replacing the damage consistency parameter $\dot{\rho}_k = \dot{\Xi}_k$ by $\rho_k \Phi_k$ for matrix and fibers respectively

$$\dot{D}_k = \begin{cases} \rho_k \Phi_k \bar{h}_k(\Xi_k, D_k) & \text{if} \quad \Phi_k > 0 \quad \text{and} \quad \mathbf{N}_k : \dot{\bar{\mathbf{C}}} > 0. \\ 0 & \text{otherwise} \end{cases} \tag{96}$$

Here ρ is the damage viscosity coefficient, Φ_k denotes the viscous damage flow function and these are defined in (85) for matrix and fibers.

5.3 Time-Dependent Softening Coupled Response

In order to extend this hyperelastic model to the case of viscoelastic and damage behaviour of the ground matrix, the isotropic contribution is now assumed to be

$$\bar{\Psi}_{\text{iso}} = [1 - D_{\text{m}}]\, \bar{\Psi}^0_{\text{m}} - \frac{1}{2} \sum_{i=1}^{n} \left[\bar{\mathbf{C}} : \mathbf{Q}^i_{\text{m}} \right], \tag{97}$$

where $[1 - D_{\text{m}}]$ is the so-called reduction factor [64] with $D_{\text{m}} \in [0, 1]$ a monotonically increasing damage internal variable, while the viscoelastic response of the material is represented by n second-order tensors $\mathbf{Q}^i_{\text{m}}$ associated to $\bar{I}_1$ which may be interpreted as non-equilibrium stresses [51].

By analogy with (97), the anisotropic contribution to the total strain energy is assumed to be

$$\bar{\Psi}_{\text{ani}} = \left[1 - D_{f_1}\right] \bar{\Psi}^0_{f_1} + \left[1 - D_{f_2}\right] \bar{\Psi}^0_{f_2} - \frac{1}{2} \sum_{i=1}^{n} \left[\bar{\mathbf{C}} : \mathbf{Q}^i_{f_1} + \bar{\mathbf{C}} : \mathbf{Q}^i_{f_2} \right], \tag{98}$$

where $D_{\text{f1,f2}} \in [0, 1]$ are the damage variables associated to the first and second families of fibers, respectively, and $\mathbf{Q}_{f_1}$ and $\mathbf{Q}_{f_2}$ in (98) are second order non-equilibrium stress tensors representing the directional viscoelastic response of the material now associated to the invariants $\bar{I}_4$ and $\bar{I}_6$, respectively.

The non-equilibrium stress tensors $\mathbf{Q}^i_{\kappa}$ are assumed to be governed by the set of rate equations

$$\dot{\mathbf{Q}}^i_{\kappa} + \frac{1}{\tau^i_{\kappa}} \mathbf{Q}^i_{\kappa} = \frac{\gamma^i_{\kappa}}{\tau^i_{\kappa}} [1 - D_{\kappa}]\, DEV \left(\frac{\partial \bar{\Psi}^0_{\kappa}}{\partial \bar{\mathbf{C}}} \right), \tag{99}$$

$$\lim_{t \to -\infty} \mathbf{Q}^i_{\kappa} = \mathbf{0}$$

where $\gamma_\kappa^i \in [0,1]$ are free energy dimensionless factors associated with the relaxation times $\tau_\kappa^i > 0$ [49]. The application of standard derivations on (99) leads to convolution representation

$$\mathbf{Q}_\kappa^i(t) = \frac{\gamma_\kappa^i}{\tau_\kappa^i}[1-D_\kappa]\int_{-\infty}^{t}\exp\left(\frac{-[t-s]}{\tau_\kappa^i}\right)DEV\left(\frac{\partial\bar{\Psi}_\kappa^0}{\partial\bar{\mathbf{C}}}\right)\mathrm{d}s. \tag{100}$$

Algorithmically, the constitutive model is appealing since Eq. (99) can be evaluated via a simple recursion relation which was originally developed for finite strains by Simo [62]. In particular, if the material state is known at a time t_n and the deformation is known at a time $t_{n+1} = t_n + \Delta t$ with $\Delta t > 0$, we may write

$$\begin{aligned}\mathbf{S}_{n+1} =& J_{n+1}p_{n+1}\mathbf{C}_{n+1}^{-1} + J_{n+1}^{-\frac{2}{3}}\sum_{k=m,f_1,f_2}\left[[1-\sum_{i=1}^{n}\gamma_{ik}][1-D_{k_{n+1}}]\bar{\mathbf{S}}_{k_{n+1}}^0\right]\\ &+ J_{n+1}^{-\frac{2}{3}}\sum_{i=1}^{n}\left[\gamma_{ik}\,DEV\left(\mathbf{H}_{n+1}^{(ik)}\right)\right],\end{aligned} \tag{101}$$

where the subscripts n and $n+1$ denote quantities evaluated at times t_n and t_{n+1} [49,52,62] and $\mathbf{H}_{n+1}^{(ik)}$ are internal algorithmic history variables defined as

$$\mathbf{H}_{n+1}^{(ik)} = \exp\left(\frac{-\Delta t}{\tau_{ik}}\right)\mathbf{H}_n^{(ik)} + \exp\left(\frac{-\Delta t}{2\tau_{ik}}\right)\left[[1-D_{k_{n+1}}]\bar{\mathbf{S}}_{k_{n+1}}^0 - [1-D_{k_n}]\bar{\mathbf{S}}_{k_n}^0\right]. \tag{102}$$

Note that the time discretization scheme (101) used for the calculation of the current value of the stress $\mathbf{S}_{n+1}$ requires the storage of 2k symmetric second-order tensors $\mathbf{H}_n^{(ik)}$ and $\bar{\mathbf{S}}_{k_n}^0$ and 2k scalars D_{k_n} and Ξ_{k_t} at the previous time $t = t_n$ at each Gauss point of the finite element mesh.

The iterative Newton procedure to solve a nonlinear finite element problem requires the determination of the consistent tangent material operator. The symmetric algorithmic material tensor which is expressed as [62]

$$\begin{aligned}\mathbf{c}_{n+1} =& \bar{\mathbf{c}}_{\mathrm{vol}\,n+1}^0 + \sum_{k=m,f_1,f_2}[[1-D_{k_{n+1}}][1-\gamma_k+\nu_k]\bar{\mathbf{c}}_{k_{n+1}}^0 +\\ &-\frac{2}{3}J_{n+1}^{-\frac{4}{3}}\sum_{i=1}^{n}\gamma_{ik}[DEV(\tilde{\mathbf{H}}_n^{(ik)})\otimes\bar{\mathbf{C}}_{n+1}^{-1} + \bar{\mathbf{C}}_{n+1}^{-1}\otimes DEV(\tilde{\mathbf{H}}_n^{(ik)}) -\\ &-[\tilde{\mathbf{H}}_n^{(ik)}:\bar{\mathbf{C}}][\mathsf{I}_{\bar{C}_{n+1}}^{-1} - \frac{1}{3}\bar{\mathbf{C}}_{n+1}^{-1}\otimes\bar{\mathbf{C}}_{n+1}^{-1}]] - \bar{\mathbf{S}}_{k_{n+1}}],\end{aligned} \tag{103}$$

where $\mathsf{I}_{\bar{C}}^{-1} = -\frac{1}{2}[C_{ik}^{-1}C_{jl}^{-1} + C_{il}^{-1}C_{jk}^{-1}]$ and

$$\bar{\mathbf{S}}_{k_{n+1}} = \begin{cases} \bar{g}'_{k_{n+1}} \bar{\mathbf{S}}^0_{k_{n+1}} \otimes \bar{\mathbf{S}}^0_{k_{n+1}} & \text{if} \quad \phi = 0 \quad \text{and} \quad \mathbf{N}_k : \dot{\mathbf{C}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{104}$$

and

$$\tilde{\mathbf{H}}^{(ik)}_n = \exp\left(\frac{-\Delta t}{\tau_{ik}}\right) \mathbf{H}^{(ik)}_n - \exp\left(\frac{-\Delta t}{2\tau_{ik}}\right)[1 - D_{k_n}]\bar{\mathbf{S}}^0_{k_n} \tag{105}$$

being $\gamma_k = \sum_{i=1}^n \gamma_{ik}$, $\nu_k = \sum_{i=1}^n \gamma_{ik} \exp\left(\frac{-\Delta t}{2\tau_{ik}}\right)$. For more detail to derivation (103) see [52].

For the reader's convenience, we have summarized the developed algorithm in the following scheme.

---

1. Database at each Gaussian point
   $\bar{\mathbf{S}}^0_{k_n}$, $\mathbf{H}^{(ik)}_n$, D_{k_n}, Ξ_{k_t} $i = 1 \ldots$ N internal viscoelastic variables and $k = m,\ f_1,\ f_2$
2. Compute the initial elastic stress tensors
   $dev[\boldsymbol{\sigma}^0_{k_{n+1}}] = \frac{1}{J_{n+1}} dev\left\{\bar{\mathbf{F}}_{n+1}[2\frac{\partial \bar{\Psi}^0_{k_{n+1}}(\bar{\mathbf{C}}_{n+1},\mathbf{M},\mathbf{N})}{\partial \bar{\mathbf{C}}_{n+1}}]\bar{\mathbf{F}}^T_{n+1}\right\}$
   $\bar{\mathbf{S}}^0_{(k)n+1} = \bar{\mathbf{F}}_{n+1}(J_{n+1} dev[\boldsymbol{\sigma}^0_{k_{n+1}}])\bar{\mathbf{F}}^T_{n+1}$
3. Compute the current equivalent measure $\Xi_{k_{n+1}} = \sqrt{2\bar{\Psi}^0_{k_{n+1}}}$
4. Check the damage criterion
   $\phi_{k_{n+1}}(\mathbf{C}_{n+1}, \Xi_{k_t}) = \sqrt{2\bar{\Psi}^0_{k_{n+1}}} - \Xi_{k_t} = \Xi_{k_{n+1}} - \Xi_{k_t} > 0$
   YES: update damage internal variables
   $1 - D_{k_{n+1}} = 1 - \xi^2[1 - \beta_k(\xi^2 - 1)]$ and $\bar{\mathbf{S}}_{k_{n+1}} = \bar{g}'_{k_{n+1}} \bar{\mathbf{S}}^0_{k_{n+1}} \otimes \bar{\mathbf{S}}^0_{k_{n+1}}$
   NO: no additional damage
   $D_{k_{n+1}} = D_{k_n}$ and $\bar{\mathbb{S}}_{k_{n+1}} = 0$.
5. Update the viscoelastic internal variables
   $\tilde{\mathbf{H}}^{(ik)}_n = \exp[\frac{-\Delta t}{\tau_{ik}}]\mathbf{H}^{(ik)}_n - \exp[\frac{-\Delta t}{2\tau_{ik}}](1 - D_k)\bar{\mathbf{S}}^0_{k_n}$
   $\mathbf{H}^{(ik)}_{n+1} = \tilde{\mathbf{H}}^{(ik)}_n + \exp[\frac{-\Delta t}{2\tau_{ik}}](1 - D_k)\bar{\mathbf{S}}^0_{k_{n+1}}$
6. Compute the Cauchy stress tensor
   $p_{n+1} = \frac{d\Psi_{\text{vol}}(J_{n+1})}{dJ}|_{n+1}$
   $\tilde{\mathbf{h}}^{(k)}_n = \sum_{i=1}^n \gamma_{ik} dev[\bar{\mathbf{F}}_{n+1}\tilde{\mathbf{H}}^{(ik)}_n\bar{\mathbf{F}}^T_{n+1}]$
   $\tilde{h}^{(k)}_n = \sum_{i=1}^n \gamma_{ik} tr[\bar{\mathbf{F}}_{n+1}\tilde{\mathbf{H}}^{(ik)}_n\bar{\mathbf{F}}^T_{n+1}]$
   $\boldsymbol{\sigma}_{n+1} = p_{n+1}\mathbf{1} + \sum_{k=m,f_1,f_2}[(1 - \gamma_k + \nu_k)(1 - D_{k_{n+1}})dev[\boldsymbol{\sigma}^0_{k_{n+1}}] + \frac{1}{J_{n+1}}\tilde{\mathbf{h}}^{(k)}_n]$
7. Compute the initial elastic modulus
   $\mathbf{c}^0_{\text{vol}\,n+1}$ and $\bar{\mathbf{c}}^0_{k_{n+1}}$
8. Introduce the viscoelastic effects
   $\bar{\mathbf{c}}_{n+1} = \sum_{k=m,f_1,f_2}[D_{k_{n+1}}(1 - \gamma_k + \nu_k)\bar{\mathbf{c}}^0_{k_{n+1}} -$
   $-\frac{2}{3J_{n+1}}[\tilde{\mathbf{h}}^{(j)}_n \otimes \mathbf{1} + \mathbf{1} \otimes \tilde{\mathbf{h}}^{(j)}_n - \tilde{h}^{(j)}_n(\mathbb{I} - \frac{1}{3}\mathbf{1} \otimes \mathbf{1})] - \bar{\mathbf{s}}_{k_{n+1}}]$ with $\bar{\mathbf{s}} = J^{-1}\boldsymbol{\chi}_*(\bar{\mathbf{S}})$
9. Compute the elastic modulus
   $\mathbf{c}_{n+1} = \mathbf{c}^0_{\text{vol}\,n+1} + \bar{\mathbf{c}}_{n+1}$

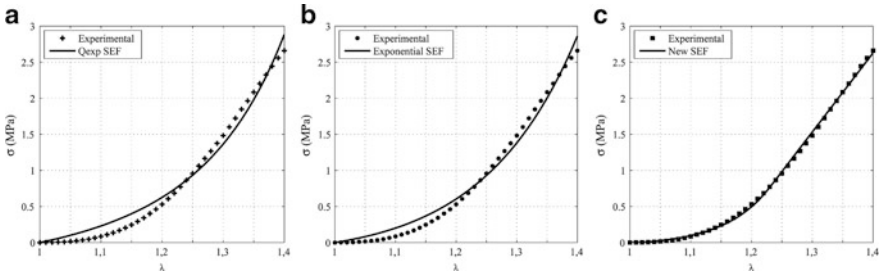


Fig. 7 Fittings of the experimental data for prolapsed tissue. (**a**) QExp SEF; (**b**) exponential SEF; (**c**) new SEF. *Source*: [41]

6 Examples

6.1 On Modeling Hyperelastic Behavior of Vaginal Tissue

Modeling of soft tissues as fiber-reinforced elastic materials on the basis of the invariant structure outlined in Sect. 2 is now well established and widely used. In this example, most used three-dimensional phenomenological models adopted in the literature for the study of elastic soft tissue are examined from a comparative point of view. Experimental data presented in [41] and the SEF (19), (24) and (20) were used.

SEF (24) has proved to reproduce the arterial behavior accurately [30]. For the case of prolapsed vaginal tissue, it was possible to fit the experimental curves with this SEF accurately, Fig. 7. A similar result was obtained when exponential SEF (19) was used. This exponential function has proved to reproduce the ligament and tendon behavior accurately [42, 56]. Unlike the two previous SEFs, (20) accounts the mechanically distinctive regions (exponential and linear) and is able to fit the experimental data very accurately showing better fitting indicators than the previous ones [41].

6.2 Knee Flexion

The geometrical data of the model developed herein were obtained by NMR (Nuclear Magnetic Resonance) for soft tissues and CT (Computerized Tomography) for bones, with images taken from a normal adult male volunteer by ZIB (Zuse-Institut Berlin). The contours of the femur, tibia, articular cartilage, menisci and ligaments (patellar tendon, anterior cruciate, posterior cruciate, medial collateral and lateral collateral) were identified using AMIRA software developed in ZIB. Tetrahedral meshes of bones, the ligaments, menisci and articular cartilages were constructed using the same software (AMIRA). A total of 450,000 elements were

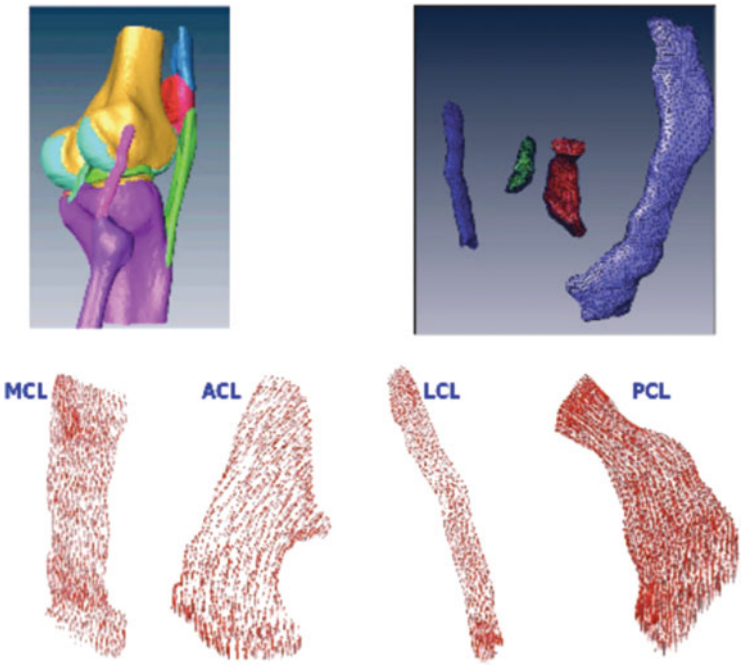


Fig. 8 Geometry and finite element model of the ligaments including their fiber orientation

used to mesh all tissue components of the knee (Fig. 8). In all the cases, we used trilinear tetrahedral elements with a full geometrically nonlinear formulation of ABAQUS.

Since bone stiffness is much higher than that of the relevant soft tissues and its influence in this study was minimal, bones were assumed to be rigid. Each bony structure (femur, tibia, fibula and patella) was therefore represented by a primary node located at its center of rotation at full extension. In the case of the femur this point was located at the midpoint of the transepicondylar line. These nodes, with six degrees of freedom, controlled the whole kinematics of each bone as rigid body [50]. Menisci and cartilage are hydrated tissues. However, in our case, and considering that the loading time of interest corresponded to that of a single leg stance, and the viscoelastic time constant of cartilage approaches 1,500 s, articular cartilage was considered to behave as a single-phase linear elastic and isotropic material with an elastic modulus of $E = 5$ MPa and a Poisson ratio of $\nu = 0.46$ [47] and similarly, menisci were also assumed to be a single-phase linear elastic and isotropic material with the following average properties: elastic modulus of $E = 59$ MPa and Poisson ratio of $\nu = 0.49$ [47]. On modelling ligaments, two important assumptions were made. First, no difference in the material behavior between the ligament body and its insertion were considered. Second, material characteristics depending on time, such as viscoelasticity, creep and relaxation were neglected due again to the high

Table 1 Material parameters for the ligament stress-free configuration

	C_1 (MPa)	C_2 (MPa)	C_3 (MPa)	C_4 (−)	C_5 (MPa)	λ^* (−)	D (MPa^{-1})
MCL	1.44	0.0	0.57	48.0	467.1	1.063	0.00126
LCL	1.44	0.0	0.57	48.0	467.1	1.063	0.00126
ACL	1.95	0.0	0.0139	116.22	535.039	1.046	0.00683
PCL	3.25	0.0	0.1196	87.178	431.063	1.035	0.0041
PT	2.75	0.0	0.065	115.89	777.56	1.042	0.00484

Table 2 % Ligament initial strains at full extension

aACL	pACL	PCL	aLCL	mLCL	pLCL	aMCL	mMCL	pMCL
0.06	0.1	0.0	0.0	0.0	0.08	0.04	0.04	0.03

ratio between the viscoelastic time constant of the material and the loading time of interest in this study. We used therefore a transversely isotropic hyperelastic model including the effect of one family of fibers presented in Sect. 2.1. The ligaments were modelled using the SEF proposed by Weiss et al. [68] where the volumetric and isochoric parts are defined by the Eqs. (27) and (17), respectively. We used the average constants obtained by Gardiner et al.[23] for the MCL in their experimental data. The LCL constants were assumed to be identical to those of the MCL. We fitted the uniaxial stress-strain curves obtained by Butler et al. [11] for ACL, PCL and PT (patellar tendon) with those obtained by Weiss's getting the associated constants that have been included in Table 1. Finally, the local fiber orientation ($\mathbf{m}_0$) was specified according to the local element geometry, see Fig. 8.

Initial strains in our model were defined from data available in literature [47] and have been included in Table 2 with the following terminology: a: anterior part of ligament; p: posterior part of ligament; m: medial part of ligament.

Boundary conditions were defined as follows, Fig. 9. The motion of each bone was controlled by the six degrees of freedom of the reference node. The position at full extension served as the starting point for the application of initial strains included in Table 2. In the first step, a 15° of flexion was applied to the tibia. After that, a combined load of 50 N in the quadriceps and 134 N anterior-posterior to the tibia were applied. The femur remained fixed in all cases.

Figure 10 shows the results obtained in different ligaments and cartilage under the quadriceps and anterior loads. A significant tensile stress appeared in the posterior part of the ACL, while a moderate tensile stress was observed in the anterior part. The obtained results also showed that the PCL was mainly in compression. The LCL was also mainly in compression except at the posterior and the femoral and tibial insertions. The LCL was tensioned due to the initial strains since during this movement it is mainly relaxed. The anterior load produced in the MCL a stress distribution similar to a shear problem, with tension in the anterior-distal and the posterior-proximal parts of the MCL. The femoral cartilage was also in compression with the minimum principal stress (maximum compression) oriented almost normal to the articular surface, showing higher stresses on the medial condyle.

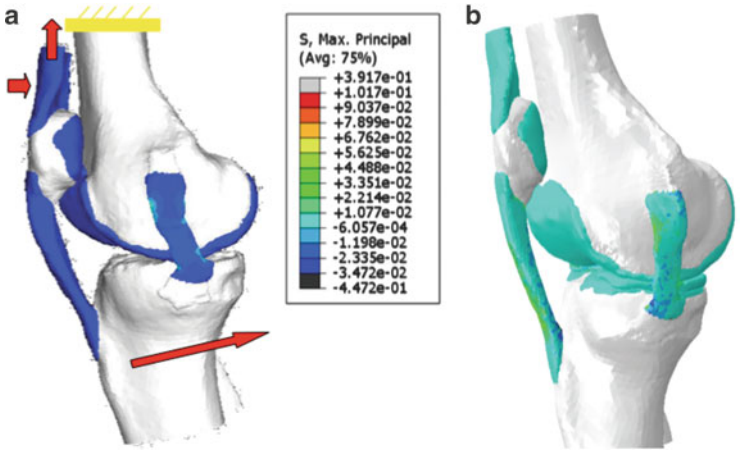


Fig. 9 Applied external loads and initial stress distribution map of the knee. (**a**) Applied external loads; (**b**) initial stress

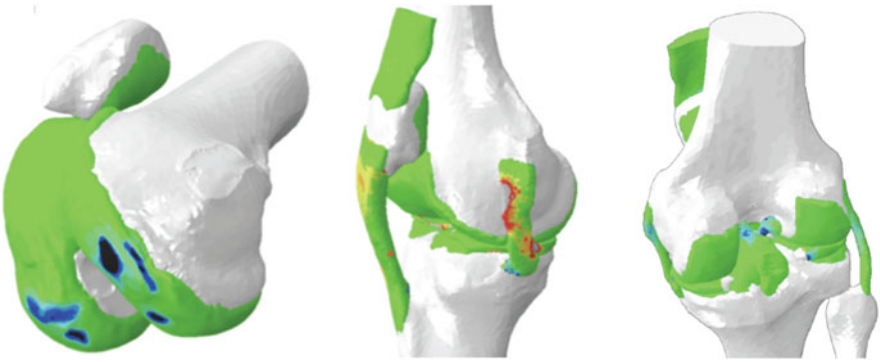


Fig. 10 Stress distribution in the ligaments and cartilage

6.3 Residual Stresses in a Real Geometry of Human Coronary Artery

In order to test the suitability of the method proposed for including the residual stresses in real arterial geometries, it was applied to a patient-specific geometry of a Left Artery Descendant (LAD) coronary artery. The model was reconstructed from a geometry obtained from Intravascular Ultrasound (IVUS) and angiography images (Fig. 11a).

The SEF presented in (27) and (24) and the material constants used were obtained from Peña et al. [48] and are written in Table 3. In order to compute the initial stress, we used the analytical opening angle solution proposed by Holzapfel et al. [30]. Therefore, the $\mathbf{F}_0$ tensor follows the expression (59) where the external and internal

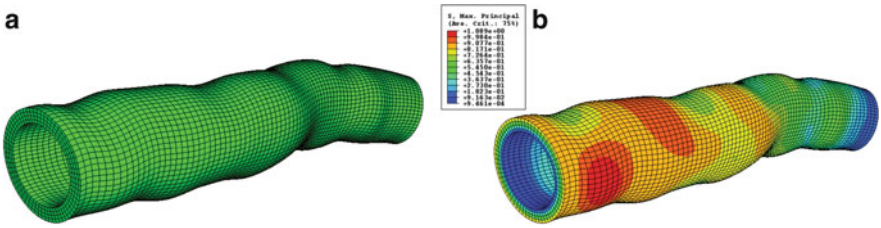


Fig. 11 (**a**) Real geometry (**b**) and initial stress. *Source*: [48]

Table 3 Material parameters for the artery (MPa)

μ	k_1	k_2	k_3	k_4	D
0.0274	$0.64 \cdot 10^{-3}$	3.54	$0.64 \cdot 10^{-3}$	3.54	1e−3

Source: [48]

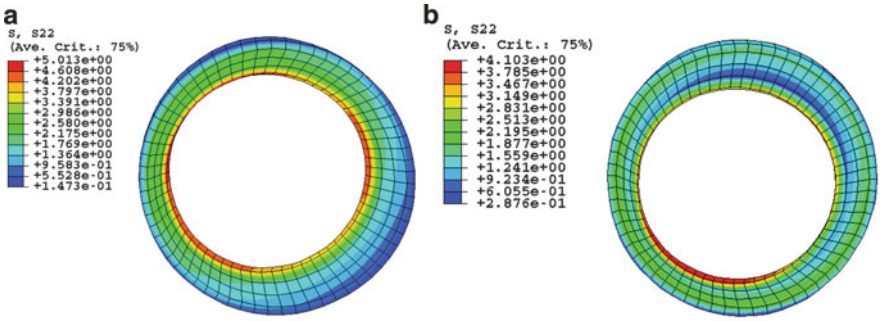


Fig. 12 Circumferential initial stresses and under internal pressure load. (**a**) No $\mathbf{F}_0$ applied; (**b**) $\mathbf{F}_0$ applied. *Source*: [48]

radius and the opening angle were $R_i = 4.25\,\text{mm}$, $R_i = 1.75\,\text{mm}$, $\alpha = 200°$, respectively. There was no experimental data for λ_z so we considered no initial strain in that direction, that is, $\lambda_z = 1$.

The effect of the internal pressure caused by the blood flow in the cylinder wall has been also analyzed. 0.0133 MPa (100 mmHg) of internal pressure was considered. As reported by several authors [16, 21], the circumferential component is much more uniform when residual stresses are taken into account. Maximal values of circumferential stress are not representative to evaluate this uniformity. When $\mathbf{F}_0$ is not included, circumferential stresses vary from 5.013 MPa in the inner layer to 0.14 MPa in the outer (Fig. 12).

6.4 Anterior Cruciate Ligament Under Different Strain Rates

To illustrate the performance of the visco-hyperelastic behaviour of ligaments and the importance of the strain-rates during their movement, a model of the human

Table 4 ACL elastic, viscoelastic and damage material parameters (MPa)

C_1	C_2	C_3	C_4	D	
1	0.0	0.4	8.1019	8.8e−3	
γ_m	τ_m	γ_{f1}	τ_{f1}		
0.31	0.15	0.69	5		
ψ^m_{min}	ψ^m_{max}	β^m	ψ^f_{min}	ψ^f_{max}	β^f
0.2946	0.4399	0.120	0.9427	1.4086	0.1538

Source: [13]

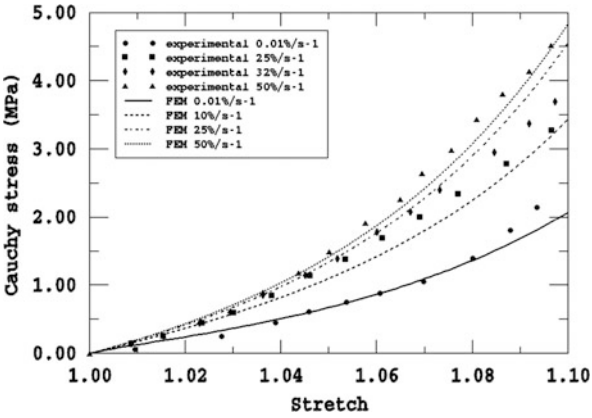


Fig. 13 Experimental results obtained and theoretical stress-strain curves at different rates of elongation for the human ACL. *Source*: [13]

anterior cruciate ligament (ACL) was constructed to simulate its behavior under a physiological anterior tibial displacement, see Fig. 14a. The surface geometries of femur and tibia were reconstructed from a set of Computer Tomography (CT) images, while for the ACL, MRI (Magnetic Resonance Images) were used [47]. Two different strain rates were applied: low ($0.012\,\%\,s^{-1}$) and high ($50\,\%\,s^{-1}$) that correspond to physiological and non-physiological strain-rates.

The elastic and viscoelastic parameters for the human ACL were fitted from published experimental data [56] and are shown in Table 4 and Fig. 13. Ligaments were attached to bone. The motion of each bone was controlled by the six degrees of freedom of its reference node. In the analyses, tibia remained fixed. The position at full extension served as the initial reference configuration. An anterior load of 134 N was applied to the femur. In this example we did not consider initial strains [48].

Maximal principal stress distributions in ACL at $0.012\,\%\,s^{-1}$ and $50\,\%\,s^{-1}$ of strain rates are presented in Fig. 14. The maximal principal stress is located in the central part of the ligament. The maximal principal stress of 7.27 MPa obtained in the central region for the higher load rate is due to the stiffening effect induced by high load rates. Under physiological strain-rates the maximal principal stress of 4.36 MPa is far from the ultimate stress.

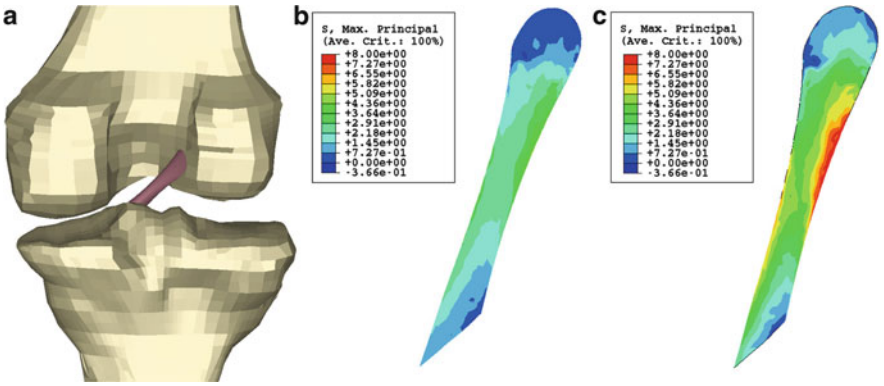


Fig. 14 Finite element model of the human ACL and maximal principal stress at low and high strain rates (MPa). (**a**) Finite element model; (**b**) 0.012 % s^{-1}; (**c**) 50 % s^{-1}. *Source*: [13]

6.5 Patellar Tendon Graft After Initial Prestress

Another clinical application of the viscoelastic model is the evolution along time of the initial prestress in bone-patellar tendon-bone grafts. Surgical reconstruction of the ACL is a common practice to treat the disability or chronical instability of knees due to ACL insufficiency [50]. The bone-patellar tendon-bone autograft remains a common practice due to its high ultimate strength and stiffness that allows for a more predictable restoration of the knee stability. Before graft fixation, an initial pretension is applied. This initial tension applied to the replacing graft significantly alters the joint kinematics. This prestress helps to provide joint stability, but a very high pretension produces an important additional stress in the graft during the knee movement. Viscoelasticity decreases the tension imposed during surgery until getting the final value after reaching equilibrium. The decrease of this initial stress can compromise the joint stability, affecting the postoperative results. In this example, we study the evolution of the initial stress in the graft.

The 3D finite element model of the graft and bone plugs is shown in (Fig. 15a). The plugs were modelled as elastic with a very high stiffness in comparison with that of the graft. The constitute law of the graft tendon was the same of the ligament [68] with the SEF (17), while bone plugs were considered to behave as a linearly elastic and isotropic material with an elastic modulus of $E = 14{,}220$ MPa and a Poisson ratio of $\nu = 0.3$ [49]. The elastic and viscoelastic parameters of the graft were obtained fitting the stress-curved obtained by Pioletti et al. [56] from the human patellar tendon (PT). These parameters are included in Table 5.

Displacements were applied to the femoral bone plug up to an initial stress of 2.93 MPa (Fig. 15b) corresponding to a pretension of about 60 N [46] and then fixed, while the tibial bone plug remained always fixed. After that, the relaxation process of the graft was computed until thermodynamic equilibrium (Fig. 15c).

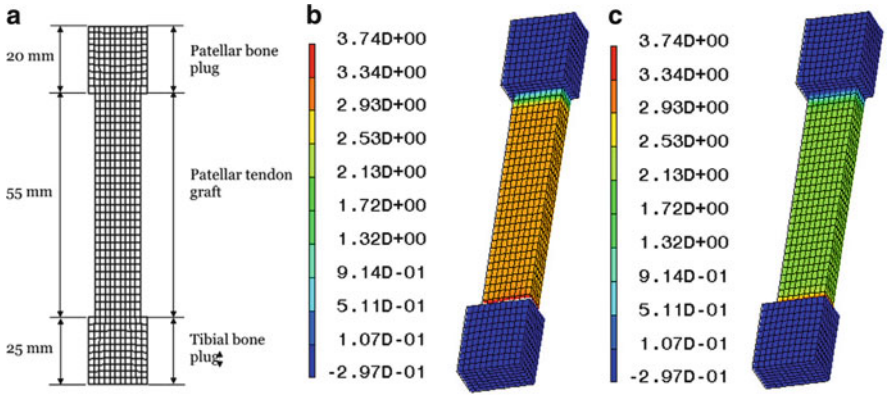


Fig. 15 Finite element model of the graft and prestress at different times (MPa). (**a**) Graft model; (**b**) $t = 0^+$ s; (**c**) $t = 1{,}000$ s. *Source*: [49]

Table 5 PT material parameters (MPa)

C_1	C_2	C_3	C_4	D	
2.7	0.0	15.3146	107.473	0.004938	
γ_{11}	τ_{11}(s)	γ_{12}	τ_{12}(s)	γ_{14}	τ_{14}(s)
0.55	10	0.55	10	0.35	150

Source: [46]

Figure 16 illustrates the evolution of the initial prestress with time. As can be observed, the initial value at time $t = 0.0^+$ decreased very fast at the beginning of the relaxation process. This results show that tension within the PT graft is reduced shortly after the fixation. For $t = 1{,}000$ s the stress decreased a 32.5 %. Graft et al. [25] showed a reduction of 30 % in the graft load when tensioned up to a strain of 2.5 % after 10 min, we tensioned up to a strain of 2.4 %. To minimize the stress relaxation response, preconditioning of the graft is usually recommended.

6.6 *Damage in Arteries After Balloon Angioplasty*

The purpose of this example is demonstrate the applicability of the viscous damage model under a more general biomedical loading condition. We will attempt to model Oktay experiments in bovine coronary arteries. The geometry of a healthy left anterior descending coronary artery (LAD) of 40 mm in length and internal and external diameters of $D_i = 2.7$ mm and $D_o = 4.5$ mm respectively, was considered, see [1]. The artery was simulated as a multi-layered composite material by considering the *media*, and *adventitia* layers without plaque.

The well-known quadratic exponential SEF for blood vessels (24) proposed by Holzapfel et al. [30] was used to model the elastic behavior of vascular tissue. The material properties for both layers are given in Table 6. The damage evolution has

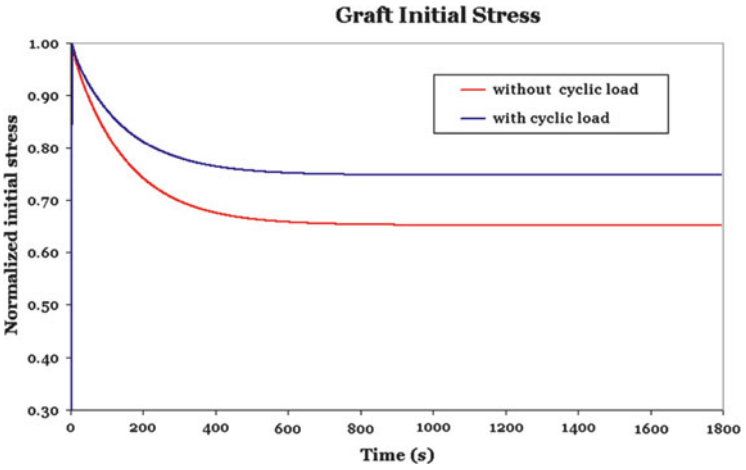


Fig. 16 Normalized initial stress evolution with and without cyclic load. *Source*: [49]

Table 6 Material and damage parameters for LAD

	μ	k_1	k_2	k_3	k_4	D
Media	2.7	5.1	15.4	5.1	15.4	0.001
Adventitia	1.4	0.64	3.54	0.64	3.54	0.001
	α_m	β_m	μ_m	Ξ_m^0		
Media	1.24	1.095	7.01	1.29		
Adventitia	0.94	1.0095	5.01	1.14		
	α_{f_1}	β_{f_1}	μ_{f_1}	$\Xi_{f_1}^0$		
Media	1.18	0.00114	11.3141	2.97		
Adventitia	1.073	0.00612	7.974	1.48		

μ, k_1, k_3 are in kPa, D and Ξ^0 are in kPa^{-1} and $\text{kPa}^{1/2}$ and the other parameters are dimensionless. *Source*: [45]

been defined by the Eq. (94). Unfortunately, there is no data in the literature that includes damage region for each layer of arterial tissue, so this example does not have a real clinical meaning. In addition, initial stress was accounted for in the simulation by imposing an opening angle of 120° by means of an initial compatible deformation gradient, as proposed by Rodríguez et al. [61]. The axial extension is restrained at both ends while allowing radial expansion.

For the balloon, we have taken a Grüntzig-type balloon catheter. The initial configuration of the balloon is taken as a cylindrical tube with external diameter, $d = 1.7$ mm, wall-thickness of 0.1 and 20 mm in length. The mechanical behavior of the balloon has been taken as orthotropic with reinforcing fibers running longitudinally and circumferentially as suggested by Rodríguez et al. [61]. Assuming symmetry conditions, only a quarter of the geometry has been considered, Fig. 17. Three steps were applied in order to simulate the angioplasty: (1) the residual stress was imposed in the model as proposed by Rodríguez et al. [61]; (2) the physiological

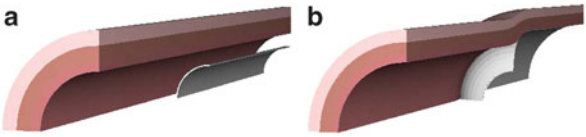


Fig. 17 Unloaded and balloon loading of the artery. (**a**) Initial unloaded configuration; (**b**) deformed configuration upon balloon inflation at 1 bar

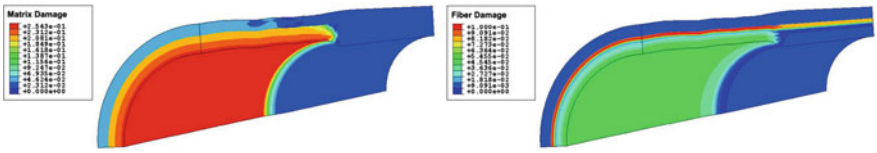


Fig. 18 Damage distribution in the arterial wall under balloon inflation. *Source*: [45]

condition where the artery was inflated up to a pressure of 13.3 kPa (100 mmHg); (3) the angioplasty as a balloon inflation up to a pressure of 100 kPa (1 bar), Fig. 17. Figure 18 shows the damage distribution in the arterial wall after balloon inflation for the matrix and fibers. Damage in the matrix occurred in the entire media and intima, as shown in Fig. 13. In the clinical context damage, known as "controlled vessel injury" occurs predominantly in the media. Regarding to the fibers, damage is developed in the adventitia layer, particularly at the media-adventitia interface. This damage distribution is due to the fiber arrangement and the particular loading at which the artery is subjected during balloon inflation. As mentioned before, the balloon induces large longitudinal and circumferential stretching in the artery which causes larger fiber deformation in the adventitia than in the media leading to larger stresses and more rapid damage of this layer.

6.7 Damage of Human Ligament Under Impact Testing

The purpose of this simulation is to demonstrate the effectiveness of the numerical algorithm and finite element implementation discussed in previous sections and the applicability of the model to simulate the structural behaviour of soft biological tissues. We reproduce, in a human medial collateral ligament (MCL), the distraction experiment. This example was previously modelled in [52] using a continuous inviscid damage model. The SEF presented in (19) was used and elastic and damage parameters for the human MCL (Fig. 19) were fitted from published experimental data [7] at 0.1 % s^{-1} of strain rate and are shown in Table 7.

Damage distributions in matrix and fibers at 0.01 and 113 mm/s of displacement rates are presented in Fig. 20. In all cases, we consider failure of the MCL when damage reached a value of 0.6 for both matrix and fibers. We can observe the effect of the strain rate into the damage behavior. At 113 mm/s of displacement

Table 7 Material, viscoelastic and damage parameters for the human MCL

a_1	a_2	a_3	a_4	D	
0.1539	0.0	0.1507	34.7929	3.986e−4	
γ_m	τ_m	γ_{f1}	τ_{f1}	γ_{f2}	τ_{f2}
0.4352	0.15	0.1500	2	–	–
Ξ^m_{min}	Ξ^m_{max}	β^m	Ξ^f_{min}	Ξ^f_{max}	β^f
0.0750	0.0932	0.120	0.3389	1.6652	0.1538

C_1, C_2, C_3, C_5 and Ξ^i are in MPa, D is in MPa^{-1}, τ_i in seconds and the rest of parameters are dimensionless. *Source*: [52]

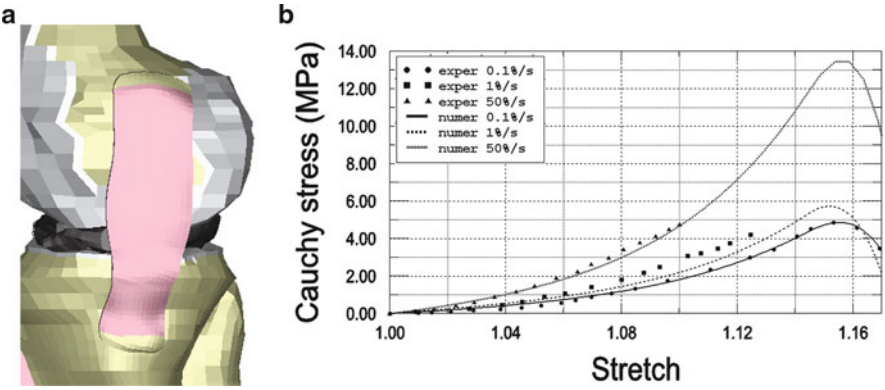


Fig. 19 Finite element model of the human MCL and stress-strain response of the human MCL. (**a**) Model; (**b**) stress-stretch response. *Source*: [52]

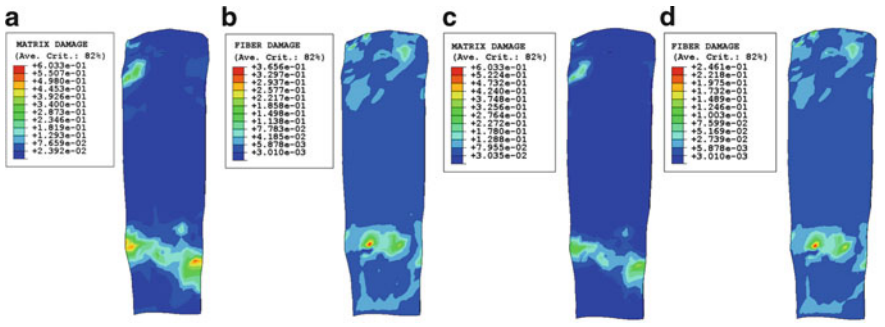


Fig. 20 Damage in a human MCL under different displacement rates. (**a**) Matrix damage at 0.01 mm/s; (**b**) fibre damage at 0.01 mm/s; (**c**) matrix damage at 113 mm/s; (**d**) fibre damage at 113 mm/s. *Source*: [52]

rate, damage in matrix and fibers was much lower than that at 0.01 mm/s. This effect is especially evident in the fibers where damage decreased from 0.36 during the quasi-static test (0.01 mm/s) to 0.24 in the impact test (113 mm/s). The peak values appeared in the ligament substance (contact region between ligament and

tibial plate) as also has been reported in previous experimental studies [17, 70]. Damage during distraction usually appears in that region.

7 Conclusion

It is well known that fibered soft biological tissues are subject to finite deformations and that their mechanical behavior is highly nonlinear, anisotropic and essentially incompressible with non-zero residual stress and in the non-physiological domain presents viscoelasticity and damage. In this article, we have provided a critical review of the fundamental aspects in modeling this kind of the materials. The application of these constitutive relationships in the context of vascular system and knee joint mechanics has been presented. The increasing effort devoted to studies of mechanical models for soft fibred tissues and the applications aimed at refining basic and clinical analysis demonstrates the vitality of the field of biomechanics [29].

Acknowledgements The authors gratefully acknowledge research support from the Spanish Ministry of Science and Technology through the research projects DPI2011-27939-C02-01, DPI2011-15551-E and DPI2010-20746-C03-01, and the CIBER-BBN initiative.

References

1. Alastrué, V.: Some inelastic problems in the modeling of blood vessels. Applications to non-physiological states and vascular surgery. Ph.D. thesis, University of Zaragoza, Spain, Division of Structural Mechanics (2008)
2. Alastrué, V., Calvo, B., Peña, E., Doblare, M.: Biomechanical modeling of refractive corneal surgery. ASME J. Biomech. Eng. **128**, 150–160 (2006)
3. Alastrué, V., Peña, E., Martínez, M.A., Doblaré, M.: Assessing the use of the "opening angle method" to enforce residual stresses in patient-specific arteries. Ann. Biomed. Eng. **35**, 1821–1837 (2007)
4. Alastrué, V., García, A., Peña, E., Rodríguez, J.F., Martínez, M.A., Doblaré, M.: Numerical framework for patient-specific computational modelling of vascular tissue. Commun. Numer. Meth. Eng. **26**, 35–51 (2007)
5. Alastrué, V., Sáez, P., Martínez, M.A., Doblaré, M.: On the use of the Bingham statistical distribution in microsphere-based constitutive models for arterial tissue. Mech. Res. Commun. **37**, 700–706 (2007)
6. Alastrué, V., Martínez, M.A., Doblaré, M., Menzel, A.: Anisotropic micro-sphere-based finite elasticity applied to blood vessel modeling. J. Mech. Phys. Solids **57**, 178–203 (2009)
7. Arnoux, P.J., Chabrand, P., Jean, M., Bonnoit, J.: A visco-hyperelastic with damage for the knee ligaments under dynamic constraints. Comp. Meth, Biomech. Biomed. Eng. **5**, 167–174 (2002)
8. Arruda, E.M., Boyce, M.C.: A three-Ddimensional constitutive model for the large stretch behavior of rubber elastic materials. J. Mech. Phys. Solids **41**(2), 389–412 (1993)
9. Bingham, C.: An antipodally symmetric distribution on the sphere. Ann. Stat. **2**(6), 1201–1225 (1974)

10. Bonet, J., Wood, R.D.: Nonlinear Continuum Mechanics for Finite Element Analysis. Cambridge University Press, Cambridge (2008)
11. Butler, D.L., Guan, Y., Kay, M., Cummings, M., Feder, S., Levy, M.: Location-dependent variations in the material properties of the anterior cruciate ligament. J. Biomech. **25**, 511–518 (1992)
12. Calvo, B., Peña, E., Martínez, M.A., Doblare, M.: An uncoupled directional damage model for fibered biological soft tissues. Formulation and computational aspects. Int. J. Numer. Meth. Eng. **69**, 2036–2057 (2007)
13. Calvo, B., Peña, E., Martínez, M.A., Doblare, M.: Computational modeling of ligaments at non-physiological situations. Int. J. Comput. Vision Biomech. IJV&B. **1**, 107–115 (2008)
14. Calvo, B., Peña, E., Martins, P., Mascarenhas, T., Doblare, M., Natal, R., Ferreira, A.: On modeling damage process in vaginal tissue. J. Biomech. **42**, 642–651 (2009)
15. Chaudhry, H.R., Bukiet, B., Davis, A., Ritter, A.B., Findley, T.: Residual stress in oscillating thoracic arteries reduce circumferential stresses and stress gradient. J. Biomech. **30**, 57–62 (1997)
16. Chuong, C.J., Fung, Y.C.: On residual stress in arteries. ASME J. Biomech. Eng. **108**, 189–192 (1986)
17. Crisco, J.J., Moore, D.C., McGovern, R.D.: Strain-rate sensityvity of the rabbit MCL diminishes at traumatic loading rates. J. Biomech. **35**, 1379–1385 (2002)
18. Dingemans, K., Teeling, P., Lagendijk, J.H., Becker, A.E.: Extracellular matrix of the human aortic media: an ultrastructural histochemical and immunohistochemical study of the adult aortic media. Anat. Rec. **258**, 1–14 (2000)
19. Flory, P.J.: Thermodynamic relations for high elastic materials. Trans. Faraday Soc. **57**, 829–838 (1961)
20. Fung, Y.C.: Biomechanics. Mechanical propeties of living tissues. Springer, New York (1993)
21. Fung, Y.C., Liu, S.Q.: Changes of zero-stress state of rat pulmonary arteries in hypoxic hypertension. J. Appl. Physiol. **70**, 2455–2470 (1991)
22. García, A., Peña, E., Martínez, M.A.: Viscoelastic properties of the passive mechanical behavior of the porcine carotid artery: influence of proximal and distal positions. Biorheology **49**, 271–288 (2012)
23. Gardiner, J.C., Weiss, J.A., Rosenberg, T.D.: Strain in the human medial collateral ligament during valgus lading of the knee. Clin. Orthop. Relat. R. **391**, 266–274 (2001)
24. Gasser, T.C., Ogden, R.W., Holzapfel, G.A.: Hyperelastic modeling of arterial layers with distributed collagen fibre orientations. J. R. Soc. Interface **3**, 15–35 (2006)
25. Graft, B.K., Vanderby, R. Jr., Ulm, M.J.: Effect of preconditioning on the viscoelastic response of primate patellar tendon. Arthroscopy **10**, 90–96 (1994)
26. Hayes, W.C., Mockros, L.F.: Viscoelastic constitutive relations for human articular cartilage. J. Appl. Physiol **18**, 562–568 (1971)
27. Herz, C.S.: Bessel functions of matrix argument. Ann. Math. **61**(3), 474–523 (1955)
28. Holzapfel, G.A., Ogden, R.W.: Constitutive modeling of passive myocardium: a structurally based framework for material characterization. Philos. Trans. A Math. Phys. Eng. Sci. **367**, 3445–3475 (2009)
29. Holzapfel, G.A., Ogden, R.W.: Constitutive modeling of arteries. Philos. Trans. A Math. Phys. Eng. Sci. **466**, 1551–1597 (2010)
30. Holzapfel, G.A., Gasser, T.C., Ogden, R.W.: A new constitutive framework for arterial wall mechanics and a comparative study of material models. J. Elasticity **61**, 1–48 (2000)
31. Holzapfel, G.A., Gasser, T.C., Stadler, M.: A structural model for the viscoelastic behaviour of arterial walls: continuum formultaion and finite element analysis. Eur. J. Mech. A Solids **21**, 441–463 (2002)
32. Holzapfel, G.A., Gasser, C.T., Sommer, G., Regitnig, P.: Determination of the layer-specific mechanical properties of human coronary arteries with non-atherosclerotic intimal thickening, and related constitutive modeling. Am. J. Physiol Heart Circ. Physiol **289**, H2048–H2058 (2005)

33. Hsu, E.W., Muzikant, A.L., Matulevicius, S.A., Penland, R.C., Henriquez, C.S.: Magnetic resonance myocardial fiber-orientation mapping with direct histological correlation. Am. J. Physiol Heart Circ. Physiol **274**, H1627–H1634 (1998)
34. Humphrey, J.D.: Mechanics of the arterial wall: review and directions. Crit. Rev. Biomed. Eng. **23**, 1–162 (1995)
35. Humphrey, J.D., Yin, F.C.P.: Constitutive relations and finite deformations of passive cardiac tissue II: stress analysis in the left ventricle. Circ. Res. **65**, 805–817 (1989)
36. Johnson, G.A., Livesay, G.A., Woo, S.L.Y., Rajagopal, K.I.R.: A single integral finite strain viscoelastic model of ligaments and tendons. ASME J. Biomech. Eng. **118**, 221–226 (1996)
37. Lanir, Y.: A structural theory for the homogeneous biaxial stress-strain relationship in flat collageneous tissues. J. Biomech. **12**, 423–436 (1979)
38. Lanir, Y.: Constitutive equations for fibrous connective tissues. J. Biomech. **16**, 1–12 (1983)
39. Lin, D.H.S., Yin, F.C.P.: A multiaxial constitutive law for mammalian left ventricular myocardium in steady-state barium contracture or tetanus. ASME J. Biomech. Eng. **120**, 504–517 (1998)
40. Marsden, J.E., Hughes, T.J.R.: Mathematical Foundations of Elasticity. Dover, New York (1994)
41. Martins, P., Peña, E., Calvo, B., Doblaré, M., Mascarenhas, T., Jorge, R.N., Ferreira, A.: Prediction of nonlinear elastic behavior of vaginal tissue: experimental results and model formulation. Comp. Meth. Biomech. Biomed. Eng. **13**(3), 327–337 (2010, in press)
42. Natali, A.N., Pavan, P.G., Carniel, E.L., Dorow, C.: A transverselly isotropic elasto-damage constitutive model for the periodontal ligament. Comp. Meth. Biomech. Biomed. Eng. **6**, 329–336 (2003)
43. Ogden, R.W.: Large deformation isotropic elasticity II: on the correlation of theory and experiment for compresible rubberlike solids. Proc. R. Soc. Lond. A Math. Phys. Sci. 328, 567–583 (1972)
44. Peña, E.: Evolution equations for the internal damage variables for soft biological fibred tissues. Mech. Res. Commun. **38**, 610–615 (2011)
45. Peña, E.: A rate dependent directional damage model for fibred materials. Application to soft biological tissues. Comp. Mech. **48**, 407–420 (2011)
46. Peña, E., Calvo, B., Martinez, M.A., Doblaré, M.: A finite element simulation of the effect of graft stiffness and graft tensioning in ACL reconstruction. C. Biomech. **20**, 636–644 (2005)
47. Peña, E., Calvo, B., Martinez, M.A., Doblaré, M.: A three-dimensional finite element analysis of the combined behavior of ligaments and menisci in the healthy human knee joint. J. Biomech. **39**(9), 1686–1701 (2006)
48. Peña, E., Calvo, B., Martínez, M.A., Doblaré, M.: On the numerical treatment of initial strains in soft biological tissues. Int. J. Numer. Meth. Eng. **68**, 836–860 (2006)
49. Peña, E., Calvo, B., Martínez, M.A., Doblaré, M.: An anisotropic visco-hyperelastic model for ligaments at finite strains: formulation and computational aspects. Int. J. Solids Struct. **44**, 760–778 (2007)
50. Peña, E., del Palomar, A.P., Calvo, B., Martínez, M.A., Doblaré, M.: Computational modeling of diarthrodial joints. Physiological, pathological and pos-surgery simulations. Arch. Comput. Method Eng. **14**(1), 47–91 (2007)
51. Peña, E., Calvo, B., Martínez, M.A., Doblaré, M.: Computer simulation of damage on distal femoral articular cartilage after meniscectomies. Comput. Biol. Med. **38**, 69–81 (2008)
52. Peña, E., Calvo, B., Martínez, M.A., Doblaré, M.: On finite strain damage of viscoelastic fibred materials: application to soft biological tissues. Int. J. Numer. Meth. Eng. **74**, 1198–1218 (2008)
53. Peña, E., Peña, J.A., Doblaré, M.: On the Mullins effect and hysteresis of fibered biological materials: a comparison between continuous and discontinuous damage models. Int. J. Solids Struct. **46**, 1727–1735 (2009)
54. Peña, E., Alastrue, V., Laborda, A., Martínez, M.A., Doblare, M.: A constitutive formulation of vascular tissue mechanics including viscoelasticity and softening behaviour. J. Biomech. **43**, 984–989 (2010)

55. Pinsky, P.M., Datye, V.: A microstructurally-based finite element model of the incised human cornea. J. Biomech. **10**, 907–922 (1991)
56. Pioletti, D.P., Rakotomanana, L., Leyvraz, P.F., Benvenuti, J.F.: Finite element model of the anterior cruciate ligament. Comp. Meth. Biomech. Biomed. Eng. (1997)
57. Provenzano, P.P., Heisey, D., Hayashi, K., Lakes, R., Vanderby, R.: Subfailure damage in ligament: a structural and cellular evaluation. J. Appl. Physiol **92**, 362–371 (2002)
58. Puso, M.A., Weiss, J.A.: Finite element implementation of anisotropic quasilinear viscoelasticity. ASME J. Biomech. Eng. **120**, 162–170 (1998)
59. Rachev, A., Hayashi, K.: Theoretical study of the effects of vascular smooth muscle contraction on strain and stress distributions in arteries. Ann. Biomed. Eng. **27**(4), 459–468 (1999)
60. Rodríguez, J.F., Cacho, F., Bea, J.A., Doblaré, M.: A stochastic-structurally based three dimensional finite-strain damage model for fibrous soft tissue. J. Mech. Phys. Solids **54**, 564–886 (2006)
61. Rodríguez, J.F., Alastrue, V., Doblaré, M.: Finite element implementation of a stochastic three dimensional finite-strain damage model for fibrous soft tissue. Comput. Methods Appl. Mech. Eng. **197**, 946–958 (2008)
62. Simo, J.C.: On a fully three-dimensional finite-strain viscoelastic damage model: formulation and computational aspects. Comput. Methods Appl. Mech. Eng. **60**, 153–173 (1987)
63. Simo, J.C., Hughes, T.J.R.: Computational Inelasticity. Springer, New York (1998)
64. Simo, J.C., Ju, J.W.: Strain- and stress-based continuum damage models. I. Formulation. Int. J. Solids Struct. **23**, 821–840 (1987)
65. Simo, J.C., Taylor, R.L., Pister, K.S.: Variational and projection methods for the volume constraint in finite deformation elasto-plasticity. Comput. Methods Appl. Mech. Eng. **51**, 177–208 (1985)
66. Spencer, A.J.M.: Theory of Invariants. In: Continuum Physics, pp. 239–253. Academic, New York (1971)
67. Viidik, A.: Structure and function of normal and healing tendons and ligaments. In: Mow, V.C., Ratchiffe, A., Woo, S.L.Y. (eds) Biomecanics of Diarthorial Joints. Springer, New York (1990)
68. Weiss, J.A., Maker, B.N., Govindjee, S.: Finite element implementation of incompressible, transversely isotropic hyperelasticity. Comput. Methods Appl. Mech. Eng. **135**, 107–128 (1996)
69. Weiss, J.A., Gardiner, J.C., Bonifasi-Lista, C.: Ligament material behavior is nonlinear, viscoelastic and rate-independent under shear loading. J. Biomech. **35**, 943–950 (1996)
70. Woo, S.L.Y., Peterson, R.H., Ohland, K.J., Sites, T.J., Danto, M.I.: The effects of strain rate on the properties of the medial collateral ligament in skeletally inmatura and mature rabbits: a biomechanical and histological study. J. Orthopaed. Res. **8**, 712–721 (1990)

Some Remarks on Avalanches Modelling: An Introduction to Shallow Flows Models

Enrique D. Fernández-Nieto and Paul Vigneaux

These notes are dedicated to D. Antonio Valle Sánchez (1930–2012). D. Antonio was the first Spanish PhD student of Jacques-Louis Lions. He can be considered as one of the founders of modern Applied Mathematics in Spain.

Abstract The main goal of these notes is to present several depth-averaged models with application in granular avalanches. We begin by recalling the classical Saint-Venant or Shallow Water equations and present some extensions like the Saint-Venant–Exner model for bedload sediment transport. The first part is devoted to the derivation of several avalanche models of Savage–Hutter type, using a depth-averaging procedure of the 3D momentum and mass equations. First, the Savage–Hutter model for aerial avalanches is presented. Two other models for partially fluidized avalanches are then described: one in which the velocities of both the fluid and the solid phases are assumed to be equal, and another one in which both velocities are unknowns of the system. Finally, a Savage–Hutter model for submarine avalanches is derived. The second part is devoted to non-newtonian models, namely viscoplastic fluids. Indeed, a one-phase viscoplastic model can also be used to simulate fluidized avalanches. A brief introduction to Rheology and plasticity is presented in order to explain the Herschel–Bulkley constitutive law. We finally present the derivation of a shallow Herschel–Bulkley model.

E.D. Fernández-Nieto (✉)
Dpto. Matemática Aplicada I, Universidad de Sevilla, Sevilla, Spain
e-mail: edofer@us.es

P. Vigneaux
Unité de Mathématiques Pures et Appliquées, Ecole Normale Supérieure de Lyon, Lyon, France
e-mail: Paul.Vigneaux@math.cnrs.fr

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_2,

1 Introduction: Shallow Water Equations

The classical shallow water equations were first derived in 1871 by Saint-Venant (see [46]). This system of equations describe the motion of a shallow layer of fluid in a channel, a lake, coastal areas, etc.

Several extensions of these classical equations have been proposed in the literature. For example, in [25] Gerbeau and Perthame propose a viscous Shallow Water model. They perform the asymptotic analysis of the Navier–Stokes equations where friction effects at the bottom have been taken into account. While in a first order approximation the viscous terms do not appear in the equations, a second order is needed to get them. In [35], a viscous one layer 2D Shallow-Water system is derived, by including a surface-tension term associated to the capillary effects at the free surface and a quadratic friction term at the bottom. These terms have been useful to prove the existence of global weak solutions in [13].

In the simple case of a rectangular channel with constant width and a fixed bottom topography (see Fig. 1), the Shallow Water equations are

$$\begin{cases} \partial_t h + \partial_x (h\,u) = 0, \\ \partial_t (hu) + \partial_x \left(hu^2 + \dfrac{1}{2} g h^2 \right) = -gh\partial_x z_b - ghS_f, \end{cases} \tag{1}$$

where x denotes the horizontal variable through the axis of the channel and t is the time variable. $u(x,t)$ and $h(x,t)$ represent the velocity and the height of the water column, respectively. g is the gravity and $z_b(x)$ the bottom topography (see Fig. 1).

The term S_f models the friction forces. In the particular case of the Manning law we have

$$S_f = \frac{g\eta^2 |u| u}{R_h^{4/3}}, \tag{2}$$

where η is the Manning's coefficient and R_h is the hydraulic radius, which can be approximated by h.

The Saint-Venant–Exner equations take into account the bed-load sediment transport. In this case, we have the Shallow Water or Saint-Venant system coupled with a continuity equation to model the evolution of the sediment layer,

$$\begin{cases} \partial_t h + \partial_x (h\,u) = 0, \\ \partial_t (hu) + \partial_x \left(hu^2 + \dfrac{1}{2} g h^2 \right) = -gh\partial_x z_b - ghS_f, \\ \partial_t z_b + \xi \partial_x q_b = 0, \end{cases} \tag{3}$$

where $\xi = 1/(1-\psi_0)$ and ψ_0 is the porosity of the sediment layer. $q_b = q_b(h,q)$ represents the solid transport discharge. The definition of the solid transport

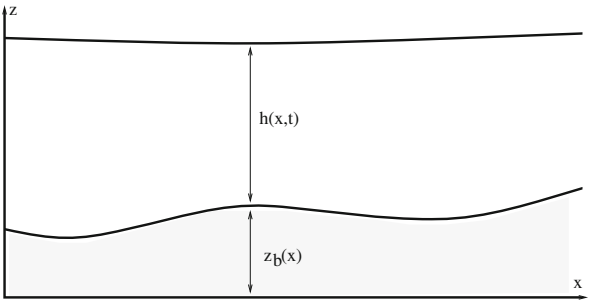


Fig. 1 Shallow water equations with a fixed bottom

discharge is set usually by empirical laws. In the Appendix some classical formulae are presented.

Far to be exhaustive, several extensions of the Shallow Water equations can be mentioned: models that take into account varying bottom topography [11, 18, 24]; models to study erosion phenomena with local coordinate variable in space and time [12] or to study flows in rotating drums [27]; models that take into account dispersive effects [33]; models for two-layer stratified flows with viscosity and capillarity [39], turbidity currents models [38], multilayer shallow water models to incorporate tridimensional effects [7, 23] . . .

In the pioneering work of Savage and Hutter [47], a shallow-water type model has been proposed to study aerial avalanches. In the following section we describe the derivation of the Savage–Hutter model. The classical Shallow Water system is the particular case of the Savage–Hutter model obtained by neglecting the Coulomb friction term. Therefore, its derivation from Navier–Stokes equations is a particular case of the general study presented in next section.

In the following sections, the derivation of several shallow water type models to study three different types of avalanches are presented. Sections 2–6 correspond to Savage–Hutter type models for aerial avalanches, partially fluidized aerial avalanches and submarine avalanches (see [21]). In Sect. 7, we present a brief introduction to Rheology and plasticity in order to explain the constitutive equation of the Herschel–Bulkley model. A depth-averaged Herschel–Bulkley model is presented in Sect. 8. This model is a one-phase approach to study solid–fluid mixtures avalanches and represents an alternative to the two-phase Coulomb approach.

2 Savage–Hutter Model for Aerial Avalanches

Numerical modelling of sub-aerial debris or snow avalanches has been extensively investigated during this last decade with application to both laboratory experiments dealing with granular flows and geological events (see for example [2, 5, 6, 12, 27, 31, 34, 50]). Most of the models devoted to gravitational granular flows describe the behavior of dry granular material following the pioneering work of Savage and Hutter (see [47]) in which a shallow water type model (i.e. thin layer approximation

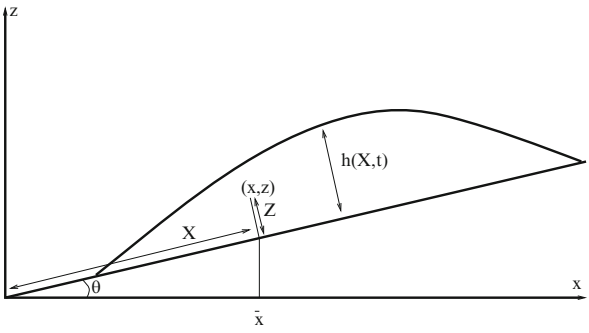


Fig. 2 Local coordinates

for a continuum medium) is derived to describe granular flows over a slopping plane based on Mohr–Coulomb considerations: a Coulomb friction is assumed to reflect the avalanche/bottom interaction and the normal stress tensor is defined by a constitutive law relating the longitudinal and the normal stresses through a proportionality factor K.

New Savage–Hutter models over a general bottom have been proposed. For example in [11], Bouchut et al., propose a Savage–Hutter type model for aerial avalanche which takes into account the curvature of the bottom. A two-layer Shallow Water type model with compressible effects has been introduced in [37] by Morales de Luna. He considers an upper compressible layer and a lower incompressible layer.

In this section, we present the derivation of the Savage–Hutter model over a plane with constant slope. First, we consider the Euler equations in Cartesian coordinates $\mathbf{X} = (x, z)$,

$$\mathbf{V} = \begin{pmatrix} u \\ w \end{pmatrix}, \qquad \nabla \cdot \mathbf{V} = 0, \tag{4}$$

$$\partial_t(\rho \mathbf{V}) + \rho \mathbf{V} \cdot \nabla_{\mathbf{X}} \mathbf{V} = -\nabla \cdot P + \rho \nabla_{\mathbf{X}}(\mathbf{g} \cdot \mathbf{X}), \tag{5}$$

where $\mathbf{g} = (0, -g)$, g being the gravity acceleration, $\mathbf{V}$ is the velocity field and ρ, the density of the granular layer. Moreover, we denote by P the negative Cauchy stress tensor, also named pressure tensor,

$$P = \begin{pmatrix} p_{xx} & p_{xz} \\ p_{zx} & p_{zz} \end{pmatrix},$$

with $p_{xz} = p_{zx}$.

Let us rewrite first the Euler equations in local coordinates (X, Z) on an inclined plain whose slope is $\tan(\theta)$ (see Fig. 2). Z is the distance between the points (x, z) and $(\bar{x}, b(\bar{x}))$, where

$$b(\bar{x}) = \tan(\theta)\, \bar{x}.$$

That is, Z is the distance to the bed, measured along the normal direction and X measures the arc length along the inclined plain. $\bar{x}$ is the x-Cartesian coordinate of the point $(X, 0)$ (see Fig. 2).

Let $h(x,t)$ be the height of the granular layer along the normal direction to the bed. The domain is

$$\{(X,Z); \quad X \in [0,L],\ 0 < Z < h(X,t)\}. \tag{6}$$

The relation between the Cartesian coordinates $\mathbf{X} = (x,z)$ and the coordinates $(\bar{x}, Z)$ is

$$\mathbf{X} = \Big(\bar{x} - Z\sin\theta, b(\bar{x}) + Z\cos\theta)\Big), \tag{7}$$

where $(\bar{x}, b(\bar{x}))$ is a point of the bed.
The following definitions will also be used:

- U and W are the tangential and normal velocities, respectively,

$$\begin{pmatrix} U \\ W \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \mathbf{V}.$$

- And $\mathscr{P}$ is the rotated Cauchy stress tensor:

$$\mathscr{P} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} P \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \mathscr{P}_{XX} & \mathscr{P}_{XZ} \\ \mathscr{P}_{ZX} & \mathscr{P}_{ZZ} \end{pmatrix}.$$

  Note that, as $p_{xz} = p_{xz}$, then $\mathscr{P}_{XZ} = \mathscr{P}_{ZX}$.

  Equations (4) and (5) are re-written in the new variables as follows:

$$\begin{cases} \partial_X(U) + \partial_Z(W) = 0, \\ \\ \rho\partial_t(U) + \rho\partial_X(U^2) + \rho\partial_Z(WU) - \rho\partial_X(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{XX}) - \partial_Z(\mathscr{P}_{XZ}), \\ \\ \rho\partial_t(W) + \rho\partial_X(U\,W) + \rho\partial_Z(W^2) - \rho\partial_Z(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{ZX}) - \partial_Z(\mathscr{P}_{ZZ}). \end{cases} \tag{8}$$

In what follows, the derivation of the model proposed by Savage and Hutter in [47] to study aerial avalanches is described following the items:

- $[\partial]$ Boundary and kinematic conditions.
- $[\tilde{A}]$ Dimensional analysis.
- $[\updownarrow]$ Hydrostatic pressure and constitutive law.
- $[M]$ Momentum conservation law.

- $[\int]$ Integration process.
- $[\hookrightarrow]$ Final system of equations.

2.1 $[\partial]$ *Boundary and Kinematic Conditions*

We denote by $\mathbf{n}^h$ the unit normal vector to the free granular surface $Z = h$ with positive vertical component, and by $\mathbf{n}^0 = (0, 1)$ the unit normal vector to the bottom ($Z = 0$).

The following kinematic condition is considered

$$\partial_t h + U|_{Z=h}\partial_X h - W|_{Z=h} = 0, \tag{9}$$

which means that the particles at the free surface are transported with velocity $(U|_{Z=h}, W|_{Z=h})$.

The following boundary conditions are imposed:

- On $Z = h$:

$$\mathbf{n}^h \cdot \mathscr{P}\mathbf{n}^h = 0 \tag{10}$$

$$\mathscr{P}\mathbf{n}^h - \mathbf{n}^h(\mathbf{n}^h \cdot \mathscr{P}\mathbf{n}^h) = \begin{pmatrix} fric_h(U) \\ 0 \end{pmatrix} \quad i = 1, 2, \tag{11}$$

  where $fric_h(U)$ is the friction term between the granular layer and the air. For the sake of simplicity we will suppose that $fric_h(U) = 0$.
- On $Z = 0$:

$$(U, W) \cdot \mathbf{n}^0 = 0 \quad \Rightarrow \quad W = 0, \tag{12}$$

$$\mathscr{P}\mathbf{n}^0 - \mathbf{n}^0(\mathbf{n}^0 \cdot \mathscr{P}\mathbf{n}^0) = \begin{pmatrix} -\mathbf{n}^0 \cdot \mathscr{P}\mathbf{n}^0 \left.\dfrac{U}{|U|}\right|_{Z=0} \tan\delta_0 \\ 0 \end{pmatrix}. \tag{13}$$

  This last condition corresponds to a Coulomb friction law, defined in terms of the angle of repose δ_0 (see [47]).

2.2 $[\tilde{A}]$ *Dimensional Analysis*

Next, a dimensional analysis of the set of Eqs. (8), the kinematic and boundary conditions is performed. The non-dimensional variables ($\tilde{\cdot}$) read:

$$
\begin{aligned}
&(X, Z, t) = (L\tilde{X}, H\tilde{Z}, (L/g)^{1/2}\tilde{t}),\\
&(U, W) = (Lg)^{1/2}(\tilde{U}, \varepsilon\tilde{W}),\\
&h = H\tilde{h},\\
&(\mathscr{P}_{XX}, \mathscr{P}_{ZZ}) = gH(\tilde{\mathscr{P}}_{XX}, \tilde{\mathscr{P}}_{ZZ}),\\
&\mathscr{P}_{XZ} = gH\mu\tilde{\mathscr{P}}_{XZ},
\end{aligned}
\tag{14}
$$

where:

- $\mu = \tan\delta_0$, δ_0 being the angle of repose in the Coulomb term.
- By L and H, we denote, respectively, the tangential and normal characteristic lengths.
- $\varepsilon = H/L$, which is supposed to be small: the Savage–Hutter model has been shown to reproduce experimental granular collapse over horizontal plane for aspect ratio $\epsilon \leq 0.5$, see [34].

Using the above change of variables, the system of Eqs. (8) is re-written as follows:

$$\partial_X(U) + \partial_Z(W) = 0, \tag{15}$$

$$\partial_t(\rho U) + \rho U\,\partial_X U + \rho W\,\partial_Z U + \rho\partial_X\left(b + Z\cos\theta + \frac{\mathscr{P}_{XX}}{\rho}\right)\varepsilon = -\mu\partial_Z(\mathscr{P}_{XZ}), \tag{16}$$

$$
\begin{aligned}
&\varepsilon\{\partial_t(\rho W) + \rho U\,\partial_X(W) + \rho W\,\partial_Z(W) + \partial_X(\mathscr{P}_{XZ})\} +\\
&+ \rho\partial_Z(b + Z\cos\theta) = -\partial_Z(\mathscr{P}_{ZZ}),
\end{aligned}
\tag{17}
$$

where tildes have been dropped for simplicity.

The kinematic condition (9) is re-written as:

$$\partial_t h + U|_{Z=h}\partial_X h - W|_{Z=h} = 0. \tag{18}$$

Finally, the boundary conditions (10)–(13) are now given by:

- On $Z = h$, we have $\mathbf{n}^h = (-\varepsilon\partial_X h, 1)/\varphi^S$ with $\varphi^S = \sqrt{1 + \varepsilon^2(\partial_X h)^2}$, then from (10) and (11) we obtain

$$-\varepsilon\partial_X h\,\mathscr{P}_{XX} + \mu\mathscr{P}_{ZX} = 0, \tag{19}$$

$$-\varepsilon\partial_X h\mu\mathscr{P}_{XZ} + \mathscr{P}_{ZZ} = 0. \tag{20}$$

- On $Z = 0$, we have $\mathbf{n}^0 = (0, 1)$, then from (12) and (13) we obtain

$$W|_{Z=0} = 0, \tag{21}$$

$$\mu \mathscr{P}_{XZ} = -\mathscr{P}_{ZZ} \left. \frac{U}{|U|} \right|_{Z=0} \tan \delta_0. \tag{22}$$

2.3 [↕] *Hydrostatic Pressure and Constitutive Law*

From (17) we obtain

$$\partial_Z(\mathscr{P}_{ZZ}) = -\rho \cos\theta + \mathscr{O}(\varepsilon). \tag{23}$$

If we integrate (23) from $Z > 0$ to h, we have, up to order ε,

$$\mathscr{P}_{ZZ} = \rho(h - Z)\cos\theta. \tag{24}$$

The following constitutive law is considered (see [47])

$$\mathscr{P}_{XX} = K\, \mathscr{P}_{ZZ},$$

where K measures the anisotropy or normal stress effects: while $K = 1$ corresponds to isotropic conditions, $K \neq 1$ makes 'overburden pressures' different from the normal stresses parallel to the basal surface. In the case of the Shallow Water equations, $K = 1$ is assumed.

The coefficient K is defined according to the motion of the granular layer (see [45]):

$$K = \begin{cases} K_{act} \text{ if } \partial_X U > 0, \\ K_{pas} \text{ if } \partial_X U < 0, \end{cases}$$

with

$$K_{act/pas} = 2\sec^2\phi \left(1 \mp (1 - \cos^2\phi \sec^2\delta_0)^{1/2} \right) - 1,$$

being ϕ the internal friction angle, defined in terms of the type of grains and size.

The definition of K can be done in different ways. For example, while in [28] Heinrich et al. consider $K = 1$, other definitions of K can be found in [29]. The effects related to the definition of K in numerical modelling of experimental and natural flows is studied in [43, 45].

Using the previous relations, we have, up to order ϵ,

$$\mathscr{P}_{XX} = K\, \mathscr{P}_{ZZ} = K\rho(h - Z)\cos\theta. \tag{25}$$

2.4 [M] *Momentum Conservation Law: With Hydrostatic Pressure and Anisotropy of the Normal Stress*

By replacing (24) and (25) in (16) and using the incompressibility equation (15), we obtain, up to second order,

$$\partial_t(\rho U) + \rho\partial_X U^2 + \rho\partial_Z(UW) + \rho\partial_X\left(b + Z\cos\theta + K(h-Z)\cos\theta\right)\varepsilon = -\mu\partial_Z(\mathscr{P}_{XZ}). \tag{26}$$

2.5 [$\int$] *Integration Process*

In this section, the mass equation (15) and the momentum equation (26) are depth-averaged in the normal direction. Let us introduce the following notation: we denote by $\bar{U}$ the average of the velocity along the normal direction:

$$\bar{U} = \frac{1}{h}\int_0^h U(X,Z)dZ.$$

We also introduce the notation:

$$\overline{U^2} = \frac{1}{h}\int_0^h U^2(X,Z)dZ.$$

If Eq. (15) is integrated from $Z = 0$ to $Z = h$, we obtain

$$0 = \partial_X(h\bar{U}) - U|_{Z=h}\partial_X h + W|_{Z=h} - W|_{Z=0}.$$

Now, using (18) and (21), the averaged mass equation is obtained

$$\partial_t h + \partial_X(h\bar{U}) = 0.$$

Let us now integrate Eq. (26) from $Z = 0$ to $Z = h$. As in the previous case, we use the kinematic condition (18) to obtain

$$\partial_t(h\bar{U}) + \partial_X(h\overline{U^2}) + \left(\int_0^h \partial_X\left(b + Z\cos\theta + K(h-Z)\cos\theta\right)dZ\right)\varepsilon = -\frac{\mu}{\rho}(\mathscr{P}_{XZ}(h) - \mathscr{P}_{XZ}(0)). \tag{27}$$

Moreover, we have

$$\int_0^h \partial_X \left(b + Z\cos\theta + (h-Z)\cos\theta K \right) dZ = h\partial_X b + \partial_X \left(\frac{h^2}{2}\cos\theta K \right).$$

By replacing this last expression in (27) we obtain the equation

$$\partial_t(h\bar{U}) + \partial_X \left(h\overline{U^2} + \varepsilon \frac{h^2}{2}\cos\theta K \right) = -\varepsilon h\partial_X b - \frac{\mu}{\rho}(\mathscr{P}_{XZ}(h) - \mathscr{P}_{XZ}(0)). \quad (28)$$

Then, the boundary conditions and the constitutive laws are used to derive $\mu\,\mathscr{P}_{XZ}(h)$ and $\mu\,\mathscr{P}_{XZ}(0)$:

- From (19), by using (24) and $\mathscr{P}_{XX} = K\mathscr{P}_{ZZ}$, we have

  $$\mu\mathscr{P}_{XZ}(h) = \varepsilon\partial_X h\mathscr{P}_{ZZ}K.$$

  In [27] Gray introduced the assumption that the Coulomb term is of order γ for some $\gamma \in (0,1)$. That is, $\mu = \tan\delta_0 = \mathscr{O}(\varepsilon^\gamma)$. Under this assumption, we have

  $$\mu\mathscr{P}_{XZ}(h) = \mathscr{O}(\varepsilon^{1+\gamma}). \quad (29)$$

- Using Eq. (22), we obtain

  $$\frac{\mu}{\rho}\mathscr{P}_{XZ}(0) = -\frac{\mathscr{P}_{ZZ}(0)}{\rho}\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0 = -h\cos\theta\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0.$$

  Therefore, assuming $\tan\delta_0 = \mathscr{O}(\varepsilon^\gamma)$, we have

  $$\frac{\mu}{\rho}\mathscr{P}_{XZ}(0) = -h\cos\theta\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}). \quad (30)$$

Finally, substituting (29) and (30) in (28), the averaged momentum equation is obtained:

$$\partial_t(h\bar{U}) + \partial_X \left(h\overline{U^2} + \varepsilon\frac{h^2}{2}\cos\theta K \right) = -\varepsilon h\partial_X b - h\cos\theta\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}).$$

2.6 [↪] *Final System of Equations*

Coming back to the original variables, using (14), neglecting terms of order $\varepsilon^{1+\gamma}$ and supposing a constant profile of the velocities, the following system is obtained:

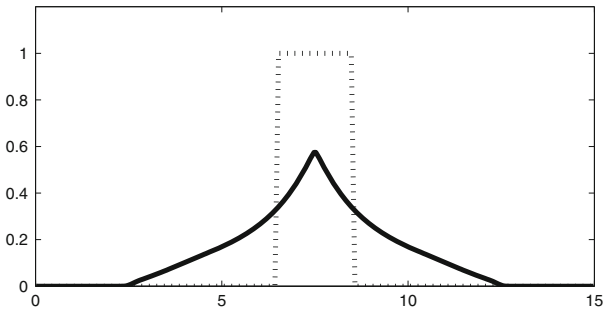


Fig. 3 Savage–Hutter model. *Dotted line*: initial profile of the granular layer. *Continuous line*: stationary profile of the granular layer

$$\begin{cases} \partial_t h + \partial_X (h\bar{U}) = 0, \\ \partial_t (h\bar{U}) + \partial_X \left(h\bar{U}^2 + g\cos\theta \dfrac{h^2}{2} K \right) = -gh\partial_X b + \mathscr{T}, \end{cases} \tag{31}$$

where $\mathscr{T}$ represents the Coulomb friction term. This term must be understood as follows:

$$\begin{aligned} &\text{If } |\mathscr{T}| \geq \sigma_c \quad \Rightarrow \quad \mathscr{T} = -gh\cos\theta \frac{\bar{U}}{|\bar{U}|} \tan\delta_0, \\ &\text{If } |\mathscr{T}| < \sigma_c \quad \Rightarrow \quad \bar{U} = 0, \end{aligned} \tag{32}$$

where $\sigma_c = gh\cos\theta\tan\delta_0$.

Let us illustrate the effects of the Coulomb friction term. We consider a test case consisting of a granular layer over a flat bottom whose initial profile is rectangular. The evolution of the layer is simulated by numerically solving System (31). Let us stress the importance of an adequate treatment of the Coulomb friction term (32) to obtain satisfactory numerical results (see for example [34]). In Fig. 3 the continuous line corresponds to the stationary profile of the granular layer for $\delta_0 = 25°$. The initial condition is represented too (dotted line). The main difference between the classical Shallow Water equations and the Savage–Hutter model is the presence of the Coulomb friction term: if a closed domain is considered and the Coulomb friction term is neglected, the stationary solution is a horizontal free surface, corresponding to water at rest.

In Fig. 4 the evolution of the granular layer surface and its discharge is represented at several times. Observe that, while at $t = 2$ the solution is stationary, at $t = 1.5$ only the front of the avalanche is still moving.

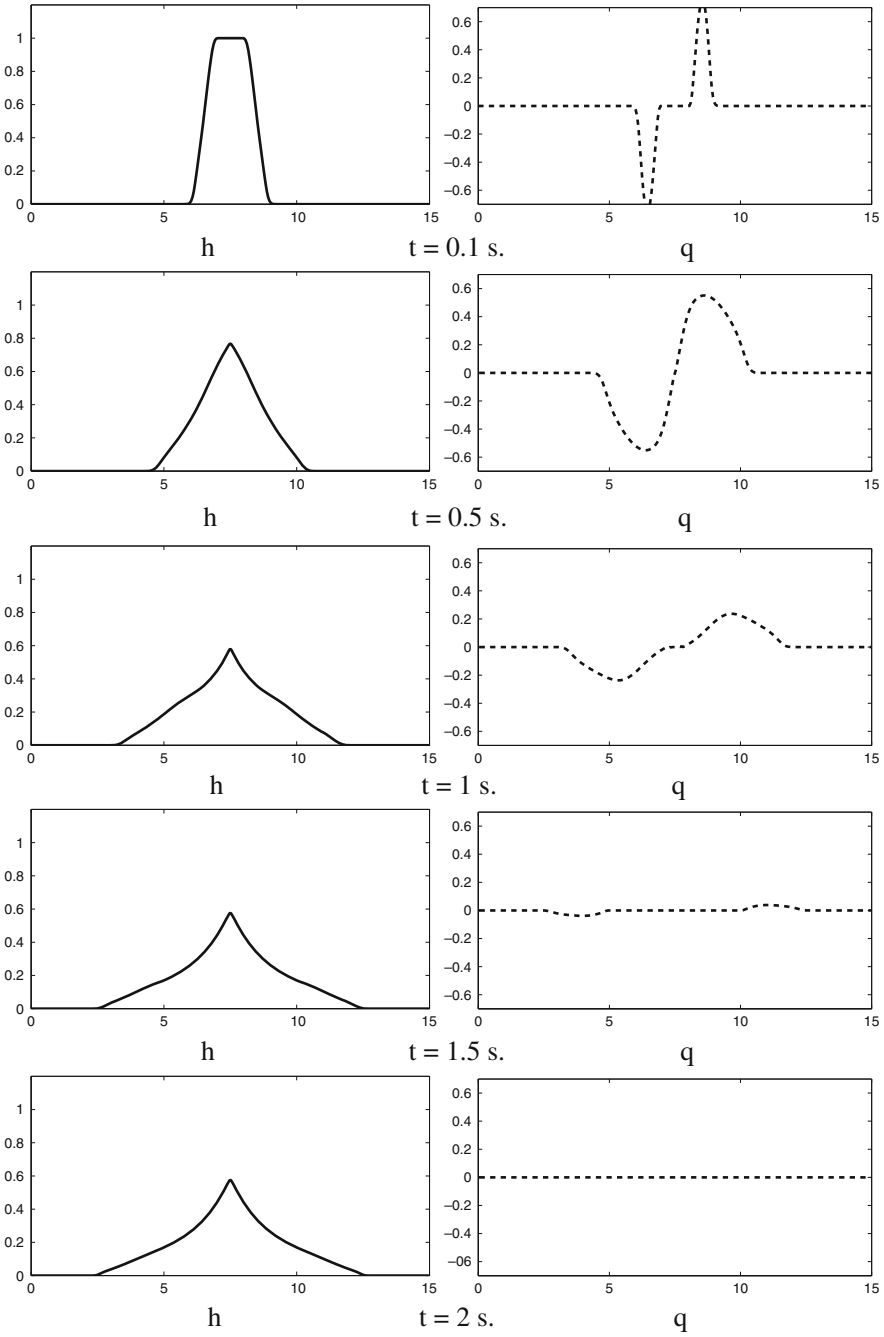


Fig. 4 Savage–Hutter model. *Left*: *continuous line*: granular free surface. *Right*: *dashed line*: discharge

3 Fluid–Solid Mixture Aerial Avalanches

In most practical applications to real debris flows, the fluid which is present in the granular material can not be neglected. Recent attempts have been developed to describe mixtures of grains and fluid in shallow water two-phase or mixture models [29, 42, 44].

The model introduced by Jackson in [30] allows to model geophysical mass flows containing a mixture of solid and fluid materials, by taking into account buoyancy effects. It is defined by the mass and momentum equations for each phase.

Let us use the following notation: subscript "s" refers to the solid phase and subscript "f" to the fluid one. The solid volume fraction is denoted by φ. The grain density, ρ_s and the fluid density, ρ_f, are supposed to be constant.

Then, the two-phase model is defined by the following mass and momentum equations:

$$\partial_t(\rho_s\varphi) + \operatorname{div}(\rho_s\varphi\mathbf{V}_s) = 0, \tag{33a}$$

$$\partial_t(\rho_f(1-\varphi)) + \operatorname{div}(\rho_f(1-\varphi)\mathbf{V}_f) = 0, \tag{33b}$$

$$\rho_s\varphi(\partial_t\mathbf{V}_s + \mathbf{V}_s\nabla\mathbf{V}_s) = -\operatorname{div}P_s + f_0 + \rho_s\varphi\nabla(\mathbf{g}\cdot\mathbf{X}), \tag{33c}$$

$$\rho_f(1-\varphi)(\partial_t\mathbf{V}_f + \mathbf{V}_f\nabla\mathbf{V}_f) = -\operatorname{div}P_f - f_0 + \rho_f(1-\varphi)\nabla(\mathbf{g}\cdot\mathbf{X}). \tag{33d}$$

Where P_s and P_f represent the stress tensors for the solid and the fluid phase, respectively. f_0 represents the averaged value of the resultant force exerted by the fluid on a solid particle.

To obtain Jackson's model, the force f_0 is decomposed into the buoyancy force f_B and all the remaining contributions f according to [4]:

$$f_0 = f_B + f = -\varphi\nabla p_f + f, \tag{34}$$

where p_f denotes the fluid pressure. The term f collects the drag force, the lift force and the virtual mass force (see [4, 30] for details). Here, we assume that f reduces to the drag force.

If we assume that the viscous forces related to the fluid are negligible, then the fluid stress tensor reduces to the pressure term:

$$\nabla\cdot P_f = \nabla p_f. \tag{35}$$

By taking this expression into (33c) and (33d), we obtain the system (33a), (33b) and

$$\rho_s \varphi(\partial_t \mathbf{V}_s + \mathbf{V}_s \nabla \mathbf{V}_s) = -\mathrm{div} P_s - \varphi \nabla p_f + f + \rho_s \nabla(\mathbf{g}\cdot\mathbf{X}), \tag{36a}$$

$$\rho_f (1-\varphi)(\partial_t \mathbf{V}_f + \mathbf{V}_f \nabla \mathbf{V}_f) = -(1-\varphi)\nabla p_f - f + \rho_f (1-\varphi)\nabla(\mathbf{g}\cdot\mathbf{X}). \tag{36b}$$

The model proposed by Pitman and Le in [44] and reformulated in [42], can be deduced following a dimensional analysis and an integration process of Jackson's model. They suppose a constant vertical profile of the velocity for the solid and the fluid phase: U_s and U_f, respectively. Pitman–Le model can be written as follows:

$$\partial_t(h\varphi) + \partial_X(h\varphi U_s) = 0; \tag{37a}$$

$$\partial_t(h(1-\varphi)) + \partial_X(h(1-\varphi)U_f) = 0; \tag{37b}$$

$$\begin{aligned} \partial_t(\varphi h U_s) + \partial_X(\varphi h U_s^2) = & -\frac{1}{2}(1-r)gh^2\cos\theta\,\partial_X\varphi \\ & - gh\cos\theta\,\varphi\,\partial_X h \\ & - gh\varphi\partial_X b \\ & + \beta h(U_f - U_s) + \mathscr{T}; \end{aligned} \tag{37c}$$

$$\begin{aligned} \partial_t((1-\varphi)hU_f) + \partial_X((1-\varphi)h\,U_f^2) = & -gh\cos\theta\,(1-\varphi)\,\partial_X h \\ & - gh\,(1-\varphi)\partial_X b \\ & - \frac{1}{r}\beta h(U_f - U_s); \end{aligned} \tag{37d}$$

where $r = \rho_f/\rho_s$, β is a friction coefficient between the phases (see [44]) and $\mathscr{T}$ is the Coulomb friction term.

Let us illustrate the influence of φ in the evolution of the avalanche. We consider first a test case consisting of a granular layer over a flat bottom whose initial profile is rectangular. The evolution of the layer is simulated by numerically solving System (37). We consider $r = 0.34$, $\delta_0 = 25°$ and $\varphi = 0.8$. In Fig. 5 the evolution of the layer is shown. The left column shows the total height, $h_s + h_f$ (continuous line). In order to make visible the evolution of the total solid volume fraction, h_s is also plotted in the figures (dotted line). Notice that at the front of the avalanche, the dotted line practically coincides with the continuous one, meaning that there is only

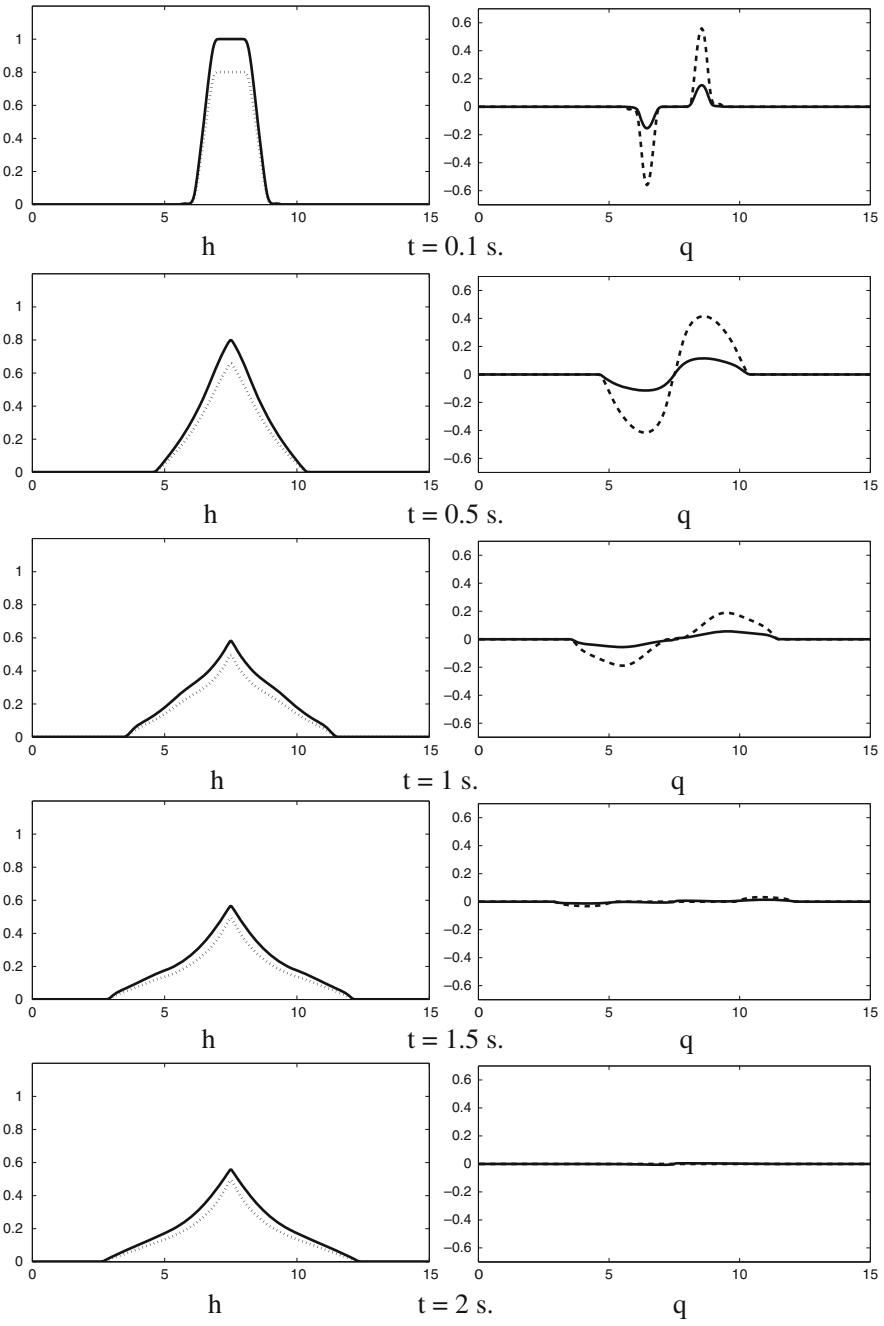


Fig. 5 Two-phase avalanche. $\varphi = 0.8$. *Left*: *continuous line*: granular free surface. *Dotted line*: h_s; *Right*: discharge. *Continuous line*: q_s. *Dashed line*: q_f

granular material near the front. The right column shows q_f and q_s. We can observe that the motion of the solid phase stops before. Figure 6 shows the evolution of the avalanche for $\varphi = 0.4$. The initial condition and the values of r and δ are the same. Let us remark that at the front the dotted line practically coincides with the horizontal axis, meaning that there is only fluid near of the front. If we compare these two simulations, we can also observe that the maximum heights and the total lengths of spreading of the avalanches are completely different. For $\varphi = 0.4$, the fluid goes out the domain.

The presence of an interstitial fluid in the avalanche, neglected in the Pitman–Le model, may have a strong influence in its evolution. The flow of fluidized avalanches can be much more complex than the ones simulated with the Pitman–Le model. For example, non-hydrostatic pressure effects, related to the pore fluid pressure, may appear. Iverson and Denlinger extended the Savage–Hutter model in [29] to study avalanches of fluidized granular masses where the pores between the grains are assumed to be filled with a fluid, under the assumption that the velocities of both phases coincide, and by including the bed pore fluid pressure as an unknown of the system. Let us study now the derivation of a simplified version of the model proposed by Iverson and Denlinger in [29] to study partially fluidized aerial avalanches (Fig. 7).

We consider a granular layer of density ρ_s and porosity ψ_0. We assume that the pores in the granular layer are filled with a fluid of density ρ_w. Then, the density of the fluidized layer is defined as

$$\rho = (1 - \psi_0)\rho_s + \psi_0 \rho_w. \tag{38}$$

As in the previous section, the model will be described in local coordinates over a plain with constant slope (see Fig. 2). Again U is the velocity parallel to the bottom; W, the velocity perpendicular to the bottom; and $\mathscr{P}$ the rotated pressure tensor.

Let us consider again the system of equations given by Euler equations in local coordinates:

$$\begin{cases} \partial_X(U) + \partial_Z(W) = 0, \\ \rho\partial_t(U) + \rho\partial_X(U^2) + \rho\partial_Z(WU) - \rho\partial_X(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{XX}) - \partial_Z(\mathscr{P}_{ZX}), \\ \rho\partial_t(W) + \rho\partial_X(UW) + \rho\partial_Z(W^2) - \rho\partial_Z(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{XZ}) - \partial_Z(\mathscr{P}_{ZZ}). \end{cases} \tag{39}$$

In a binary mixture model the pressure tensor of the mixture is given by

$$\mathscr{P} = \mathscr{P}^s + \mathscr{P}^f - \rho_w(U_f - U_b)\otimes(U_f - U_b) - \rho_s(U_s - U_b)\otimes(U_s - U_b),$$

where U_w is the velocity of the fluid phase, U_s is the velocity of the solid phase and $U_b = (\rho_w U_w + \rho_w U_s)/(\rho_s + \rho_w)$ is the barycentric velocity.

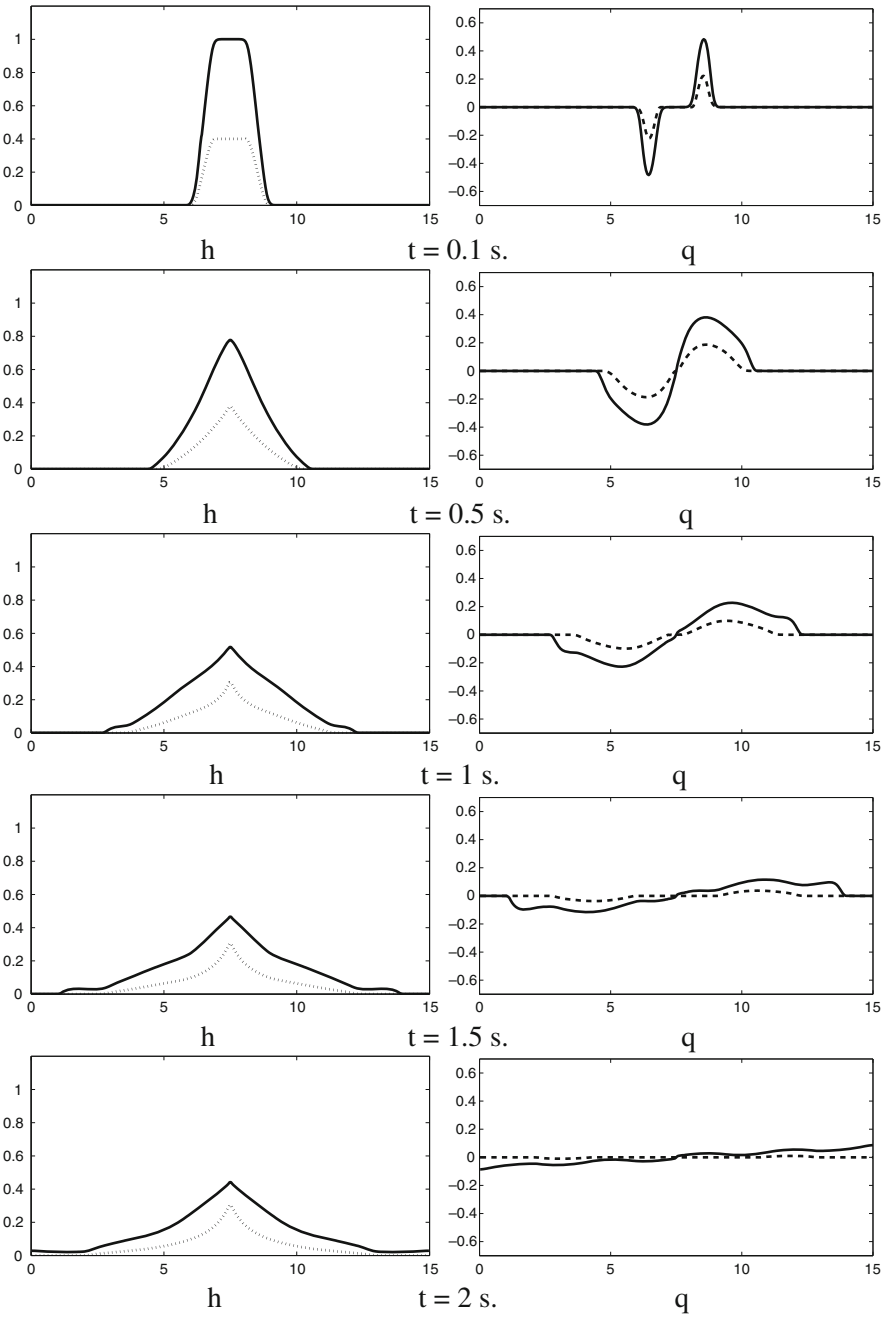


Fig. 6 Two-phase avalanche. $\varphi = 0.4$. *Left*: *continuous line*: granular free surface. *Dotted line*: h_s; *Right*: discharge. *Continuous line*: q_s. *Dashed line*: q_f

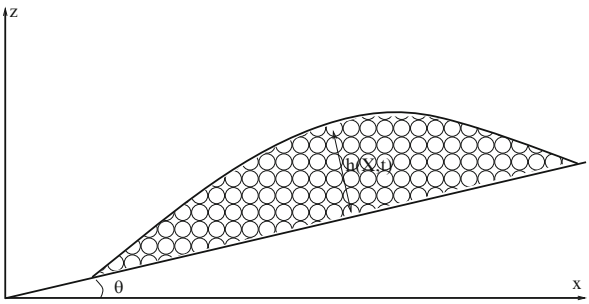


Fig. 7 Partially fluidized avalanches

In order to model the evolution of the granular layer using the Euler equations, we suppose, following [29], that the velocity of the fluid in the pores and the grains are the same, $U_s = U_f = U$. Then $\mathscr{P}$ can be written as

$$\mathscr{P} = \mathscr{P}^s + \mathscr{P}^f,$$

where $\mathscr{P}^s$ and $\mathscr{P}^f$ are the pressure tensor of the solid phase (grains) and the fluid phase.

The derivation of the model follows the same items as in the previous section.

3.1 [∂] *Boundary and Kinematic Conditions*

Let us denote again by $\mathbf{n}^h$ the unit normal vector to the free granular surface $Z = h$ with positive vertical component and by $\mathbf{n}^0 = (0, 1)$ the unit normal vector to the bottom ($Z = 0$). The kinematic condition is defined by (9). For the boundary conditions, the only difference is the definition of the Coulomb friction term. The following boundary conditions are imposed:

- On $Z = h$: $\mathscr{P}\mathbf{n}^h = 0$.
- On $Z = 0$: the non-penetration condition $W = 0$ and the following Coulomb friction law are imposed:

$$\mathscr{P}\mathbf{n}^0 - \mathbf{n}^0(\mathbf{n}^0 \cdot \mathscr{P}\mathbf{n}^0) = \begin{pmatrix} -\mathbf{n}^0 \cdot (\mathscr{P} - \mathscr{P}^f)\mathbf{n}^0 \left.\dfrac{U}{|U|}\right|_{Z=0} \tan\delta_0 \\ 0 \end{pmatrix}.$$

In this last condition, the difference between the stress tensor and the fluid stress tensor, $\mathscr{P} - \mathscr{P}^f$, is used to take into account the buoyancy effects, since $\mathscr{P}^s = \mathscr{P} - \mathscr{P}^f$.

3.2 [$\tilde{A}$] *Dimensional Analysis*

The same non-dimensional variables as in the previous section, defined in (14), are considered. Then,

- The system of equations is defined by (15)–(17).
- The kinematic condition by (18).
- The boundary condition on $Z = h$ by (11) and (20).
- The boundary condition on $Z = 0$ is different from (22). In this case we have

$$\mu \mathscr{P}_{XZ} = -(\mathscr{P}_{ZZ} - \mathscr{P}^f_{ZZ}) \left. \frac{U}{|U|} \right|_{Z=0} \tan \delta_0, \tag{40}$$

  and the non-penetration condition $W = 0$.

3.3 [↕] *Hydrostatic Pressure and Constitutive Law*

The main difference between the Savage–Hutter model presented in the previous section and the model for partially fluidized avalanches appears in this item.

From (17) we obtain

$$\partial_Z(\mathscr{P}_{ZZ}) = -\rho \cos\theta + \mathscr{O}(\varepsilon). \tag{41}$$

If we integrate (41) from $Z > 0$ to $Z = h$, we have, up to order ε,

$$\mathscr{P}_{ZZ} = \rho(h - Z)\cos\theta. \tag{42}$$

But, as $\mathscr{P} = \mathscr{P}^s + \mathscr{P}^f$, we have

$$\mathscr{P}^s_{ZZ} + \mathscr{P}^f_{ZZ} = \mathscr{P}_{ZZ} = \rho \cos\theta (h - Z). \tag{43}$$

In order to consider the anisotropy of the solid phase the following constitutive conditions are again considered (see for example [29, 45]):

$$\mathscr{P}^s_{XX} = K\, \mathscr{P}^s_{ZZ}, \quad \mathscr{P}^f_{XX} = \mathscr{P}^f_{ZZ},$$

where K measures the anisotropy or normal stress effects in the solid phase (see previous section).

The difference appears in this step because, in order to impose these two constitutive conditions, an expression for both $\mathscr{P}^s_{ZZ}$ and $\mathscr{P}^f_{ZZ}$ has to be known. But only the expression of the total pressure $\mathscr{P}_{ZZ}$ is known.

In order to model the flow of a grain-fluid mixture, Iverson and Denlinger (see [29]) assume a linear profile of the normal stress $\mathscr{P}^f_{ZZ}$, which is consistent with Eq. (43). Moreover they suppose that $\mathscr{P}^f_{ZZ}$ takes its maximum value at $Z = 0$ and is proportional to the pressure in absence of the granular phase. They suppose

$$\mathscr{P}^f_{ZZ}(Z) = \lambda \rho \cos\theta (h - Z), \tag{44}$$

being λ a parameter of the model. In this case, by (43) we have

$$\mathscr{P}^s_{ZZ}(Z) = (1 - \lambda)\rho \cos\theta (h - Z). \tag{45}$$

Remark 1. In [29], the authors propose not to set λ as a fixed parameter in time. Instead, they propose to rewrite the model in terms of the pore fluid pressure $p_{bed} = \lambda \rho h \cos\theta$. Then, they assume that the evolution of p_{bed} can be described by a convection-diffusion equation. For the sake of simplicity in these notes, λ is considered as a fixed parameter. Let us remark that we can set $\lambda = \psi_0$, the porosity of the layer.

3.4 [M] *Momentum Conservation Law: With Hydrostatic Pressure and Anisotropy of the Normal Stress of the Solid Phase*

By replacing (44) and (45) in (16) and using the incompressibility equation (15), we obtain up to second order

$$\partial_t(U) + \partial_X U^2 + \partial_Z(UW) + \partial_X\Big(b + Z\cos\theta + (h - Z)\cos\theta(\lambda + K(1 - \lambda))\Big)\varepsilon$$
$$= -\frac{\mu}{\rho}\partial_Z(\mathscr{P}_{XZ}). \tag{46}$$

3.5 [$\int$] *Integration Process*

As in the previous section, let us define:

$$\bar{U} = \frac{1}{h}\int_0^h U(X, Z)dZ \quad \text{and} \quad \overline{U^2} = \frac{1}{h}\int_0^h U^2(X, Z)dZ.$$

As there is no difference in the integration process for the mass equation, we focus here on the momentum equation.

Let us integrate Eq. (46) from $Z = 0$ to $Z = h$ and use the kinematic conditions (18) to obtain

$$\partial_t(h\bar{U}) + \partial_X(h\overline{U^2}) + \left(\int_0^h \partial_X\Big(b + Z\cos\theta + (h-Z)\cos\theta(\lambda + K(1-\lambda))\Big)dZ\right)\varepsilon$$
$$= -\frac{\mu}{\rho}(\mathscr{P}_{XZ}(h) - \mathscr{P}_{XZ}(0)). \tag{47}$$

Moreover,

$$\int_0^h \partial_X\Big(b + Z\cos\theta + (h-Z)\cos\theta(\lambda + K(1-\lambda))\Big)dZ = h\partial_X b$$
$$+ \partial_X\left(\frac{h^2}{2}\cos\theta(\lambda + K(1-\lambda))\right).$$

Then, we obtain the following averaged momentum conservation law,

$$\partial_t(h\bar{U}) + \partial_X\left(h\overline{U^2} + \varepsilon\frac{h^2}{2}\cos\theta(\lambda + K(1-\lambda))\right) = -\varepsilon h\partial_X b - \frac{\mu}{\rho}(\mathscr{P}_{XZ}(h) - \mathscr{P}_{XZ}(0)). \tag{48}$$

Now, we can use the boundary conditions and the constitutive laws to derive $\mu\,\mathscr{P}_{XZ}(h)$ and $\mu\,\mathscr{P}_{XZ}(0)$:

- Using (19) and the constitutive laws $\mathscr{P}^s_{XX} = K\,\mathscr{P}^s_{ZZ}$, $\mathscr{P}^f_{XX} = \mathscr{P}^f_{ZZ}$ we have

$$\mu\mathscr{P}_{XZ}(h) = \varepsilon\partial_X h\mathscr{P}^s_{ZZ}(K-1).$$

  If we suppose again that μ is of order γ for some $\gamma \in (0,1)$, ($\mu = \tan\delta_0 = \mathscr{O}(\varepsilon^\gamma)$) then

$$\mu\mathscr{P}_{XZ}(h) = \mathscr{O}(\varepsilon^{1+\gamma}). \tag{49}$$

- Using Eq. (40), we obtain

$$\mu\mathscr{P}_{XZ}(0) = -(\mathscr{P}_{ZZ}(0) - \mathscr{P}^f_{ZZ}(0))\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0.$$

  Now, using (42) and (43) we have

$$(\mathscr{P}_{ZZ}(0) - \mathscr{P}^f_{ZZ}(0)) = \rho h\cos\theta(1-\lambda) + \mathscr{O}(\varepsilon).$$

Then,

$$\frac{\mu}{\rho}\mathscr{P}_{XZ}(0) = -h\cos\theta(1-\lambda)\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}). \tag{50}$$

Finally, substituting (49) and (50) in (48), the averaged momentum equation is obtained

$$\partial_t(h\bar{U}) + \partial_X\left(h\overline{U^2} + \varepsilon\frac{h^2}{2}\cos\theta(\lambda + K(1-\lambda))\right)$$

$$= -\varepsilon h\partial_X b - h(1-\lambda)\cos\theta\left.\frac{U}{|U|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}).$$

3.6 [↪] *Final System of Equations*

Coming back to the original variables, using (14), neglecting the terms of order $\varepsilon^{1+\gamma}$ and supposing a constant profile of the velocities, the following system is obtained

$$\begin{cases} \partial_t h + \partial_X(h\bar{U}) = 0, \\ \partial_t(h\bar{U}) + \partial_X\left(h\bar{U}^2 + g\cos\theta\dfrac{h^2}{2}(\lambda + K(1-\lambda))\right) = -gh\partial_X b + \mathscr{T}, \end{cases} \tag{51}$$

where $\mathscr{T}$ is the Coulomb friction term. In this model this term must be understood as follows:

$$\begin{aligned} &\text{If } |\mathscr{T}| \geq \sigma_c \quad \Rightarrow \quad \mathscr{T} = -gh(1-\lambda)\cos\theta\frac{\bar{U}}{|\bar{U}|}\tan\delta_0, \\ &\text{If } |\mathscr{T}| < \sigma_c \quad \Rightarrow \quad \bar{U} = 0, \end{aligned}$$

where $\sigma_c = gh(1-\lambda)\cos\theta\tan\delta_0$.

4 Comparison with Pitman–Le Model

Let us remark that if anisotropy is taken into account in the deduction of the solid phase momentum equation in the Pitman–Le model, Eqs. (37c) and (37d) will read as follows:

$$\partial_t(\varphi h U_s) + \partial_X(\varphi h U_s^2 + K g \cos\theta \varphi \frac{h^2}{2}) = \frac{r}{2} g h^2 \cos\theta\, \partial_X \varphi - g h \varphi \partial_X b + \beta h (U_f - U_s) + \mathscr{T}; \tag{52a}$$

$$\partial_t((1-\varphi) h U_f) + \partial_X((1-\varphi) h\, U_f^2 + g \cos\theta (1-\varphi)\frac{h^2}{2}) = g \frac{h^2}{2} \cos\theta \partial_X (1-\varphi) - g h\, (1-\varphi) \partial_X b - \frac{1}{r} \beta h (U_f - U_s). \tag{52b}$$

Let us now consider that $U_s = U_f = U$, that is, the assumption considered in the deduction of the Iverson–Denlinger model and let us define

$$\rho = \varphi \rho_s + (1-\varphi)\rho_f.$$

Then, from (37a) and (37b) we obtain

$$\partial_t(\rho h) + \partial_X(\rho h U) = 0.$$

And from (52a) and (52b) we obtain that

$$\partial_t(\rho h U) + \partial_X(\rho h U^2 + g \cos\theta \frac{h^2}{2}(\rho + \rho_s \varphi (K-1))) = -g h \rho \partial_X b + \mathscr{T}. \tag{53}$$

Note that in the two-phase model the pressure of the fluid phase evaluated at $Z = 0$ is

$$p_{bed}^{PL} = g \rho_f \cos\theta (1-\varphi) h,$$

while in the model proposed by Iverson and Denlinger the pressure at the bottom of the fluid phase is assumed to be:

$$p_{bed} = \lambda \rho g \cos\theta h.$$

If we set

$$\psi = \frac{\rho_f}{\rho}(1-\varphi),$$

we have

$$p_{bed}^{PL} = \psi \rho g \cos\theta h.$$

On the other hand, we have

$$\rho + \rho_s \varphi (K-1) = \rho - \rho_s + \psi \frac{\rho}{r} + K(\rho_s - \psi \frac{\rho}{r}).$$

If we define finally λ by

$$\rho\lambda = \rho - \rho_s + \psi \frac{\rho}{r},$$

we have the equality

$$\rho + \rho_s \varphi (K-1) = \rho\lambda + K(\rho - \rho\lambda).$$

Then, we can rewrite system (53) as

$$\partial_t(\rho h U) + \partial_X(\rho h U^2 + g\cos\theta \frac{h^2}{2}(\rho\lambda + K(\rho - \rho\lambda))) = -gh\rho\partial_X b + \mathscr{T}.$$

That is, we have the same structure as the momentum equation of (51) for the Iverson–Denlinger model. This implies a relation between the hypothesis considered in [29] and the two-phase model when the velocities of the two phases coincide:

$$p_{bed}^{pl} = p_{bed} r + (1-\varphi)(\rho_s - \rho_f) r g h \cos\theta.$$

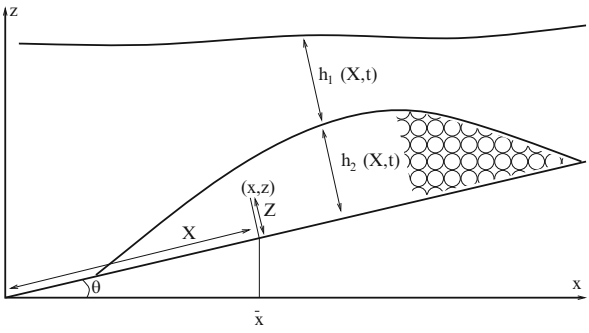
Let us remark finally that, while in the Pitman–Le model the pressure of the phases are considered to be hydrostatic, the inclusion of the pressure at the bed in terms of the parameter λ can be understood as a way to introduce a deviation from the hydrostatic pressure law in the Iverson–Denlinger model.

5 Submarine Avalanches

In this section, we present a simplified version of the two-layer Savage–Hutter type model proposed in [21], with application to submarine avalanches and tsunami waves generated by them.

Submarine avalanches or landslides have been poorly studied compared to their subaerial counterparts. This is however a key issue in geophysics. Indeed, submarine granular flows driven by gravity participate in the evolution of the sea bottom and in particular of the continental margins. They also represent a threat to submarine

Fig. 8 Submarine avalanches



infrastructures, especially for the oil or port industries as well as to many sea shore inhabitants due to the potential tsunamis that can be triggered by such landslides.

In the model derived in this section, index 1 refers to the upper layer, composed of an homogeneous inviscid fluid of constant density ρ_1. Index 2 refers to the lower layer, composed of a granular material of density ρ_s and porosity ψ_0 (see Fig. 8). The pores of the granular layer are assumed to be filled with the fluid of the upper layer. Accordingly, the density of layer 2 is given by:

$$\rho_2 = (1 - \psi_0)\rho_s + \psi_0\rho_1. \tag{54}$$

We consider the incompressible Euler equations, with unknowns

$$\mathbf{V}_i = \begin{pmatrix} u_i \\ w_i \end{pmatrix}, \quad i = 1, 2,$$

being u_i and w_i, the horizontal and vertical velocity components of each layer, respectively. Then, the incompressible Euler equations can be written as

$$\text{div}\mathbf{V}_i = 0, \; i = 1, 2, \tag{55}$$

$$\rho_i \partial_t \mathbf{V}_i + \rho_i \mathbf{V}_i \nabla \mathbf{V}_i = -\text{div} P_i + \rho_i \nabla(\mathbf{g} \cdot \mathbf{X}), \quad i = 1, 2, \tag{56}$$

where $P_i, i = 1, 2$, represent the pressure tensor of each layer

$$P_i = \begin{pmatrix} p_{i,x\,x} & p_{i,x\,z} \\ p_{i,z\,x} & p_{i,z\,z} \end{pmatrix}, i = 1, 2,$$

with $p_{i,x\,z} = p_{i,z\,x}, \rho_i, i = 1, 2$, the densities of each layer, $\mathbf{X} = (x, z)$, the Cartesian coordinates and $\mathbf{g} = (0, -g)$, the gravity.

In order to model the evolution of the granular layer using the Euler equations, on the one hand we suppose following [29] (see Sect. 3) that the velocity of the fluid in the pores of the second layer coincides with that of the grains. On the other hand, P_2 is assumed to be decomposed as

$$P_2 = P_2^s + P_2^f,$$

where P_2^s and P_2^f are the pressure tensor of the solid phase (grains) and the fluid phase, respectively.

Next, a change of variables is performed: local variables over a non-erodible bottom defined by $z = b(x)$ are considered. X denotes the arc's length of the bottom and Z is measured orthogonally to the bottom (see Fig. 8 and Sect. 2).

In what follows, we denote by h_1 and h_2 the thickness of the fluid and granular layers, respectively, measured orthogonally to the bottom (see Fig. 8), by $S = h_1 + h_2$ the free water surface. The details of this change of variables have been given in Sect. 2. Equations (55)–(56) are re-written in the new variables as follows:

$$\begin{cases} \partial_X(U_i) + \partial_Z(W_i) = 0, \quad i = 1,2, \\ \\ \rho_i \partial_t(U_i) + \rho_i \partial_X(U_i^2) + \rho_i \partial_Z(W_i U_i) - \rho_i \partial_X(\mathbf{g} \cdot \mathbf{X}) \\ \quad = -\partial_X(\mathscr{P}_{i\,XX}) - \partial_Z(\mathscr{P}_{i\,XZ}) \quad i = 1,2, \\ \\ \rho_i \partial_t(W_i) + \rho_i \partial_X(U_i\, W_i) + \rho_i \partial_Z(W_i^2) - \rho_i \partial_Z(\mathbf{g} \cdot \mathbf{X}) \\ \quad = -\partial_X(\mathscr{P}_{i\,ZX}) - \partial_Z(\mathscr{P}_{i\,ZZ}) \quad i = 1,2, \end{cases} \tag{57}$$

where U_i, $i = 1,2$, represent the velocity parallel to the bottom and W_i, $i = 1,2$, the perpendicular one. The pressure tensor $\mathscr{P}_i$ is defined by

$$\mathscr{P}_i = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} P_i \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \mathscr{P}_{i,XX} & \mathscr{P}_{i,XZ} \\ \mathscr{P}_{i,ZX} & \mathscr{P}_{i,ZZ} \end{pmatrix}.$$

Remember that, as $p_{i,xz} = p_{i,xz}$, then $\mathscr{P}_{i,XZ} = \mathscr{P}_{i,ZX}$.

Moreover, let us recall that ρ_1 is the density of the fluid and that ρ_2 is defined by (54). θ is the angle between the tangent vector of the bottom and the horizontal axis (see Fig. 8).

5.1 [∂] *Boundary and Kinematic Conditions*

Let $\mathbf{n}^S$, $\mathbf{n}^{h_2}$ and $\mathbf{n}^0 = (0,1)$ be the unit normal vector to the free water surface $Z = S$ ($S = h_1 + h_2$), to the interface ($Z = h_2$) and to the bottom ($Z = 0$), respectively.

The following kinematic conditions are considered

$$\partial_t S + U_1|_{Z=S} \partial_X S - W_1|_{Z=S} = 0, \tag{58}$$

$$\partial_t h_2 + U_i|_{Z=h_2} \partial_X h_2 - W_i|_{Z=h_2} = 0, \quad i = 1,2. \tag{59}$$

Equation (59) assumes that no water is exchanged between the two layers, which is a simplification of the model.

Finally, the following boundary conditions are imposed:

- On $Z = S$:

$$\mathscr{P}_1 \mathbf{n}^S = 0. \tag{60}$$

- On $Z = h_2$:

$$\mathbf{n}^{h_2} \cdot (\mathscr{P}_1 - \mathscr{P}_2)\mathbf{n}^{h_2} = 0 \tag{61}$$

$$\mathscr{P}_i \mathbf{n}^{h_2} - \mathbf{n}^{h_2}(\mathbf{n}^{h_2} \cdot \mathscr{P}_i \mathbf{n}^{h_2}) = \begin{pmatrix} fric(U_1, U_2) \\ 0 \end{pmatrix} \quad i = 1, 2, \tag{62}$$

where $fric(U_1, U_2)$ is a friction term between the layers.

- On $Z = 0$:

$$(U_2, W_2) \cdot \mathbf{n}^0 = 0 \quad \Rightarrow \quad W_2 = 0, \tag{63}$$

$$\mathscr{P}_2 \mathbf{n}^0 - \mathbf{n}^0(\mathbf{n}^0 \cdot \mathscr{P}_2 \mathbf{n}^0) = \begin{pmatrix} -\mathbf{n}^0 \cdot (\mathscr{P}_2 - \mathscr{P}_1)\mathbf{n}^0 \left.\dfrac{U_2}{|U_2|}\right|_{Z=0} \tan \delta_0 \\ 0 \end{pmatrix}. \tag{64}$$

Let us remark that the term $(\mathscr{P}_2 - \mathscr{P}_1)$ in the Coulomb friction law in Eq. (64) is used again in order to take into account the buoyancy effects.

5.2 $[\tilde{A}]$ *Dimensional Analysis*

Next, a dimensional analysis of the set of Eqs. (57), the kinematic and boundary conditions is performed. The non-dimensional variables ($\tilde{\cdot}$) read:

$$\begin{aligned}
&(X, Z, t) = (L\tilde{X}, H\tilde{Z}, (L/g)^{1/2}\tilde{t}), \\
&(U_i, W_i) = (Lg)^{1/2}(\widetilde{U_i}, \varepsilon\widetilde{W_i}),\ i = 1, 2, \\
&h_i = H\widetilde{h_i},\ i = 1, 2, \\
&(\mathscr{P}_{i\,XX}, \mathscr{P}_{i\,ZZ}) = \rho_i g H(\tilde{\mathscr{P}}_{i\,XX}, \tilde{\mathscr{P}}_{i\,ZZ}),\ i = 1, 2, \\
&\mathscr{P}_{i\,XZ} = \rho_i g H \mu_i \tilde{\mathscr{P}}_{i\,XZ}, \quad i = 1, 2,
\end{aligned} \tag{65}$$

where $\mu_1 = 1$, $\mu_2 = \tan\delta_0$, δ_0 being the angle of repose in the Coulomb term (see [47]). By L and H we denote, respectively, the characteristic tangential and normal lengths. We suppose a shallow domain, so $\varepsilon = H/L$ is supposed to be small.

Using this change of variable, the system of Eqs. (57) are rewritten as (tildes are omitted):

$$\partial_X(U_i) + \partial_Z(W_i) = 0, \quad i = 1,2, \tag{66}$$

$$\partial_t(\rho_i U_i) + \rho_i U_i \partial_X U_i + \rho_i W_i \partial_Z U_i + \rho_i \partial_X(b + Z\cos\theta + \frac{\mathscr{P}_{i\,XX}}{\rho_i})\varepsilon$$
$$= -\mu_i \partial_Z(\mathscr{P}_{i\,XZ}) \quad i = 1,2, \tag{67}$$

$$\varepsilon\{\partial_t(\rho_i W_i) + \rho_i U_i \partial_X(W_i) + \rho_i W_i \partial_Z(W_i) + \partial_X(\mathscr{P}_{i\,XZ})\}+$$
$$+ \rho_i \partial_Z(b + \cos\theta Z) = -\partial_Z(\mathscr{P}_{i\,ZZ}) \quad i = 1,2. \tag{68}$$

The kinematic conditions (58)–(59) are rewritten as:

$$\partial_t S + U_1 \partial_X S - W_1 = 0|_{Z=S}, \quad \partial_t h_2 + U_i \partial_X h_2 - W_i = 0|_{Z=h_2}, \quad i = 1,2. \tag{69}$$

Finally, the boundary conditions (60)–(64) are now given by:

- On $Z = S$, we have $\mathbf{n}^S = (-\varepsilon\partial_X S, 1)/\varphi^S$ with $\varphi^S = \sqrt{1 + \varepsilon^2(\partial_X S)^2}$, then from (60) we obtain

$$-\varepsilon\partial_X S\, \mathscr{P}_{1\,XX} + \mu_1 \mathscr{P}_{1\,ZX} = 0, \tag{70}$$

$$-\varepsilon\partial_X S \mu_1 \mathscr{P}_{1\,XZ} + \mathscr{P}_{1\,ZZ} = 0. \tag{71}$$

- On $Z = h_2$, we have $\mathbf{n}^{h_2} = (-\varepsilon\partial_X h_2, 1)/\varphi^{h_2}$ with $\varphi^{h_2} = \sqrt{1 + \varepsilon^2(\partial_X h_2)^2}$, then from (61) and (62) we obtain

$$\mathscr{P}_{1\,ZZ} = \mathscr{P}_{2\,ZZ} + \mathscr{O}(\varepsilon), \tag{72}$$

$$-\varepsilon\mathscr{P}_{i\,XX}\partial_X h_2 + \mu_i \mathscr{P}_{i\,XZ} = -(\mathbf{n}^{h_2}\mathscr{P}_i\mathbf{n}^{h_2})(\varepsilon\partial_X h_2) + fric(U_1, U_2), \; i = 1,2, \tag{73}$$

$$-\varepsilon\mu_i \mathscr{P}_{i\,ZX}\partial_X h_2 + \mathscr{P}_{i\,ZZ} = (\mathbf{n}^{h_2}\mathscr{P}_i\mathbf{n}^{h_2}) \quad i = 1,2. \tag{74}$$

- On $Z = 0$, we have $\mathbf{n}^0 = (0,1)$, then from (63) and (64) we obtain

$$W_2 = 0, \tag{75}$$

$$\mu_2 \mathscr{P}_{2\,XZ} = -(\mathscr{P}_{2\,ZZ} - \mathscr{P}_{1\,ZZ})\left.\frac{U_2}{|U_2|}\right|_{Z=0} \tan\delta_0. \tag{76}$$

5.3 [↕] *Hydrostatic Pressure and Constitutive Law*

From (68) we obtain

$$\partial_Z(\mathscr{P}_{1\,ZZ}) = -\rho_1 \cos\theta + \mathscr{O}(\varepsilon), \tag{77}$$

$$\partial_Z(\mathscr{P}_{2\,ZZ}) = -\rho_2 \cos\theta + \mathscr{O}(\varepsilon). \tag{78}$$

If we integrate (77) from $Z \geq h_2$ to S, we have, up to order ε,

$$\mathscr{P}_{1\,ZZ} = \rho_1 (S - Z)\cos\theta, \tag{79}$$

therefore, $\mathscr{P}_{1\,ZZ}(h_2) = \rho_1 h_1 \cos\theta$. Using this last expression, taking into account (72) and integrating (78) from $Z > 0$ to h_2, we have

$$\mathscr{P}^s_{2\,ZZ} + \mathscr{P}^f_{2\,ZZ} = \mathscr{P}_{2\,ZZ} = \rho_1 h_1 \cos\theta + \rho_2 \cos\theta (h_2 - Z), \tag{80}$$

up to first order. This last equation defines the total pressure, $\mathscr{P}_{2\,ZZ}$, perpendicular to the bottom. The constitutive relation for both the grains and the fluid, i. e. $\mathscr{P}^s_{2\,ZZ}$ and $\mathscr{P}^f_{2\,ZZ}$, are required to close the model. The following relations are considered:

$$\mathscr{P}_{1\,XX} = \mathscr{P}_{1\,ZZ}, \quad \mathscr{P}^s_{2\,XX} = K\,\mathscr{P}^s_{2\,ZZ}, \quad \mathscr{P}^f_{2\,XX} = \mathscr{P}^f_{2\,ZZ}, \tag{81}$$

where K measures the anisotropy or normal stress effects in the solid phase (see Sect. 2).

The same difficulty found in Sect. 3 related to the definition of $\mathscr{P}^s_{2\,ZZ}$ and $\mathscr{P}^f_{2\,ZZ}$ appears here. The assumptions considered there can be adapted. We suppose

$$\mathscr{P}^f_{2\,ZZ}(Z) = \lambda_1 \rho_1 h_1 \cos\theta + \lambda_2 \rho_1 \cos\theta (h_2 - Z), \tag{82}$$

where λ_1 and λ_2 are two parameters. Moreover, by (80), we have

$$\mathscr{P}^s_{2\,ZZ}(Z) = \rho_1 h_1 \cos\theta (1 - \lambda_1) + \cos\theta (h_2 - Z)(\rho_2 - \lambda_2 \rho_1). \tag{83}$$

Remark 2. Note that if (82) and (83) are evaluated in $Z = h_2$, we obtain

$$\mathscr{P}^f_{2\,ZZ}(h_2) = \lambda_1 \rho_1 h_1 \cos\theta, \qquad \mathscr{P}^s_{2\,ZZ}(h_2) = \rho_1 h_1 \cos\theta (1 - \lambda_1);$$

Then, λ_1 controls the distribution of the pressure at the interface between the two phases of the second layer:

- A possible choice is to set $\lambda_1 = \lambda_2 = \psi_0$, where ψ_0 is the porosity of the second layer.
- We can rewrite the model in terms of $p_{bed} = \lambda_1 \rho_1 h_1 \cos\theta + \lambda_2 \rho_1 h_2 \cos\theta$, the pore-fluid basal pressure. In this case, a more sophisticated model can be defined

by coupling it with a convection-diffusion equation (as proposed in [29]), as it has been mentioned in the case of partially fluidized avalanches.

By taking into account the constitutive closure equations (81) we deduce the following expression of $\mathscr{P}_{2\,XX} = \mathscr{P}^s_{2\,XX} + \mathscr{P}^f_{2\,XX}$,

$$\begin{aligned}\mathscr{P}_{2\,XX} &= K\,\mathscr{P}^s_{2\,ZZ} + \mathscr{P}^f_{2\,ZZ}\\ &= h_1\cos\theta\rho_1(\lambda_1 + K(1-\lambda_1)) + (h_2 - Z)\cos\theta(\lambda_2\rho_1 + K(\rho_2 - \lambda_2\rho_1)).\end{aligned} \tag{84}$$

5.4 [M] *Momentum Conservation Laws: With Hydrostatic Pressure and Anisotropy of the Normal Stress of the Solid Phase of the Submerged Sediment Layer*

By replacing (79) and (84) in (5.2) and using the incompressibility equation (66), we obtain up to second order

$$\partial_t(\rho_1 U_1) + \rho_1\partial_X U_1^2 + \rho_1\partial_Z(U_1W_1) + \rho_1\partial_X(b + S\cos\theta)\varepsilon = -\mu_1\partial_Z(\mathscr{P}_{1\,XZ}), \tag{85}$$

and

$$\begin{aligned}\partial_t(\rho_2 U_2) + \rho_2\partial_X U_2^2 + \rho_2\partial_Z(U_2W_2) + \rho_2\partial_X\Big(b + Z\cos\theta + \frac{1}{\rho_2}[h_1\cos\theta\rho_1(\lambda_1 + K(1-\lambda_1))\\ + (h_2 - Z)\cos\theta(\lambda_2\rho_1 + K(\rho_2 - \lambda_2\rho_1))]\Big)\varepsilon = -\mu_2\partial_Z(\mathscr{P}_{2\,XZ}).\end{aligned} \tag{86}$$

5.5 [$\int$] *Integration Process*

In this section, Eqs. (66), (85) and (86) are depth-averaged along the normal direction to the topography. Let us introduce the following notation: we denote by $\bar{U}_i$, $i = 1, 2$ the velocities of each layer averaged along the normal direction to the basal surface:

$$\bar{U}_1 = \frac{1}{h_1}\int_{h_2}^{S} U_1(X, Z)dZ, \qquad \bar{U}_2 = \frac{1}{h_2}\int_0^{h_2} U_2(X, Z)dZ.$$

We also define:

$$\overline{U_1^2} = \frac{1}{h_1}\int_{h_2}^{S} U_1^2(X, Z)dZ, \qquad \overline{U_2^2} = \frac{1}{h_2}\int_0^{h_2} U_2^2(X, Z)dZ.$$

If Eq. (66) is integrated from $Z = h_2$ to $Z = S$, we obtain

$$0 = \partial_X(h_1\bar{U_1}) - U_1(S)\partial_X S + W_1(S) + U_1(h_2)\partial_X h_2 - W_1(h_2).$$

And using the kinematic condition (69), the following equation is derived:

$$\partial_t h_1 + \partial_X(h_1\bar{U_1}) = 0.$$

Analogously, by integrating (66) between $Z = 0$ and $Z = h_2$ we obtain

$$0 = \partial_X(h_2\bar{U_2}) - U_2(h_2)\partial_X h_2 + W_2(h_2) - W_2(0),$$

and, using the kinematic condition (69) and the boundary condition (75), the following equation is obtained:

$$\partial_t h_2 + \partial_X(h_2\bar{U_2}) = 0.$$

If (85) is integrated from $Z = h_2$ to $Z = S$, we obtain

$$\rho_1\partial_t(h_1\bar{U_1}) + \rho_1\partial_X(h_1\overline{U_1^2}) - \rho_1 U_1(S)[\partial_t(S) + U_1(S)\partial_X S - W_1(S)]$$

$$+\rho_1 U_1(h_2)[\partial_t h_2 + U_1(h_2)\partial_X h_2 - W_1(h_2)] + \rho_1\left(\int_{h_2}^{S}\left(\partial_X(b + S\cos\theta)\right)dZ\right)\varepsilon$$

$$= -\mu_1(\mathscr{P}_{1XZ}(S) - \mathscr{P}_{1XZ}(h_2)). \tag{87}$$

The expressions of $\mathscr{P}_{1XZ}(S)$ and $\mathscr{P}_{1XZ}(h_2)$ are now derived using the boundary conditions and the constitutive laws:

- Using (70) and (79) and the relation $\mathscr{P}_{1XX} = \mathscr{P}_{1ZZ}$ the following expression is obtained:

$$\mu_1\mathscr{P}_{1ZX}(S) = -\varepsilon\mathscr{P}_{1XX}(S)\partial_X S = -\varepsilon\mathscr{P}_{1ZZ}(S)\partial_X S = 0 + \mathscr{O}(\varepsilon^2). \tag{88}$$

- Using (73), we have

$$\mu_1\mathscr{P}_{1XZ}(h_2) + \varepsilon\partial_X h_2(\mathscr{P}_{1ZZ} - \mathscr{P}_{1XX}) = fric(U_1, U_2) + \mathscr{O}(\varepsilon^2).$$

  Therefore, applying the constitutive law for the fluid layer, that is, $\mathscr{P}_{1XX} = \mathscr{P}_{1ZZ}$, the following equality is derived:

$$\mu_1\mathscr{P}_{1XZ}(h_2) = fric(U_1, U_2) + \mathscr{O}(\varepsilon^2). \tag{89}$$

Using the kinematic condition (69), Eq. (87) and the expressions obtained for $\mu_1\mathscr{P}_{1XZ}(S)$ (88) and for $\mu_1\mathscr{P}_{1XZ}(h_2)$ (89), we obtain

$$\rho_1\partial_t(h_1\bar{U}_1)+\rho_1\partial_X(h_1\overline{U_1^2})+\rho_1\left(\int_{h_2}^{S}\partial_X(b+S\cos\theta)dZ\right)\varepsilon = fric(U_1,U_2)+\mathscr{O}(\varepsilon^2).$$

Next, the integral term in the last equality can be computed as follows:

$$\int_{h_2}^{S}\partial_X(b+S\cos\theta)dZ = h_1\partial_X b+\partial_X\left(\frac{h_1^2}{2}\cos\theta\right)+h_1\partial_X(\cos\theta h_2).$$

Finally, we obtain the equation

$$\rho_1\partial_t(h_1\bar{U}_1)+\rho_1\partial_X\left(h_1\overline{U_1^2}+\varepsilon\frac{h_1^2}{2}\cos\theta\right) = \varepsilon\rho_1\left(-h_1\partial_X b-h_1\partial_X(\cos\theta h_2)\right) + fric(U_1,U_2)+\mathscr{O}(\varepsilon^2).$$

Let us now integrate Eq. (86) from $Z=0$ to $Z=h_2$. As in the previous case, we use the kinematic conditions (69) to obtain

$$\rho_2\partial_t(h_2\bar{U}_2)+\rho_2\partial_X(h_2\overline{U_2^2})+\rho_2\left(\int_0^{h_2}\partial_X\left(b+Z\cos\theta+\frac{1}{\rho_2}[h_1\cos\theta\rho_1(\lambda_1+K(1-\lambda_1))\right.\right.$$

$$\left.\left.+(h_2-Z)\cos\theta(\lambda_2\rho_1+K(\rho_2-\lambda_2\rho_1))]\right)dZ\right)\varepsilon = -\mu_2(\mathscr{P}_{2XZ}(h_2)-\mathscr{P}_{2XZ}(0)). \tag{90}$$

Let us introduce the densities ratio

$$r=\frac{\rho_1}{\rho_2},$$

where ρ_1 is the density of the fluid and ρ_2 is defined by (54). We obtain

$$\int_0^{h_2}\partial_X\left(b+Z\cos\theta+\frac{1}{\rho_2}[h_1\cos\theta\rho_1(\lambda_1+K(1-\lambda_1))\right.$$

$$\left.+(h_2-Z)\cos\theta(\lambda_2\rho_1+K(\rho_2-\lambda_2\rho_1))]\right)dZ = h_2\partial_X b$$

$$+rh_2(K(1-\lambda_1)+\lambda_1)\partial_X(h_1\cos\theta)$$

$$+\partial_X\left(\frac{h_2^2}{2}\cos\theta(r\lambda_2+K(1-r\lambda_2))\right).$$

Replacing this last expression in (90) and dividing by ρ_2, the following equation is obtained:

$$\partial_t(h_2\bar{U}_2) + \partial_X\left(h_2\overline{U_2^2} + \varepsilon\frac{h_2^2}{2}\cos\theta(r\lambda_2 + K(1-r\lambda_2))\right) = -\varepsilon h_2\partial_X b$$

$$-\varepsilon r h_2(\lambda_1 + K(1-\lambda_1))\partial_X(h_1\cos\theta) - \frac{\mu_2}{\rho_2}(\mathscr{P}_{2XZ}(h_2) - \mathscr{P}_{2XZ}(0)). \tag{91}$$

Just like in the previous case, the boundary conditions and the constitutive laws are used to derive $\mu_2\,\mathscr{P}_{2XZ}(h_2)$ and $\mu_2\,\mathscr{P}_{2XZ}(0)$:

- Using (73) and $\mathscr{P}^s_{2XX} = K\mathscr{P}^s_{2ZZ}$, $\mathscr{P}^f_{2XX} = \mathscr{P}^f_{2XX}$, we have

  $$\mu_2\mathscr{P}_{2XZ}(h_2) = fric(U_1, U_2) + \mu_2\varepsilon\partial_X h_2\mathscr{P}^s_{2ZZ}(K-1).$$

  We suppose again that $\mu_2 = \tan\delta_0 = \mathscr{O}(\varepsilon^\gamma)$ with $\gamma \in (0,1)$. Under this assumption, we have

  $$\mu_2\mathscr{P}_{2XZ}(h_2) = fric(U_1, U_2) + \mathscr{O}(\varepsilon^{1+\gamma}). \tag{92}$$

- Using Eq. (76), we obtain

  $$\mu_2\mathscr{P}_{2XZ}(0) = -(\mathscr{P}_{2ZZ}(0) - \mathscr{P}_{1ZZ}(0))\left.\frac{U_2}{|U_2|}\right|_{Z=0}\tan\delta_0.$$

  Now, using (79) and (80) we have

  $$(\mathscr{P}_{2ZZ}(0) - \mathscr{P}_{1ZZ}(0)) = h_2\cos\theta(\rho_2 - \rho_1) + \mathscr{O}(\varepsilon).$$

  Therefore, assuming $\tan\delta_0 = \mathscr{O}(\varepsilon^\gamma)$, we have

  $$\mu_2\mathscr{P}_{2XZ}(0) = -(\rho_2 - \rho_1)h_2\cos\theta\left.\frac{U_2}{|U_2|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}). \tag{93}$$

Finally, substituting (92) and (93) in (91), we derive the averaged momentum equation for the second layer

$$\partial_t(h_2\bar{U}_2) + \partial_X\left(h_2\overline{U_2^2} + \varepsilon\frac{h_2^2}{2}\cos\theta(r\lambda_2 + K(1-r\lambda_2))\right)$$

$$= -\varepsilon h_2\partial_X b - \varepsilon r h_2(\lambda_1 + K(1-\lambda_1))\partial_X(h_1\cos\theta)$$

$$-\frac{1}{\rho_2}fric(U_1, U_2) - ((1-r)h_2\cos\theta + h_2\bar{U}_2^2 d_X\theta)\left.\frac{U_2}{|U_2|}\right|_{Z=0}\tan\delta_0 + \mathscr{O}(\varepsilon^{1+\gamma}).$$

5.6 [↪] *Final System of Equations*

Reverting to the original variables [see (65)], neglecting the terms of order $\varepsilon^{1+\gamma}$ and supposing a constant profile of the velocities we obtain the following system:

$$\begin{cases} \partial_t h_1 + \partial_X (h_1 \bar{U}_1) = 0; \\ \\ \partial_t (h_1 \bar{U}_1) + \partial_X (h_1 \bar{U}_1^2 + g \dfrac{h_1^2}{2} \cos\theta) = \\ \qquad = -g h_1 \partial_X b - g \cos\theta h_1 \partial_X h_2 + \dfrac{1}{\rho_1} fric(U_1, U_2); \\ \\ \partial_t h_2 + \partial_X (h_2 \bar{U}_2) = 0; \\ \\ \partial_t (h_2 \bar{U}_2) + \partial_X \left(h_2 \bar{U}_2^2 + g \dfrac{h_2^2}{2} \cos\theta (r\lambda_2 + K(1 - r\lambda_2)) \right) = \\ \qquad = -g h_2 \partial_X b - r g \cos\theta h_2 (\lambda_1 + K(1 - \lambda_1)) \partial_X h_1 - \dfrac{1}{\rho_2} fric(U_1, U_2) + \mathscr{T}; \end{cases} \tag{94}$$

where by $\mathscr{T}$, we denote the Coulomb friction term. Again, this term must be understood as follows:

$$\begin{aligned} &\text{If } |\mathscr{T}| \geq \sigma_c \quad \Rightarrow \quad \mathscr{T} = -(g(1-r)h_2 \cos\theta) \frac{\bar{U}_2}{|\bar{U}_2|} \tan\delta_0, \\ &\text{If } |\mathscr{T}| < \sigma_c \quad \Rightarrow \quad \bar{U}_2 = 0, \end{aligned}$$

where $\sigma_c = g((1-r)h_2 \cos\,\theta \tan\delta_0$. Recall that

$$r = \frac{\rho_1}{\rho_2},$$

where ρ_1 is the density of the fluid and ρ_2 is defined in (54). We can define the friction term between the layers $fric(U_1, U_2)$ under the following structure

$$fric(U_1, U_2) = -\mathscr{K}_{in} \cdot (\bar{U}_1 - \bar{U}_2), \quad \text{with } \mathscr{K}_{in} = \rho_1 K_{in} |\bar{U}_1 - \bar{U}_2|,$$

being K_{in} a positive constant.

In Fig. 9, an example of application of the model is presented. As initial condition a rectangular granular layer is imposed at the middle of the domain. The water layer is initially at rest and its free surface is flat. As we can see in the different times shown in the figure, the avalanche produced by the granular layer interacts with the fluid and some waves appear. The stationary solution reached consists of water at rest with a flat free surface over a granular layer in equilibrium. This simulation has

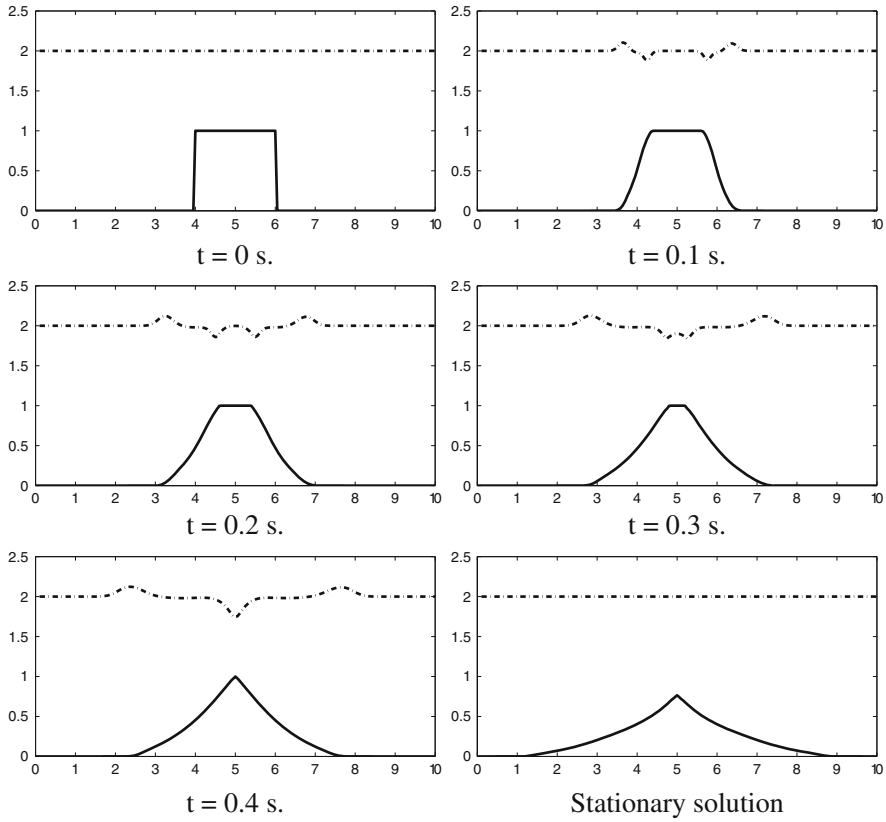


Fig. 9 Submarine avalanche test. *Continuous line*: granular free surface; *Dashed-dotted line*: water free surface

been obtained by numerically solving System (94) with the finite volume method introduced in [21]. See also [32], where an application of the model to the case of tsunamis in the Alboran Sea is studied.

6 Entropy Inequality and Stationary Solutions of Savage–Hutter Type Models

In this section we state without proof a result concerning the entropy inequality and the stationary solutions of the Savage–Hutter type models presented in previous sections for aerial, fluid–solid mixture and submarine avalanches. The following result can be proved for the submarine avalanche model:

Theorem 1. *System (94) has the following properties:*

(*i*) *It admits an entropy dissipation inequality,*

$$\begin{aligned}
&\partial_t\left(\frac{r\Lambda_1 h_1\bar{U}_1^2+h_2\bar{U}_2^2}{2}+gb(r\Lambda_1 h_1+h_2)+gcos\,\theta\frac{r\Lambda_1 h_1^2+\Lambda_2 h_2^2}{2}\right.\\
&\left.+gcos\,\theta\, r\Lambda_1 h_1 h_2\right)\\
&+\partial_X\left(r\Lambda_1 h_1\bar{U}_1\left(\frac{\bar{U}_1^2}{2}+gb+cos\,\theta(h_1+h_2)\right)+h_2\bar{U}_2\right.\\
&\left.\left(\frac{\bar{U}_2{}^2}{2}+gb+gcos\,\theta(r\Lambda h_1+\Lambda_2 h_2)\right)\right)\\
&\quad\le -rK_{in}|\bar{U}_1-\bar{U}_2|(U_2-U_1)(U_2-\Lambda_1 U_1)-g((1-r)h_2 cos\,\theta\\
&\qquad+h_2 d_X\theta(\bar{U}_2^2-\frac{gh_2 cos\,\theta}{2}))|\bar{U}_2|tan\,\delta_0\\
&\qquad+g\frac{h_2^2}{2}U_2(1-\Lambda_2)sin\,\theta\partial_X\theta.
\end{aligned}$$

where

$$\Lambda_1=\lambda_1+K\,(1-\lambda_1),\quad \Lambda_2=r\,\lambda_2+K\,(1-r\,\lambda_2).$$

(*ii*) *It has the family of steady state solutions:*

$$\bar{U}_1=0,\quad \bar{U}_2=0, \tag{95}$$

$$b+(h_1+h_2)cos\,\theta=cst, \tag{96}$$

$$|(\Lambda_2-r\Lambda_1)\partial_x(b+h_2 cos\,\theta)+(1-\Lambda_2)(\partial_X b-\frac{h_2}{2}sin\,\theta\partial_x\theta)|\le(1-r)tan\,\delta_0, \tag{97}$$

corresponding to water at rest over a stationary granular layer. □

Note that the models presented in Sects. 2 and 3 can be seen as particular cases of this one. They can be obtained as follows:

- Savage–Hutter model: set $h_1=0, \bar{U}_1=0, \lambda_2=0, r=1, \lambda_1=1$.
- Iverson–Denlinger model: set $h_1=0, \bar{U}_1=0, \lambda_2=\lambda, r=1, \lambda_1=1$.

Therefore, the properties presented in Theorem 1 are also valid for the Savage–Hutter and Iverson–Denlinger models.

Note that the stationary solutions defined by (97) correspond to situations in which the free surface of the granular layer is in equilibrium with the internal friction angle. The angle δ_0 can be measured in laboratory experiments.

7 Rheology of Complex Avalanches

Several differential models have been proposed in the literature to describe sediment mixtures: a review is presented in [2]. A possible approximation is given by the model presented in Sect. 3, based on a two-phase approach and a friction law proportional to the normal stress and the tangent of the internal friction angle. Another possibility is the use of visco-plastic models. They represent an approximation of the rheological behaviors of complex flows, such as debris flows, lava flows and snow avalanches.

In this section a brief introduction to non-Newtonian fluids is first given in order to motivate the definition of the stress tensor corresponding to the Herschel–Bulkley model. This model can be used to study debris flows, fluid–solid mixture avalanches.

As in the previous section we consider local coordinates on a plane slope with angle θ (see Sect. 2 for details on the notation). Let us denote the velocity vector as

$$\mathbf{U} = \begin{pmatrix} U \\ W \end{pmatrix}.$$

Let us remember that the general system of Eqs. (4) and (5) can be re-written in the new variables as follows:

$$\begin{cases} \partial_X(U) + \partial_Z(W) = 0, \\ \rho\partial_t(U) + \rho\partial_X(U^2) + \rho\partial_Z(WU) - \rho\partial_X(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{XX}) - \partial_Z(\mathscr{P}_{XZ}), \\ \rho\partial_t(W) + \rho\partial_X(U\,W) + \rho\partial_Z(W^2) - \rho\partial_Z(\mathbf{g}\cdot\mathbf{X}) = -\partial_X(\mathscr{P}_{ZX}) - \partial_Z(\mathscr{P}_{ZZ}), \end{cases} \tag{98}$$

where the density ρ is assumed to be constant.

Let us remark that, although in these notes we are working with the negative stress tensor $\mathscr{P}$, it is also usual to write the system of equations in terms of σ, the positive Cauchy stress tensor, where

$$\mathscr{P} = -\sigma.$$

The stress tensor is defined as the sum of the pressure component and the viscous one (cf. [9]),

$$\sigma = -pI + \sigma',$$

where I is the identity matrix. σ' is called the deviatoric part of σ. Note that we can also write,

$$\mathscr{P} = pI - \sigma'.$$

Let us use the notation:

$$\sigma' = \begin{pmatrix} \sigma'_{XX} & \sigma'_{XZ} \\ \sigma'_{ZX} & \sigma'_{ZZ} \end{pmatrix}.$$

A fluid is said to be Newtonian if σ' is proportional to the rate of deformation tensor $D(\mathbf{U})$, where

$$D(\mathbf{U}) = \nabla \mathbf{U} + \nabla \mathbf{U}^t = \begin{pmatrix} 2\partial_X U & \partial_X W + \partial_Z U \\ \partial_X W + \partial_Z U & 2\partial_Z W \end{pmatrix}.$$

Then for a Newtonian fluid, such as water, we have

$$\sigma' = \eta D(\mathbf{U}),$$

where η is the viscosity coefficient, depending on the material. In theses notes, we suppose that η is a constant value.

Let us consider the case of a uniform flow such that

$$D(\mathbf{U}) = \begin{pmatrix} 0 & \partial_Z U \\ \partial_Z U & 0 \end{pmatrix} \quad \sigma' = \begin{pmatrix} 0 & \sigma' \\ \sigma' & 0 \end{pmatrix},$$

where $\sigma' = \sigma_{XZ} = \sigma_{ZX}$ is the shear stress. In this case the relation which characterizes the fluid as Newtonian can be easily represented as a straight line in the plane $(\partial_Z U) - (\sigma')$ with slope η (see Fig. 10).

The behavior of the flows of materials like honey, corn flour or paint cannot be modelled with such a linear relation. Moreover, the concept of Newtonian fluid is an idealization: there are always nonlinear relations between the shear stress and the shear rate. The study of the deviation (from the linear law) of σ' as a function of $D(\mathbf{U})$ belongs to the field called *Rheology*. The term *Rheology* is due to Bingham in 1929. It comes from the Greek "$\rho\varepsilon\omega$"—"to flow". It is related to the study of deformation and flow of *complex fluids*.

Years around 1900s saw a significant increase of activity on these subjects, including authors like Maxwell (1868), Boltzmann (1877), Bingham, Blair, Reiner, Herschel–Bulkley, Weissenberg (all between 1900 and 1930, see [48]). Then Rheology became a research field of intense activity.

Fluids in which η is a function of $D(\mathbf{U})$, $\eta = \eta(D(\mathbf{U}))$, i.e.

$$\sigma' = \eta(D(\mathbf{U}))\, D(\mathbf{U}).$$

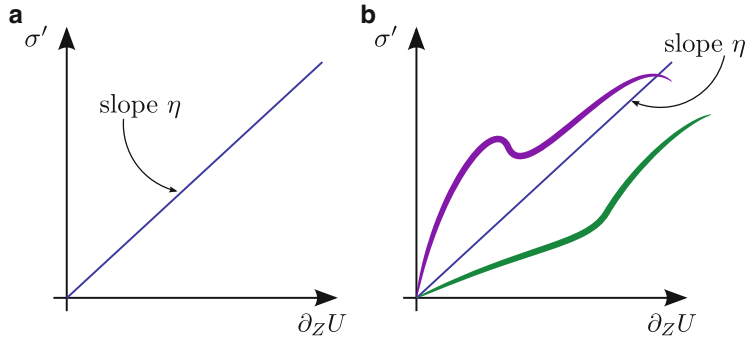
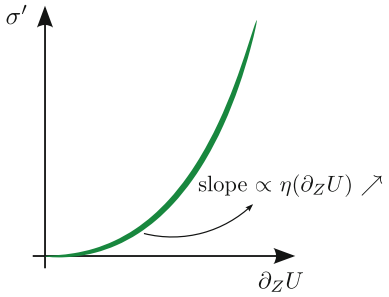


Fig. 10 (**a**) Representation of a Newtonian fluid in the plane $(\sigma') - (\partial_Z U)$ with viscosity η; (**b**) generalized Newtonian fluids

(see Fig. 10) are called *Generalized Newtonian fluids*; cf [41]). For example:

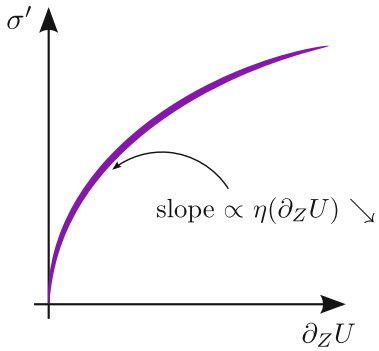
- Corn flour is a material whose behavior is fluid when it is gently mixed but it becomes very viscous if it is strongly mixed. That is, "viscosity " increases with "shear". Such materials are **shear-thickening**.

  In the case of a uniform shear-thickening flow, we have $\sigma' = \eta(\partial_Z U)\partial_Z U$, where $\eta(\partial_Z U)$ is an increasing function of $\partial_Z U$. They can be represented in the $(\sigma') - (\partial_Z U)$ plane as follows:



- There are many other materials, like paint, whose viscosity decreases with shear. These materials are **shear-thinning**.

  In the case of a uniform shear-thinning flow, we have $\sigma' = \eta(\partial_Z U)\partial_Z U$, where $\eta(\partial_Z U)$ is a decreasing function of $\partial_Z U$. They can be represented in the $(\sigma') - (\partial_Z U)$ plane as follows:



Shear-thinning and shear-thickening fluids can be modelled using **power-law fluids**. The viscosity of power-law fluid is defined by $\eta(D(\mathbf{U})) = \bar{\eta}\,|D(\mathbf{U})|^{n-1}$, for some positive constant $\bar{\eta}$ and $n \geq 0$. For these, we have

$$\sigma' = \bar{\eta}\,|D(\mathbf{U})|^{n-1} D(\mathbf{U}),$$

Let us remark that:

- If $n < 1$ the material is shear-thinning.
- If $n = 1$ the fluid is Newtonian and $\bar{\eta}$ is the constant viscosity.
- If $n > 1$ the fluid is shear-thickening.

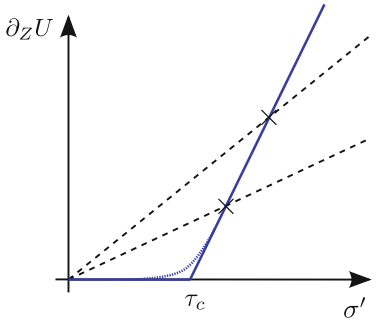
Nevertheless, the flow of some materials cannot be modeled by a power-law model. This is the case of clay, snow or lava that only flow when the shear stress is bigger than a critical value. These materials are example of what we can call "threshold" fluids. Below a stress τ_c the material present a rigid behavior but above τ_c the material begins to flow. They are visco-plastic materials. Bingham defined *Plasticity* as follows (see [10]):

> We may now define **plasticity** as a property of solids in virtue of which they hold their shape permanently under the action of small shearing stresses but they are readily deformed, worked or molded, under somewhat larger stresses.

Bingham law follows this property and it depends on a threshold shear stress τ_c. σ' is defined as follows:

$$\begin{cases} |\sigma'| < \tau_c & \text{if } |D(\mathbf{U})| = 0 \\ \sigma' = \eta D(\mathbf{U}) + \tau_c \frac{D(\mathbf{U})}{|D(\mathbf{U})|} & \text{if } |D(\mathbf{U})| \neq 0. \end{cases} \tag{99}$$

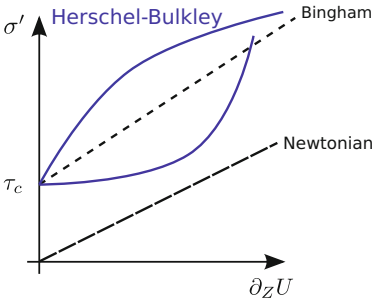
Note that this definition implies that, in the case that $|D(\mathbf{U})| = 0$, we only know that σ' is bounded by τ_c. That is, σ' is a multivalued function in this case. It is easier to understand this law by considering the inverse function. Let us suppose that we perform two experiments of a uniform flow for a plastic material and we measure the shear rate $\partial_Z U$ in terms of the shear stress σ'. We have marked with crosses in the following figure two points corresponding to the measures:



With black dashed line we plot the law corresponding to the case of Newtonian fluids, that is, straight lines passing by the point of measure and by $(0, 0)$. Let us remark that if the fluid is Newtonian then it is enough to look for only one experiment in order to measure its viscosity. Remember that the viscosity in a Newtonian fluid is the inverse of the slope of such a straight line. Performing a second experiment is a way to see if the fluid is Newtonian. If we have a graph as the one of previous figure in which the second measure does not lead to a point in the same straight line, then the fluid is not Newtonian.

Conversely, if we consider a straight line passing through these two points, we obtain the value of the shear threshold τ_c, as the point at which this line cuts the horizontal axe. Actually, measurements show that the real behavior of plastic materials does not follows exactly this straight line. They follow a curve that can be seen as a regularization of the corner around the point $(\tau_c, 0)$ (dashed-blue line in previous figure). The model proposed by Bingham, defined by (99), corresponds to the graph defined by the union of the two blue straight lines in the previous figure.

The general case combines *power-law* and *plasticity*. This is the Herschel–Bulkley constitutive equation. For the case of uniform flow it can be represented in the $(\sigma') - (\partial_Z U)$ plane as follows:



Herschel–Bulkley model is characterized by the following stress tensor:

$$\mathscr{P} = pI - \sigma', \tag{100}$$

where

$$\begin{cases} \sigma' = \tau_c \frac{D(\mathbf{U})}{|D(\mathbf{U})|} + \bar{\eta}|D(\mathbf{U})|^{n-1} D(\mathbf{U}) \text{ if } |D(\mathbf{U})| \neq 0, \\ |\sigma'| \leq \tau_c \qquad\qquad\qquad\qquad \text{if } |D(\mathbf{U})| = 0 \end{cases}$$

with $n > 0$. In the case of avalanches, we have $n \in (0, 1)$. In [3] the dam-break problem for visco-plastic Herschel–Bulkley fluids down a sloping flume is investigated and laboratory data are presented.

8 A Shallow Herschel–Bulkley Model for Fluid–Solid Mixture Avalanches

In this section, a shallow Herschel–Bulkley model is deduced (see [1, 14]). One of the difficulties of Herschel–Bulkley model is that, when $|D(\mathbf{U})| = 0$, only a bound of the stress tensor is known. Then, we cannot obtain a shallow Herschel–Bulkley model following the same steps as in the previous sections. Several types of shallow Herschel–Bulkley models have been proposed in the literature. For example in [8, 22] shallow visco-plastic models have been proposed in the case of nearly steady uniform regime. That is, the reference velocity for the asymptotic analysis is defined in terms of a stationary solution where the viscous contribution matches the gravity acceleration. Such a type of models are only valid for $\theta \neq 0$. In these notes, we present another type of shallow model, which corresponds to the inertial regime, where inertial and pressure-gradient terms are of the same magnitude.

As mentioned, we cannot follows exactly the same steps as in the derivation of Savage–Hutter. Basically, we cannot reproduce the items $[M]$: the momentum conservation law and $[\int]$: integration process. Nevertheless, note that the integration process in the derivation of the Savage–Hutter model is equivalent to consider the variational formulation of the model with test functions that do not depend on the vertical variable Z. Following this idea, we can consider first the variational formulation of Herschel–Bulkley model, which has the form of a variational inequality and then consider test functions that do not depend on the variable Z.

Thus, the derivation of the shallow Herschel–Bulkley model is done following the items:

- $[\partial]$ Boundary and kinematic conditions.
- $[\int M \cdot \psi(X, Z) \geq 0]$ Momentum conservation law in variational form.
- $[\tilde{A}]$ Dimensional analysis.
- $[\int M \cdot \psi(X) \geq 0]$ Variational inequality for test functions independent on Z.
- $[\hookrightarrow]$ Final system of equations.

8.1 [∂] *Boundary and Kinematic Conditions*

Let us remember that $\mathbf{n}^h$ is the unit normal vector to the free granular surface $Z = h$ with positive vertical component, and $\mathbf{n}^0 = (0, 1)$, the unit normal vector to the bottom ($Z = 0$), denoted as Γ_b.

The kinematic condition is considered at the free surface

$$\partial_t h + U|_{Z=h}\partial_X h - W|_{Z=h} = 0, \tag{101}$$

And the following boundary conditions are imposed:

- On $Z = h$:

$$\mathbf{n}^h \cdot \mathscr{P}\mathbf{n}^h = 0 \tag{102}$$

$$\mathscr{P} \cdot \mathbf{n}^h - \mathbf{n}^h(\mathbf{n}^h \cdot \mathscr{P}\mathbf{n}^h) = \begin{pmatrix} fric_h(U) \\ 0 \end{pmatrix}, \tag{103}$$

  where $fric_h(U)$ is the friction term between the granular layer and the air. For simplicity, we will suppose that $fric_h(U) = 0$.
- On $Z = 0$:

$$(U, W) \cdot \mathbf{n}^0 = 0 \quad \Rightarrow \quad W = 0, \tag{104}$$

$$\mathscr{P}\mathbf{n}^0 - \mathbf{n}^0(\mathbf{n}^0 \cdot \mathscr{P}\mathbf{n}^0) = \begin{pmatrix} -\alpha\, U \\ 0 \end{pmatrix}. \tag{105}$$

  That is, we consider a simple linear friction law between the material and the bottom. This is one of the main differences between the model derived in this section and the Savage–Hutter one: while in the Savage–Hutter model the Coulomb friction law controls the yielding of the material, in the shallow Herschel–Bulkley model this effect is due to the stress tensor definition and, in particular, to the rigidity coefficient τ_c (also called *yield stress*).

8.2 [$\int M \cdot \psi(X, Z) \geq 0$] *Momentum Conservation Law in Variational Form*

In this item, we write the variational formulation of the system defined by (98)–(100). By the definition of the stress tensor σ' we obtain a variational inequality (see [19]).

Let us suppose that the domain filled by the avalanche, $\Omega(t)$, can be written as follows:

$$\Omega(t) = \{(X, Z) \in \mathbb{R}^2;\ X \in [0, L];\ Z \in [0, h(X, t)]\}.$$

By supposing that $h(X, t)$, the height of the avalanche, is a bounded function in space and time we may consider

$$\mathscr{W}(t) = \big\{\Psi = (\psi, \varphi);\ \psi, \varphi \in W^{1,1+n}(\Omega(t)) \ /\quad \psi|_{X=0} = \psi|_{X=L} = 0,\ \varphi|_{Z=0} = 0\big\}.$$

Then, we look for the solution $\mathbf{U}(t, \cdot) \in \mathscr{W}(t)$ and $p(t, \cdot) \in L^{(1+n)'}(\Omega(t))$ satisfying for every $\Psi = (\psi, \varphi) \in \mathscr{W}(t)$ and for every $q \in L^{(1+n)'}(\Omega(t))$:

- The incompressibility condition:

$$\int_{\Omega(t)} q\, \mathrm{div}_{(X,Z)}\mathbf{U} dX dZ = 0, \qquad \forall q \in L^{(1+n)'}(\Omega(t)).$$

- And the momentum variational inequality:

$$\begin{aligned}
&\int_{\Omega(t)} \rho\Big(\frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U}\cdot\nabla_{(X,Z)})\mathbf{U}\Big)\cdot(\Psi - \mathbf{U}) dX dZ \\
&- \int_{\Omega(t)} \rho(b + Z\cos\theta + \frac{p}{\rho})(\mathrm{div}_{(X,Z)}\Psi - \mathrm{div}_{(X,Z)}\mathbf{U}) dX dZ \\
&+ \int_{\Omega(t)} \frac{\bar{\eta}}{2}|D(\mathbf{U})|^{n-1} D(\mathbf{U}) : (D(\Psi) - D(\mathbf{U})) dX dZ \\
&+ \frac{\tau_c}{2}\int_{\Omega(t)} (|D(\Psi)| - |D(\mathbf{U})|)\, dX dZ \\
&+ \int_{\Gamma_b(t)} \alpha U(\psi - U) d\gamma \geq 0 \qquad \forall\, \Psi \in \mathscr{W}(t)
\end{aligned} \tag{106}$$

with $n \in (0, 1)$.

Remark 3. The inequality in the weak formulation of the momentum equation is a consequence of the weak formulation of the rigidity term in the stress tensor. If we consider the case $|D(\mathbf{U})| \neq 0$, we have

$$\begin{aligned}
-\int_{\Omega(t)} \mathrm{div}_{(X,Z)}(\sigma')(\Psi - \mathbf{U}) dX dZ &= \frac{1}{2}\int_{\Omega(t)} \sigma' : D(\Psi - \mathbf{U}) dX dZ \\
&= \frac{1}{2}\int_{\Omega(t)} \tau_c \frac{D(\mathbf{U})}{|D(\mathbf{U})|} : (D(\Psi) - D(\mathbf{U})) dX dZ \\
&+ \frac{1}{2}\int_{\Omega(t)} \bar{\eta}|D(\mathbf{U})|^{n-1} D(\mathbf{U}) : (D(\Psi) - D(\mathbf{U})) dX dZ.
\end{aligned}$$

Moreover,

$$\tau_c \int_{\Omega(t)} \frac{D(\mathbf{U})}{|D(\mathbf{U})|} : (D(\Psi) - D(\mathbf{U}))dXdZ$$
$$= \tau_c \int_{\Omega(t)} \frac{D(\mathbf{U}) : D(\Psi)}{|D(\mathbf{U})|} dXdZ - \tau_c \int_{\Omega(t)} |D(\mathbf{U})|dXdZ$$

and,

$$\tau_c \int_{\Omega(t)} \frac{D(\mathbf{U}) : D(\Psi)}{|D(\mathbf{U})|} dXdZ \leq \tau_c \int_{\Omega(t)} |D(\Psi)|dXdZ.$$

Observe that this is also true for the case $|D(U)| = 0$. Therefore, a solution of the problem in differential form is a solution of the variational inequality. Nevertheless, it is not trivial to prove rigorously that the solution of the variational inequality is a solution of the differential problem. Although, formally it is possible to deduce the differential system from the variational inequality. One can consider as test functions $\Psi = \mathbf{U} + \lambda\mathbf{V}$, $\Psi = \mathbf{U} - \lambda\mathbf{V}$ and to study the limit when λ tends to zero. □

Let us develop the variational inequality (106) in terms of the components of the vector. As $\Psi = (\psi, \varphi)$ and $\mathbf{U} = (U, W)$, we have

$$\begin{aligned}
&\int_{\Omega(t)} \rho(\partial_t(U) + \rho U\partial_X U + \rho W\partial_Z U)(\psi - U)dXdZ \\
&+ \int_{\Omega(t)} \rho(\partial_t(W) + \rho U\partial_X W + \rho W\partial_Z W)(\varphi - W)dXdZ \\
&- \int_{\Omega(t)} \rho(b + Z\cos\theta + \frac{p}{\rho})(\partial_X(\psi - U) + \partial_Z(\varphi - W))dXdZ \\
&+ \int_{\Omega(t)} \bar{\eta}|D(\mathbf{U})|^{n-1}\Big(2\partial_X(U)\partial_X(\psi - U) \\
&+ (\partial_X(W) + \partial_Z(U))(\partial_X(\varphi - W) + \partial_Z(\psi - U)) + 2\partial_Z(W)\partial_Z(\varphi - W)\Big)dXdZ \\
&+ \tau_c \int_{\Omega(t)} \Big(\sqrt{(\partial_X\psi)^2 + \frac{1}{2}(\partial_X\varphi + \partial_Z\psi)^2 + (\partial_Z\varphi)^2} \\
&- \sqrt{(\partial_X U)^2 + \frac{1}{2}(\partial_X W + \partial_Z U)^2 + (\partial_Z W)^2}\Big)dXdZ \\
&+ \int_{\Gamma_b} \alpha U(\psi - U)d\gamma \geq 0
\end{aligned} \tag{107}$$

where

$$|D(\mathbf{U})|^{n-1} = \left(4(\partial_X U)^2 + 2(\partial_X W + \partial_Z U)^2 + 4(\partial_Z W)^2\right)^{(n-1)/2}.$$

8.3 $[\tilde{A}]$ Dimensional Analysis

Next, a dimensional analysis of the set of Eqs. (98), the kinematic and boundary conditions is performed. The non-dimensional variables ($\tilde{\cdot}$) read:

$$\begin{aligned}
&(X, Z, t) = (L\tilde{X}, H\tilde{Z}, (L/g)^{1/2}\tilde{t}),\\
&(U, W) = (Lg)^{1/2}(\tilde{U}, \varepsilon\tilde{W}),\\
&h = H\tilde{h},\\
&\alpha = \varepsilon(Lg)^{1/2}\tilde{\alpha},\\
&p = gH\tilde{p},\\
&\tau_c = gH\widetilde{\tau_c},\\
&\bar{\eta} = Hg^{(1-\frac{n}{2})}L^{\frac{n}{2}}\,\tilde{\bar{\eta}}.
\end{aligned} \tag{108}$$

Then,

$$\sigma' = gH\widetilde{\sigma'} = gH\begin{pmatrix} \widetilde{\sigma'}_{\tilde{X}\tilde{X}} & \widetilde{\sigma'}_{\tilde{X}\tilde{Z}} \\ \widetilde{\sigma'}_{\tilde{X}\tilde{Z}} & \widetilde{\sigma'}_{\tilde{Z}\tilde{Z}} \end{pmatrix}.$$

With

$$\begin{cases} \widetilde{\sigma'} = \widetilde{\tau_c}\frac{\tilde{D}_\varepsilon(\tilde{\mathbf{U}})}{|\tilde{D}_\varepsilon(\tilde{\mathbf{U}})|} + \tilde{\nu}\tilde{D}_\varepsilon(\tilde{\mathbf{U}}) & \text{if } |\tilde{D}_\varepsilon(\tilde{\mathbf{U}})| \neq 0, \\ |\widetilde{\sigma'}| \leq \widetilde{\tau_c} & \text{if } |\tilde{D}_\varepsilon(\tilde{\mathbf{U}})| = 0. \end{cases}$$

And

$$\tilde{D}_\varepsilon(\tilde{\mathbf{U}}) = \begin{pmatrix} 2\partial_{\tilde{X}}\tilde{U} & \varepsilon\partial_{\tilde{X}}\tilde{W} + \dfrac{1}{\varepsilon}\partial_{\tilde{Z}}\tilde{U} \\ \varepsilon\partial_{\tilde{X}}\tilde{W} + \dfrac{1}{\varepsilon}\partial_{\tilde{Z}}\tilde{U} & 2\partial_{\tilde{Z}}\tilde{W} \end{pmatrix}.$$

Using the above change of variables, the system of Eqs. (98) can be also rewritten as follows (tildes are again omitted):

$$\partial_X(U) + \partial_Z(W) = 0, \tag{109}$$

$$\partial_t(\rho U) + \rho U\partial_X U + \rho W\partial_Z U + \rho\partial_X(b + Z\cos\theta + \frac{p}{\rho})\varepsilon = \varepsilon\partial_X(\sigma'_{XX}) + \partial_Z(\sigma'_{XZ}), \tag{110}$$

$$\begin{aligned}\varepsilon\{\partial_t(\rho W) + \rho U\partial_X(W) + \rho W\partial_Z(W) - \partial_X(\sigma'_{XZ})\} + \\ + \rho\partial_Z(b + \cos\theta Z) = -\partial_Z(p) + \partial_Z(\sigma'_{ZZ}).\end{aligned} \tag{111}$$

For the momentum variational inequality (107), we also consider the following non-dimensional test functions:

$$\Psi = (\psi, \varphi) = (Lg)^{1/2}(\tilde{\psi},\ \varepsilon\tilde{\varphi}).$$

The variational inequality (107) is rewritten as (tildes are omitted again):

$$\begin{aligned}
&\int_{\Omega(t)} \rho(\partial_t(U) + \rho U\partial_X U + \rho W\partial_Z U)(\psi - U)dXdZ \\
&+\varepsilon^2\int_{\Omega(t)} \rho(\partial_t(W) + \rho U\partial_X W + \rho W\partial_Z W)(\varphi - W)dXdZ \\
&+\int_{\Omega(t)} \rho\varepsilon(\partial_X b(\psi - U) + \cos\theta(\varphi - W))dXdZ \\
&-\int_{\Omega(t)} \rho\varepsilon\frac{p}{\rho}(\partial_X(\psi - U) + \partial_Z(\varphi - W))dXdZ \\
&+\int_{\Omega(t)} \bar{\eta}|D_\varepsilon(\mathbf{U})|^{n-1}\Big(2\varepsilon\partial_X(U)\partial_X(\psi - U) + \varepsilon(\varepsilon\partial_X(W) + \frac{1}{\varepsilon}\partial_Z(U))(\varepsilon\partial_X(\varphi - W) \\
&+\frac{1}{\varepsilon}\partial_Z(\psi - U)) \\
&\qquad + 2\varepsilon\partial_Z(W)\partial_Z(\varphi - W)\Big)dXdZ \\
&+\tau_c\int_{\Omega(t)} \varepsilon\Big(\sqrt{(\partial_X\psi)^2 + \frac{1}{2}(\varepsilon\partial_X\varphi + \frac{1}{\varepsilon}\partial_Z\psi)^2 + (\partial_Z\varphi)^2} \\
&-\sqrt{(\partial_X U)^2 + \frac{1}{2}(\varepsilon\partial_X W + \frac{1}{\varepsilon}\partial_Z U)^2 + (\partial_Z W)^2}\Big)dXdZ \\
&+\int_{\Gamma_b} \alpha U(\psi - U)d\gamma \geq 0
\end{aligned} \tag{112}$$

where

$$|D_\varepsilon(\mathbf{U})|^{n-1} = \left(4(\partial_X U)^2 + 2(\varepsilon\partial_X W + \frac{1}{\varepsilon}\partial_Z U)^2 + 4(\partial_Z W)^2\right)^{(n-1)/2}.$$

Let us also remark that for the inclined plane case considered in these notes we have

$$\partial_X b = \sin\theta.$$

8.4 $[\int M \cdot \psi(X) \geq 0]$ *Variational Inequality for Test Functions Independent of Z*

In this section, we obtain the mass and momentum equations of a Shallow Herschel–Bulkley model. To obtain it, we neglect the second order terms ($\mathcal{O}(\varepsilon^2)$) and we consider test functions which are independent of Z.

Let us remark that to consider test functions independent of Z is analogous to depth average the mass and momentum equations. In fact, we can see that if $\tau_c = 0$, then we have a variational equality for the momentum conservation and the procedure described below is another way of deriving the Shallow Water equations. And, if the Coulomb friction law is considered at the bottom, the Savage–Hutter model deduced in Sect. 2 is recovered.

First, note that if $q \in L^{(1+n)'}(\Omega(t))$ is independent of Z then

$$\int_{\Omega(t)} q \,\mathrm{div}_{(X,Z)}\mathbf{U} = \int_0^L q(X)\left(\int_0^h \mathrm{div}_{(X,Z)}\mathbf{U} dZ\right)dX = 0.$$

By using the kinematic conditions, we obtain

$$\int_0^L q(X)\Big(\partial_t h + \partial_X(hU)\Big)dX = 0, \quad \forall q \in L^{(1+n)'}([0,L]).$$

This gives a different way to obtain the mass conservation equation:

$$\partial_t h + \partial_X(hU) = 0.$$

Let us now consider test functions $\Psi = (\psi, \varphi)$ where ψ is independent of Z. Analogously to previous sections, we assume that the velocity parallel to the bottom U is independent of Z. Then, if we neglect second order terms ($\mathcal{O}(\varepsilon^2)$) in (112), we obtain

$$
\begin{aligned}
&\int_0^L \rho h(\partial_t(U) + \rho U \partial_X U)(\psi - U) dX dZ \\
&+ \int_{\Omega(t)} \rho\varepsilon(\partial_X b(\psi - U) + \cos\theta(\varphi - W)) dX dZ \\
&- \int_{\Omega(t)} \rho\varepsilon \frac{p}{\rho}(\partial_X(\psi - U) + \partial_Z(\varphi - W)) dX dZ \\
&+ \int_{\Omega(t)} \bar{\eta}\, 2^{2n-2}\Big((\partial_X U)^2 + (\partial_Z W)^2\Big)^{n-1}\Big(2\varepsilon\partial_X(U)\partial_X(\psi - U) \\
&+ 2\varepsilon\partial_Z(W)\partial_Z(\varphi - W)\Big) dX dZ \\
&+ \tau_c \int_{\Omega(t)} \varepsilon\Big(\sqrt{(\partial_X\psi)^2 + (\partial_Z\varphi)^2} - \sqrt{(\partial_X U)^2 + (\partial_Z W)^2}\Big) dX dZ \\
&+ \int_0^L \alpha U(\psi - U) d\gamma \geq 0.
\end{aligned}
\tag{113}
$$

Moreover, by using the incompressibility condition, by choosing also test functions with zero divergence whose vertical component vanishes at the bottom—to be consistent with boundary condition (104)—we have

$$
W = -Z\partial_X U, \quad \text{and} \quad \varphi = -Z\partial_X\psi. \tag{114}
$$

By using (114) we get:

$$
\int_{\Omega(t)} (\partial_X b(\psi - U) + \cos\theta(\varphi - W)) dX dZ = \int_0^L \Big(h\partial_X b + \partial_X\big(\frac{h^2}{2}\cos\theta\big)\Big)(\psi - U) dX.
$$

Finally, using this last equality and (114), we obtain from (113):

$$
\begin{aligned}
&\int_0^L \rho h(\partial_t(U) + \rho U \partial_X U)(\psi - U) dX dZ \\
&+ \int_0^L \rho\Big(h\partial_X b + \partial_X\big(\frac{h^2}{2}\cos\theta\big)\Big)(\psi - U) dX \\
&+ \int_0^L 2^{3n-1}\bar{\eta}|\partial_X U|^{n-1}\varepsilon\partial_X(U)\partial_X(\psi - U) dX \\
&+ 2\tau_c \int_0^L \varepsilon h(|\partial_X\psi| - |\partial_X U|) dX \\
&+ \int_0^L \alpha U(\psi - U) d\gamma \geq 0.
\end{aligned}
\tag{115}
$$

8.5 [↪] *Final System of Equations*

Coming back to the original variables and using (108), we obtain the following system:

- Mass conservation:

$$\int_0^L \Big(\partial_t h + \partial_X(hU)\Big) q(X)dX = 0, \quad \forall q \in L^{(1+n)'}([0,L]). \tag{116}$$

- Momentum variational inequality: $\forall \psi \in W^{1,1+n}([0,L])$,

$$\begin{aligned}
&\int_0^L \rho\Big(h\partial_t(U) + \rho hU\partial_X U + gh\partial_X b + g\partial_X(\frac{h^2}{2}\cos\theta)\Big)(\psi - U)dXdZ \\
&+ \int_0^L 2^{3n-1}\bar{\eta}|\partial_X U|^{n-1}\partial_X(U)\partial_X(\psi - U)dX \\
&+2\tau_c \int_0^L h(|\partial_X \psi| - |\partial_X U|)dX \\
&+ \int_0^L \alpha U(\psi - U)d\gamma \geq 0.
\end{aligned} \tag{117}$$

Note that the first line corresponds to convection and pressure terms in Shallow Water systems and the second one to viscous effects. The third line contains the terms associated to the rigidity properties of the material. Last line of previous equation correspond to the bottom friction term.

Let us remark that (116) and (117) corresponds to the weak formulation of the following partial differential system:

$$\begin{cases} \partial_t h + \partial_X(hU) = 0, \\ \\ h\Big(\partial_t U + U\partial_x U + g(b + h\cos\theta)\Big) + \alpha U - \partial_X(h\sigma') = 0 \end{cases} \tag{118}$$

where

$$\begin{cases} \sigma' = 2^{3n-1}\bar{\eta}|\partial_X U|^{n-1}\partial_X U + 2\tau_c \dfrac{\partial_X U}{|\partial_X U|} & \text{if } |\partial_x U| \neq 0 \\ \\ |\sigma'| \leq 2\tau_c & \text{if } |\partial_x U| = 0. \end{cases} \tag{119}$$

As mentioned before, it is easy to see that if $\tau_c = 0, n = 1$ and the linear friction law is replaced by the Coulomb friction law (13) then the Savage–Hutter model deduced in Sect. 2—with standard viscous terms—is obtained.

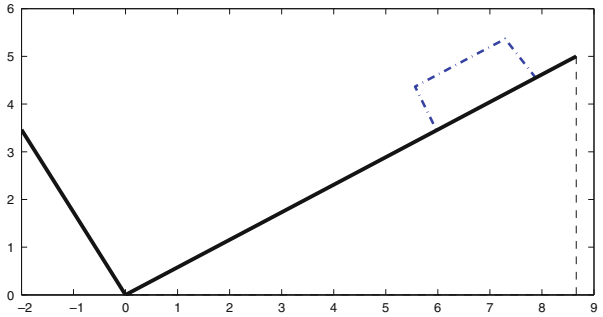


Fig. 11 Initial condition. *Dashed-dotted line*: free surface of the viscoplastic material. *Continuous line*: bottom

We refer to [1] for the discretization of this shallow-Herschel–Bulkley model. Let us present an example. As initial condition a rectangular layer is considered on a closed domain with a plain with a slope of 30° (see Fig. 11).

Figure 12 shows the evolution of the avalanche of the visco-plastic material corresponding to $n = 1$, $\tau_c = 4$, $\bar{\eta} = 10^{-2}$, $\alpha = 10^{-2}$. The left column shows the evolution of the free surface at times $t \in \{ 0.4, 1, 2.4, 4 \}$ s. The right column shows the velocity profile. This simulation shows a typical behavior of visco-plastic fluids: at the beginning it moves as a rigid body and then it starts to flow as a viscous fluid. Indeed, notice that for $t = 0.4$ and $t = 1$ s the velocity profile is nearly constant on all the domain filled by the avalanche, but this is no more the case for $t = 2.4$ s. For $t = 4$ s. the material is at rest. These different behaviors are due to the definition of the stress tensor (119).

Appendix: Bed-Load Sediment Transport Formulae

In this appendix we present several possible definitions of the solid transport discharge, q_b, that allow one to close the Saint-Venant Exner system (see Sect. 1).

The study of the definition of the solid transport discharge can be seen as a deterministic problem or a probabilistic one. For example, deterministic methods have been proposed by Meyer-Peter & Müller [36] and probabilistic methods by Einstein [20].

In general, the models take into account the fact that motion of the granular sediment begins when the shear stress (τ) is bigger than a certain critical shear stress (τ_c). Moreover, shear stress can be written in terms of the hydrodynamic unknowns h and u by

$$\tau = \gamma R_h |S_f|. \tag{120}$$

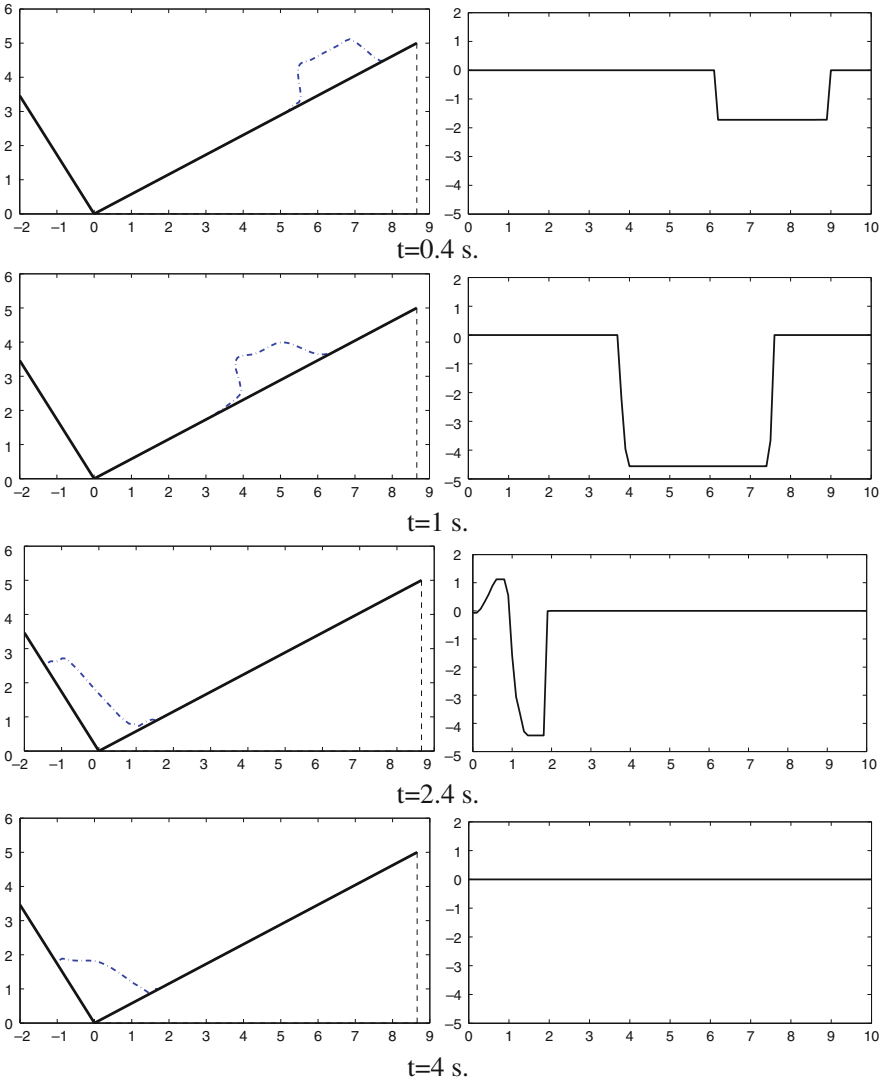


Fig. 12 Complex avalanche: Herschel–Bulkley model. *Left*: free surface; *Right*: velocity

Here S_f is defined by (2) and γ is the specific weight of fluid $\gamma = g\rho_w$, where ρ_w is the water density.

Shear stress appears usually in non-dimensional form in the formula of q_b. If τ_* and τ_{*c} represent the non-dimensional shear stress and the critical shear stress, respectively, then

$$\tau_* = \frac{\tau}{(\gamma_s - \gamma)d}, \quad \tau_{*c} = \frac{\tau_c}{(\gamma_s - \gamma)d}. \tag{121}$$

Here d is the sediment grain size and γ_s is the specific sediment weight $\gamma_s = g\rho_s$, where ρ_s is the sediment density.

Using (120) and (121), τ_* can be written as a function of the specific gravity or the relative density of fluids $r = \rho_s/\rho_w$.

$$\tau_* = \frac{g\eta^2 u^2}{(r-1)dR_h^{1/3}}.$$

To determine τ_{*c} many experiments have been performed in different works. Concretely, Shields proposed the well-known Shields-diagram (cf. [40], p. 107).

Some usual formulae for rivers are the following:

- Grass (see [26]) proposed the following formula for the solid transport discharge,

$$q_b = A_g u\,|u|^{m_g - 1}, \ 1 \leqslant m_g \leqslant 4,$$

  where the constant A_g (s^2/m) must take into account the grain diameter and the kinematic viscosity. It is usually obtained by experimental data. The usual value of exponent m_g is set to $m_g = 3$.
- Meyer-Peter & Müller (see [36]) developed one of the most popular formulae for the solid transport discharge,

$$q_b = \sqrt{(r-1)gd^3}\,\mathrm{sgn}\,(u) 8\,(\tau_* - \tau_{*c})^{3/2},$$

  where τ_{*c} usually is set to 0.047.
- Van Rijn (see [49]) developed the following formula for the solid transport discharge,

$$q_b = \sqrt{(r-1)gd^3}\frac{0.005}{C_D^{1.7}}\left(\frac{d}{h}\right)^{0.2}\tau_*^{1/2}\,\mathrm{sgn}\,(u)\left(\tau_*^{1/2} - \tau_{*c}^{1/2}\right)^{2.4},$$

  where C_D is the drag coefficient.
- Nielsen (see [40]) developed the following formula

$$q_b = \sqrt{(r-1)gd^3}\,\mathrm{sgn}\,(u) 12\sqrt{\tau_*}(\tau_* - \tau_{*c}).$$

  In this case the usual value of τ_{*c} is set equal to $\tau_{*c} = 0.05$.

All these formulae have a range of application which depends on the grain size, the slope of the bottom, the Froude number and the relative density r. For example, the M-P&M formula can be applied if $0.4 \leq d \leq 29\,\text{mm}$, the slope of the bottom is smaller than 0.02 and $1.25 \leq r \leq 4.2$. For more details see [15–17].

□

Acknowledgements The first author would like to thanks the organizers of the Jacques-Louis Lions Spanish-French school for the invitation. The second author would like to thank the Institute of Mathematics of the University of Seville (IMUS) for the financial support to work on the numerical analysis of models for visco-plastic avalanches.

References

1. Acary-Robert, C., Fernández-Nieto, E., Narbona-Reina, G., Vigneaux, P.: A well-balanced finite volume-augmented Lagrangian method for an integrated Herschel-Bulkley model. J. Sci. Comput. **53**, 608–641 (2012)
2. Ancey, C.: Plasticity and geophysical flows: a review. J. Non-Newtonian Fluid Mech. **142**, 4–35 (2007)
3. Ancey, C., Cochard, S.: The dam-break problem for Herschel-Bulkley viscoplastic fluids down steep flumes. J. Non-Newtonian Fluid Mech. **158**, 18–35 (2009)
4. Anderson, T.B., Jackson, R.: A fluid mechanical description of fluidized beds. Ind. Eng. Chem. Fundam. **6**, 527–539 (1967)
5. Aradian, A., Raphael, E., de Gennes, P.G.: Surface flow of granular materials: a short introduction to some recent models. C. R. Phys. **3**, 187–196 (2002)
6. Aranson, I.S., Tsimring, L.S.: Continuum theory of partially fluidized granular flows. Phys. Rev. E. **65**, 061303 (2002)
7. Audusse, E., Bristeau, M.-O., Perthame, B., Sainte-Marie, J.: A multilayer Saint-Venant system with mass exchanges for shallow water flows. Derivation and numerical validation. ESAIM Math. Model. Numer. Anal. **45**, 169–200 (2011)
8. Balmforth, N.J., Craster, R.V., Rust, A.C., Sassi, R.: Viscoplastic flow over an inclined surface. J. Non-Newtonian Fluid Mech. **139**, 103–127 (2006)
9. Batchelor, G.K.: An Introduction to Fluid Dynamics. Cambridge University Press, New Delhi (2000)
10. Bingham, E.C.: Fluidity and Plasticity. Mc Graw-Hill, New York (1922)
11. Bouchut, F., Mangeney-Castelnau, A., Perthame, B., Vilotte, J.P.: A new model of Saint Venant and Savage-Hutter type for gravity driven shallow flows. C. R. Acad. Sci. Paris Ser. I **336**, 531–536 (2003)
12. Bouchut, F., Fernández-Nieto, E.D., Mangeney, A., Lagree, P.Y.: On new erosion models of Savage-Hutter type for avalanches. Acta Mecha. **199**, 181–208 (2008)
13. Bresch, D., Desjardins, B.: Existence of global weak solutions for a 2D viscous shallow water equations and convergence to the quasi-geostrophic model. Commun. Math. Phys. **238**, 211–223 (2003)
14. Bresch, D., Fernandez-Nieto, E.D., Ionescu, I.R., Vigneaux, P.: Augmented Lagrangian method and compressible visco-plastic flows: Applications to shallow dense avalanches. In: Advances in Mathematical Fluid Mechanics, pp. 57–89. Birkhauser, Basel (2010)
15. Castro-Díaz, M.J., Fernández-Nieto, E.D., Ferreiro, A.: Sediment transport models in Shallow Water equations and numerical approach by high order finite volume methods. Comput. Fluids **37**, 299–316 (2008)

16. Castro-Díaz, M.J., Fernández-Nieto, E.D., Ferreiro, A., Parés, C.: Two-dimensional sediment transport models in shallow water equations: a second order finite volume approach on unstructured meshes. Comput. Meth. App. Mech. Eng. **198**, 2520–2538 (2009)
17. Cordier, S., Le, M., Morales de Luna, T.: Bedload transport in shallow water models: why splitting (may) fail, how hyperbolicity (can) help. Adv. Water Resour. **34**, 980–989 (2011)
18. Dressler, R.F.: New nonlinear shallow equations with curvature. J. Hydraul. Res. **16**, 205–22 (1978)
19. Duvaut, G., Lions, J.-L.: Inequalities in Mechanics and Physics. Springer, Berlin (1976)
20. Einstein, H.A.: The bed load function for sediment transport in open channel flows. Technical Bulletin no. 1026. U.S. Department of Agriculture, Soil Conservation Service, Washington, DC (1950)
21. Fernandez-Nieto, E.D., Bouchut, F., Bresch, D., Castro-Díaz, M.J., Mangeney, A.: A new Savage-Hutter type model for submarine avalanches and generated tsunami. J. Comput. Phys. **227**, 7720–7754 (2008)
22. Fernández-Nieto, E.D., Noble, P., Vila, J.P.: Shallow Water equations for Non-Newtonian fluids. J. Non-Newtonian Fluid Mech. **165**, 712–732 (2010)
23. Fernández-Nieto, E.D., Koné, E.H., Chacón, T.: A multilayer method for the hydrostatic Navier-Stokes equations: a particular weak solution. J. Sci. Comput. (2013). doi: 10.1007/s10915-013-9802-0
24. Ferrari, S., Saleri, F.: A new two-dimensional shallow water model including pressure effects and slow varying bottom topography. Math. Model. Numer. Anal. **38**, 211–234 (2004)
25. Gerbeau, J., Perthame, B.: Derivation of viscous Saint-Venant system for laminar shallow water: numerical validation. Discrete Contin. Dyn. Syst. Ser. B **1**, 89–102 (2001)
26. Grass, A.J.: Sediments transport by waves and currents. SERC London Centre for Marine Technology, Report No. FL29 (1981)
27. Gray, J.M.N.T.: Granular flow in partially filled slowly rotating drums. J. Fluid Mech. **441**, 1–29 (2001)
28. Heinrich, P., Piatanesi, A., Hébert, H.: Numerical modelling of tsunami generation and propagation from submarine slumps: the 1998 Papua New Guinea event. Geophys. J. Int. **145**, 97–11 (2001)
29. Iverson, R.M., Denlinger, R.P.: Flow of variably fluidized granular masses across three-dimensional terrain. J. Geophys. Res. **106**, 537–552 (2001)
30. Jackson, R.: The Dynamics of Fluidized Particles. Cambridge Monographs on Mechanics. Cambridge University Press, New York (2000)
31. Khakhar, D.V., Orpe, A.V., Andresén, P., Ottino, J.M.: Surface flow of granular materials: model and experiments in heap formation. J. Fluid Mech. **441**, 225–264 (2001)
32. Macías, J., Fernández-Salas, L.M., González-Vida, J.M., Vázquez, J.T., Castro Días, M.J., Bárcenas, P., del Río, P., Díaz, V., Morales de Luna, T., de la Asunción, M., Parés, C.: Deslizamientos Submarinos y Tsunamis en el Mar de Alborán. Un ejemplo de modelización, vol. 6. Instituto Español de Oceanografía, Spain (2012)
33. Madsen, P.A., Bingham, H.B., Schaffer, H.A.: Boussinesq-type formulations for fully nonlinear and extremely dispersive water waves: derivation and analysis. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. **459**, 1075–104 (2003)
34. Mangeney-Castelnau, A., Bouchut, F., Vilotte, T.P., Lajeneusse, E., Aubertin, A., Pirulli, M.: On the use of Saint-Venant equations to simulate the spreading of a granular mass. J. Geophys. Res. **110**, B09103 (2005)
35. Marche, F.: Theoretical and numerical study of Shallow Water models. Application to Nearshore hydrodynamics. Thesis of the University of Bordeaux, France (2005)
36. Meyer-Peter, E., Müller, R.: Formulas for bed-load transport. Report on 2nd Meeting of International Association for Hydraulic Research, 39–64. Stockholm (2005)
37. Morales de Luna, T.: A Saint Venant model for gravity driven shallow water flows with variable density and compressibility effects. Math. Comput. Model. **47**, 436–444 (2008)

38. Morales de Luna, T., Castro Díaz, M.J., Parés Madroñal, C., Fernández Nieto, E.D.: On a shallow water model for the simulation of turbidity currents. Commun. Comput. Phys. **6**(4), 848–882 (2009)
39. Narbona, G., Zabsonre, J., Fernandez-Nieto, E.D., Bresch, D.: Derivation of a bilayer model for shallow water equations with viscosity: numerical validation. Comput. Model. Eng. Sci. **43**, 27–71 (2009)
40. Nielsen, P.: Coastal bottom boundary layers and sediment transport. In: Advanced Series on Ocean Engineering, vol. 4. World Scientific Publishing, Singapore (1992)
41. Oswald, P.: Rheophysics: The Deformation and Flow of Matter. Cambridge University Press, New York (2009)
42. Pelanti, M., Bouchut, F., Mangeney, A.: A roe-type scheme for two-phase Shallow granular flows with bottom topography. Math. Model. Numer. Anal. **42**, 851–885 (2008)
43. Pirulli, M., Bristeau, M.O., Mangeney, A., Scavia, C.: The effect of the earth pressure coefficients on the runout of granular material. Environ. Model. Softw. **22**, 1437–1454 (2007)
44. Pitman, E.B., Le, L.: A two-fluid model for avalanche and debris flows. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **363**,1573–1601 (2005)
45. Pudasaini, S., Hutter, K.: Avalanche Dynamics. Springer, New York (2007)
46. Saint-Venant, A.J.C.: Théorie du mouvement non-permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit. C. R. Acad. Sci. Paris **73**, 147–54 (1871)
47. Savage, S.B., Hutter, K.: The dynamics of avalanches of granular materials from initiation to run-out. Acta Mech. **86**, 201–223 (1991)
48. Tanner, R.I., Walters, K.: Rheology: An Historical Perspective. Elsevier, Amsterdam (1998)
49. Van Rijn, L.C.: Sediment transport (III): bed forms and alluvial roughness. J. Hydraul. Div. Proc. ASCE **112**, 1733–1754 (1984)
50. Wieland, M., Gray, J.M.N.T., Hutter, K.: Channelized free-surface flow of cohesionless granular avalanches in a chute with shallow lateral curvature. J. Fluid Mech. **392**, 73–100 (1999)

Introduction to Stochastic Calculus and to the Resolution of PDEs Using Monte Carlo Simulations

Emmanuel Gobet

Abstract I give a pedagogical introduction to Brownian motion, stochastic calculus introduced by Itô in the fifties, following the elementary (at least not too technical) approach by Föllmer [Seminar on Probability, XV (Univ. Strasbourg, Strasbourg, 1979/1980) (French), pp. 143–150. Springer, Berlin, 1981]. Based on this, I develop the connection with linear and semi-linear parabolic PDEs. Then, I provide and analyze some Monte Carlo methods to approximate the solution to these PDEs. This course is aimed at master students, Ph.D. students and researchers interesting in the connection of stochastic processes with PDEs and their numerical counterpart. The reader is supposed to be familiar with basic concepts of probability (say first chapters of the book *Probability essentials* by Jacod and Protter [Probability Essentials, 2nd edn. Springer, Berlin, 2003]), but no a priori knowledge on martingales and stochastic processes is required.

1 The Brownian Motion and Related Processes

1.1 A Brief History of Brownian Motion

Historically, the Brownian motion (BM in short) is associated with the analysis of motions which time evolution is so disordered that it seems difficult to forecast their evolution, even in a very short time interval. It plays a central role in the theory of random processes, because in many theoretical and applied problems, the Brownian motion (or the diffusion processes that are built from Brownian motion) provides simple limit models on which many calculations can be made.

E. Gobet (✉)
Centre de Mathématiques Appliquées, Ecole Polytechnique and CNRS, Route de Saclay, 91128 Palaiseau Cedex, France
e-mail: emmanuel.gobet@polytechnique.edu

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_3,

In 1827, the English botanist Robert Brown (1773–1858) first described the erratic motion of fine organic particles in suspension in a gas or a fluid. At the nineteenth century, after him, several physicists had admitted that this motion is very irregular and does not seem to admit a tangent; thus one could not speak of his speed, nor apply the laws of mechanics to him! In 1900 [4], Louis Bachelier (1870–1946) introduced the Brownian motion to model the dynamics of the stock prices, but his approach then is forgotten until the sixties... His Ph.D. thesis, *Théorie de la spéculation*, is the starting point of modern finance.

But Physics is the field at the beginning of the twentieth century which is at the origin of great interest for this process. In 1905, Albert Einstein (1879–1955) built a probabilistic model to describe the motion of a diffusive particle: he found that the law of the particle position at the time t, given the initial state x, admits a density which satisfies the *heat equation*, and actually it is *Gaussian*. Its theory is then quickly confirmed by experimental measurements of satisfactory diffusion constants. The same year as Einstein, a discrete version of the Brownian motion is proposed by the Polish physicist Smoluchowski using random walks.

In 1923, Norbert Wiener (1894–1964) built rigorously the random function that is called *Brownian motion*; he established in particular that the trajectories are continuous. By 1930, while following an idea of Paul Langevin, Ornstein and Uhlenbeck studied the Gaussian random function which bears their name and which seems to be the stationary or mean-reverting equivalent model associated to the Brownian motion.

It is the beginning of a very active theoretical research in Mathematics. Paul Lévy (1886–1971) discovered then, with other mathematicians, many properties of the Brownian motion [55] and introduced a first form of the stochastic differential equations, the study of which is later systematized by K. Itô (1915–2008). His work is gathered in a famous treaty published in 1948 [44] which is usually referred to as *Itô stochastic calculus*.

But History knows sometimes incredible bounces. Indeed in 2000, the French Academy of Science opened a manuscript remained sealed since 1940 pertaining to the young mathematician Doeblin (1915–1940), a French telegraphist died during the German offensive. Doeblin was already known for his remarkable achievements in the theory of probability due to his works on the stable laws and the Markov processes. This sealed manuscript gathered in fact his recent research, written between November 1939 and February 1940: it was actually related to his discovery (before Itô) of the stochastic differential equations and their relations with the Kolmogorov partial differential equations. Perhaps the Itô stochastic calculus could have been called Doeblin stochastic calculus...

1.2 The Brownian Motion and Its Paths

In the following, we study the basic properties of the Brownian motion and its paths.

1.2.1 Definition and Existence

The very erratic path which is a specific feature of the Brownian motion is in general associated with the observation that the phenomenon, although very disordered, has a certain time homogeneity, i.e. the origin date does not have importance to describe the time evolution. These properties underly the next definition.

Definition 1 (of Standard Brownian Motion). A *standard Brownian motion* is a random process $\{W_t; t \geq 0\}$ with continuous paths, such that

- $W_0 = 0$.
- The time increment $W_t - W_s$ with $0 \leq s < t$ has the Gaussian law,[1] with zero mean and variance equal $(t - s)$.
- For any $0 = t_0 < t_1 < t_2 \ldots . < t_n$, the increments $\{W_{t_{i+1}} - W_{t_i}; 0 \leq i \leq n-1\}$ are independent[2] random variables.

There are important remarks following from the definition.

1. The state W_t of the system at time t is distributed as a Gaussian r.v. with mean 0 and variance t (increasing as time gets larger). Its probability density is

$$\mathbb{P}(W_t \in [x, x + \mathrm{d}x]) = g(t, x)\mathrm{d}x = \frac{1}{\sqrt{2\pi t}} \exp(-x^2/2t)\mathrm{d}x. \tag{1}$$

2. With probability 95 %, we have $|W_t| \leq 1.96\sqrt{t}$ (see Fig. 1) for a given time t. However, it may occur that W goes out this confidence interval.
3. The random variable W_t, as the sum of its increments, can be decomposed as a sum of independent Gaussian r.v.: this property serves as a basis from the further stochastic calculus.

Theorem 1. *The Brownian motion exists!*

Proof. There are different constructive ways to prove the existence of Brownian motion. Here, we use a Fourier based approach (proposed by Wiener), showing that W can be represented as a superposition of Gaussian signals. Also, we use a

[1] A Gaussian random variable X (see [46]) with mean μ and variance $\sigma^2 > 0$ (often denoted by $\mathcal{N}(\mu, \sigma^2)$) is the r.v. with density

$$g_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}], \quad x \in \mathbb{R}.$$

If $\sigma^2 = 0$, $X = \mu$ with probability 1. Moreover, for any $u \in \mathbb{R}$, $\mathbb{E}(e^{uX}) = e^{u\mu + \frac{1}{2}u^2\sigma^2}$.

[2] Two random variables X_1 and X_2 are independent if and only if $\mathbb{E}(f(X_1)g(X_2)) = \mathbb{E}(f(X_1))\mathbb{E}(g(X_2))$ for any bounded functions f and g. This extends similarly to a vector.

Fig. 1 Simulation of a Brownian motion with the 95 %-confidence interval curves $f_{\pm}(t) = \pm 2\sqrt{t}$

equivalent characterization of Brownian motion as a Gaussian process[3] with zero mean and covariance function $\mathbb{C}\mathrm{ov}(W_t, W_s) = \min(s,t) = s \wedge t$.

Let $(G_m)_{m\geq 0}$ be a sequence of independent Gaussian r.v. with zero mean and unit variance and set

$$W_t = \frac{t}{\sqrt{\pi}} G_0 + \sqrt{\frac{2}{\pi}} \sum_{m\geq 1} \frac{\sin(mt)}{m} G_m.$$

We now show that W is a Brownian motion on $[0,\pi]$; then it is enough to concatenate and sum up such independent processes to get finally a Brownian motion defined on $\mathbb{R}^+$. We sketch the proof of our statement on W. First, the series is *a.s.*[4] convergent since this is a Cauchy sequence in L_2: indeed, thanks to the independence of the Gaussian random variables, we have

$$\| \sum_{m_1 \leq m \leq m_2} \frac{\sin(mt)}{m} G_m \|_{L_2}^2 = \sum_{m_1 \leq m \leq m_2} \frac{\sin^2(mt)}{m^2} \leq \sum_{m_1 \leq m} \frac{1}{m^2} \underset{m_1 \to +\infty}{\longrightarrow} 0.$$

[3] $(X_1, \ldots, X_n)$ is a Gaussian vector if and only if for any $(\lambda_i)_{1\leq i\leq n} \in \mathbb{R}^n$, $\sum_{i=1}^n \lambda_i X_i$ has a Gaussian distribution. Independent Gaussian random variables form a Gaussian vector. A process $(X_t)_t$ is Gaussian if $(X_{t_1}, \ldots, X_{t_n})$ is a Gaussian vector for any times $(t_1, \ldots, t_n)$ and any n. A Gaussian process is characterized by its mean $m(t) = \mathbb{E}(X_t)$ and its covariance function $K(s,t) = \mathbb{C}\mathrm{ov}(X_s, X_t)$.

[4] We recall that "an event A occurs *a.s.*" (*almost surely*) if $\mathbb{P}(\omega : \omega \in A) = 1$ or equivalently if $\{w : w \notin A\}$ is a set of zero probability measure.

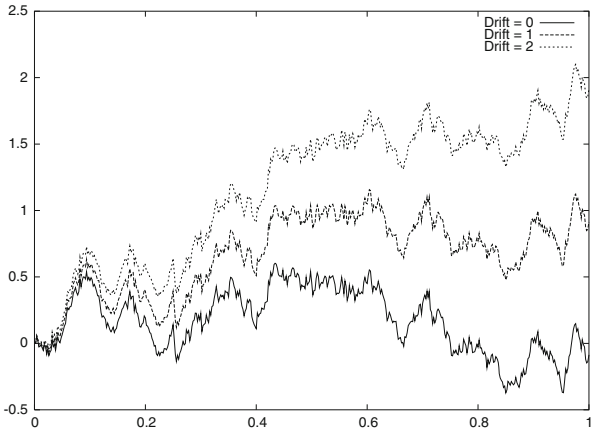


Fig. 2 Arithmetic Brownian motion with different drift parameters

The partial sum has a Gaussian distribution, thus the *a.s.* limit[5] too. The same argument gives that W is a Gaussian process. It has zero mean and its covariance is the limit of the covariance of partial sums: thus

$$\mathbb{C}\mathrm{ov}(W_t, W_s) = \frac{ts}{\pi} + \frac{2}{\pi} \sum_{m \geq 1} \frac{\sin(mt)}{m} \frac{\sin(ms)}{m}.$$

The above series is equal to $\min(s,t)$ for $(s,t) \in [0,\pi]^2$, by a standard computation of the Fourier coefficients of the function $t \in [-\pi, \pi] \mapsto \min(s,t)$ (for s fixed). The proof of continuity of W is based on the uniform convergence of the function series along some appropriate subsequences, which we do not detail (see [45, pp. 21–22]). □

In many applications, it is useful to consider non standard Brownian motion.

Definition 2 (of Arithmetic Brownian Motion). An *arithmetic Brownian motion* (ABM in short) is a random process $\{X_t; t \geq 0\}$ where $X_t = x_0 + bt + \sigma W_t$ and

- W is a standard Brownian motion.
- $x_0 \in \mathbb{R}$ is the starting value of X.
- $b \in \mathbb{R}$ is the drift parameter.
- $\sigma \in \mathbb{R}$ is the diffusion parameter.

Usually, σ can be taken non-negative due to the symmetry of Brownian motion (see Proposition 1). X is still a Gaussian process, which position X_t at time t is distributed as $\mathcal{N}(x_0 + bt, \sigma^2 t)$ (Fig. 2).

[5] Here, we use the following standard result: let $(X_n)_{n \geq 1}$ be a sequence of random variables, each having the Gaussian distribution with mean μ_n and variance σ_n^2. If the distribution of X_n converges, then (μ_n, σ_n^2) converge to (μ, σ^2), and the limit distribution is Gaussian with mean μ and variance σ^2. We recall that if X_n converges *a.s.*, then it also converges in distribution.

1.2.2 First Easy Properties of the Brownian Path

Proposition 1. *Let $\{W_t; t \in \mathbb{R}^+\}$ a standard Brownian motion.*

i) SYMMETRY PROPERTY*: $\{-W_t; t \in \mathbb{R}^+\}$ is a standard Brownian motion.*

ii) SCALING PROPERTY*: for any $c > 0$, $\{W^c_t; t \in \mathbb{R}^+\}$ is a standard Brownian motion where*

$$W^c_t = c^{-1} W_{c^2 t}. \tag{2}$$

iii) TIME REVERSAL*: for any fixed T, $\hat{W}^T_t = W_T - W_{T-t}$ defines a standard Brownian motion on $[0, T]$.*

iv) TIME INVERSION*: $\{\hat{W}_t = tW_{1/t}, t > 0, \quad \hat{W}_0 = 0\}$ is a standard Brownian motion.*

The scaling property is important and illustrates the fractal feature of Brownian motion path: ε times W_t behaves like a Brownian motion at time $\varepsilon^2 t$.

Proof. It is a direct verification of the Brownian motion definition, related to independent, stationary and Gaussian increments. The continuity is also easy to verify, except for the case *iv)* at time 0. For this, we use that $\lim_{t \to 0^+} tW_{1/t} = \lim_{s \to +\infty} \frac{W_s}{s} = 0$, see Proposition 7. □

1.3 Time-Shift Invariance and Markov Property

Previously, we have studied simple spatial transformation of Brownian motion. We now consider time-shifts, by first considering deterministic shifts.

Proposition 2 (Invariance by a Deterministic Time-Shift). *The Brownian Motion shifted by $h \geq 0$, given by $\{\bar{W}^h_t = W_{t+h} - W_h; t \in \mathbb{R}^+\}$, is another Brownian motion, independent of the Brownian Motion stopped at h, $\{W_s; s \leq h\}$.*

In other words, $\{W_{t+h} = W_h + \bar{W}^h_t; t \in \mathbb{R}^+\}$ is a Brownian motion starting from W_h. The above property is associated to the *weak Markov property* which states (possibly applicable to other processes) that *the distribution of W after h conditionally on the past up to time h depends only on the present value W_h.*

Proof. The Gaussian property of $\bar{W}^h$ is clear.

The independent increments of W induce those of $\bar{W}^h$.

It remains to show the independence w.r.t. the past up to h, i.e. the sigma-field generated by $\{W_s; s \leq h\}$, or equivalently w.r.t. the sigma-field generated by $\{W_{s_1}, \dots W_{s_N}\}$ for any $0 \leq s_1 \leq \cdots \leq s_N \leq h$. The independence of increments of W ensures that $(\bar{W}^h_{t_1}, \bar{W}^h_{t_2} - \bar{W}^h_{t_1}, \cdots, \bar{W}^h_{t_k} - \bar{W}^h_{t_{k-1}}) = (W_{t_1+h} - W_h, \cdots, W_{t_k+h} - W_{t_{k-1}+h})$ is independent of $(W_{s_1}, W_{s_2} - W_{s_1}, \cdots, W_{s_j} - W_{s_{j-1}})$. Then $(\bar{W}^h_{t_1}, \bar{W}^h_{t_2}, \cdots, \bar{W}^h_{t_k})$ is independent of $\{W_s; s \leq h\}$. □

As a consequence, we can derive a nice symmetry result making the connection between the maximum of Brownian motion monitored along a finite time grid $t_0 = 0 < t_1 < \cdots < t_N = T$ and that of W_T only.

Proposition 3. *For any $y \geq 0$, we have*

$$\mathbb{P}[\sup_{i \leq N} W_{t_i} \geq y] \leq 2\mathbb{P}[W_T \geq y] = \mathbb{P}[|W_T| \geq y]. \tag{3}$$

Proof. The equality at the r.h.s. comes from the symmetric distribution of W_T. Now we show the inequality on the left. Denote by t_y^* the first time t_j when W reaches the level y. Notice that $\{\sup_{i \leq N} W_{t_i} \geq y\} = \{t_y^* \leq T\}$ and $\{t_y^* = t_j\} = \{W_{t_i} < y, \forall i < j, W_{t_j} \geq y\}$. For each $j < N$, the symmetry of Brownian increments gives $\mathbb{P}[W_T - W_{t_j} \geq 0] = \frac{1}{2}$. Since the shifted Brownian motion $(\bar{W}_t^{t_j} = \bar{W}_{t_j+t} - W_{t_j} : t \in \mathbb{R}^+)$ is independent of $(W_s : s \leq t_j)$, we have

$$\begin{aligned}
&\tfrac{1}{2}\ \mathbb{P}[\sup_{i \leq N} W_{t_i} \geq y] = \frac{1}{2}\mathbb{P}[t_y^* \leq T] = \frac{1}{2}\sum_{j=0}^{N} \mathbb{P}[t_y^* = t_j] \\
&= \frac{1}{2}\mathbb{P}[W_{t_i} < y, \forall i < N, W_T \geq y] + \sum_{j=0}^{N-1} \mathbb{P}[W_{t_i} < y, \forall i < j, W_{t_j} \geq y]\mathbb{P}[W_T - W_{t_j} \geq 0] \\
&= \frac{1}{2}\mathbb{P}[W_{t_i} < y, \forall i < N, W_T \geq y] + \sum_{j=0}^{N-1} \mathbb{P}[W_{t_i} < y, \forall i < j, W_{t_j} \geq y, W_T - W_{t_j} \geq 0] \\
&\leq \mathbb{P}[W_{t_i} < y, \forall i < N, W_T \geq y] + \sum_{j=0}^{N-1} \mathbb{P}[W_{t_i} < y, \forall i < j, W_{t_j} \geq y, W_T \geq y] \\
&= \mathbb{P}[t_y^* \leq T, W_T \geq y] = \mathbb{P}[W_T \geq y].
\end{aligned}$$

At the two last lines, we have used $\{W_{t_j} \geq y, W_T - W_{t_j} \geq 0\} \subset \{W_{t_j} \geq y, W_T \geq y\}$ and $\{W_T \geq y\} \subset \{t_y^* \leq T\}$. □

Taking a grid with time step T/N with $N \to +\infty$, we have $\sup_{i \leq N} W_{t_i} \uparrow \sup_{0 \leq t \leq T} W_{t_i}$. Then, we can pass to the limit (up to some probabilistic convergence technicalities) in the inequality (3) to get

$$\mathbb{P}[\sup_{0 \leq t \leq T} W_t \geq y] \leq \mathbb{P}[|W_T| \geq y]. \tag{4}$$

Actually, the inequality (4) is an equality: it is proved later in Proposition 5.

Now, our aim is to extend Proposition 2 to the case of stochastic time-shifts h. Without extra assumption on h, the result is false in general: a counter-example is the last passage time of W at zero before the time 1 ($L = \sup\{t \leq 1 : W_t = 0\}$), which does not satisfy the property. Indeed, since $(W_{s+L} - W_s)_{s \geq 0}$ do not vanish

a.s. at short time (due to the definition of L), the marginal distribution can not be Gaussian and the time-shifted process can not be a Brownian motion.

The right class for extension is the class of stopping times, defined as follows.

Definition 3 (Stopping Time). A stopping time is non-negative random variable U (taking possibly the value $+\infty$), such that for any $t \geq 0$, the event $\{U \leq t\}$ depends only on the Brownian motion values $\{W_s; s \leq t\}$.

The stopping time is discrete if it takes only a countable set of values $(u_1, \cdots, u_n, \cdots)$.

In other words, it suffices to observe the Brownian motion until time t to know whether or not the event $\{U \leq t\}$ occurs. Of course, deterministic times are stopping times. A more interesting example is the first hitting time of a level $y > 0$

$$T_y = \inf\{t > 0; W_t \geq y\};$$

it is a stopping time, since $\{T_y \leq t\} = \{\exists s \leq t, W_s = y\}$ owing to the continuity of W. Observe that the counter-example of last passage time L is not a stopping time.

Proposition 4. *Let U be a stopping time. On the event $\{U < +\infty\}$, the Brownian motion shifted by $U \geq 0$, i.e. $\{\bar{W}_t^U = W_{t+U} - W_U; t \in \mathbb{R}^+\}$, is a Brownian motion independent of $\{W_t; t \leq U\}$.*

This result is usually referred to as the *strong Markov property*.

Proof. We show that for any $0 \leq t_1 < \cdots < t_k$, any $0 \leq s_1 < \cdots < s_l$, any $(x_1, \cdots, x_k)$ and any measurable sets $(B_1, \cdots, B_{l-1})$, we have

$$\begin{aligned}
&\mathbb{P}(\bar{W}_{t_1}^U < x_1, \cdots, \bar{W}_{t_k}^U < x_k, W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U < +\infty) \\
&= \mathbb{P}(W'_{t_1} < x_1, \cdots, W'_{t_k} < x_k)\mathbb{P}(W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U < +\infty), \quad (5)
\end{aligned}$$

where W' is a Brownian motion independent of W. We begin with the easier case where U is a discrete stopping time valued in $(u_n)_{n\geq 1}$: then

$$\begin{aligned}
&\mathbb{P}(\bar{W}_{t_1}^U < x_1, \cdots, \bar{W}_{t_k}^U < x_k, W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U < +\infty) \\
&= \sum_n \mathbb{P}(\bar{W}_{t_1}^U < x_1, \cdots, \bar{W}_{t_k}^U < x_k, W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U, U = u_n) \\
&= \sum_n \mathbb{P}(\bar{W}_{t_1}^{u_n} < x_1, \cdots, \bar{W}_{t_k}^{u_n} < x_k, W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U, U = u_n) \\
&= \sum_n \mathbb{P}(W'_{t_1} < x_1, \cdots, W'_{t_k} < x_k)\mathbb{P}(W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U, U = u_n) \\
&= \mathbb{P}(W'_{t_1} < x_1, \cdots, W'_{t_k} < x_k)\mathbb{P}(W_{s_1} \in B_1, \cdots, W_{s_{l-1}} \in B_{l-1}, s_l \leq U < +\infty)
\end{aligned}$$

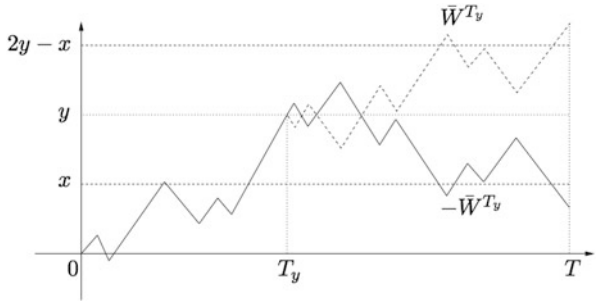


Fig. 3 Brownian motion $(W_{T_y+t} = \bar{W}_t^{T_y} + y : t \in \mathbb{R}^+)$ starting from y and its symmetric path

applying at the last equality but one the time-shift invariance with deterministic shift u_n. For the general case for U, we apply the result to the discrete stopping time $U_n = \frac{[nU]+1}{n}$, and then pass to the limit using the continuity of W. □

1.4 Maximum, Behavior at Infinity, Path Regularity

We apply the strong Markov property to identify the law of the Brownian motion maximum.

Proposition 5 (Symmetry Principle). *For any $y \geq 0$ and any $x \leq y$, we have*

$$\mathbb{P}[\sup_{t\leq T} W_t \geq y; W_T \leq x] = \mathbb{P}[W_T \geq 2y - x], \tag{6}$$

$$\mathbb{P}[\sup_{t\leq T} W_t \geq y] = \mathbb{P}[|W_T| \geq y] = 2\int_{\frac{y}{\sqrt{T}}}^{+\infty} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \mathrm{d}x. \tag{7}$$

Proof. Denote by $T_y = \inf\{t > 0 : W_t \geq y\}$ and $+\infty$ if the set is empty. Observe that T_y is a stopping time and that $\{\sup_{t\leq T} W_t \geq y; W_T \leq x\} = \{T_y \leq T; W_T \leq x\}$. By Proposition 4, on $\{T_y \leq T\}$, $(W_{T_y+t} = \bar{W}_t^{T_y} + y : t \in \mathbb{R}^+)$ is a Brownian motion starting from y, independent of $(W_s : s \leq T_y)$. By symmetry (see Fig. 3), the events $\{T_y \leq T, W_T < x\}$ and $\{T_y \leq T, W_T > 2y-x\}$ has the same probability. But for $x \leq y$, we have $\{T_y \leq T, W_T > 2y - x\} = \{W_T > 2y - x\}$ and the first result is proved.

For the second result, take $y = x$ and write $\mathbb{P}[\sup_{t\leq T} W_t \geq y] = \mathbb{P}[\sup_{t\leq T} W_t \geq y, W_T > y] + \mathbb{P}[\sup_{t\leq T} W_t \geq y, W_T \leq y] = \mathbb{P}[W_T > y] + \mathbb{P}[W_T \geq y] = 2\mathbb{P}(W_T \geq y) = \mathbb{P}(|W_T| \geq y)$. □

As a consequence of the identification of the law of the maximum up to a fixed time, we prove that the range of Brownian motion becomes $\mathbb{R}$ at time goes to infinity.

Proposition 6. *With probability 1, we have*

$$\limsup_{t\to+\infty} W_t = +\infty, \qquad \liminf_{t\to+\infty} W_t = -\infty.$$

Proof. For $T \geq 0$, set $M_T = \sup_{t\leq T} W_t$. As $T \uparrow +\infty$, it defines a sequence of increasing r.v., thus converging *a.s.* to a limit r.v. M_∞. Applying twice the monotone convergence theorem, we obtain

$$\begin{aligned}\mathbb{P}[M_\infty = +\infty] = \lim_{y\uparrow+\infty} \mathbb{P}[M_\infty > y] &= \lim_{y\uparrow+\infty} \Big(\lim_{T\uparrow+\infty} \mathbb{P}[M_T > y]\Big) \\ &= \lim_{y\uparrow+\infty} \Big(\lim_{T\uparrow+\infty} \mathbb{P}[|W_T| \geq y]\Big) = 1\end{aligned}$$

using (7). This proves that $\limsup_{t\to+\infty} W_t = +\infty$ *a.s.* and a symmetry argument gives the liminf. □

However, the increasing rate of W is sublinear as time goes to infinity.

Proposition 7. *With probability 1, we have*

$$\lim_{t\to+\infty} \frac{W_t}{t} = 0.$$

Proof. The strong law of large numbers yields that $\frac{W_n}{n} = \frac{1}{n}\sum_{i=1}^n (W_i - W_{i-1})$ converges *a.s.* to $\mathbb{E}(W_1) = 0$. The announced result is thus proved along the sequence of integers. To fill the gaps between integers, set $\tilde{M}_n = \sup_{n<t\leq n+1}(W_t - W_n)$ and $\tilde{M}'_n = \sup_{n<t\leq n+1}(W_n - W_t)$: due to Proposition 5, $\tilde{M}_n$ and $\tilde{M}'_n$ have the same distribution as $|W_1|$. Then, the Chebyshev inequality writes

$$\mathbb{P}(|\tilde{M}_n| + |\tilde{M}'_n| \geq n^{3/4}) \leq 2\frac{\mathbb{E}(|\tilde{M}_n|^2) + \mathbb{E}(|\tilde{M}'_n|^2)}{n^{3/2}} = 4n^{-3/2},$$

implying that $\sum_{n\geq 0} \mathbb{P}(|\tilde{M}_n| + |\tilde{M}_n| \geq n^{3/4}) < +\infty$. Thus, by Borel–Cantelli's lemma, we obtain that with probability 1, for n large enough $|\tilde{M}_n| + |\tilde{M}'_n| < n^{3/4}$, i.e. $\frac{\tilde{M}_n}{n}$ and $\frac{\tilde{M}'_n}{n}$ both converge *a.s.* to 0. □

By time inversion, $\hat{W}_t = tW_{1/t}$ is another Brownian motion: the $\hat{W}$-growth in infinite time gives an estimate on W at 0, which writes

$$+\infty = \limsup_{t\to+\infty} |\hat{W}_t| = \limsup_{s\to 0^+} \frac{|W_s - W_0|}{s}$$

which shows that W is not differentiable at time 0. By time-shift invariance, this is also true at any given time t. The careful reader may notice that the set of full probability measure depends on t and it is unclear at this stage if a single full set is available for any t, i.e. if

$$\mathbb{P}(\exists t_0 \text{ such that } t \mapsto W_t \text{ is differentiable at } t_0) = 0.$$

Actually, the above result holds true and it is due to Paley–Wiener–Zygmund (1933). The following result is of comparable nature: we claim that *a.s.* there does not exist any interval on which W is monotone.

Proposition 8 (Nowhere Monotonicity). *We have*

$$\mathbb{P}(t \mapsto W_t \text{is monotone on an interval}) = 0.$$

Proof. Define $M^{\uparrow}_{s,t} = \{\omega : u \mapsto W_u(\omega) \text{ is increasing on the interval}]s,t[\}$ and $M^{\downarrow}_{s,t}$ similarly. Observe that

$$M = \{t \mapsto W_t \text{ is monotone on the interval}\} = \bigcup_{s,t\in\mathbb{Q},0\le s<t} (M^{\uparrow}_{s,t} \cup M^{\downarrow}_{s,t}),$$

and since this is a countable union, it is enough to show $\mathbb{P}(M^{\uparrow}_{s,t}) = \mathbb{P}(M^{\downarrow}_{s,t}) = 0$ to conclude $\mathbb{P}(M) \le \sum_{s,t\in\mathbb{Q},0\le s<t}[\mathbb{P}(M^{\uparrow}_{s,t}) + \mathbb{P}(M^{\downarrow}_{s,t})] = 0$. For fixed n, set $t_i = s + i(t-s)/n$, then

$$\mathbb{P}(M^{\uparrow}_{s,t}) \le \mathbb{P}(W_{t_{i+1}} - W_{t_i} \ge 0, 0 \le i < n) = \prod_{i=0}^{n-1} \mathbb{P}(W_{t_{i+1}} - W_{t_i} \ge 0) = \frac{1}{2^n},$$

leveraging the symmetric distribution of the increments. Taking now n large gives $\mathbb{P}(M^{\uparrow}_{s,t}) = 0$. We argue similarly for $\mathbb{P}(M^{\downarrow}_{s,t}) = 0$. □

In view of this lack of smoothness, it seems impossible to define differential calculus along the paths of Brownian motion. However, as it will be further developed, Brownian motion paths enjoy a nice property of finite quadratic variations, which serves to build an appropriate stochastic calculus.

There are much more to tell about the properties of Brownian motion. We mention few extra properties without proof:

- HOLDER REGULARITY: for any $\rho \in (0, \frac{1}{2})$ and any deterministic $T > 0$, there exists a *a.s.* finite r.v. $C_{\rho,T}$ such that

$$\forall\, 0 \le s,t \le T, \quad |W_t - W_s| \le C_{\rho,T}|t-s|^{\rho}.$$

- LAW OF ITERATED LOGARITHM: setting $h(t) = \sqrt{2t \log\log t^{-1}}$, we have

$$\limsup_{t\downarrow 0} \frac{W_t}{h(t)} = 1 \quad a.s. \quad \text{and} \quad \liminf_{t\downarrow 0} \frac{W_t}{h(t)} = -1 \quad a.s.$$

- ZEROS OF BROWNIAN MOTION: the set $\chi = \{t \geq 0 : W_t = 0\}$ of the zeros of W is closed, unbounded, with null Lebesgue measure and it has no isolated points.

1.5 The Random Walk Approximation

Another algorithmic way to build a Brownian motion consists in rescaling a random walk. This is very simple and very useful for numerics: it leads to the so-called tree methods and it has some connections with finite differences in PDEs.

Consider a sequence $(X_i)_i$ of independent random variables with Rademacher distribution: $\mathbb{P}(X_i = \pm 1) = \frac{1}{2}$. Then

$$S_n = \sum_{i=1}^{n} X_i$$

defines a random walk on $\mathbb{Z}$. Like Brownian motion, it is a process with stationary independent increments, but it is not Gaussian. Actually S_n has a binomial distribution:

$$\mathbb{P}(S_n = -n + 2k) = \mathbb{P}(k \text{ rises}) = 2^{-n}\binom{n}{k}.$$

A direct computation shows that $\mathbb{E}(S_n) = 0$ and $\mathbb{V}\mathrm{ar}(S_n) = n$. When we rescale the walk and we let n go towards infinity, we observe however that due the Central Limit Theorem, the distribution of $\frac{S_n}{\sqrt{n}}$ converges to the Gaussian law with zero mean and unit variance. The fact that it is equal to the law of W_1 is not a coincidence, since it can be justified that the full trajectory of the suitably rescaled random walk converges towards that of a Brownian motion, see Fig. 4. This result is known as *Donsker theorem*, see for instance [12] for a proof.

Proposition 9. *Define $(Y_t^n)_t$ as the piecewise constant process*

$$Y_t^n = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} X_i. \tag{8}$$

The distribution of the process $(Y_t^n)_t$ converges to that of a Brownian motion $(W_t)_t$ as $n \to +\infty$, i.e. for any continuous functional

$$\lim_{n\to\infty} \mathbb{E}(\Phi(Y_t^n : t \leq 1)) = \mathbb{E}(\Phi(W_t : t \leq 1)).$$

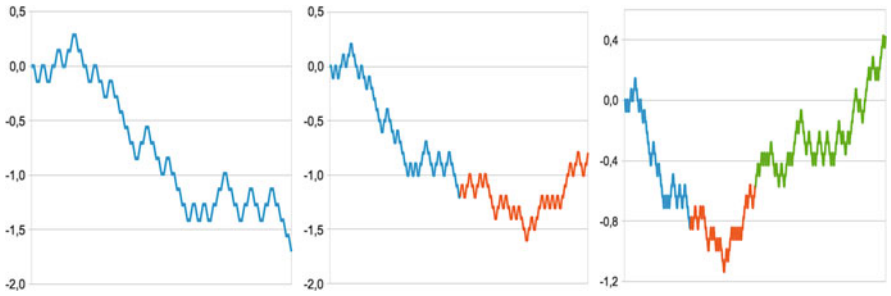


Fig. 4 The random walk rescaled in time and space. *From left to right*: the process Y^n for $n = 50,100,200$. The pieces of path with same color are built with the same X_i

The last result gives a simple way to evaluate numerically expectations of functionals of Brownian motion. It is the principle of the so-called *binomial tree methods*.

Link with Finite Difference Scheme. The random walk can be interpreted as a explicit FD scheme for the heat equation. We anticipate a bit on the following where the connection between Brownian motion and heat equation will be more detailed.

For $t = \frac{i}{n}$ ($i \in \{0, \dots, n\}$) and $x \in \mathbb{R}$, set

$$u^n(t,x) = \mathbb{E}\Big(f\big(x + Y^n_{\frac{i}{n}}\big)\Big).$$

The independence of $(X_i)_i$ gives

$$\begin{aligned} u^n\big(\frac{i}{n}, x\big) &= \mathbb{E}\big(f(x + Y^n_{\frac{i-1}{n}} + \frac{X_i}{\sqrt{n}})\big) \\ &= \frac{1}{2}u^n\big(\frac{i-1}{n}, x + \frac{1}{\sqrt{n}}\big) + \frac{1}{2}u^n\big(\frac{i-1}{n}, x - \frac{1}{\sqrt{n}}\big), \\ \frac{u^n\big(\frac{i}{n}, x\big) - u^n\big(\frac{i-1}{n}, x\big)}{\frac{1}{n}} &= \frac{1}{2}\frac{u^n\big(\frac{i-1}{n}, x + \frac{1}{\sqrt{n}}\big) - 2u^n\big(\frac{i-1}{n}, x\big) + u^n\big(\frac{i-1}{n}, x - \frac{1}{\sqrt{n}}\big)}{\big(\frac{1}{\sqrt{n}}\big)^2}. \end{aligned}$$

Thus, u^n related to the expectation of the random walk can be read as an explicit FD scheme of the heat equation $\partial_t u(t,x) = \frac{1}{2}\partial^2_{xx} u(t,x)$ and $u(0,x) = f(x)$, with time step $\frac{1}{n}$ and space step $\frac{1}{\sqrt{n}}$.

1.6 Other Stochastic Processes

We present other one-dimensional processes, with continuous trajectories, which derive from the Brownian motion.

1. Geometric Brownian motion: this model is popular in finance to model stocks and other assets by a positive process.
2. Ornstein–Uhlenbeck process: it has important applications in physics, mechanics, economy and finance to model stochastic phenomena exhibiting mean-reverting features (like spring endowed with random forces, interest-rates or inflation, ...).
3. Stochastic differential equations: it gives the more general framework.

1.6.1 Geometric Brownian Motion

Definition 4. A Geometric Brownian Motion (GBM in short) with deterministic initial value $S_0 > 0$, drift coefficient μ and diffusion coefficient σ, is a process $(S_t)_{t\geq 0}$ defined by

$$S_t = S_0 e^{(\mu-\frac{1}{2}\sigma^2)t+\sigma W_t}, \tag{9}$$

where $\{W_t; t \geq 0\}$ is a standard Brownian motion.

As the argument in the exponential has a Gaussian distribution, the random variable S_t (with t fixed) is known as Lognormal.

This is a process with continuous trajectories, which takes strictly positive values. The Geometric Brownian motion is often used as a model of asset price (see Samuelson [65]): this choice is justified on the one hand, by the positivity of S and on the other hand, by the simple Gaussian properties of its returns:

- The returns $\log(S_t) - \log(S_s)$ are Gaussian with mean $(\mu - \frac{1}{2}\sigma^2)(t-s)$ and variance $\sigma^2(t-s)$.
- For all $0 < t_1 < t_2 \ldots .. < t_n$, the relative increments $\{\frac{S_{t_i+1}}{S_{t_i}}; 0 \leq i \leq n-1\}$ are independent.

The assumption of Gaussian returns is not valid in practice but this model still serves as a proxy for more sophisticated models.

Naming μ the drift parameter may be surprising at first sight since it appears in the deterministic component as $(\mu - \frac{1}{2}\sigma^2)t$. Actually, a computation of expectation gives easily

$$\mathbb{E}(S_t) = S_0 e^{(\mu-\frac{1}{2}\sigma^2)t}\mathbb{E}(e^{\sigma W_t}) = S_0 e^{(\mu-\frac{1}{2}\sigma^2)t} e^{\frac{1}{2}\sigma^2 t} = S_0 e^{\mu t}.$$

The above equality gives the interpretation to μ as a mean drift term: $\mu = \frac{1}{t}\log[\mathbb{E}(S_t)/S_0]$.

1.6.2 Ornstein–Uhlenbeck Process

Let us return to physics and to the Brownian motion by Einstein in 1905. In order to propose a more adequate modeling of the phenomenon of particles diffusion, we introduce the process of Ornstein–Uhlenbeck and its principal properties.

So far we have built the Brownian motion like a model for a microscopic particle in suspension in a liquid subjected to thermal agitation. An important criticism made with this modeling concerns the assumption that displacement increments are independent and they do not take into account the effects of the particle speed due to particle inertia.

Let us denote by m the particle mass and by $\dot{X}(t)$ its speed. Owing to Newton's second law, the momentum change $m\dot{X}(t+\delta(t)) - m\dot{X}(t)$ is equal to the resistance $-k\dot{X}(t)\delta t$ of the medium during time δt, plus the momentum change due to molecular shocks, that we assume to be with stationary independent increments and thus associated with a Brownian motion. The process thus modeled is called sometimes the *physical Brownian motion*. The equation for the increments becomes

$$m\,\delta[\dot{X}(t)] = -k\dot{X}(t)\delta t + m\sigma\delta W_t.$$

Trajectories of the Brownian motion being not differentiable, the equation has to be read in an integral form

$$m\dot{X}(t) = m\dot{X}(0) - \int_0^t k\dot{X}(s)\mathrm{d}s + m\sigma W_t.$$

$\dot{X}(t)$ is thus solution of the linear stochastic differential equation (known as Langevin equation)

$$V_t = v_0 - a\int_0^t V_s\mathrm{d}s + \sigma W_t$$

where $a = \frac{k}{m}$. If $a = 0$, we recover an arithmetic Brownian motion and to avoid this trivial reduction, we assume $a \neq 0$ in the sequel. However, the existence of solution is not clear since W is not differentiable. To overcome this difficulty, set $Z_t = V_t - \sigma W_t$: that leads to the new equation

$$Z_t = v_0 - a\int_0^t (Z_s + \sigma W_s)\mathrm{d}s,$$

which is now a linear ordinary differential equation that can be solved path by path. The variation of parameter method gives the representation of the unique solution of this equation like

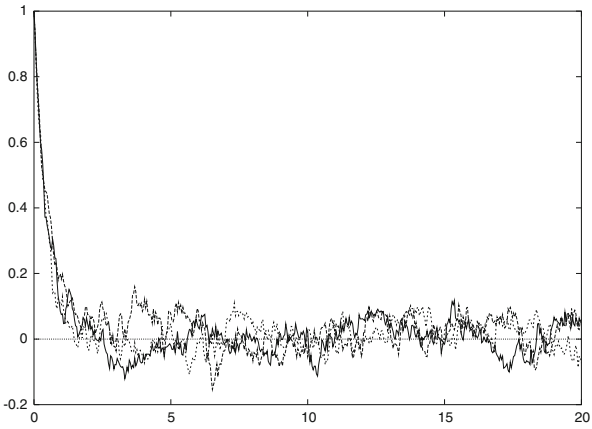


Fig. 5 Ornstein–Uhlenbeck paths with $V_0 = 1$, $a = 2$ and $\sigma = 0.1$

$$Z_t = v_0 e^{-at} - \sigma \int_0^t a e^{-a(t-s)} W_s \mathrm{d}s.$$

The initial solution is thus

$$V_t = v_0 e^{-at} + \sigma W_t - \sigma \int_0^t a e^{-a(t-s)} W_s \mathrm{d}s. \tag{10}$$

Using stochastic calculus, we will derive later (see Sect. 3.3) another convenient representation of V as follows:

$$V_t = v_0 e^{-at} + \sigma \int_0^t e^{-a(t-s)} \mathrm{d}W_s \tag{11}$$

using a stochastic integral not yet defined. From (10), assuming that v_0 is deterministic, we can show the following properties (see also Sect. 3.3).

- For a given t, V_t has a Gaussian distribution: indeed, as the limit of a Riemann sum, it is the *a.s.* limit of a Gaussian r.v., see footnote 5 page 111.
- More generally, V is a Gaussian process.
- Its mean is $v_0 e^{-a\,t}$, its covariance function $\mathbb{C}\mathrm{ov}(V_t, V_s) = e^{-a(t-s)} \frac{\sigma^2}{2a}(1 - e^{-2as})$ for $t > s$.

Observe that for $a > 0$, the Gaussian distribution of V_t converges to $\mathcal{N}(0, \frac{\sigma^2}{2a})$ as $t \to +\infty$: it does not depend anymore on v_0 and illustrates the mean-reverting feature of this model, see Fig. 5.

1.6.3 Stochastic Differential Equations and Euler Approximations

The previous example gives the generic form of a Stochastic Differential Equation, that generalizes the usual Ordinary Differential Equations $x_t' = b(x_t)$ or in integral form $x_t = x_0 + \int_0^t b(x_s)\mathrm{d}s$.

Definition 5. Let $b, \sigma : x \in \mathbb{R} \mapsto \mathbb{R}$ be two functions, respectively the drift and the diffusion coefficient. A Stochastic Differential Equation (SDE in short) with parameter (b, σ) and initial value x is a stochastic process $(X_t)_{t\geq 0}$ solution of

$$X_t = x + \int_0^t b(X_s)\mathrm{d}s + \int_0^t \sigma(X_s)\mathrm{d}W_s, \quad t \geq 0,$$

where $(W_t)_t$ is a standard Brownian motion.

A slightly more general definition (not considered here) could include the case of time-dependent coefficients $b(t, x)$ and $\sigma(t, x)$, the subsequent analysis would be quite similar. In the definition above, we use a stochastic integral $\int_0^t \ldots \mathrm{d}W_s$ which has not yet been defined: it will be explained in the next section. For the moment, the reader needs to know that in the simplest case where σ is constant, we simply have $\int_0^t \sigma(X_s)\mathrm{d}W_s = \sigma W_t$. The previous examples fit this setting:

- The arithmetic Brownian motion corresponds to $b(x) = b$ et $\sigma(x) = \sigma$.
- The Ornstein–Uhlenbeck process corresponds to $b(x) = -ax$ et $\sigma(x) = \sigma$.

Taking σ to be non constant allows for more general situations and more flexible models. Instead of discussing now the important issues of existence and uniqueness to such SDE, we rather consider natural approximations of them, namely the Euler scheme (which is the direct extension of Euler scheme for ODEs).

Definition 6. Let (b, σ) be given drift and diffusion coefficients. The Euler scheme associated to the SDE with coefficients (b, σ), initial value x and time step h, is defined by

$$\begin{cases} X_0^h = x, \\ X_t^h = X_{ih}^h + b(X_{ih}^h)(t - ih) + \sigma(X_{ih}^h)(W_t - W_{ih}), \quad i \geq 0, t \in (ih, (i+1)h]. \end{cases} \tag{12}$$

In other words, X^h is a piecewise arithmetic Brownian motion with coefficients on the interval $(ih, (i+1)h]$ computed according to the functions (b, σ) evaluated at X_{ih}^h. In general, the law of X_t^h is not known analytically: at most, we can give explicit representations using an induction of the time-step. On the other hand, as it will be seen further, the random simulation of X^h at time $(ih)_{i\geq 0}$ is easily performed by simulating the independent Brownian increments $(W_{(i+1)h} - W_{ih})$. The accuracy of the approximation of X by X^h is expected to get improved as h goes to 0.

Complementary References. See [48, 57, 63].

2 Feynman–Kac Representations of PDE Solutions

Our purpose in this section is to make the connection between the expectations of functionals of Brownian motion and the solution of second order linear parabolic partial differential equations (PDE in short): this leads to the well-known Feynman–Kac representations. We extend this point of view to other simple processes introduced before.

2.1 The Heat Equations

2.1.1 Heat Equation in the Whole Space

Let us return to the law of $x + W_t$, the Gaussian density of which is

$$g(t, x, y) := g(t, y - x) = \frac{1}{\sqrt{2\pi t}} \exp(-(y - x)^2/2t),$$

often called in this context the fundamental solution of the heat equation. One of the key properties is *the property of convolution*

$$g(t + s, x, y) = \int_{\mathbb{R}} g(t, x, z) g(s, z, y) \mathrm{d}z \tag{13}$$

which says in an analytical language that $x + W_{t+s}$ is the sum of the independent Gaussian variables $x + W_t$ and $W_{t+s} - W_t$. A direct calculation on the density shows that the Gaussian density is solution to the *heat equation* w.r.t. the two variables x and y

$$\begin{cases} g'_t(t, x, y) = \frac{1}{2} g''_{yy}(t, x, y), \\ g'_t(t, x, y) = \frac{1}{2} g''_{xx}(t, x, y). \end{cases} \tag{14}$$

This property is extended to a large class of functions built from the Brownian motion.

Theorem 2 (Heat Equation with Cauchy Initial Boundary Condition). *Let f be a bounded*[6] *measurable function. Consider the function*

$$u(t, x, f) = \mathbb{E}[f(x + W_t)] = \int_{\mathbb{R}} g(t, x, y) f(y) \mathrm{d}y :$$

[6]This growth condition can be relaxed into $|f(x)| \leq C \exp\left(\frac{|x|^2}{2\alpha}\right)$ for any x, for some positive constants C and α: in that case, the smoothness of the function u is satisfied for $t < \alpha$ only.

the function u is infinitely continuously differentiable in space and time for $t > 0$ and solves the heat equation

$$u'_t(t,x,f) = \frac{1}{2}u''_{xx}(t,x,f), \quad u(0,x,f) = f(x). \tag{15}$$

Equation (15) is the *heat equation with initial boundary condition* (Cauchy problem, see [22]).

Proof. Standard Gaussian estimates allow to differentiate u w.r.t. t or x by differentiating under the integral sign: then, we have

$$u'_t(t,x,f) = \int_{\mathbb{R}} g'_t(t,x,y)f(y)dy = \int_{\mathbb{R}} \frac{1}{2}g''_{xx}(t,x,y)f(y)dy = \frac{1}{2}u''_{xx}(t,x,f).$$

□

When the function considered is regular, another formulation can be given to this relation, which will play a significant role in the following.

Proposition 10. *If f is of class $\mathscr{C}^2_b$ (bounded and twice continuously differentiable with bounded derivatives),*[7] *we have*

$$u'_t(t,x,f) = u(t,x,\frac{1}{2}f''_{xx}),$$

or equivalently using a probabilistic viewpoint

$$\mathbb{E}[f(x+W_t)] = f(x) + \int_0^t \mathbb{E}\big[\frac{1}{2}f''_{xx}(x+W_s)\big]\mathrm{d}s. \tag{16}$$

Proof. Write $u(t,x,f) = \mathbb{E}[f(x+W_t)] = \int_{\mathbb{R}} g(t,0,y)f(x+y)\mathrm{d}y = \int_{\mathbb{R}} g(t,x,z)f(z)\mathrm{d}z$ and differentiate under the integral sign: it gives

$$u''_{xx}(t,x,f) = \int_{\mathbb{R}} g(t,0,y)f''_{xx}(x+y)\mathrm{d}y = u(t,x,f''_{xx}) = \int_{\mathbb{R}} g''_{xx}(t,x,z)f(z)\mathrm{d}z,$$

$$u'_t(t,x,f) = \int_{\mathbb{R}} g'_t(t,x,z)f(z)\mathrm{d}z = \frac{1}{2}\int_{\mathbb{R}} g''_{xx}(t,x,z)f(z)\mathrm{d}z = \frac{1}{2}u(t,x,f''_{xx}),$$

using at the first line two integration by parts and at the second line the heat equation satisfied by g. Then the probabilistic representation (16) easily follows:

[7] Here again, the boundedness could be relaxed to some exponential growth.

$$\mathbb{E}[f(x+W_t)] - f(x) = u(t,x,f) - u(0,x,f) = \int_0^t u_t'(s,x,f)\mathrm{d}s$$
$$= \int_0^t u(s,x,\frac{1}{2}f_{xx}'')\mathrm{d}s = \int_0^t \mathbb{E}\big[\frac{1}{2}f_{xx}''(x+W_s)\big]\mathrm{d}s.$$

□

2.1.2 Heat Equation in an Interval

We now extend the previous results in two directions: first, we allow the function f to also depend smoothly on time and second, the final time t is replaced by a stopping time U. The first extension is straightforward and we state it without proof.

Proposition 11. *Let f be a function of class $\mathscr{C}_b^{1,2}$ (bounded, once continuously differentiable in time, twice in space, with bounded derivatives): we have*

$$\mathbb{E}[f(t,x+W_t)] = f(0,x) + \int_0^t \mathbb{E}[f_t'(s,x+W_s) + \frac{1}{2}f_{xx}''(s,x+W_s)]\mathrm{d}s$$
$$= f(0,x) + \mathbb{E}\left[\int_0^t (f_t'(s,x+W_s) + \frac{1}{2}f_{xx}''(s,x+W_s))\mathrm{d}s\right]. \tag{17}$$

The second equality readily follows from Fubini's theorem to invert $\mathbb{E}$ and time integral: this second form is more suitable for an extension to stochastic times t.

Theorem 3. *Let f be a function of class $\mathscr{C}_b^{1,2}$, we have*

$$\mathbb{E}[f(U,x+W_U)] = f(0,x) + \mathbb{E}\left[\int_0^U (f_t'(s,x+W_s) + \frac{1}{2}f_{xx}''(s,x+W_s))\mathrm{d}s\right] \tag{18}$$

for any bounded[8] *stopping time U.*

The above identity between expectations is far to be obvious to establish by hand since the law of U is quite general and an analytical computation is out of reach. This level of generality on U is quite interesting for applications: it provides a powerful tool to determine the distribution of hitting times, to show how often multidimensional Brownian motion visits a given point or a given set. Regarding

[8] Meaning that for a deterministic positive constant C, $\mathbb{P}(U \le C) = 1$.

this lecture, it gives a key tool to derive probabilistic representations of heat equation with Dirichlet boundary conditions.

Proof. Let us start by giving alternatives of the relation (17). We observe that it could have been written with a random initial condition X_0, like for instance

$$\begin{aligned}&\mathbb{E}[\mathbf{1}_{A_0} f(t, X_0 + W_t)]\\ &= \mathbb{E}\left[\mathbf{1}_{A_0} f(0, X_0) + \mathbf{1}_{A_0} \int_0^t (f_t'(s, X_0 + W_s) + \frac{1}{2} f_{xx}''(s, X_0 + W_s))\mathrm{d}s\right],\end{aligned}$$

with W independent of X_0 and where the event A_0 depends on X_0. Similarly, using the time-shifted Brownian motion $\{\bar{W}_t^u = W_{t+u} - W_u; t \in \mathbb{R}^+\}$ that is independent of the initial condition $x + W_u$ (Proposition 2), it leads to

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{A_u} f(t + u, x + W_u + \bar{W}_t^u)] = \mathbb{E}\Big[&\mathbf{1}_{A_u} f(u, x + W_u) +\\ &\mathbf{1}_{A_u} \int_0^t (f_t'(u + s, x + W_u + \bar{W}_s^u) + \frac{1}{2} f_{xx}''(u + s, x + W_u + \bar{W}_s^u))\mathrm{d}s\Big]\end{aligned}$$

for any event A_u depending only of the values $\{W_s : s \leq u\}$, or equivalently

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{A_u} f(t + u, x + W_{t+u})] = \mathbb{E}\Big[&\mathbf{1}_{A_u} f(u, x + W_u)\\ &+\mathbf{1}_{A_u} \int_u^{t+u} (f_t'(s, x + W_s) + \frac{1}{2} f_{xx}''(s, x + W_s))\mathrm{d}s\Big].\end{aligned}$$

Set $M_t = f(t, x + W_t) - f(0, x) - \int_0^t (f_t'(s, x + W_s) + \frac{1}{2} f_{xx}''(s, x + W_s))\mathrm{d}s$: our aim is to prove $\mathbb{E}(M_U) = 0$. Observe that the preliminary computation has shown that

$$\mathbb{E}(\mathbf{1}_{A_u}(M_{t+u} - M_u)) = 0 \tag{19}$$

for $t \geq 0$. In particular, taking $A_u = \Omega$ we obtain that the expectation $\mathbb{E}(M_t)$ is constant[9] w.r.t. t.

Now, consider first that U is a discrete stopping time valued in $\{0 = u_0 < u_1 < \cdots < u_n = T\}$: then

$$\mathbb{E}(M_U) = \sum_{k=0}^{n-1} \mathbb{E}(M_{U \wedge u_{k+1}} - M_{U \wedge u_k}) = \sum_{k=0}^{n-1} \mathbb{E}(\mathbf{1}_{U > u_k}(M_{u_{k+1}} - M_{u_k})) = 0$$

[9] Actually, (19) proves that M is a martingale and the result to be proved is related to the *optional sampling theorem*.

by applying (19) since $\{U \le u_k\}$ does depend only of $\{W_s : s \le u_k\}$ (by definition of a stopping time). □

Second, for a general stopping time (bounded by T), we take $U_n = \frac{[nU]+1}{n}$ which is a stopping time converging to U: since $(M_t)_{0\le t\le T}$ is bounded and continuous, the dominated convergence theorem gives $0 = \mathbb{E}(M_{U_n}) \underset{n\to\infty}{\longrightarrow} \mathbb{E}(M_U)$. □

As a consequence, we now make explicit the solutions of the heat equation in an interval and with initial condition: it is a partial generalization[10] of Theorem 2, which characterized them in the whole space. The introduction of (non-homogeneous) boundary conditions of Dirichlet type is connected to the passage time of the Brownian motion.

Corollary 1 (Heat Equation with Cauchy–Dirichlet Boundary Condition). *Consider the PDE*

$$\begin{cases} u'_t(t,x) = \frac{1}{2}u''_{xx}(t,x), & \text{for } t > 0 \text{ and } x \in]a,b[, \\ u(0,x) = f(0,x) & \text{for } t = 0 \text{ and } x \in [a,b], \\ u(t,x) = f(t,x) & \text{for } x = a \text{ or } b, \text{ with } t \ge 0. \end{cases}$$

If a solution u of class $C_b^{1,2}([0,T]\times[a,b])$ exists, then it is given by

$$u(t,x) = \mathbb{E}[f(t-U, x+W_U)]$$

where $U = T_a \wedge T_b \wedge t$ (using the previous notation for the first passage time T_y at the level y for the Brownian motion starting at x, i.e. $(x+W_t)_{t\ge 0}$).

Proof. First, extend smoothly the function u outside the interval $[a,b]$ in order to apply previous results. The way to extend is unimportant since u and its derivatives are only evaluated inside $[a,b]$. Clearly U is a bounded (by t) stopping time. Apply now the equality (18) to the function $(s,y) \mapsto u(t-s,y) = v(s,y)$ of class $C_b^{1,2}([0,t]\times\mathbb{R})$, satisfying $v'_s(s,y) + \frac{1}{2}v''_{yy}(s,y) = 0$ for $(s,y) \in [0,t]\times[a,b]$. We obtain

$$\mathbb{E}[v(U, x+W_U)] = v(0,x) + \mathbb{E}\left[\int_0^U (v'_s(s, x+W_s) + \frac{1}{2}v''_{yy}(s, x+W_s))\mathrm{d}s\right] = v(0,x),$$

since for $s \le U$, $(s, x+W_s) \in [0,t]\times[a,b]$. To conclude, we easily check that $v(0,x) = u(t,x)$ and $v(U, x+W_U) = f(t-U, x+W_U)$. □

[10] Indeed, the result gives uniqueness and not the existence.

2.1.3 A Probabilistic Algorithm to Solve the Heat Equation

To illustrate our purpose, we consider a toy example regarding the numerical evaluation of $u(t,x) = \mathbb{E}(f(x+W_t))$ using random simulations, in order to discuss main ideas underlying to Monte Carlo methods. Actually, the arguments below apply also to $u(t,x) = \mathbb{E}[f(t-U, x+W_U)]$ with $U = T_a \wedge T_b \wedge t$, although there are some extra significant issues in the simulation of (U, W_U).

For the notational simplicity, denote by X the random variable inside the expectation to compute, that is $X = f(x+W_t)$ in our toy example. As a difference with a PDE method (based on finite differences or finite elements), a standard Monte Carlo method provides an approximation of $u(t,x)$ at a given point (t,x), without evaluating the values at other points. Actually, this fact holds because the PDE u is linear; in Sect. 5 related to non-linear PDEs, the situation is different.

The Monte Carlo method is able to provide a convergent, tractable approximation of $u(t,x)$, with a priori error bounds, under two conditions.

1. An arbitrary large number of independent realizations of X can be generated (denote them by $(X_i)_{i\geq 1}$): in our toy example, this is straightforward since it requires only the simulation of W_t which is distributed as a Gaussian r.v. $\mathcal{N}(0,t)$ and then we have to compute $X = f(x+W_t)$. The independence of simulations is achieved by using a *good* generator of random numbers, like the excellent *Mersenne Twister*[11] generator.
2. Additionally, X which is already integrable ($\mathbb{E}|X| < +\infty$) is assumed to be square integrable: $\mathbb{V}\text{ar}(X) < +\infty$.

Then, by the law of large numbers, we have

$$\overline{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i \underset{M\to+\infty}{\longrightarrow} \mathbb{E}(X); \tag{20}$$

hence the empirical mean of simulations of X provides a convergent approximation of the expectation $\mathbb{E}(X)$. As a difference with PDE methods where some stability conditions may be required (like the Courant–Friedrichs–Lewy condition), the above Monte Carlo method does not require any extra condition to converge: it is *unconditionally convergent*. The extra moment condition is used to define a priori error bounds on the *statistical error*: the approximation error is controlled by means of the Central Limit Theorem

$$\lim_{M\to+\infty} \mathbb{P}\left(\sqrt{\frac{M}{\mathbb{V}\text{ar}(X)}}\left(\overline{X}_M - \mathbb{E}(X)\right) \in [a,b]\right) = \mathbb{P}(G \in [a,b]),$$

[11] http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html.

where G is a centered unit variance Gaussian r.v. Observe that the error bounds are stochastic: we can not do better than arguing that with probability $\mathbb{P}(G \in [a,b])$, the unknown expectation (asymptotically as $M \to +\infty$) belongs to the interval

$$\left[\overline{X}_M - b\sqrt{\frac{\mathbb{V}\mathrm{ar}(X)}{M}}, \overline{X}_M - a\sqrt{\frac{\mathbb{V}\mathrm{ar}(X)}{M}}\right].$$

This is known as a confidence interval at level $\mathbb{P}(G \in [a,b])$. The larger a and b, the larger the confidence interval, the higher the confidence probability.

To obtain a fully explicit confidence interval, one may replace $\mathbb{V}\mathrm{ar}(X)$ by its estimator using the same simulations:

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \approx \frac{M}{M-1}\left(\frac{1}{M}\sum_{i=1}^{M} X_i^2 - \overline{X}_M^2\right) := \sigma_M^2.$$

The factor $M/(M-1)$ plays the role of unbiasing[12] the value $\mathbb{V}\mathrm{ar}(X)$, although it is not a big deal for M large ($M \geq 100$). Anyway, we can prove that the above conditional intervals are asymptotically unchanged by taking the empirical variance σ_M^2 instead of $\mathbb{V}\mathrm{ar}(X)$. Gathering these different results and seeking a symmetric confidence interval $-a = b = 1.96$ and $\mathbb{P}(G \in [a,b]) \approx 95\,\%$, we obtain the following: with probability 95 %, approximatively for M large enough, we have

$$\mathbb{E}(X) \in \left[\overline{X}_M - 1.96\frac{\sigma_M}{\sqrt{M}}, \overline{X}_M + 1.96\frac{\sigma_M}{\sqrt{M}}\right]. \tag{21}$$

The symmetric confidence interval at level 99 % is given by $-a = b = 2.58$. Since a Monte Carlo method provides random evaluations of $\mathbb{E}(X)$, different program runs will give different results (as a difference with a deterministic method which systematically has the same output) which seems uncomfortable: that is why it is important to produce a confidence interval. This is also very powerful and useful to have at hand a numerical method able to state that the error is at most of `xxx` with high probability.

The confidence interval depends on

- The confidence level $\mathbb{P}(G \in [a,b])$, chosen by the user.
- The number of simulations: improving the accuracy by a factor 2 requires 4 times more simulations.
- The variance $\mathbb{V}\mathrm{ar}(X)$ or its estimator σ_M^2, which depends on the problem to handle (and not much on M as soon as M is large). This variance can be very different from one problem to another: on Fig. 6, see the width of confidence intervals for two similar computations. There exist variance reduction techniques

[12] Indeed, we can show that $\mathbb{E}(\sigma_M^2) = \mathbb{V}\mathrm{ar}(X)$.

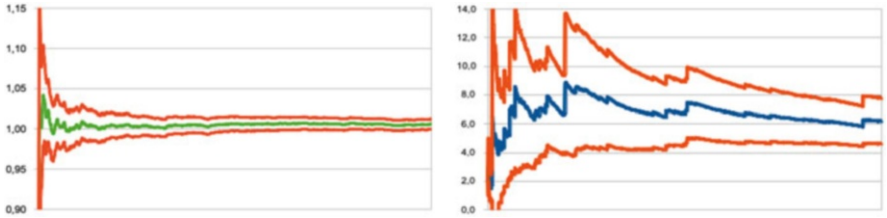


Fig. 6 Monte Carlo computations of $\mathbb{E}(e^{G/10}) = e^{\frac{1}{2}\frac{1}{10^2}} \approx 1.005$ on *the left* and $\mathbb{E}(e^{2G}) = e^{\frac{1}{2}2^2} \approx 7.389$ on *the right*, where G is a Gaussian r.v. with zero mean and unit variance. The empirical mean and the symmetric 95 %-confidence intervals are plotted w.r.t. the number of simulations

able to significantly reduce this factor in order to provide thinner confidence intervals while maintaining the same computational cost.

Another advantage of such a Monte Carlo algorithm is the simplicity of code, consisting of one loop on the number of simulations; within this loop, the empirical variance should be simultaneously computed. However, the simulation procedure of X can be delicate in some situations, see Sect. 4.

At last, we focus our discussion on the impact of the dimension of the underlying PDE, which has been equal to 1 so far. Consider now a state variable in $\mathbb{R}^d$ ($d \geq 1$) and a heat equation with Cauchy initial condition in dimension d; (15) becomes

$$u'_t(t,x,f) = \frac{1}{2}\Delta u(t,x,f), \quad u(0,x,f) = f(x), \quad t > 0, x \in \mathbb{R}^d, \tag{22}$$

where $\Delta = \sum_{i=1}^d \partial^2_{x_i x_i}$ stands for the Laplacian in $\mathbb{R}^d$. Using similar arguments as in dimension 1, we check that

$$u(t,x,f) = \int_{\mathbb{R}^d} \frac{1}{(2\pi t)^{d/2}} \exp(-|y-x|^2/2t) f(y) \mathrm{d}y = \mathbb{E}[f(x+W_t)]$$

where $W = \begin{pmatrix} W_1 \\ \vdots \\ W_d \end{pmatrix}$ is a d-dimensional Brownian motion, i.e. each W_i is a one-dimensional Brownian motion and the d components are independent (Fig. 7).

- The Monte Carlo computation of $u(t,x)$ is then achieved using independent simulations of $X = f(x+W_t)$: the accuracy is then of order $1/\sqrt{N}$ and the computational effort is $N \times d$. Thus, the dimension has a very low effect on the complexity of the algorithm.

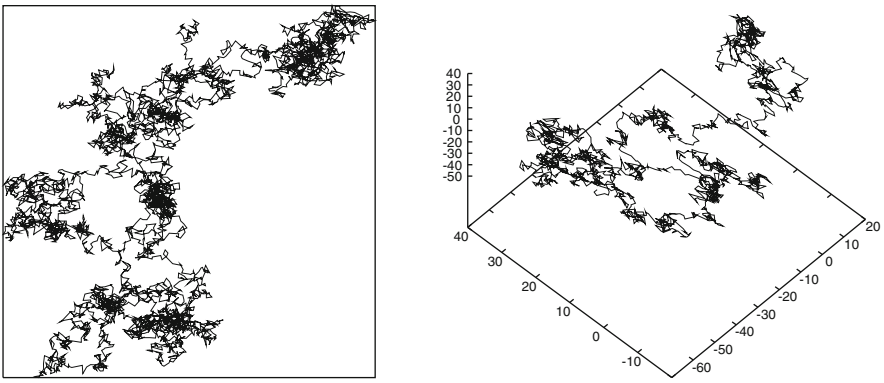


Fig. 7 Brownian motion in dimension 2 and 3

- As a comparison with a PDE discretization scheme, to achieve an accuracy of order $1/N$, we essentially[13] need N points in each spatial direction and it follows that the resulting linear system to invert is of size N^d: thus, without going into full details, it is clear that the computational cost to achieve a given accuracy depends much on the dimension d. And the situation becomes less and less favourable as the dimension increases. Also, the memory required to run a PDE algorithm increases exponentially with the dimension, as a difference with a Monte Carlo approach.

It is commonly admitted that a PDE approach is more suitable and efficient in dimension 1 and 2, whereas a Monte Carlo procedure is more adapted for higher dimensions. On the other hand, a PDE-based method computes a global approximation of u (at any point (t, x)), while a Monte Carlo scheme gives a pointwise approximation only. The probabilistic approach can be directly used for Parallel Computing, each processor being in charge of a bunch of simulations at a given point (t, x).

2.2 PDE Associated to Other Processes

We extend the Feynman–Kac representation for the Brownian motion to the Arithmetic Brownian Motion and the Ornstein–Uhlenbeck process.

[13] In fact, it generally depends on the regularity of u.

2.2.1 Arithmetic Brownian Motion

First consider the Arithmetic Brownian motion defined by $\{X_t^x = x+bt+\sigma W_t, t \geq 0\}$. The distribution of X_t is Gaussian with mean $x+bt$ and variance $\sigma^2 t$: we assume in the following that $\sigma \neq 0$ which ensures that its density exists and is given by

$$g_{b,\sigma^2}(t,x,y) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp -\frac{(y-x-bt)^2}{2\sigma^2 t} = g(\sigma^2 t, x+bt, y)$$
$$= g(\sigma^2 t, x, y-bt).$$

Denote by $L_{b,\sigma^2}^{\text{ABM}}$ the second order operator

$$L_{b,\sigma^2}^{\text{ABM}} = \frac{1}{2}\sigma^2 \partial_{xx}^2 + b\partial_x \,, \tag{23}$$

also called *infinitesimal generator*[14] of X. A direct computation using the heat equation for $g(t,x,y)$ gives

$$\partial_t g_{b,\sigma^2}(t,x,y) = \frac{1}{2}\sigma^2 g''_{xx}(\sigma^2 t, x+bt, y) + b g'_x(\sigma^2 t, x+bt, y) = L_{b,\sigma^2}^{\text{ABM}} g_{b,\sigma^2}(t,x,y).$$

Hence, multiplying by $f(y)$ and integrating over $y \in \mathbb{R}$, we obtain the following representation that generalizes Theorem 2.

Theorem 4. *Let f be a bounded measurable function. The function*

$$u_{b,\sigma^2}(t,x,f) = \mathbb{E}[f(X_t^x)] = \int_{\mathbb{R}} g_{b,\sigma^2}(t,x,y) f(y) \mathrm{d}y \tag{24}$$

solves

$$\begin{cases} u'_t(t,x,f) = L_{b,\sigma^2}^{ABM} u(t,x,f) = \frac{1}{2}\sigma^2 u''_{xx}(t,x,f) + b u'_x(t,x,f), \\ u(0,x,f) = f(x). \end{cases} \tag{25}$$

The extension of Propositions 10 and 11 follows the arguments used for the BM case.

Proposition 12. *If $f \in \mathscr{C}_b^{1,2}$ and U is a bounded stopping time (including deterministic time), then*

[14] This labeling comes from the infinitesimal decomposition of $\mathbb{E}(f(X_t))$ as time is small, $\partial_t \mathbb{E}(f(X_t))|_{t=0} = L_{b,\sigma^2}^{\text{ABM}} f(x)$, see Proposition 12.

$$\mathbb{E}[f(U, X_U^x)] = f(0,x) + \mathbb{E}\Big[\int_0^U [L_{b,\sigma^2}^{\text{ABM}} f(s, X_s^x) + f_t'(s, X_s^x)]\mathrm{d}s\Big].$$

Theorem 4 gives the Feynman–Kac representation of the Cauchy problem written w.r.t. the second order operator $L_{b,\sigma^2}^{\text{ABM}}$. When Dirichlet boundary conditions are added, Corollary 1 extends as follows, using Proposition 12.

Corollary 2. *Assume the existence of a solution u of class $C_b^{1,2}([0,T]\times[a,b])$ to the PDE*

$$\begin{cases} u_t'(t,x,f) = L_{b,\sigma^2}^{\text{ABM}} u(t,x,f), & \text{for } t>0 \text{ and } x \in]a,b[, \\ u(0,x,f) = f(0,x) & \text{for } t=0 \text{ and } x\in[a,b], \\ u(t,x,f) = f(t,x) & \text{for } x = a \text{ or } b, \text{ with } t\geq 0. \end{cases}$$

Then it is given by

$$u(t,x) = \mathbb{E}[f(t-U^x, X_{U^x}^x)]$$

where $U^x = \inf\{s>0 : X_s^x \notin]a,b[\} \wedge t$ is the first exit time from the interval $]a,b[$ by the process X^x before t.

As for the standard heat equation, this representation naturally leads to a probabilistic algorithm to compute the PDE solution, by empirical mean of independent simulation of $f(t-U^x, X_{U^x}^x)$.

2.2.2 Ornstein–Uhlenbeck Process

Now consider the process solution to $V_t^x = x - a\int_0^t V_s^x \mathrm{d}s + \sigma W_t$: we emphasize in our notation the dependence w.r.t. the initial value $V_0 = x$. We define an appropriate second order operator

$$L_{a,\sigma^2}^{\text{OU}} g(t,x) = \frac{1}{2}\sigma^2 g_{xx}''(t,x) - axg_x'(t,x)$$

which plays the role of the infinitesimal generator for the Ornstein–Uhlenbeck process. We recall that the Gaussian distribution of V_t^x has mean xe^{-at} and variance $\frac{\sigma^2}{2a}(1-e^{-2at})$, the density of which at y (assuming $\sigma \neq 0$ for the existence) is

$$p(t,x,y) = g(v_t, xe^{-at}, y).$$

Using the heat equation satisfied by g, we easily derive that

$$p_t'(t,x,y) = \frac{1}{2}\sigma^2 p_{xx}''(t,x,y) - axp_x'(t,x,y) = L_{a,\sigma^2}^{\text{OU}} p(t,x,y),$$

from which we deduce the PDE satisfied by $u(t,x,f) = \mathbb{E}[f(V_t^x)]$. Incorporating Dirichlet boundary conditions is similar to the previous cases. We state the related results without extra details.

Theorem 5. *Let f be a bounded measurable function. The function*

$$u(t,x,f) = \mathbb{E}[f(V_t^x)] = \int_{\mathbb{R}} p(t,x,y) f(y) \mathrm{d}y$$

solves

$$\begin{cases} u_t'(t,x,f) = L_{a,\sigma^2}^{OU} u(t,x,f), \\ u(0,x,f) = f(x). \end{cases}$$

Proposition 13. *If $f \in \mathscr{C}_b^{1,2}$ and U is a bounded stopping time, then*

$$\mathbb{E}[f(U, V_U^x)] = f(0,x) + \mathbb{E}\Big[\int_0^U [L_{a,\sigma^2}^{OU} f(s, V_s^x) + f_t'(s, V_s^x)] \mathrm{d}s\Big].$$

Corollary 3. *Assume the existence of a solution u of class $C_b^{1,2}([0,T] \times [a,b])$ to the PDE*

$$\begin{cases} u_t'(t,x,f) = L_{a,\sigma^2}^{OU} u(t,x,f), & \text{for } t > 0 \text{ and } x \in]a,b[, \\ u(0,x,f) = f(0,x) & \text{for } t = 0 \text{ and } x \in [a,b], \\ u(t,x,f) = f(t,x) & \text{for } x = a \text{ or } b, \text{ with } t \geq 0. \end{cases}$$

Then u is given by

$$u(t,x) = \mathbb{E}[f(t - U^x, V_{U^x}^x)]$$

where $U^x = \inf\{s > 0 : V_s^x \notin]a,b[\} \wedge t$.

2.2.3 A Natural Conjecture for Stochastic Differential Equations

The previous examples serve as a preparation for more general results, relating the dynamics of a process and its Feynman–Kac representation. Denote X^x the solution (whenever it exists) to the Stochastic Differential Equation

$$X_t^x = x + \int_0^t b(X_s^x) \mathrm{d}s + \int_0^t \sigma(X_s^x) \mathrm{d}W_s, \quad t \geq 0.$$

In view of the results in simpler models, we announce the following facts.

1. Set $L^X_{b,\sigma^2} g = \frac{1}{2}\sigma^2(x) g''_{xx} + b(x) g'_x$.
2. $u(t,x) = \mathbb{E}(f(X_t^x))$ solves

$$u'_t(t,x) = L^X_{b,\sigma^2} u(t,x), \quad u(0,x) = f(x).$$

3. If $f \in \mathscr{C}_b^{1,2}$ and U is a bounded stopping time, then

$$\mathbb{E}[f(U, X_U^x)] = f(0,x) + \mathbb{E}\Big[\int_0^U [L^X_{b,\sigma^2} f(s, X_s^x) + f'_t(s, X_s^x)]\mathrm{d}s\Big].$$

4. If u of class $C_b^{1,2}([0,T] \times [a,b])$ solves the PDE

$$\begin{cases} u'_t(t,x) = L^X_{b,\sigma^2} u(t,x), & \text{for } t > 0 \text{ and } x \in]a,b[, \\ u(0,x) = f(0,x) & \text{for } t = 0 \text{ and } x \in [a,b], \\ u(t,x) = f(t,x) & \text{for } x = a \text{ or } b, \text{ with } t \geq 0, \end{cases}$$

then it is given by $u(t,x) = \mathbb{E}[f(t - U^x, X^x_{U^x})]$ where $U^x = \inf\{s > 0 : X_s^x \notin]a,b[\} \wedge t$.

The above result could be extended to PDE with a space variable in $\mathbb{R}^d$ ($d \geq 1$) by considering a $\mathbb{R}^d$-valued SDE: it would be achieved by replacing W by a d-dimensional standard Brownian motion, by having a drift coefficient $b : \mathbb{R}^d \mapsto \mathbb{R}^d$ and a diffusion coefficient $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d \otimes \mathbb{R}^d$, a reward function $f : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}$, by replacing the interval $[a,b]$ by a domain D in $\mathbb{R}^d$ and defining U^x as the first exit time by X^x from that domain. Then the operator L would be a linear parabolic second order operator of the form

$$L^X_{b,\sigma\sigma^\top} g = \frac{1}{2} \sum_{i,j=1}^d [\sigma\sigma^\top]_{i,j}(x) \partial^2_{x_i x_j} g + \sum_{i=1}^d b_i(x) \partial_{x_i} g,$$

where $\top$ denotes the transpose. We could also add a zero-order term in $L^X_{b,\sigma\sigma^\top}$, by considering a discounting factor for f; we do not develop further this extension.

The next section provides stochastic calculus tools, that allow to show the validity of these Feynman–Kac type results, under some appropriate smoothness and growth assumptions on b, σ, f. To allow non smooth f or Dirichlet boundary conditions, we may additionally assume a non-degeneracy condition on $L^X_{b,\sigma\sigma^\top}$ (like ellipticity condition $|\sigma\sigma^\top(x)| \geq \frac{1}{c}$ for some $c > 0$).

Complementary References. See [1, 15, 20, 22, 23, 48].

3 The Itô Formula

One achievement of Itô's formula is to go from an infinitesimal time-decomposition in expectation like

$$\mathbb{E}[f(t, x + W_t)] - f(0, x) = \int_0^t \mathbb{E}[f_t'(s, x + W_s) + \frac{1}{2} f_{xx}''(s, x + W_s)]\mathrm{d}s$$

(see (17)) to a pathwise infinitesimal time-decomposition of

$$f(t, x + W_t) - f(0, x).$$

Since Brownian motion paths are not differentiable, it is hopeless to apply standard differential calculus based on usual first order Taylor formula. Instead of this, we go up to the second order, taking advantage of the fact that W has a finite quadratic variation. The approach presented below is taken from the nice paper *Calcul d'Itô sans probabilité* by Föllmer [19], which does not lead to the most general and deepest approach but it has the advantage of light technicalities and straightforward arguments compared to the usual tough arguments using L_2-spaces and isometry (see for instance [48] or [63] among others).

3.1 Quadratic Variation

3.1.1 Notations and Definitions

Brownian increments in a small interval $[t, t + h]$ are centered Gaussian r.v. with variance h, which thus behave like $\sqrt{h}$. The total variation does not exist, because the trajectories are not differentiable, but the quadratic variation has interesting properties.

To avoid convergence technicalities, we consider particular time subdivisions.

Definition 7 (Dyadic Subdivision of Order n). Let n be an integer. The subdivision of $\mathbb{R}^+$ defined by $\mathbb{D}_n = \{t_0 < \cdots < t_i < \cdots\}$ where $t_i = i2^{-n}$ is called the dyadic subdivision of order n. The subdivision step is $\delta_n = 2^{-n}$.

Definition 8 (Quadratic Variation). The quadratic variation of a Brownian motion W associated with the dyadic subdivision of order n is defined, for $t \geq 0$, by

$$V_t^n = \sum_{t_i \leq t} (W_{t_{i+1}} - W_{t_i})^2. \tag{26}$$

3.1.2 Convergence

Then there is the following remarkable result.

Proposition 14 (Pointwise Convergence). *With probability 1, we have*

$$\lim_{n\to\infty} V_t^n = t$$

for any $t \in \mathbb{R}^+$.

Had W been differentiable, the limit of V^n would be equal to 0.

Proof. First let us show the *a.s.* convergence for a fixed time t, and denote by $n(t)$ the index of the dyadic subdivision of order n such that $t_{n(t)} \le t < t_{n(t)+1}$. Then observe that $V_t^n - t = \sum_{j=0}^{n(t)} Z_j + (t_{n(t)+1} - t)$ where $Z_j = (W_{t_{j+1}} - W_{t_j})^2 - (t_{j+1} - t_j)$. The term $t_{n(t)+1} - t$ converges to 0 as the subdivision step shrinks to 0. The random variables Z_j are independent, centered, square integrable (since the Gaussian law of $W_{t_{j+1}} - W_{t_j}$ has finite fourth moments): additionally, the scaling property of Proposition 1 ensures that $\mathbb{E}(Z_j^2) = C_2(t_{j+1} - t_j)^2$ for a positive constant C_2. Thus

$$\mathbb{E}\left(\sum_{j=0}^{n(t)} Z_j\right)^2 = \sum_{j=0}^{n(t)} \mathbb{E}\left(Z_j^2\right) = \sum_{j=0}^{n(t)} C_2(t_{j+1} - t_j)^2 \le C_2(T+1)\delta_n.$$

This proves the L_2-convergence of $\sum_{j=0}^{n(t)} Z_j$ towards 0.

Moreover we obtain $\sum_{n\ge1} \mathbb{E}\left(\sum_{j=0}^{n(t)} Z_j\right)^2 < \infty$, i.e. the random series $\sum_{n\ge1}\left(\sum_{j=0}^{n(t)} Z_j\right)^2$ has a finite expectation, whence *a.s.* finite and consequently its general term converges *a.s.* to 0. This shows that for any fixed t, $V_t^n \to t$ except on a negligible set N_t.

We now extend the result to any time: first the set $N = \cup_{t\in\mathbb{Q}^+} N_t$ is still negligible because the union of negligible sets is taken on a countable family. For an arbitrary t, take two monotone sequences of rational numbers $r_p \uparrow t$ and $s_p \downarrow t$ as $p \to +\infty$. Since $t \mapsto V_t^n$ is increasing for fixed n, we deduce, for any $\omega \notin N$

$$r_p = \lim_{n\to\infty} V_{r_p}^n(\omega) \le \liminf_{n\to\infty} V_t^n(\omega) \le \limsup_{n\to\infty} V_t^n(\omega) \le \lim_{n\to\infty} V_{s_p}^n(\omega) = s_p.$$

Passing to the limit in p gives the result. □

As a consequence, we obtain the formula giving the infinitesimal decomposition of W_t^2.

Proposition 15 (A First Itô Formula). *Let W be a standard Brownian motion. With probability 1, we have for any $t \ge 0$*

$$W_t^2 = 2\int_0^t W_s \mathrm{d}W_s + t \tag{27}$$

where the stochastic integral $\int_0^t W_s dW_s$ *is the* a.s. *limit of* $\sum_{t_i \le t} W_{t_i}(W_{t_{i+1}} - W_{t_i})$, *along the dyadic subdivision.*

For usual C^1-function $f(t)$, we have $f^2(t) - f^2(0) = 2\int_0^t f(s)\mathrm{d}f(s)$: the extra term t in (27) is intrinsically related to Brownian motion paths.

Proof. Adopting once again the notation with $n(t)$, we have

$$\begin{aligned} W_t^2 &= W_t^2 - W_{t_{n(t)+1}}^2 + \sum_{t_i \le t}(W_{t_i+1}^2 - W_{t_i}^2) \\ &= W_t^2 - W_{t_{n(t)+1}}^2 + \sum_{t_i \le t}(W_{t_{i+1}} - W_{t_i})^2 + 2\sum_{t_i \le t} W_{t_i}(W_{t_{i+1}} - W_{t_i}). \end{aligned}$$

The first term at the r.h.s. tends towards 0 by continuity of the Brownian paths. The second term is equal to V_t^n and converges towards t. Consequently, the third term at the right-hand side *must* converge *a.s.* towards a term that we call *stochastic integral* and that we denote by $2\int_0^t W_s \mathrm{d}W_s$. □

The random function V_t^n, as a function of t, is increasing and can be associated to the cumulative distribution function of the positive discrete measure

$$\sum_{i \ge 0}(W_{t_{i+1}} - W_{t_i})^2 \delta_{t_i}(.) = \mu^n(.)$$

satisfying $\mu_n(f) = \sum_{i \ge 0} f(t_i)(W_{t_{i+1}} - W_{t_i})^2$.

The convergence of cumulative distribution function of $\mu^n(.)$ (Proposition 14) can then be extended to integrals of continuous functions (possibly random as well). It is the purpose of the following result which is of deterministic nature.

Proposition 16 (Convergence as a Positive Measure). *For any continuous function* f, *with probability 1 we have*

$$\lim_{n\to\infty} \sum_{t_i \le t} f(t_i)(W_{t_{i+1}} - W_{t_i})^2 = \int_0^t f(s)\mathrm{d}s$$

for any $t \ge 0$.

The proof is standard: the result first holds for functions of the form $f(s) = \mathbf{1}_{]r_1,r_2]}(s)$, then for piecewise constant functions, at last for continuous functions by simple approximations.

3.2 The Itô Formula for Brownian Motion

Differential calculus extends to other functions than $x \to x^2$. To the usual classical formula with functions that are smooth in time, a term should be added, due to the non-zero quadratic variation.

Theorem 6 (Itô Formula). *Let $f \in \mathscr{C}^{1,2}(\mathbb{R}^+ \times \mathbb{R}, \mathbb{R})$. Then with probability 1, we have $t \geq 0$*

$$f(t, x + W_t) = f(0, x) + \int_0^t f'_x(s, x + W_s)\, \mathrm{d}W_s + \int_0^t f'_t(s, x + W_s)\, \mathrm{d}s + \frac{1}{2} \int_0^t f''_{xx}(s, x + W_s)\, \mathrm{d}s. \quad (28)$$

The term $\mathscr{I}_t(f) = \int_0^t f'_x(s, x+W_s)\mathrm{d}W_s$ is called the stochastic integral of $f'_x(s, x+W_s)$ w.r.t. W and it is the a.s. *limit of*

$$\mathscr{I}_t^n(f, W) = \sum_{t_i \leq t} f'_x(t_i, x + W_{t_i})(W_{t_{i+1}} - W_{t_i})$$

taken along the dyadic subdivision of order n.

The reader should compare the equality (28) with (17) to see that, under the extra assumptions that f is bounded with bounded derivatives, we have proved that the stochastic integral $\mathscr{I}_t(f)$ is centered:

$$\mathbb{E}\left(\int_0^t f'_x(s, x + W_s)\, \mathrm{d}W_s\right) = 0. \quad (29)$$

This explains how we can expect to go from (28) to (17):

1. Apply Itô formula.
2. Take expectation.
3. Prove that the stochastic integral is centered.

This is an interesting alternative proof to the property satisfied by the Gaussian kernel, which is difficult to extend to more general (non Gaussian) process.

Proof. As before, let us introduce the index $n(t)$ such that $t_{n(t)} \leq t < t_{n(t)+1}$; then we can write

$$\begin{aligned} f(t\ , x + W_t) = f(0, x) &+ [f(t, x + W_t) - f(t_{n(t)+1}, x + W_{t_{n(t)+1}})] \\ &+ \sum_{t_i \leq t} [f(t_{i+1}, x + W_{t_{i+1}}) - f(t_i, x + W_{t_{i+1}})] \\ &+ \sum_{t_i \leq t} [f(t_i, x + W_{t_{i+1}}) - f(t_i, x + W_{t_i})]. \end{aligned}$$

- The second term of the r.h.s. $[f(t, x + W_t) - f(t_{n(t)+1}, x + W_{t_{n(t)+1}})]$ converges to 0 by continuity of $f(t, x + W_t)$.
- The third term is analyzed by means of the first order Taylor formula:

$$f(t_{i+1}, x + W_{t_{i+1}}) - f(t_i, x + W_{t_{i+1}}) = f_t'(\tau_i, x + W_{t_{i+1}})(t_{i+1} - t_i)$$

for $\tau_i \in]t_i, t_{i+1}[$. The uniform continuity of $(W_s)_{0\le s\le t+1}$ ensures that $\sup_i |f_t'(\tau_i, x + W_{t_{i+1}}) - f_t'(t_i, x + W_{t_i})| \to 0$: thus $\lim_{n\to\infty} \sum_{t_i \le t} f_t'(\tau_i, x + W_{t_{i+1}})(t_{i+1} - t_i)$ equals to

$$\lim_{n\to\infty} \sum_{t_i \le t} f_t'(t_i, x + W_{t_i})(t_{i+1} - t_i) = \int_0^t f_t'(s, x + W_s)\mathrm{d}s.$$

- A second order Taylor formula allows to write the fourth term: $f(t_i, x + W_{t_{i+1}}) - f(t_i, x + W_{t_i})$ equals

$$f_x'(t_i, x + W_{t_i})(W_{t_{i+1}} - W_{t_i}) + \frac{1}{2} f_{xx}''(t_i, x + \xi_i)(W_{t_{i+1}} - W_{t_i})^2$$

where $\xi_i \in (W_{t_i}, W_{t_{i+1}})$. Similarly to before, $\sup_i |f_{xx}''(t_i, x + \xi_i) - f_{xx}''(t_i, x + W_{t_i})| = \epsilon_n \to 0$ and it leads to

$$\left| \sum_{t_i \le t} (f_{xx}''(t_i, x + \xi_i) - f_{xx}''(t_i, x + W_{t_i}))(W_{t_{i+1}} - W_{t_i})^2 \right| \le \epsilon_n V_t^n,$$

$$\lim_{n\to\infty} \sum_{t_i \le t} f_{xx}''(t_i, x + W_{t_i})(W_{t_{i+1}} - W_{t_i})^2 = \int_0^t f_{xx}''(s, x + W_s)\mathrm{d}s,$$

by applying Proposition 16.

Observe that in spite of the non-differentiability of W, $\sum_{t_i \le t} f_x'(t_i, x + W_{t_i})(W_{t_{i+1}} - W_{t_i})$ is necessarily convergent as a difference of convergent terms. □

Interestingly, we obtain a representation of the random variable $f(x + W_t)$ as a stochastic integral, in terms of the derivatives of solution u to the heat equation

$$u_t'(t, x) = \frac{1}{2} u_{xx}''(t, x), \quad u(0, x) = f(x).$$

Corollary 4. *Assume that $u \in \mathscr{C}_b^{1,2}([0, T] \times \mathbb{R})$. We have*

$$f(x + W_T) = u(T, x) + \int_0^T u_x'(T - s, x + W_s)\mathrm{d}W_s. \tag{30}$$

Proof. Apply the Itô formula to $v(t,x) = u(T-t,x)$ (which satisfies $v'_t(t,x) + \frac{1}{2}v''_{xx}(t,x) = 0$) at time T. This gives $f(x+W_T) = u(0,x+W_T) = u(T,x) + \int_0^T u'_x(T-s,x+W_s)\mathrm{d}W_s$. □

This representation formula leads to important remarks.

- If the above stochastic integral has zero expectation (as for the examples presented before), taking the expectation shows that

$$u(T,x) = \mathbb{E}(f(x+W_T)),$$

  recovering the Feynman–Kac representation of Theorem 2.
- Then, the above representation writes, setting $\Psi = f(x+W_T)$,

$$\Psi = \mathbb{E}(\Psi) + \int_0^T h_s \mathrm{d}W_s.$$

  Actually, a similar stochastic integral representation theorem holds in a larger generality on the form of Ψ, since any bounded[15] functional of $(W_t)_{0\le t\le T}$ can be represented as its expectation plus a stochastic integral: the process h is not tractable in general, whereas here it is explicitly related to the derivative of u along the Brownian path.
- Assume $u \in \mathscr{C}_b^{1,2}([0,T]\times\mathbb{R})$ imposes that $f \in \mathscr{C}_b^2(\mathbb{R})$ which is too strong for many applications: however, the assumptions on u can be relaxed to handle bounded measurable function f, because the heat equation is immediately smoothing out the initial condition. The proof of this extension involves extra stochastic calculus technicalities that we do not develop.

3.3 Wiener Integral

In general, it is not possible to make explicit the law of the stochastic integral $\int_0^t f'_x(s,x+W_s)\mathrm{d}W_s$, except in a situation where $f'_x(s,x) = h(s)$ is independent of x and square integrable. In that case, $\int_0^t h(s)\mathrm{d}W_s$ is distributed as a Gaussian r.v. The resulting stochastic integral is called *Wiener integral*. We sum up its important properties.

Proposition 17 (Wiener Integral and Integration by Parts). *Let $f : [0,T] \mapsto \mathbb{R}$ be a continuously differentiable function, with bounded derivatives on $[0,T]$.*

1. With probability 1, for any $t \in [0,T]$ we have

[15] Integrability is the right assumption.

$$\int_0^t f(s)\mathrm{d}W_s = f(t)W_t - \int_0^t W_s f'(s)\mathrm{d}s. \tag{31}$$

2. *The process* $\{\int_0^t f(s)\mathrm{d}W_s\,; t \in [0,T]\}$ *is a continuous Gaussian process, with zero mean and with a covariance function*

$$\mathbb{C}\mathrm{ov}(\int_0^t f(u)\mathrm{d}W_u, \int_0^s f(u)\mathrm{d}W_u) = \int_0^{s\wedge t} f^2(u)\mathrm{d}u. \tag{32}$$

3. *For another function g satisfying the same assumptions, we have*

$$\mathbb{C}\mathrm{ov}(\int_0^t f(u)\mathrm{d}W_u, \int_0^s g(u)\mathrm{d}W_u) = \int_0^{s\wedge t} f(u)g(u)\mathrm{d}u. \tag{33}$$

Proof. The first item is a direct application of Theorem 6 to the function $(t,x) \mapsto f(t)x$.

For any coefficients $(\alpha_i)_{1\le i\le N}$ and times $(T_i)_{1\le i\le N}$, $\sum_{i=1}^N \alpha_i \int_0^{T_i} f(u)\mathrm{d}W_u$ is a Gaussian r.v. since it can written as a limit of Gaussian r.v. of the form $\sum_j \beta_j (W_{t_{j+1}} - W_{t_j})$: thus, $\{\int_0^t f(s)\mathrm{d}W_s\,; t \in [0,T]\}$ is a Gaussian process. Its continuity is obvious in view of (31). Its expectation is the limit of the expectation of $\sum_{t_i\le t} f(t_i)[W_{t_{i+1}} - W_{t_i}]$, thus equal to 0. The covariance is the limit of the covariance

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}(\sum_{t_i\le t} f(t_i)[W_{t_{i+1}} - W_{t_i}], \sum_{t_i\le s} f(t_i)[W_{t_{i+1}} - W_{t_i}]) \\
&= \sum_{t_i\le t, t_j\le s} f(t_i)f(t_j)\mathbb{C}\mathrm{ov}(W_{t_{i+1}} - W_{t_i}, W_{t_{i+1}} - W_{t_i}) \\
&= \sum_{t_i\le t, t_j\le s} f(t_i)f(t_j)\delta_{i,j}(t_{i+1} - t_i) \underset{n\to+\infty}{\longrightarrow} \int_0^{s\wedge t} f^2(u)\mathrm{d}u.
\end{aligned}$$

The second item is proved. The last item is proved similarly. □

As a consequence, going back to the Ornstein–Uhlenbeck process (Sect. 1.6.2), we can complete the proof of its representation (11) using a stochastic integral, starting from (10). For this apply the result below to the function $f(s) = e^{-a(t-s)}$ (t fixed): it gives $\int_0^t e^{-a(t-s)}\mathrm{d}W_s = W_t - a\int_0^t e^{-a(t-s)}W_s\mathrm{d}s$. It leads to

$$V_t = v_0 e^{-at} + \sigma \int_0^t e^{-a(t-s)}\mathrm{d}W_s. \tag{34}$$

Then the Gaussian property from Proposition 17 gives that the variance of V_t is equal to $\sigma^2 \int_0^t e^{-2a(t-s)}\mathrm{d}s = \frac{\sigma^2}{2a}(1 - e^{-2at})$.

3.4 Itô Formula for Other Processes

The reader should have noticed that the central property for the proof of Theorem 6 is that the Brownian motion has a finite quadratic variation. Thus, the Itô formula can directly be extended to processes X which enjoy the same property.

3.4.1 The One-Dimensional Case

In this paragraph, we first consider scalar processes. The multidimensional extension is made afterwards.

Definition 9 (Quadratic Variation of a Process). A continuous process X has a finite quadratic variation if for any $t \geq 0$, the limit

$$V_t^n = \sum_{t_i \leq t} (X_{t_{i+1}} - X_{t_i})^2 \tag{35}$$

along the dyadic subdivision of order n, exists *a.s.* and is finite. We denote this limit by $\langle X \rangle_t$ and it is usually called the *bracket* of X at time t.

If $X = W$ is a Brownian motion, we have $\langle X \rangle_t = t$. More generally, it is easy to check that $\langle X \rangle$ is increasing and continuous. We associate to it a positive measure and this extends Proposition 16 to X.

Proposition 18. *For any continuous function f, with probability 1 for any $t \geq 0$ we have*

$$\lim_{n \to \infty} \sum_{t_i \leq t} f(t_i)(X_{t_{i+1}} - X_{t_i})^2 = \int_0^t f(s) \mathrm{d}\langle X \rangle_s.$$

Theorem 6 becomes

Theorem 7 (Itô Formula for X). *Let $f \in \mathscr{C}^{1,2}(\mathbb{R}^+ \times \mathbb{R}, \mathbb{R})$ and X be with finite quadratic variation. With probability 1, for any $t \geq 0$ we have*

$$\begin{aligned} f(t, X_t) = f(0, X_0) &+ \int_0^t f_x'(s, X_s)\, \mathrm{d}X_s + \int_0^t f_t'(s, X_s)\, \mathrm{d}s \\ &+ \frac{1}{2} \int_0^t f_{xx}''(s, X_s)\, \mathrm{d}\langle X \rangle_s, \end{aligned} \tag{36}$$

where $\int_0^t f_x'(s, X_s) \mathrm{d}X_s$ is the stochastic integral of $f_x'(s, X_s)$ w.r.t. X and it is the a.s. *limit of $\sum_{t_i \leq t} f_x'(t_i, X_{t_i})(X_{t_{i+1}} - X_{t_i})$ along dyadic subdivision of order n.*

Often, the Itô formula is written formally under a differential form

$$\mathrm{d}f(t,X_t) = f'_x(t,X_t)\,\mathrm{d}X_t + f'_t(t,X_t)\,\mathrm{d}t + \frac{1}{2}f''_{xx}(t,X_t)\,\mathrm{d}\langle X\rangle_t.$$

We now provide hand-made tools to compute the bracket of X in practice.

Proposition 19 (Computation of the Bracket). *Let A and M two continuous processes such that A has a finite variation*[16] *and M has a finite quadratic variation:*

1. $\langle A\rangle_t = 0$.
2. *If* $X_t = x + M_t$, *then* $\langle X\rangle_t = \langle M\rangle_t$.
3. *If* $X_t = \lambda M_t$, *then* $\langle X\rangle_t = \lambda^2\langle M\rangle_t$.
4. *If* $X_t = M_t + A_t$, *then* $\langle X\rangle_t = \langle M\rangle_t$.
5. *If* $X_t = f(A_t, M_t)$ *with* $f \in C^1$, *then* $\langle X\rangle_t = \int_0^t [f'_m(A_s, M_s)]^2 \mathrm{d}\langle M\rangle_s$.

The proof is easy and it uses deterministic arguments based on the definition of $\langle X\rangle$, we skip it. Item (5) shows that the class of processes with finite quadratic variation is stable by smooth composition. The following examples are important.

Example 1 (Arithmetic Brownian Motion). ($X_t = x + bt + \sigma W_t$): we have

$$\langle X\rangle_t = \langle \sigma W\rangle_t = \sigma^2\langle W\rangle_t = \sigma^2 t.$$

Itô's formula becomes

$$\begin{aligned}\mathrm{d}f(t,X_t) &= (f'_t(t,X_t) + f'_x(t,X_t)b + \frac{1}{2}f''_{xx}(t,X_t)\sigma^2)\mathrm{d}t + f'_x(t,X_t)\sigma\mathrm{d}W_t \\ &:= (f'_t(t,X_t) + L^{\mathtt{ABM}}_{b,\sigma^2} f(t,X_t))\mathrm{d}t + f'_x(t,X_t)\sigma\mathrm{d}W_t. \qquad (37)\end{aligned}$$

An important example is associated to $f(x) = \exp(x)$:

$$\mathrm{d}[\exp(X_t)] = \exp(X_t)(b + \frac{1}{2}\sigma^2)\mathrm{d}t + \exp(X_t)\sigma\mathrm{d}W_t. \qquad (38)$$

Example 2 (Geometric Brownian Motion). ($S_t = S_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W_t}$): we have

$$\langle S\rangle_t = \int_0^t \sigma^2 S_s^2 \mathrm{d}s.$$

From (38), we obtain a linear equation for the dynamics of S,

[16] That is the sum of $\sum_{t_i \le t} |A_{t_{i+1}} - A_{t_i}|$ exists and is finite, for instance A is continuously differentiable.

$$\mathrm{d}S_t = S_t\mu\mathrm{d}t + S_t\sigma\mathrm{d}W_t$$

also written $\frac{\mathrm{d}S_t}{S_t} = \mu\mathrm{d}t + \sigma\mathrm{d}W_t$ putting an emphasize of the financial interpretation as returns. The Itô formula writes

$$\begin{aligned}\mathrm{d}f(t,S_t) &= (f_t'(t,S_t) + f_x'(t,S_t)S_t\mu + \frac{1}{2}f_{xx}''(t,S_t)S_t^2\sigma^2)\mathrm{d}t + f_x'(t,S_t)\sigma S_t\mathrm{d}W_t \\ &:= (f_t'(t,S_t) + L^{\mathrm{GBM}}_{\mu,\sigma^2}f(t,S_t))\mathrm{d}t + f_x'(t,S_t)\sigma S_t\mathrm{d}W_t. \end{aligned} \tag{39}$$

Example 3 (Ornstein–Uhlenbeck Process). ($V_t = v_0 - a\int_0^t V_s\mathrm{d}s + \sigma W_t$): we have

$$\langle V\rangle_t = \sigma^2 t.$$

The Itô formula follows

$$\begin{aligned}\mathrm{d}f(t,V_t) &= (f_t'(t,V_t) - af_x'(t,V_t)V_t + \frac{1}{2}\sigma^2 f_{xx}''(t,V_t))\mathrm{d}t + f_x'(t,V_t)\sigma\mathrm{d}W_t \\ &:= (f_t'(t,V_t) + L^{\mathrm{OU}}_{a,\sigma^2}f(t,V_t))\mathrm{d}t + f_x'(t,V_t)\sigma\mathrm{d}W_t. \end{aligned} \tag{40}$$

Example 4 (Euler Scheme Defined in (12)). ($X_t^h = X_{ih}^h + b(X_{ih}^h)(t-ih) + \sigma(X_{ih}^h)(W_t - W_{ih})$ for $i \geq 0, t \in (ih,(i+1)h]$). Since X^h is an arithmetic Brownian motion on each interval $(ih,(i+1)h]$, we easily obtain

$$\langle X^h\rangle_t = \int_0^t \sigma^2(\varphi(s), X^h_{\varphi(s)})\mathrm{d}s$$

where $\varphi(t) = ih$ for $t \in (ih,(i+1)h]$. The Itô formula writes

$$\begin{aligned}\mathrm{d}f(t,X_t^h) = (f_t'(t,X_t^h) &+ b(X^h_{\varphi(t)})f_x'(t,X_t^h) + \frac{1}{2}\sigma^2(X^h_{\varphi(t)})f_{xx}''(t,X_t^h))\mathrm{d}t \\ &+ f_x'(t,X_t^h)\sigma(X^h_{\varphi(t)})\mathrm{d}W_t. \end{aligned} \tag{41}$$

3.4.2 The Multidimensional Case

We briefly expose the situation when $X = (X_1,\ldots,X_d)$ takes values in $\mathbb{R}^d$. The main novelty consists in considering the cross quadratic variation defined by the limit (assuming its existence, along dyadic subdivision) of

$$\langle X_k, X_l\rangle_t^n = \sum_{t_i\leq t}(X_{k,t_{i+1}} - X_{k,t_i})(X_{l,t_{i+1}} - X_{l,t_i}) \underset{n\to+\infty}{\longrightarrow} \langle X_k, X_l\rangle_t. \tag{42}$$

We list basic properties.

Properties 8

1. SYMMETRY: $\langle X_k, X_l\rangle_t = \langle X_l, X_k\rangle_t$.
2. USUAL BRACKET: $\langle X_k, X_k\rangle_t = \langle X_k\rangle_t$.
3. POLARIZATION: $\langle X_k, X_l\rangle_t = \frac{1}{4}\left(\langle X_k + X_l\rangle_t - \langle X_k - X_l\rangle_t\right)$.
4. *$\langle\cdot,\cdot\rangle_t$ is bilinear.*
5. *For any continuous function f, we have*

$$\lim_{n\to\infty}\sum_{t_i\le t} f(t_i)(X_{k,t_{i+1}} - X_{k,t_i})(X_{l,t_{i+1}} - X_{l,t_i}) = \int_0^t f(s)\mathrm{d}\langle X_k, X_l\rangle_s.$$

6. *Let $X_{1,t} = f(A_{1,t}, M_{1,t})$ and $X_{2,t} = g(A_{2,t}, M_{2,t})$, where the variation (resp. quadratic variation) of $A = (A_1, A_2)$ (resp. $M = (M_1, M_2)$) is finite, and let f and g be two $\mathscr{C}^1$-functions: we have*

$$\langle X_1, X_2\rangle_t = \int_0^t f'_m(A_{1,s}, M_{1,s})g'_m(A_{2,s}, M_{2,s})\mathrm{d}\langle M_1, M_2\rangle_s.$$

   In particular, $\langle A_1 + M_1, A_2 + M_2\rangle_t = \langle M_1, M_2\rangle_t$.
7. *Let W_1 and W_2 be two independent Brownian motions: then*

$$\langle W_1, W_2\rangle_t = 0.$$

Proof. The statements (1)–(6) are easy to check from the definition or using previous arguments. The statement (7) is important and we give details: use the polarization identity

$$\langle W_1, W_2\rangle_t = \frac{1}{4}\left(\langle W_1 + W_2\rangle_t - \langle W_1 - W_2\rangle_t\right).$$

We observe that both $\frac{1}{\sqrt{2}}(W_1 + W_2)$ and $\frac{1}{\sqrt{2}}(W_1 - W_2)$ are Brownian motions, since each one is a continuous Gaussian process with the right covariance function. Thus, $\langle\frac{1}{\sqrt{2}}(W_1 + W_2)\rangle_t = \langle\frac{1}{\sqrt{2}}(W_1 - W_2)\rangle_t = t$ and the result follows. □

The Itô formula naturally extends to this setting.

Theorem 9 (Multidimensional Itô Formula). *Let $f \in \mathscr{C}^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d, \mathbb{R})$ and X be a continuous d-dimensional process with finite quadratic variation. Then, with probability 1, for any $t \ge 0$ we have*

$$f(t, X_t) = f(0, X_0) + \sum_{k=1}^{d}\int_0^t f'_{x_k}(s, X_s)\,\mathrm{d}X_{k,s} + \int_0^t f'_t(s, X_s)\,\mathrm{d}s + \frac{1}{2}\sum_{k,l=1}^{d}\int_0^t f''_{x_k,x_l}(s, X_s)\,\mathrm{d}\langle X_k, X_l\rangle_s$$

where the sum of stochastic integrals are defined as before.

In particular, the integration by parts formula writes

$$X_{1,t}X_{2,t} = X_{1,0}X_{2,0} + \int_0^t X_{1,s}\mathrm{d}X_{2,s} + \int_0^t X_{2,s}\mathrm{d}X_{1,s} + \langle X_1, X_2\rangle_t.$$

For two independent Brownian motions, we recover the usual deterministic formula (because $\langle W_1, W_2\rangle_t = 0$), but in general, formulas are different because of the quadratic variation.

3.5 More Properties on Stochastic Integrals

So far, we have defined some specific stochastic integrals, those appearing in deriving a Itô formula and which have the form

$$\int_0^t f'_x(s, X_s)\mathrm{d}X_s = \lim_{n\to+\infty} \sum_{t_i \le t} f'_x(t_i, X_{t_i})(X_{t_{i+1}} - X_{t_i}), \tag{43}$$

the limit being taken along dyadic subdivision. Also, we have proved that if f has bounded derivatives and $X = W$ is a Brownian motion, the above stochastic integral must have zero-expectation [see equality (29)]. Moreover, we also have established that in the case of deterministic integrand (Wiener integral), the second moment of the stochastic integral is explicit and given by

$$\mathbb{E}(\int_0^t h_s\mathrm{d}W_s)^2 = \int_0^t h_s^2\mathrm{d}s.$$

The aim of this paragraph is to provide extensions of the above properties on the two first moments to more general integrands, under some suitable boundedness or integrability conditions.

3.5.1 Heuristic Arguments

In view of the previous construction, there is a natural candidate to be the stochastic integral $\int_0^t h_s dW_s$. When h is piecewise constant process (called *simple process*), that is $h_s = h_{t_i}$ if $s \in [t_i, t_{i+1}]$ for a given deterministic time grid $(t_i)_i$, we set

$$\int_0^t h_s\mathrm{d}W_s = \sum_{t_i \le t} h_{t_i}(W_{t\wedge t_{i+1}} - W_{t_i}), \tag{44}$$

Without extra assumptions on the stochasticity of h, it is not clear why its expectation equals 0. This property should come from the centered Brownian increments $W_{t\wedge t_{i+1}} - W_{t_i}$ and their independence to h_{t_i} so that

$$\mathbb{E}(\int_0^t h_s \mathrm{d}W_s) = \sum_{t_i \leq t} \mathbb{E}(h_{t_i})\mathbb{E}(W_{t\wedge t_{i+1}} - W_{t_i}) = 0.$$

To validate this computation, we shall assume that h_t depends only the Brownian Motion W before t and it is integrable. To go to the second moment, assume additionally that h is square integrable: then

$$\begin{aligned}
&\mathbb{E}|\int_0^t h_s \mathrm{d}W_s|^2 \\
&= 2\sum_{t_i<t_j\leq t} \mathbb{E}(h_{t_i}h_{t_j}(W_{t\wedge t_{i+1}} - W_{t_i}))\mathbb{E}(W_{t\wedge t_{j+1}} - W_{t_j}) + \sum_{t_i\leq t} \mathbb{E}(h_{t_i}^2)\mathbb{E}|W_{t\wedge t_{i+1}} - W_{t_i}|^2 \\
&= \sum_{t_i\leq t} \mathbb{E}(h_{t_i}^2)(t\wedge t_{i+1} - t_i) = \mathbb{E}(\int_0^t h_s^2 \mathrm{d}s). \qquad (45)
\end{aligned}$$

This equality should be read as an isometry property (usually referred to as *Itô isometry*), on which we can rely an extension of the stochastic integral of simple process to more general process. At this point, we should need to enter into measurability considerations to describe what "h_t *depends only the Brownian Motion* W *before* t" means at the most general level. It goes far beyond this introductory lecture: for the exposure of the general theory, see for instance [48] or [63].

For most of the examples considered in this lecture, we can restrict to *very good integrands*, in the sense that a integrand h is *very good* if

1. $(h_t)_t$ is continuous or piecewise continuous (as for simple processes).
2. For a given t, h_t is a continuous functional of $(W_s : s \leq t)$.
3. It is square integrable in time and ω: $\mathbb{E}(\int_0^t h_s^2 \mathrm{d}s) < +\infty$ for any t.

This setting ensures that we can define stochastic integrals for very good integrands as the L_2-limit of stochastic integrals for simple integrands: indeed, a Cauchy sequence $(h_n)_n$ in $L_2(\mathrm{d}t \otimes \mathrm{d}\mathbb{P})$ gives a Cauchy sequence $(\int_0^t h_{n,s}\mathrm{d}W_s)_n$ in $L_2(\mathbb{P})$ due to the isometry (45).

3.5.2 General Results

We collect here all the stochastic integration results needed in this lecture.

Theorem 10. *Let h be a very good integrand. Then the stochastic integral $\int_0^t h_s \mathrm{d}W_s$ is such that*

1. *It is the L_2 limit of $\sum_{t_i\leq t} h_{t_i}(W_{t\wedge t_{i+1}} - W_{t_i})$ along time subdivision which time step goes to 0.*
2. *It is centered: $\mathbb{E}(\int_0^t h_s \mathrm{d}W_s) = 0$.*

3. It is square integrable: $\mathbb{E}|\int_0^t h_s \mathrm{d}W_s|^2 = \mathbb{E}(\int_0^t h_s^2 \mathrm{d}s)$.
4. For two very good integrands h_1 *and* h_2*, we have*

$$\mathbb{E}\Big[(\int_0^t h_{1,s}\mathrm{d}W_s)(\int_0^t h_{2,s}\mathrm{d}W_s)\Big] = \mathbb{E}(\int_0^t h_{1,s}h_{2,s}\mathrm{d}s).$$

Beyond the t-by-t construction, actually the full theory gives a construction for any t simultaneously, proving additionally time continuity property, general centering property (*martingale* property), tight L_p-estimates on the value at time t and the extrema until time t (Burkholder–Davis–Gundy inequalities) and so on... For multidimensional W and h, the construction should be understood componentwise. Another fruitful extension is to allow t to be a bounded stopping time, similarly to the discussion we have made in the proof of Theorem 3.

Another interesting part in the theory is devoted to the existence and uniqueness of solution to Stochastic Differential Equations (also known as diffusion processes). The easiest setting is to assume globally Lipschitz coefficients[17]: it is similar to the ODE framework, and the proof is also based on the Picard fixed-point argument. We state the results without proof.

Theorem 11. *Let* W *be a* d*-dimensional standard Brownian motion.*

Assume that the functions $b : \mathbb{R}^d \mapsto \mathbb{R}^d$ *and* $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d \otimes \mathbb{R}^d$ *are globally Lipschitz. Then, for any initial condition* $x \in \mathbb{R}^d$*, there exists a unique*[18] *continuous solution* $(X_t^x)_{t\geq 0}$ *valued in* $\mathbb{R}^d$ *which satisfies*

$$X_t^x = x + \int_0^t b(X_s^x)\mathrm{d}s + \int_0^t \sigma(X_s^x)\mathrm{d}W_s, \tag{46}$$

with $\sup_{0\leq t\leq T}\mathbb{E}|X_t^x|^2 < +\infty$ *for any given* $T \in \mathbb{R}^+$.
The continuous process X^x *has a finite quadratic variation given by*

$$\langle X_k^x, X_l^x\rangle_t = \int_0^t [\sigma\sigma^\top]_{k,l}(X_s^x)\mathrm{d}s, \quad 1 \leq k, l \leq d. \tag{47}$$

Observe that this general result includes all the model considered before, such as Arithmetic and Geometric Brownian Motion, Ornstein–Uhlenbeck processes, here stated in a possibly multidimensional framework.

Complementary References. See [48, 63].

[17] Leading to the notion of *strong* solution; the case of non-smooth coefficients is much more delicate and related to *weak* solutions, see [67].

[18] Up to a set of zero probability measure.

4 Monte Carlo Resolutions of Linear PDEs Related to SDEs

Probabilistic methods to solve PDEs have become very popular during the two last decades. They are usually not competitive compared to deterministic methods in low dimension, but for higher dimension they provide very good alternative schemes. In the sequel, we give a brief introduction to the topics, relying on the material presented in the previous sections. We start with linear parabolic PDEs, with Cauchy–Dirichlet boundary conditions. Next section is devoted to semi-linear PDEs.

4.1 Second Order Linear Parabolic PDEs with Cauchy Initial Condition

4.1.1 Feynman–Kac Formulas

We start with a verification Theorem generalizing Theorems 2, 4, 5 to the case of general SDEs. We incorporate a source term g.

Theorem 12. *Under the assumptions of Theorem 11, let X^x be the solution (46) starting from $x \in \mathbb{R}^d$ and set*

$$L^X_{b,\sigma\sigma^\top} = \frac{1}{2}\sum_{i,j=1}^{d} [\sigma\sigma^\top]_{i,j}(x)\partial^2_{x_i x_j} + \sum_{i=1}^{d} b_i(x)\partial_{x_i}.$$

Assume there is a solution $u \in \mathscr{C}_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d, \mathbb{R})$ to the PDE

$$\begin{cases} u'_t(t,x) = L^X_{b,\sigma\sigma^\top} u(t,x) + g(x), \\ u(0,x) = f(x) \end{cases} \tag{48}$$

for two given functions $f, g : \mathbb{R}^d \to \mathbb{R}$. Then u is given by

$$u(t,x) = \mathbb{E}[f(X_t^x) + \int_0^t g(X_s^x)\mathrm{d}s]. \tag{49}$$

Proof. Let t be fixed. We apply the general Itô formula (Theorem 9) to the process X^x and to the function $v : (s,y) \mapsto u(t-s,y)$: it gives

$$\mathrm{d}v(s, X_s^x) = \left[v'_s(s, X_s^x) + L^X_{b,\sigma\sigma^\top} v(s, X_s^x)\right]\mathrm{d}s + Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s \tag{50}$$

$$= -g(X_s^x)\mathrm{d}s + Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s \tag{51}$$

where $Dv := (\partial_{x_1} v, \ldots, \partial_{x_d} v)$. Observe that the integrand $h_s = Dv(s, X_s^x)\sigma(X_s^x)$ is very good, since v has bounded derivatives, σ has a linear growth, and X_s has bounded second moments, locally uniformly in s: thus, the stochastic integral $\int_0^t Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s$ has zero expectation. Hence, applying the above decomposition between $s = 0$ and $s = t$ and taking the expectation, it gives

$$\mathbb{E}(f(X_t^x)) = \mathbb{E}(v(t, X_t^x)) = v(0, x) - \mathbb{E}(\int_0^t g(X_s^x)\mathrm{d}s) = u(t, x) - \mathbb{E}(\int_0^t g(X_s^x)\mathrm{d}s).$$

We are done. □

Smoothness assumptions on u are satisfied in f, g are smooth enough. If not, and if a uniform ellipticity condition is met on $\sigma\sigma^\top$, the fundamental solution of the PDE is smoothing the data and the result can be extended. However, the derivatives blow up as time goes to 0, and more technicalities are necessary to justify the same stochastic calculus computations. The fundamental solution $p(t, x, y)$ has a simple probabilistic interpretation: it is the density of X_t^x at y. Indeed, identify $\mathbb{E}[f(X_t^x) + \int_0^t g(X_s^x)\mathrm{d}s]$ with

$$u(t, x) = \int_{\mathbb{R}^d} p(t, x, y) f(y)\mathrm{d}y + \int_0^t \int_{\mathbb{R}^d} p(s, x, y) g(y) dy\, \mathrm{d}s.$$

4.1.2 Monte Carlo Schemes

Since $u(t, x)$ is represented as an expectation, it allows the use of a Monte Carlo method to numerically compute the solution. The difficulty is that in general, X can not be simulated perfectly accurately, only an approximation on a finite time-grid can be simply and efficiently produced. Namely we use the Euler scheme with time step $h = t/N$:

$$\begin{cases} X_0^{x,h} = x, \\ X_s^{x,h} = X_{ih}^{x,h} + b(X_{ih}^{x,h})(s - ih) + \sigma(X_{ih}^{x,h})(W_s - W_{ih}), \quad i \geq 0, s \in (ih, (i+1)h]. \end{cases} \tag{52}$$

Observe that to get $X_t^{x,h}$, we do not need to sample the continuous path of $X^{x,h}$ (as difficult as having a continuous path of a Brownian motion): in fact, we only need to compute $X_{ih}^{x,h}$ iteratively for $i = 0$ to $i = N$. Each time iteration requires to sample d new independent Gaussian increments $W_{k,(i+1)h} - W_{k,ih}$, centered with variance h: it is straightforward. The computational cost is essentially equal to $C(d)N$ where the constant depends on the dimension (coming from d-dimensional vector and matrix computations).

As an approximation of the expectation of $\mathscr{E}(f, g, X^x) = f(X_t^x) + \int_0^t g(X_s^x)\mathrm{d}s$, we take the expectation

$$\mathscr{E}(f,g,X^{x,h}) = f(X_{Nh}^{x,h}) + \sum_{i=0}^{N-1} g(X_{ih}^{x,h})h, \tag{53}$$

a random variable of which we sample M independent copies, that are denoted by $\{\mathscr{E}(f,g,X^{x,h,m}) : 1 \le m \le M\}$. Then, the Monte Carlo approximation, based on this sample of M Euler schemes with time step h, is

$$\frac{1}{M}\sum_{m=1}^{M} \mathscr{E}(f,g,X^{x,h,m}) = u(t,x) + \underbrace{\frac{1}{M}\sum_{m=1}^{M} \mathscr{E}(f,g,X^{x,h,m}) - \mathbb{E}(\mathscr{E}(f,g,X^{x,h}))}_{\text{statistical error } \mathrm{Err}_{\text{stat.}}(h,M)}$$
$$+ \underbrace{\mathbb{E}(\mathscr{E}(f,g,X^{x,h})) - u(t,x)}_{\text{discretization error } \mathrm{Err}_{\text{disc.}}(h)}. \tag{54}$$

The first error contribution is due to the sample of finite size: the larger M, the better the accuracy. As mentioned in Sect. 2.1.3, once renormalized by $\sqrt{M}$, this error is still random and its distribution is closed to the Gaussian distribution with zero mean and variance $\mathbb{V}\mathrm{ar}(\mathscr{E}(f,g,X^{x,h}))$: the latter still depends on h but very little, since it is expected to be close to $\mathbb{V}\mathrm{ar}(\mathscr{E}(f,g,X^{x}))$.

The second error contribution is related to the time discretization effect: the smaller the time h, the better the accuracy. In the sequel (Sect. 4.1.3), we theoretically estimate this error in terms of h, and proves that it is of order h (even equivalent to) under some reasonable and fairly general assumptions.

What Is the Optimal Tuning of $h \to 0$ and $M \to +\infty$? An easy complexity analysis shows that the computational effort is $\mathscr{C}_e = C(d)Mh^{-1}$. Observe that the rate does not depend on the dimension d, as a difference with a PDE method, but on the other hand, the solution is computed only at single point (t,x). The squared quadratic error is equal to

$$[\mathrm{Err}_2(h,M)]^2 := \mathbb{E}\Big[\frac{1}{M}\sum_{m=1}^{M} \mathscr{E}(f,g,X^{x,h,m}) - u(t,x)\Big]^2$$
$$= \frac{\mathbb{V}\mathrm{ar}(\mathscr{E}(f,g,X^{x,h}))}{M} + \Big[\mathbb{E}(\mathscr{E}(f,g,X^{x,h})) - u(t,x)\Big]^2.$$

Only the first factor $\mathbb{V}\mathrm{ar}(\mathscr{E}(f,g,X^{x,h}))$ can be estimated with the same sample, for M large, and it depends little of h. Say that the second term is equivalent to $(Ch)^2$ as $h \to 0$, with $C \neq 0$. Then, three asymptotic situations occur:

1. If $M \gg h^{-2}$, the statistical error becomes negligible and $\frac{1}{M}\sum_{m=1}^{M} \mathscr{E}(f,g,X^{x,h,m}) - u(t,x) \sim Ch$. The computational effort is $\mathscr{C}_e \gg h^{-3}$ and thus $\mathrm{Err}_2(h,M) \gg \mathscr{C}_e^{-1/3}$. Deriving a confidence interval as in Sect. 2.1.3 is meaningless, we face with the discretization error only.

2. If $M \ll h^{-2}$, the discretization error becomes negligible and the distribution of $\sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M}\mathscr{E}(f,g,X^{x,h,m}) - u(t,x)\right)$ converges to that a Gaussian r.v. *centered* with variance $\mathbb{V}\mathrm{ar}(\mathscr{E}(f,g,X^x))$ (that can be asymptotically computed using the M- sample). Thus, we can derive confidence intervals: setting $\sigma^2_{h,M}$ the empirical variance of $\mathscr{E}(f,g,X^{x,h})$, with probability 95 % we have

$$u(t,x) \in \Big[\frac{1}{M}\sum_{m=1}^{M}\mathscr{E}(f,g,X^{x,h,m}) - 1.96\frac{\sigma_{h,M}}{\sqrt{M}}, \frac{1}{M}\sum_{m=1}^{M}\mathscr{E}(f,g,X^{x,h,m}) + 1.96\frac{\sigma_{h,M}}{\sqrt{M}}\Big].$$

Regarding the computational effort, we have $\mathscr{C}_e \gg M^{3/2}$ and thus $\mathrm{Err}_2(h,M) \gg \mathscr{C}_e^{-1/3}$.

3. If $M \sim ch^{-2}$, both statistical and discretization errors have the same magnitude and one can still derive a asymptotic confidence interval, but it is no more centered (as in $M \ll h^{-2}$) and unfortunately, the bias is not easily estimated on the fly. The problem is that the bias is of same magnitude as the size of the confidence interval, thus it reduces the interest of having such a priori statistical error estimate. Here, $\mathrm{Err}_2(h,M) = O(\mathscr{C}_e^{-1/3})$.

Summing up by considering the ability of having or not on-line error estimates and by optimizing the final accuracy w.r.t. the computational effort, the second case $M = h^{-2+\varepsilon}$ (for a small $\varepsilon > 0$) may be the most attractive since it achieves (almost) the best accuracy w.r.t. the computational effort and gives a centered confidence interval (and therefore tractable and meaningful error bounds).

4.1.3 Convergence of the Euler Scheme

An important issue is to analyze the impact of time discretization of SDE. This dates back to the end of eighties, see [68] among others. The result below gives a mathematical justification of the use of the Euler scheme as an approximation for the distribution of the SDE.

Theorem 13. *Assume that b and σ are $\mathscr{C}_b^2$, let X^x be the solution (46) starting from $x \in \mathbb{R}^d$ and let $X^{h,x}$ be its Euler scheme defined in (52). Assume that $u(t,x) = \mathbb{E}[f(X_t^x) + \int_0^t g(X_s^x)\mathrm{d}s]$ is a $\mathscr{C}_b^{2,4}([0,T]\times\mathbb{R}^d,\mathbb{R})$-function solution of the PDE of Theorem 12. Then,*

$$\mathbb{E}\Big[f(X_{Nh}^{x,h}) + \sum_{i=0}^{N-1} g(X_{ih}^{x,h})h\Big] - \mathbb{E}\Big[f(X_t^x) + \int_0^t g(X_s^x)\mathrm{d}s\Big] = O(h).$$

Proof. Denote by $\mathrm{Err}_{\mathrm{disc.}}(h)$ the above discretization error. As in Theorem 12, we use the function $v : (s,y) \mapsto u(t-s,y)$ (for a fixed t) and we apply the Itô formula to $X^{h,x}$ (Theorem 9): it gives (setting $Dv := (\partial_{x_1}v,\ldots,\partial_{x_d}v)$)

$$
\begin{aligned}
\mathrm{d}v(s,X_s^{h,x}) &= \Big[v_s'(s,X_s^{h,x}) + \frac{1}{2}\sum_{i,j=1}^{d}[\sigma\sigma^\top]_{i,j}(X_{\varphi(s)}^{h,x})\partial^2_{x_ix_j}v(s,X_s^{h,x}) \\
&\qquad + \sum_{i=1}^{d} b_i(X_{\varphi(s)}^{h,x})\partial_{x_i}v(s,X_s^{h,x})\Big]\mathrm{d}s + Dv(s,X_s^{h,x})\sigma(X_{\varphi(s)}^{h,x})\mathrm{d}W_s. \\
&= \Big[\frac{1}{2}\sum_{i,j=1}^{d}\big([\sigma\sigma^\top]_{i,j}(X_{\varphi(s)}^{h,x}) - [\sigma\sigma^\top]_{i,j}(X_s^{h,x})\big)\partial^2_{x_ix_j}v(s,X_s^{h,x}) \\
&\qquad + \sum_{i=1}^{d}\big(b_i(X_{\varphi(s)}^{h,x}) - b_i(X_s^{h,x})\big)\partial_{x_i}v(s,X_s^{h,x}) - g(X_s^{h,x})\Big]\mathrm{d}s \\
&\quad + Dv(s,X_s^{h,x})\sigma(X_{\varphi(s)}^{h,x})\mathrm{d}W_s
\end{aligned}
$$

where at the second equality, we have used the PDE solved by v at (s, X_s^x). Then, by taking the expectation (it removes the stochastic integral term because the integrand is very good), we obtain

$$
\begin{aligned}
\mathrm{Err.}_{\mathrm{disc.}}(h) &= \mathbb{E}\Big[v(Nh,X_{Nh}^{x,h}) + \sum_{i=1}^{N} hg(X_{ih}^{x,h})\Big] - v(0,x) \\
&= \mathbb{E}\Big(\int_0^t\Big[\frac{1}{2}\sum_{i,j=1}^{d}\big([\sigma\sigma^\top]_{i,j}(X_{\varphi(s)}^{h,x}) - [\sigma\sigma^\top]_{i,j}(X_s^{h,x})\big)\partial^2_{x_ix_j}v(s,X_s^{h,x})\Big]\mathrm{d}s\Big) \\
&\quad + \mathbb{E}\Big(\int_0^t\Big[\sum_{i=1}^{d}\big(b_i(X_{\varphi(s)}^{h,x}) - b_i(X_s^{h,x})\big)\partial_{x_i}v(s,X_s^{h,x})\Big]\mathrm{d}s\Big) \\
&\quad + \mathbb{E}\Big(\int_0^t\big[g(X_{\varphi(s)}^{h,x}) - g(X_s^{h,x})\big]\mathrm{d}s\Big).
\end{aligned}
$$

The global error is represented as a summation of local errors. For instance, let us estimate the first term related to $\sigma\sigma^\top$: apply once again the Itô formula on the interval $[kh, s] \subset [kh, (k+1)h]$ and to the function $(s,y) \mapsto \big([\sigma\sigma^\top]_{i,j}(X_{\varphi(s)}^{h,x}) - [\sigma\sigma^\top]_{i,j}(y)\big)\partial^2_{x_ix_j}v(s,y)$. It gives raise to a time integral between $kh = \varphi(s)$ and s and a stochastic integral that vanishes in expectation. Proceed similarly for the other contributions with b and g. Finally we obtain a representation formula of the form

$$
\mathrm{Err.}_{\mathrm{disc.}}(h) = \sum_{\alpha:0\le|\alpha|\le 4}\mathbb{E}\Big(\int_0^t\int_{\varphi(s)}^{s}\partial_x^{|\alpha|}v(r,X_r^{h,x})l_\alpha\big(X_{\varphi(r)}^{h,x},X_r^{h,x}\big)\mathrm{d}r\mathrm{d}s\Big)
$$

where the summation is made on differentiation multi-indices of length smaller than 4, where l_α are functions depending on b, σ, g and their derivatives up to order 2,

and where l_α has at most a linear growth w.r.t its two variables. Taking advantage of the boundedness of the derivatives of v, we easily complete the proof.

Observe that, by strengthening the assumptions and by going a bit further in the analysis, we could establish an expansion w.r.t. h. □

The previous assumption on u implies that $f \in \mathscr{C}_b^4$ and $g \in \mathscr{C}_b^2$, which is too strong in practice. The extension to non smooth f is much more difficult and we have to take advantage of the smoothness coming from the non-degenerate distribution of X or X^h. We may follow the same types of computations, mixing PDE techniques and stochastic arguments, see [6]. But this is a pure stochastic analysis approach (Malliavin calculus) which provides the extension under the minimal non-degeneracy assumption (i.e. only stated at the initial point x), see [38]. We state the result without proofs.

Theorem 14. *Assume that b and σ are $\mathscr{C}_b^\infty$, let X^x be the solution (46) starting from $x \in \mathbb{R}^d$ and let $X^{h,x}$ be its Euler scheme defined in (52). Assume additionally that $\sigma\sigma^\top(x)$ is invertible. Then, for any bounded measurable function f, we have*

$$\mathbb{E}\Big[f(X_t^{x,h})\Big] - \mathbb{E}\Big[f(X_t^x)\Big] = O(h).$$

In the same reference [38], the result is also proved for hypoelliptic system, where the hypoellipticity holds only at the starting point x. On the other hand, without such a degeneracy condition and for non smooth f (like Heaviside function), the convergence may fail.

The case of coefficients b and σ with low regularity or exploding behavior is still an active fields of research.

4.1.4 Sensitivities

If in addition we are interested by computing derivatives of $u(t,x)$ w.r.t. x or other model parameters, this is still possible using Monte Carlo simulations. For the sake of simplicity, in our discussion we focus on the gradient of u w.r.t. x. Essentially, two approaches are known.

Resimulation Method. The derivative is approximated using the finite difference method

$$\partial_{x_i} u(t,x) \approx \frac{u(t,x+\varepsilon e_i) - u(t,x-\varepsilon e_i)}{2\varepsilon}$$

where $e_i = (0,\dots,0,\underset{i^{th}}{1},0,\dots)$, and ε is small. Then, each value function is approximated by its Monte Carlo approximation given in (54). However, we have to be careful in generating the Euler scheme starting from $x+\varepsilon e_i$ and $x-\varepsilon e_i$: its sampling should use the same Brownian motion increments, that is

$$\partial_{x_i} u(t,x) \approx \frac{1}{M} \sum_{m=1}^{M} \frac{\mathscr{E}(f,g,X^{x+\varepsilon e_i,h,m}) - \mathscr{E}(f,g,X^{x-\varepsilon e_i,h,m})}{2\varepsilon}. \tag{55}$$

Indeed, for an infinite sample ($M = +\infty$), it does not have any impact on the statistical error whether or not we use the *same driving noise*, but for finite M, this trick likely maintains a smaller statistical error. Furthermore, the optimal choice of h, M and ε is an important issue, but here results are different according to the regularity of f and g, we do not go into details.

Likelihood Method. To avoid the latter problems of selecting the appropriate value of the finite difference parameter ε, we may prefer another Monte Carlo estimator of $\partial_{x_i} u(t,x)$, which consists in appropriately weighting the output. When g equals 0, it takes the following form

$$\partial_{x_i} u(t,x) \approx \frac{1}{M} \sum_{m=1}^{M} f(X_t^{x,h,m}) H_t^{x,h,m} \tag{56}$$

where $H_t^{x,h,m}$ is simultaneously generated with the Euler scheme and does not depend on f. The advantage of this approach is to avoid the possibly delicate choice of the perturbation parameter ε and it is valid for any function f: thus, it may reduce much the computational time, if many sensitivities are required for the same model. On the other hand, the confidence interval may be larger than that of the resimulation method.

We now provide the formula for the weight H (known as Bismut–Elworthy–Li formula). It uses the tangent process, which is the (well-defined, see [52]) derivative of $x \mapsto X_t^x$ w.r.t. x and which solves

$$DX_t^x := Y_t^x = \mathrm{Id} + \int_0^t Db(X_s^x)\, Y_s^x \,\mathrm{d}s + \sum_{j=1}^{d} \int_0^t D\sigma_j(X_s^x)\, Y_s^x \,\mathrm{d}W_{j,s} \tag{57}$$

where σ_j is the j-th column of the matrix σ.

Theorem 15. *Assume that b and σ are $\mathscr{C}_b^2$-functions, that $u \in \mathscr{C}^{1,2}([0,T]\times\mathbb{R}^d,\mathbb{R})$ solves the PDE (48), and that σ is invertible with a uniformly bounded inverse σ^{-1}. We have*

$$Du(t,x) = \mathbb{E}\left(\frac{f(X_t^x)}{t} \left[\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s \right]^\top \right).$$

Proof. First, we recall the decomposition (51) obtained from Itô formula, using $v(s,y) = u(t-s,y)$:

$$\begin{cases} v(r, X_r^x) = v(0,x) + \int_0^r Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s, \quad \forall 0 \le r \le t, \\ f(X_t^x) = v(t, X_t^x). \end{cases} \tag{58}$$

Second, taking expectation, it gives $v(0,x) = u(t,x) = \mathbb{E}(v(r, X_r^x))$ for any $r \in [0,T]$. By differentiating w.r.t. x, we obtain a nice relation letting the expectation constant in time (actually deeply related to martingale property):

$$Dv(0,x) = \mathbb{E}(Dv(r, X_r^x) Y_r^x), \quad \forall 0 \le r \le t.$$

Thus, we deduce

$$\begin{aligned} Du(t,x) = Dv(0,x) &= \mathbb{E}\left(\frac{1}{t}\int_0^t Dv(s, X_s^x) Y_s^x \, \mathrm{d}s\right) \\ &= \mathbb{E}\left(\frac{1}{t}\left[\int_0^t Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s\right]\left[\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s\right]^\top\right) \\ &= \mathbb{E}\left(\frac{v(t, X_t^x) - v(0,x)}{t}\left[\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s\right]^\top\right) \\ &= \mathbb{E}\left(\frac{f(X_t^x)}{t}\left[\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s\right]^\top\right) \end{aligned}$$

using Theorem 10 at the second and fourth equality, (58) at the third one. □

In view of the above assumptions of u, implicitly the function f is smooth. However, under the current ellipticity condition, u is still smooth even if f is not; since the formula does depend on f and not on its derivatives, it is standard to extend the formula to any bounded function f (without any regularity assumption).

The Monte Carlo evaluation of $Du(t,x)$ easily follows by independently sampling $\frac{f(X_t^x)}{t}\left[\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s\right]^\top$ and taking the empirical mean. The exact simulation is not possible and once again, we may use an Euler-type scheme, with time step h:

- The dimension-augmented Stochastic Differential Equation (X^x, Y^x) is approximated using the Euler scheme.
- We use a simple-approximation of the stochastic integral

$$\int_0^t [\sigma^{-1}(X_s^x) Y_s^x]^\top \mathrm{d}W_s = \sum_{i=0}^{N-1} [\sigma^{-1}(X_{ih}^{x,h}) Y_{ih}^{x,h}]^\top (W_{(i+1)h} - W_{ih}).$$

The analysis of discretization error is more intricate than for $\mathbb{E}(f(X_t^{x,h}) - f(X_t^x))$: nevertheless, the error is still of magnitude h (the convergence order is 1 w.r.t. h, as proved in [38]).

Theorem 16. *Under the setting of Theorem 14, for any bounded measurable function f, we have*

$$\mathbb{E}\left(\frac{f(X_t^{x,h})}{t}\sum_{i=0}^{N-1}\left[[\sigma^{-1}(X_{ih}^{x,h})Y_{ih}^{x,h}]^\top(W_{(i+1)h}-W_{ih})\right]^\top\right)-Du(t,x)=O(h).$$

4.1.5 Other Theoretical Estimates in Small Time

The representation formula of Theorem 15 is the starting point for getting accurate probabilistic estimates on the derivatives of the underlying PDE as time is small, in terms of the fractional smoothness of $f(X_t^x)$ which is related to the decay of

$$\|f(X_t^x)-\mathbb{E}(f(X_{t-s}^y))|_{y=X_s^x}\|_{L_2}\quad \text{as } s\to t.$$

The derivatives are measured in weighted L_2-norms and surprisingly, the above results are equivalence results [36]; we are not aware of such results using PDE arguments.

Theorem 17. *Under the setting*[19] *of Theorem 14, let t be fixed, for $0<\theta\le 1$ and a bounded f, the following assertions are equivalent:*

i) *For some $c\ge 0$, $\mathbb{E}|f(X_t^x)-\mathbb{E}(f(X_{t-s}^y))|_{y=X_s^x}|^2\le c(t-s)^\theta$ for $0\le s\le t$.*
ii) *For some $c\ge 0$, $\mathbb{E}|Du(t-s,X_s^x)|^2\le \frac{c}{(t-s)^{1-\theta}}$ for $0\le s<t$.*
iii) *For some $c\ge 0$, $\int_0^s \mathbb{E}|D^2u(t-r,X_r^x)|^2\mathrm{d}r\le \frac{c}{(t-s)^{1-\theta}}$ for $0\le s<t$.*

If $0<\theta<1$, it is also equivalent to:

iv) *For some $c\ge 0$, $\mathbb{E}|D^2u(t-s,X_s^x)|^2\le \frac{c}{(t-s)^{2-\theta}}$ for $0\le s<t$.*

Theorem 18. *Under the setting of Theorem 14, let t be fixed, for $0<\theta<1$ and a bounded f, the following assertions are equivalent:*

i) $\int_0^t (t-s)^{-\theta-1}\mathbb{E}|f(X_t^x)-\mathbb{E}(f(X_{t-s}^y))|_{y=X_s^x}|^2\mathrm{d}s<+\infty$.
ii) $\int_0^t (t-s)^{-\theta}\mathbb{E}|Du(t-s,X_s^x)|^2\mathrm{d}s<+\infty$.
iii) $\int_0^t (t-s)^{1-\theta}\mathbb{E}|D^2u(t-s,X_s^x)|^2\mathrm{d}s<+\infty$.

4.2 The Case of Dirichlet Boundary Conditions and Stopped Processes

4.2.1 Feynman–Kac Formula

In view of Corollary 1, the natural extension of Theorem 12 in the case of Dirichlet boundary condition is the following. We state the result without source term to simplify. The proof is similar and we skip it.

[19] To simplify the exposure.

Theorem 19. *Let D be a bounded domain of $\mathbb{R}^d$. Under the setting of Theorem 12, assume there is a solution $u \in \mathscr{C}_b^{1,2}([0,T] \times \overline{D}, \mathbb{R})$ to the PDE*

$$\begin{cases} u'_t(t,x) = L^X_{b,\sigma\sigma^\top} u(t,x), \quad \text{for } (t,x) \in]0,+\infty[\times D, \\ u(0,x) = f(0,x), \quad \text{for } x \in D, \\ u(t,x) = f(t,x), \quad \text{for } (t,x) \in \mathbb{R}^+ \times \partial D, \end{cases} \tag{59}$$

for a given function $f : \mathbb{R}^+ \times \overline{D} \to \mathbb{R}$. Then u is given by

$$u(t,x) = \mathbb{E}[f(t - \tau^x \wedge t, X^x_{\tau^x \wedge t})] \tag{60}$$

for $x \in D$, where $\tau^x = \inf\{s > 0 : X^x_s \notin D\}$ is the first exit time from D by X.

4.2.2 Monte Carlo Simulations

Performing a Monte Carlo algorithm in this context is less easy since we have to additionally simulate the exit time of X. A simple approach consists in discretizing X using the Euler scheme with time step h, and then taking for the exit time

$$\tau^{x,h} = \inf\{ih > 0 : X^{x,h}_{ih} \notin D\}.$$

It does not require any further computations than those needed to generate $(X^{x,h}_{ih}, 0 \le i \le N)$. But, the discretization error worsens much since it becomes of magnitude $\sqrt{h}$. Actually, even if the values of $(X^{x,h}_{ih}, 0 \le i \le N)$ are generated without error (like in Brownian motion case or other simple processes), the convergence order is still $\frac{1}{2}$ w.r.t. h [27]. The deterioration of the discretization error really comes from the high irregularity of Brownian motion paths (and SDE paths): even if two successive points $X^{x,h}_{ih}$ and $X^{x,h}_{(i+1)h}$ are close to the boundary but inside the domain, a discrete monitoring scheme does not detect the exit while a continuous Brownian motion-like path would likely exit from the domain between ih and $(i+1)h$. Moreover, it gives a systematic (in mean) underestimation of the true exit time. To overcome this lack of accuracy, there are several improved schemes.

- The Brownian bridge technique consists in simulating the exit time of local arithmetic Brownian motion [corresponding to the local dynamics of Euler scheme, see (12)]. For simple domain like half-space, the procedure is explicit and tractable, this is related the explicit knowledge of the distribution of the Brownian maximum, see Proposition 5. For smooth domain, we can approximate locally the domain by half-spaces. This improvement allows to recover the order 1 for the convergence, see [27, 28]. For non smooth domains (including corners for instance) and general SDEs, providing an accurate scheme and performing

its error analysis is still an open issue; for heuristics and numerical experiments, see [29] for instance.

- The boundary shifting method consists in shrinking the domain to compensate the systematic bias in the simulation of the discrete exit time. Very remarkably, there is a universal elementary rule to make the domain smaller:

  > locally at a point y close to the boundary, move the boundary inwards by a quantity proportional to $c_0\sqrt{h}$ times the norm of the diffusion coefficient in the normal direction.

  The constant c_0 is equal to the mean of the asymptotic overshoot of the Gaussian random walk as the ladder height goes to infinity: it can be expressed using the zeta function

$$c_0 = -\frac{\zeta(\frac{1}{2})}{\sqrt{2\pi}} = 0.5826\ldots.$$

  This procedure strictly improves the order $\frac{1}{2}$ of the discrete procedure, but it is still an open question whether the convergence order is 1, although numerical experiments corroborates this fact.

The result is stated as follows, see [37].

Theorem 20. *Assume that the domain D is bounded and has a $\mathscr{C}^3$-boundary, that b, σ are $\mathscr{C}_b^2$ and $f \in \mathscr{C}_b^{1,2}$. Let $n(y)$ be the unit inward normal vector to the boundary ∂D at the closest*[20] *point to y on the boundary. Set*

$$\hat{\tau}^{x,h} = \inf\left\{ih > 0 : X_{ih}^{x,h} \notin D \text{ or } d(X_{ih}^{x,h}, \partial D) \leq c_0\sqrt{h}\left|n^\top \sigma\right|(X_{ih}^{x,h})\right\}.$$

Then, we have

$$\mathbb{E}[f(t - \hat{\tau}^{x,h} \wedge t, X^x_{\hat{\tau}^{x,h}\wedge t})] - \mathbb{E}[f(t - \tau^x \wedge t, X^x_{\tau^x \wedge t})] = o(\sqrt{h}).$$

Observe that this improvement is very cheap regarding the computational cost. It can be extended (regarding to the numerical scheme and its mathematical analysis) to a source term, to time-dependent domain and to stationary problems (elliptic PDE).

Complementary References. See [2, 13, 26, 49, 53, 64] for general references. For reflected processes and Neumann boundary conditions, see [10, 28]. For variance reduction techniques, see [34, 47, 58]. For domain decomposition, see [35, 62]. This list is not exhaustive.

[20] Uniquely defined if y is close to the boundary.

5 Backward Stochastic Differential Equations and Semi-linear PDEs

The link between PDEs and stochastic processes have been developed since several decades and more recently, say in the last 20 years, researchers have paid attention to the probabilistic interpretation of non-linear PDEs, and in particular semi-linear PDEs. These PDEs are connected to non-linear processes, called Backward Stochastic Differential Equations (BSDE in short). In this section, we define these equations, firstly introduced by Pardoux and Peng [60], and give their connection with PDEs. Finally, we present a Monte Carlo algorithm to simulate them, using empirical regressions: it has the advantage to suit well the case of multidimensional problems, with a great generality on the type of semi-linearity.

These equations have many fruitful applications in stochastic control theory and mathematical finance, where they usually provide elegant proofs to characterize the solution to optimal investment problems for instance; for the related applications, we refer to reader to [17, 18]. Regarding the semi-linear PDE point of view, the applications are reaction-diffusion equations in chemistry [24], evolution of species in population biology [51,66], Hodgkin–Huxley model in neuroscience [43], Allen–Cahn equation for phase transition in physics... see the introductive course [30] and references therein. For other non-linear equations with connections with stochastic processes, see the aforementioned reference.

5.1 Existence of BSDE and Feynman–Kac Formula

5.1.1 Heuristics

As a difference with a Stochastic Differential Equation defined by (46) where the initial condition is given and the dynamics is imposed, a Backward SDE is defined through a random *terminal condition* ξ at a fixed terminal T and a dynamics imposed by a *driver* g. It takes the form

$$Y_t = \xi + \int_t^T g(s, Y_s, Z_s)\mathrm{d}s - \int_t^T Z_s \mathrm{d}W_s \tag{61}$$

where we write the integrals between t and T to emphasize the backward point of view: ξ should be thought as a stochastic target to reach at time T. A solution to (61) is the couple (Y, Z): without extra conditions, the problem has an infinite number of solutions and thus is ill-posed. For instance, if $g \equiv 0$ and $\xi = f(W_T)$: taking $c \in \mathbb{R}$, a solution is $Z_t = c$ and $Y_t = \xi + c(W_T - W_t)$, thus uniqueness fails. In addition to integrability properties (appropriate L_2-spaces) that we do not detail, an important condition is that the solution does not anticipate the future of Brownian motion, i.e. the solution Y_t depends on the Brownian Motion W up to t, and similarly to Z: we

informally say that *the solution is adapted to* W. In a stochastic control problem, this adaptedness constraint is natural since it states that the value function or the decision can not be made in advance to the flow of information given by W. Observe that in the uniqueness counter-example, Y is not adapted to W since Y_t depends on the Brownian motion on $[0, T]$ and not only on $[0, t]$.

Taking the conditional expectation in (61) gives

$$Y_t = \mathbb{E}\Big(\xi + \int_t^T g(s, Y_s, Z_s)\mathrm{d}s \Big| W_s : s \le t\Big), \tag{62}$$

because the stochastic integral (built with Brownian increments after t) is centered conditionally on the Brownian motion up to time t. Of course, this rule is fully justified by the stochastic calculus theory. Since Y_t in (62) is adapted to W, it should be the right solution (if unique); then, Z serves as a control to make the equation (61) valid (with Y adapted).

5.1.2 Feynman–Kac Formula

The connection with PDE is possible when the terminal condition is a function of a (forward) SDE: this case is called *Markovian BSDE*. Additionally, the driver may depend also on this SDE as $g(s, X_s, Y_s, Z_s)$ for a deterministic function g. We now proceed by a verification theorem. To allow a more natural presentation as backward system, we choose to write the semi-linear PDE with a terminal condition at time T instead of an initial condition at time 0.

Theorem 21. *Let $T > 0$ be given. Under the assumptions of Theorem 11, let X^x be the solution (46) starting from $x \in \mathbb{R}^d$, assume there is a solution $v \in \mathscr{C}_b^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ to the semi-linear PDE*

$$\begin{cases} v'_t(t, x) + L^X_{b,\sigma\sigma^\top} v(t, x) + g(t, x, v(t, x), Dv(t, (x)\sigma(x)) = 0, \\ v(T, x) = f(x), \end{cases} \tag{63}$$

for two given functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : [0, T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R} \otimes \mathbb{R}^d) \to \mathbb{R}$. Then, $Y_t^x = v(t, X_t^x)$ and $Z_t^x = [Dv\,\sigma](t, X_t^x)$ solves the BSDE

$$Y_t^x = f(X_T^x) + \int_t^T g(s, X_s^x, Y_s^x, Z_s^x)\mathrm{d}s - \int_t^T Z_s^x \mathrm{d}W_s. \tag{64}$$

Proof. The Itô formula (50) applied to v and X^x gives

$$\begin{aligned} \mathrm{d}v(s, X_s^x) &= \big[v'_s(s, X_s^x) + L^X_{b,\sigma\sigma^\top} v(s, X_s^x)\big]\mathrm{d}s + Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s \\ &= -g(s, X_s^x, v(s, X_s^x), [Dv\,\sigma](s, X_s^x))\mathrm{d}s + Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s, \end{aligned}$$

which writes between $s = t$ and $s = T$:

$$v(T, X_T^x) = v(t, X_t^x) - \int_t^T g(s, X_s^x, v(s, X_s^x), [Dv\,\sigma](s, X_s^x))\mathrm{d}s + \int_t^T Dv(s, X_s^x)\sigma(X_s^x)\mathrm{d}W_s.$$

Since $v(T, .) = f(.)$, we complete the proof by identification. □

In particular, at time 0 where $X_0^x = x$, we obtain $Y_0^x = v(0, x)$ and in view of (62), it gives a Feynman–Kac representation to v:

$$v(0, x) = \mathbb{E}\Big(f(X_T^x) + \int_t^T g(s, X_s^x, Y_s^x, Z_s^x)\mathrm{d}s\Big). \tag{65}$$

As in case of linear PDEs, the assumption of uniform smoothness on v up to T is too strong to include the case of non-smooth terminal function f. But with an extra ellipticity condition, as for the heat equation, the solution becomes smooth immediately away from T (see [21]) and a similar verification could be checked under milder conditions.

The above Backward SDE (64) is coupled to a Forward SDE, but the latter is not coupled to the BSDE. Another interesting extension is to allow the coupling in both directions by having the coefficients of X dependent on v, i.e. $b(x)$ and $\sigma(x)$ become functions of $x, v(t, x), Dv(t, (x)$. The resulting process is called a Forward Backward Stochastic Differential Equations and is related to Quasi-Linear PDEs, where the operator $L^X_{b,\sigma\sigma^\top}$ also depends on v and Dv, see [56].

5.1.3 Other Existence Results Without PDE Framework

So far, only Markovian BSDEs are presented but from the probabilistic point of view, the Markovian structure is not required to define a solution: what is really crucial is the ability to represent a random variable built from $(W_s : s \le T)$ as a stochastic integral w.r.t. the Brownian motion. This point has been discussed in Corollary 4. Then, in the simple case where g is Lipschitz w.r.t. y, z, (Y, Z) are built by means of an usual fixed point procedure in suitable L_2-norms and of this stochastic integral representation. We now state a more general existence and uniqueness result for BSDE, which is valid without any underlying (finite-dimensional) semi-linear PDE, we omit the proof.

Theorem 22. *Let $T > 0$ be fixed and assume the assumptions of Theorem 11 for the existence of X and that*

- *The terminal condition $\xi = f(X_s : s \le T)$ is a square integrable functional of the stochastic process $(X_s : s \le T)$.*

- *The measurable function* $g : [0, T] \times \mathbb{R}^d \times \mathbb{R} \times (\mathbb{R} \otimes \mathbb{R}^d)$ *is uniformly Lipschitz in* (y, z)*:*

$$|g(t, x, y_1, z_1) - g(t, x, y_2, z_2)| \leq C_g(|y_1 - y_2| + |z_1 - z_2|),$$

  uniformly in (t, x).
- *The driver is square integrable at* $(y, z) = (0, 0)$*:* $\mathbb{E}(\int_0^T g^2(t, X_t, 0, 0)dt) < +\infty$.

Then, there exists a unique solution (Y, Z), adapted *and in* L_2*-spaces, to*

$$Y_t = f(X_s : s \leq T) + \int_t^T g(s, X_s, Y_s, Z_s)\mathrm{d}s - \int_t^T Z_s \mathrm{d}W_s.$$

Many works have been done in the last decade to go beyond the case of Lipschitz driver, which may be too stringent for some applications. In particular, having g with quadratic growth in Z is particularly interesting in exponential utility maximization problem (the non-linear PDE term is quadratic in $|Dv|$). This leads to quadratic BSDEs (see for instance [50]). A simple example of such BSDEs can be cooked up from heat equation and Brownian motion. Namely from Corollary 4, for a smooth function f with compact support, set $u(t, x) = \mathbb{E}(\exp(f(x + W_t)))$ and $v(t, y) = u(1 - t, y)$, so that

$$\begin{aligned}
\exp(f(W_1)) &= u(1, 0) + \int_0^1 u'_x(1 - s, W_s)\mathrm{d}W_s, \\
u(1 - t, W_t) &= u(1, 0) + \int_0^t u'_x(1 - s, W_s)\mathrm{d}W_s, \\
v(t, W_t) &= \exp(f(W_1)) - \int_t^1 v'_x(s, W_s)\mathrm{d}W_s,
\end{aligned}$$

and by setting $Y_t = \log(v(t, W_t))$ and $Z_t = v'_x(t, W_t)/Y_t$, we obtain

$$Y_t = f(W_1) + \int_t^1 \frac{1}{2} Z_s^2 \mathrm{d}s - \int_t^1 Z_s \mathrm{d}W_s,$$

which is the simplest quadratic BSDE.

5.2 Time Discretization and Dynamic Programming Equation

5.2.1 Explicit and Implicit Schemes

To perform the simulation, a first stage may be the derivation of a discretization scheme, written backwardly in time (*backward dynamic programming equation*).

For the further analysis, assume that the terminal condition is of the form $\xi = f(X_T)$ where X is standard (forward) SDE.

Consider a time grid with N time steps $\pi = \{0 = t_0 < \cdots < t_i < \cdots < t_N = T\}$, with possibly non uniform time step, and set $|\pi| = \max_i(t_{i+1} - t_i)$. We will suppose later that $|\pi| \to 0$.

We write $\Delta_i = t_{i+1} - t_i$ and $\Delta W_i = W_{t_{i+1}} - W_{t_i}$. Writing the Eq. (64) between times t_i and t_{i+1}, we have

$$Y_{t_i} = Y_{t_{i+1}} + \int_{t_i}^{t_{i+1}} g(s, X_s, Y_s, Z_s)\mathrm{d}s - \int_{t_i}^{t_{i+1}} Z_s \mathrm{d}W_s.$$

Then, by applying simple approximations for $\mathrm{d}s$ and $\mathrm{d}W_s$ integrals and by replacing X by a Euler scheme computed along the grid π (and denoted X^π), we may define the discrete BSDE as

$$(Y^\pi_{t_i}, Z^\pi_{t_i}) = \underset{(Y,Z)\in L_2(\mathscr{F}^\pi_{t_i})}{\arg\min}\ \mathbb{E}(Y^\pi_{t_{i+1}} + \Delta_i g(t_i, X^\pi_{t_i}, Y, Z) - Y - Z\Delta W_i)^2$$

with the initialization $Y^\pi_T = f(X^\pi_T)$ at $i = N$, where $L_2(\mathscr{F}^\pi_{t_i})$ stands for the set of random variables (with appropriate dimension) that are square integrable and depend on the Brownian motion increments $(\Delta W_j : j \le i-1)$. The latter property is the measurability w.r.t. the sigma-field $\mathscr{F}^\pi_{t_i}$ generated by $(\Delta W_j : j \le i-1)$.

Then, a direct computation using the properties of Brownian increments gives

$$\begin{cases} Y^\pi_T = f(X^\pi_T), \\ Z^\pi_{t_i} = \dfrac{1}{\Delta_i}\mathbb{E}(Y^\pi_{t_{i+1}}\Delta W_i^\top | \mathscr{F}^\pi_{t_i}), \quad i < N \\ Y^\pi_{t_i} = \mathbb{E}(Y^\pi_{t_{i+1}} + \Delta_i g(t_i, X^\pi_{t_i}, Y^\pi_{t_i}, Z^\pi_{t_i})|\mathscr{F}^\pi_{t_i}), \quad i < N. \end{cases} \tag{66}$$

This is the *implicit scheme* since the arguments of the function at the r.h.s. depend on the quantity $Y^\pi_{t_i}$ to compute on the l.h.s. Nevertheless, since g is uniformly Lipschitz in y, it is not difficult to show that the Dynamic Programming Equation (DPE in short) (66) is well-defined for $|\pi|$ small enough and that $Y^\pi_{t_i}$ can be computed using a Picard iteration procedure.

It is easy to turn the previous scheme into an explicit scheme and therefore, to avoid this extra Picard procedure. It writes

$$\begin{cases} Y^\pi_T = f(X^\pi_T), \\ Z^\pi_{t_i} = \dfrac{1}{\Delta_i}\mathbb{E}(Y^\pi_{t_{i+1}}\Delta W_i^\top | \mathscr{F}^\pi_{t_i}), \quad i < N \\ Y^\pi_{t_i} = \mathbb{E}(Y^\pi_{t_{i+1}} + \Delta_i g(t_i, X^\pi_{t_i}, Y^\pi_{t_{i+1}}, Z^\pi_{t_i})|\mathscr{F}^\pi_{t_i}), \quad i < N. \end{cases} \tag{67}$$

In our personal experience on numerics, we have not observed a significant outperformance of one scheme on another. Moreover, from the theoretical point

of view, both schemes exhibit the same rates of convergence w.r.t. $|\pi|$, at least when the driver is Lipschitz.

The explicit scheme is the simplest one, and this is the one that we recommend in practice.

5.2.2 Time Discretization Error

Define the measure of the quadratic error

$$\mathscr{E}(Y^\pi - Y, Z^\pi - Z) = \max_{0\le i\le N} \mathbb{E}|Y^\pi_{t_i} - Y_{t_i}|^2 + \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}|Z^\pi_{t_i} - Z_t|^2 \mathrm{d}t.$$

Although not explicitly mentioned in the previous existence results on BSDE, this type of norm is appropriate to perform the fixed point argument in the proof of Theorem 22. We now state an error estimate [33], in order to show the convergence of the DPE to the BSDE.

Theorem 23. *For a Lipschitz driver w.r.t.* (x, y, z) *and* $\frac{1}{2}$*-Hölder w.r.t.* t*, there is a constant* C *independent on* π *such that we have*

$$\mathscr{E}(Y^\pi - Y, Z^\pi - Z) \le C\Big(|\pi| + \sup_{i\le N} \mathbb{E}|X^\pi_{t_i} - X_{t_i}|^2 + \mathbb{E}|f(X^\pi_T) - f(X_T)|^2 + \sum_{i=0}^{N-1} \frac{1}{\Delta_i} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \mathbb{E}|Z_t - Z_s|^2 \mathrm{d}s\, \mathrm{d}t\Big).$$

Let us discuss on the nature and the magnitude of different error contributions.

- First, we face the *strong approximation error* of the forward SDE by its Euler scheme. Here we rather focus on convergence of paths (in L_2-norms), whereas in Sect. 4.1.3, we have studied the convergence of expectations of function of X^π_T towards those of X_T. Anyway, the problem is now well-understood: under a Lipschitz condition on b and σ, we can prove $\sup_{i\le N} \mathbb{E}|X^\pi_{t_i} - X_{t_i}|^2 = O(|\pi|)$.
- Second, we should ensure a good *strong approximation* of the terminal conditions: if f is Lipschitz continuous, it readily follows from the previous term and $\mathbb{E}|f(X^\pi_T) - f(X_T)|^2 = O(|\pi|)$. For non Lipschitz f, there are partial answers, see [3].
- Finally, the last contribution $\sum_{i=0}^{N-1} \frac{1}{\Delta_i} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \mathbb{E}|Z_t - Z_s|^2 \mathrm{d}s\, \mathrm{d}t$ is related to the *L_2-regularity of Z* (or equivalently of the gradient of the semi-linear PDE along the X-path) and it is intrinsic to the BSDE-solution. For smooth data, Z has the same regularity of Brownian paths and this error term is $O(|\pi|)$. For non smooth f (but under ellipticity condition on X), the L_2-norm of Z_t blows up as $t \to T$ and the rate $|\pi|$ usually worsens: for instance for $f(x) = 1_{x\ge 0}$, it becomes $N^{-\frac{1}{2}}$ for uniform time-grid.

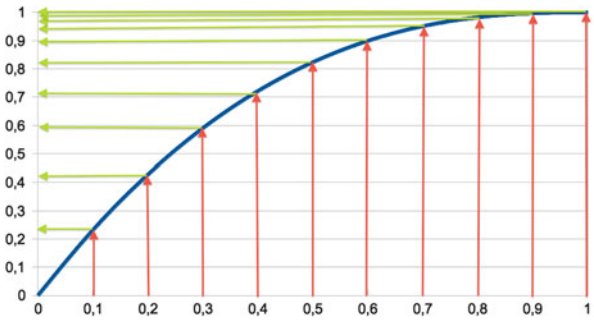


Fig. 8 On the horizontal axis, uniform grid. On the vertical axis, the grid $(t_k^{\bar\theta} : 0 \le k \le N)$, with $T = 1$

The analysis is very closely related to the fractional smoothness of $f(X_T)$ briefly discussed in Sect. 4.1.5, see also [25]. Choosing an appropriate grid of the form (see Fig. 8)

$$t_k^{\bar\theta} = T - T(1 - k/N)^{1/\bar\theta} \quad (\bar\theta \in (0, 1])$$

compensates this blow-up (for appropriate value of $\bar\theta$) and enables to retrieve the rate N^{-1}.

Actually in [31], it is shown that the upper bounds in Theorem 23 can be refined for smooth data, to finally obtain that the main error comes from *strong approximation error* on the forward component. This is an incentive to accurately approximate the SDE in L_2-sense.

5.2.3 Towards the Resolution of the Dynamic Programming Equation

The effective implementation of the explicit scheme (67) requires the iterative computations of conditional expectations: this is discussed in the next paragraphs.

Prior to this, we make some preliminary simplifications for the sake of conciseness. First, we consider the case of g independent of z,

$$g(t, x, y, z) = g(t, x, y),$$

therefore we only approximate Y^π; the general case is detailed in [39,54]. Second, it can be easily seen that it is enough to take the conditioning w.r.t. $X_{t_i}^\pi$ instead of $\mathscr{F}_{t_i}^\pi$, because of the Markovian property of X^π along the grid π and of the independent Brownian increments. Thus, (67) becomes

$$\begin{cases} Y_T^\pi = f(X_T^\pi), \\ Y_{t_i}^\pi = \mathbb{E}(Y_{t_{i+1}}^\pi + \Delta_i g(t_i, X_{t_i}^\pi, Y_{t_{i+1}}^\pi) | X_{t_i}^\pi), \quad i < N. \end{cases} \tag{68}$$

The same arguments apply to assert that for a (measurable) deterministic function y_i^π we have

$$y_i^\pi(X_{t_i}^\pi) = Y_{t_i}^\pi. \tag{69}$$

Therefore, simulating Y^π is equivalent to the computation of the functions y_i^π for any i and the simulation of the process X^π.

5.3 Approximation of Conditional Expectations Using Least-Squares Method

5.3.1 Empirical Least-Squares Problem

We adopt the point of view of conditional expectation as a *projection operator in* L_2. This is not the only possible approach, but it has the advantages (as it will be seen later)

1. To be much flexible w.r.t. the knowledge on the model for X (or X^π): only independent simulations of X^π are required (which is straightforward to perform).
2. To be little demanding on the assumptions on the underlying stochastic model: in particular, no ellipticity nor degeneracy condition are required, it could also include jumps (corresponding to PDE with a non-local Integro-Differential operator).
3. To provide robust theoretical error estimates, which allow to optimally tune the convergence parameters.
4. To be possibly adaptive to the data (*data-driven* scheme).

We recall that if a scalar random variable R (called the *response*) is square integrable, the conditional expectation of R given another possibly multidimensional r.v. O (called the *observation*) is given by

$$\mathbb{E}(R|O) = \underset{m(O) \text{ s.t. } m(.) \text{ is a meas. funct. with } \mathbb{E}|m(O)|^2<+\infty}{\text{Arg min}} \mathbb{E}|R - m(O)|^2.$$

This is a least-squares problem in infinite dimension, also called *regression problem*. Usually in this context of BSDE simulation, none of the distributions of O, R or (O, R) is known in analytical and tractable form: thus an exact computation of $\mathbb{E}(R|O)$ is hopeless. The difficulty remains unchanged if we approximate the regression function

$$m(\cdot) = \mathbb{E}(R|O = \cdot)$$

on a finite dimensional functions basis. Alternatively, we can rely on independent simulations of (O, R) to compute an empirical version of m. This is the approach subsequently developed.

The basis functions are $(\phi_k(.))_{1\le k\le K}$ and we assume that $\mathbb{E}|\phi_k(O)|^2 < +\infty$ for any k. We emphasize that

we can not assume that $(\phi_k(O))_{1\le k\le K}$ *forms an orthonormal basis in* L_2,

since in our setting, the distribution of O is not explicit. Using this finite dimensional approximation, we anticipate to unfortunately retrieve the curse of dimensionality: the larger the dimension d of O, the larger the required K for a good accuracy of m, the larger the complexity.

We compute the coefficients on the basis by solving a *empirical least-squares problem*

$$(\alpha_k^M)_k = \arg\min_{\alpha\in\mathbb{R}^K} \frac{1}{M}\sum_{i=1}^{M}(R_i - \sum_{k=1}^{K}\alpha_k\phi_k(O_i))^2,$$

where $(O_i, R_i)_{1\le i\le M}$ are independent simulations of the couple (O, R). Then, for the approximation of m, we set

$$\tilde{m}_M(.) = \sum_{k=1}^{K}\alpha_k^M\phi_k(.).$$

To efficiently compute the coefficients $(\alpha_k^M)_k$, we might use a SVD decomposition to account for instability issues, see [41].

5.3.2 Model-Free Error Estimates

Without extra assumptions on the model, we can derive model-free error estimates, see [42].

Theorem 24. *Assume that*

- $R = m(O) + \epsilon$ *with* $\mathbb{E}(\epsilon|O) = 0$.[21]
- $(O_1, R_1), \cdots, (O_M, R_M)$ *are independent copies of* (O, R).
- $\sigma^2 = \sup_x \mathbb{V}\mathrm{ar}(R|O = x) < +\infty$.
- *Let* K *be a finite positive integer and* Φ *be the linear vector space spanned by some functions* $(\phi_1, \ldots \phi_K)$, *with* $dim(\Phi) \le K$.[22]

[21] Meaning that $m(O) = \mathbb{E}(R|O)$.

[22] There may be some colinearities within $(\phi_j)_{1\le j\le K}$.

Denote by μ^M the empirical measure associated to $(O_1, \cdots, O_M)$, μ the probability measure of O and by $|\phi|_M^2 = \frac{1}{M}\sum_{i=1}^M \phi^2(O_i)$ the empirical L_2-measure of ϕ w.r.t. μ^M, and set:

$$\tilde{m}_M(.) = \arg\min_{\phi\in\Phi} \frac{1}{M}\sum_{i=1}^M |\phi(O_i) - R_i|^2. \tag{70}$$

Then

$$\mathbb{E}(|\tilde{m}_M - m|_M^2) \leq \sigma^2 \frac{K}{M} + \min_{\phi\in\Phi} |\phi - m|_{L_2(\mu)}^2.$$

The first term in the r.h.s. above is interpreted as a *statistical error*[23] *term* (due to a finite sample to compute the empirical coefficients), while the second term is an *approximation error of the functions class*[24] (due to finite-dimensional vector space). The first term converges to 0 as $M \to +\infty$ but it blows up if $K \to +\infty$, while the second one converges to 0 as $K \to +\infty$ (as least if Φ asymptotically spans all the functions in $L_2(\mu)$). This bias-variance decomposition shows that there is a necessary trade-off between K and M to ensure a convergent approximation. Without this right balance, the approximation (70) may be not convergent. Furthermore, the parameter tuning can also be made optimally.

In the quoted reference [42], the space Φ could also depend on the simulations (data-driven approximation spaces).

Proof. Assume that

$$\mathbb{E}\Big(|\tilde{m}_M - m|_M^2 \big| O_1, \cdots, O_M\Big) \leq \sigma^2 \frac{K}{M} + \min_{\phi\in\Phi} |\phi - m|_M^2. \tag{71}$$

Then, the announced result directly follows by taking expectations and observing that

$$\mathbb{E}\big(\min_{\phi\in\Phi} |\phi - m|_M^2\big) \leq \min_{\phi\in\Phi} \mathbb{E}(|\phi - m|_M^2) = \min_{\phi\in\Phi} |\phi - m|_{L_2(\mu)}^2.$$

We now prove (71). As far as computations conditionally on $O_1, \cdots, O_M$ are concerned, without loss of generality we can assume that $(\phi_1, \ldots \phi_{K_M})$ is an orthonormal family in $L_2(\mu^M)$, with possibly $K_M \leq K$:

$$\frac{1}{M}\sum_{i=1}^M \phi_k(O_i)\phi_l(O_i) = \delta_{k,l}.$$

[23] Also called variance term.

[24] Squared bias term.

Consequently, the solution $\arg\min_{\phi\in\Phi}\frac{1}{M}\sum_{i=1}^{M}|\phi(O_i)-R_i|^2$ is given by

$$\tilde{m}_M(.)=\sum_{j=1}^{K_M}\alpha_j\phi_j(.)\quad\text{with}\quad\alpha_j=\frac{1}{M}\sum_{i=1}^{M}\phi_j(O_i)R_i.$$

Now, set $\mathbb{E}^*(.)=\mathbb{E}(.|O_1,\cdots,O_M)$. Then, observe that $\mathbb{E}^*(\tilde{m}_M(.))$ is the least-squares solution to $\min_{\phi\in\Phi}\frac{1}{M}\sum_{i=1}^{M}|\phi(O_i)-m(O_i)|^2=\min_{\phi\in\Phi}|\phi-m|_M^2$. Indeed:

- On the one hand, the above least-squares solution is given by $\sum_{j=1}^{K_M}\alpha_j^*\phi_j(.)$ with $\alpha_j^*=\frac{1}{M}\sum_{i=1}^{M}\phi_j(O_i)m(O_i)$.
- On the other hand, $\mathbb{E}^*(\tilde{m}_M(.))=\sum_{j=1}^{K_M}\mathbb{E}^*(\alpha_j)\phi_j(.)$ and $\mathbb{E}^*(\alpha_j)=\frac{1}{M}\sum_{i=1}^{M}\phi_j(O_i)\mathbb{E}^*(R_i)=\frac{1}{M}\sum_{i=1}^{M}\phi_j(O_i)\mathbb{E}(m(O_i)+\epsilon_i|O_1,\cdots,O_M)=\alpha_j^*$.

Thus, by the Pythagoras theorem, we obtain

$$\begin{aligned}|\tilde{m}_M-m|_M^2&=|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2+|\mathbb{E}^*(\tilde{m}_M)-m|_M^2,\\ \mathbb{E}^*|\tilde{m}_M-m|_M^2&=\mathbb{E}^*|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2+|\mathbb{E}^*(\tilde{m}_M)-m|_M^2\\&=\mathbb{E}^*|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2+\min_{\phi\in\Phi}|\phi-m|_M^2.\end{aligned}$$

Since $(\phi_j)_j$ is orthonormal in $L_2(\mu_M)$, we have $|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2=\sum_{j=1}^{K_M}|\alpha_j-\mathbb{E}^*(\alpha_j)|^2$. Since $\alpha_j-\mathbb{E}^*(\alpha_j)=\frac{1}{M}\sum_{i=1}^{M}\phi_j(O_i)(R_i-m(O_i))$, we obtain

$$\begin{aligned}\mathbb{E}^*|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2&=\sum_{j=1}^{K_M}\frac{1}{M^2}\mathbb{E}^*\sum_{i,l=1}^{M}\phi_j(O_i)\phi_j(O_l)(R_i-m(O_i))(R_l-m(O_l)\\&=\sum_{j=1}^{K_M}\frac{1}{M^2}\sum_{i=1}^{M}\phi_j^2(O_i)\mathbb{V}\mathrm{ar}(R_i|O_i)\end{aligned}$$

taking advantage that the $(\epsilon_i)_i$ conditionally on $(O_1,\cdots O_M)$ are centered. This proves

$$\mathbb{E}^*|\tilde{m}_M-\mathbb{E}^*(\tilde{m}_M)|_M^2\le\sigma^2\sum_{j=1}^{K_M}\frac{1}{M^2}\sum_{i=1}^{M}\phi_j^2(O_i)=\sigma^2\frac{K_M}{M}\le\sigma^2\frac{K}{M}.$$

The proof of (71) is complete. □

5.3.3 Least-Squares Method for Solving Discrete BSDE

We now apply the previous empirical least-squares method to numerically solve the DPE (68). For simplicity of exposure, we consider here only uniform time grids with N time steps, $\Delta_i = T/N$. In addition to assumptions of Theorem 23, we assume that the terminal condition $f(.)$ is bounded: then, we can easily establish the following result.

Proposition 20. *Under these assumptions, the function $y_i^\pi(.)$ defined in (69) is bounded by a constant $C_\star$, which is independent on N and i.*

Actually, $C_\star$ can be given explicitly in terms of the data. To force the stability in the iterative computations of conditional expectations (68), we truncate the numerical solution at the level $C_\star$ using the soft thresholding

$$[\psi]_{C_\star} := -C_\star \vee \psi \wedge C_\star.$$

Algorithm for Approximating $y_k^\pi(\cdot)$. At each time index $0 \leq k \leq N-1$, we consider a vector space Φ_k spanned by basis functions $p_k(\cdot)$, which are understood as vectors of K_k functions. The final approximation of $y_k^\pi(\cdot)$ has the form

$$y_k^{\pi,M}(\cdot) = [\alpha_k^M \cdot p_k(\cdot)]_{C_\star}.$$

The coefficients α_k^M are computed with M independent simulations of $(X_{t_k}^\pi)_k$, that are denoted by $\{(X_{t_k}^{\pi,m})_k\}_{1\leq m\leq M}$: this single set of simulated paths are used to compute all the coefficients at once. This is done as follows:

- Initialization : for $k = N$, take $y_N^\pi(\cdot) = f(\cdot)$.
- Iteration : for $k = N-1, \cdots, 0$, solve the least-squares problem

$$\alpha_k^M = \arg\min_{\alpha\in\mathbb{R}^{K_k}} \sum_{m=1}^{M} |y_{k+1}^{\pi,M}(X_{t_{k+1}}^{\pi,m}) + \Delta_k g(t_k, X_{t_k}^{\pi,m}, y_{k+1}^{\pi,M}(X_{t_{k+1}}^{\pi,m})) - \alpha\cdot p_k(X_{t_k}^{\pi,m})|^2$$

and define $y_k^{\pi,M}(\cdot) = [\alpha_k^M \cdot p_k(\cdot)]_{C_\star}$.

Error Analysis. We now turn to the error estimates. The analysis combines the BSDE techniques (a priori estimates using stochastic calculus), regression tools as those exposed in Sect. 5.3.2, but there is a slight difference which actually requires a significant improvement in the arguments. Since we use a single set of independent paths, the "responses" $(y_{k+1}^{\pi,M}(X_{t_{k+1}}^{\pi,m}))_{1\leq m\leq M}$ are not independent, because of their dependence through the function $y_{k+1}^{\pi,M}$. To overcome this interdependence issue in the proof, we shall replace the random function $y_{k+1}^{\pi,M}$ by a deterministic neighbor: of course, there is a complexity cost to cover the different function spaces in order to provide close neighbors, and the covering numbers are well controlled using the Vapnik–Chervonenkis dimension, when the function spaces are *bounded*

(Proposition 20). This is the technical reason why we consider bounded functions. We now state a result regarding the global error, see [30, Theorem VIII.3.4] for full details.

Theorem 25. *Under the previous notations and assumptions, there is a constant $C > 0$ (independent on N) such that we have*

$$\max_{0\le k\le N} \mathbb{E}|Y^{\pi}_{t_k} - y^{\pi,M}_k(X^{\pi}_{t_k})|^2 \le C \sum_{k=0}^{N-1} \Bigg\{ N \underbrace{\frac{K_k}{M}}_{\text{statistical error}} + \underbrace{\min_{\phi\in\Phi_k} \mathbb{E}|y^{\pi}_k(X^{\pi}_{t_k}) - \phi(X^{\pi}_{t_k})|^2}_{\text{approximation error of functions class}} \Bigg\}$$

$$+ C \max_{0\le k\le N} \underbrace{\sqrt{\frac{K_k \log(M)}{M}}}_{\text{interdependence error}} .$$

When the Z-component has to be approximated as well, the estimates are slightly modified, see [54] for details.

Parameter Tuning. We conclude this analysis by providing an example of how to choose appropriately the parameters N, K_k and M. Assume that the value function y^{π} is Lipschitz continuous, uniformly in N (which usually follows from a Lipschitz terminal condition). Our objective is to achieve a global error of order $\varepsilon = \frac{1}{N}$ for $\max_{0\le k\le N} \mathbb{E}|Y^{\pi}_{t_k} - y^{\pi,M}_k(X^N_{t_k})|^2$, i.e. the same error magnitude than the time-discretization error.

For the vector spaces Φ_k, we consider those generated by functions that are constant on disjoint hypercubes of small edge. Since X^{π} has exponential moments, we can restrict the partitioning to a compact set of $\mathbb{R}^d$ with size $c\log(N)$ in any direction, and the induced error is smaller than N^{-1} provided that c is large enough. If the edge of the hypercube is like N^{-1}, the vector spaces have dimension $K_k \sim N^d$ up to logarithmic factors: then, the terms from *approximation error of functions class* are $O(N^{-2})$ and they sum up to give a contribution $O(N^{-1})$ as required. A quick inspection of the upper bounds of Theorem 25 shows that the highest constraint on M comes from the statistical error: we obtain $M \sim cN^{3+d}$, up to logarithmic terms. The complexity of the scheme is of order NM (still neglecting the log terms), because the computation of all regression coefficients at a given date has a computational cost $O(M\log(N))$ due to our specific choice of function basis. Hence, the global complexity is

$$\mathscr{C} \sim \varepsilon^{-\frac{1}{4+d}}$$

up to logarithmic terms. Not surprisingly, the convergence order deteriorates as the dimension increases, this is the curse of dimensionality. Had the value function been smoother, we would have used local polynomials and the convergence order would have been improved: the smoother the functions, the better the convergence rate.

In practice, the algorithm has been performed on a computer up to dimension $d = 10$ with satisfactory results and rather short computational times (less than 1 min). There are several possible improvements to this basic version of the algorithm.

- We can use variance reduction techniques, see [8, 9].
- Instead of writing the DPE between t_i and t_{i+1}, it can be written between t_i and T: it has the surprising effect (mathematically justified) to reduce the propagation of errors in the DPE. This scheme is called MDP scheme (for Multi step forward Dynamic Programming equation) and it is studied in [39].

Complementary References. For theoretical aspects, see [16, 56, 59, 61]; for applications, see [17, 18]; for numerics, see [5, 7, 11, 14, 32, 40, 54, 69]. This list is not exhaustive.

Acknowledgements The author's research is partly supported by the Chair *Financial Risks* of the Risk Foundation and the *Finance for Energy Market Research Centre*.

References

1. Achdou, Y., Pironneau, O.: Computational Methods for Option Pricing. SIAM Series. Frontiers in Applied Mathematics, Philadelphia (2005)
2. Asmussen, S., Glynn, P.W.: Stochastic Simulation: Algorithms and Analysis. Stochastic Modelling and Applied Probability, vol. 57. Springer, New York (2007)
3. Avikainen, R.: On irregular functionals of SDEs and the Euler scheme. Financ. Stoch. **13**, 381–401 (2009)
4. Bachelier, L.: Théorie de la spéculation. Ph.D. thesis, Ann. Sci. École Norm. Sup. (1900)
5. Bally, V., Pagès, G.: Error analysis of the optimal quantization algorithm for obstacle problems. Stoch. Process. Appl. **106**(1), 1–40 (2003)
6. Bally, V., Talay, D.: The law of the Euler scheme for stochastic differential equations: I. Convergence rate of the distribution function. Probab. Theory Relat. Fields **104**(1), 43–60 (1996)
7. Bender, C., Denk, R.: A forward scheme for backward SDEs. Stoch. Process. Their Appl. **117**(12), 1793–1823 (2007)
8. Bender, C., Steiner, J.: Least-squares Monte Carlo for BSDEs. In: Carmona, R., Del Moral, P., Hu, P., Oudjane, N. (eds.) Numerical Methods in Finance. Series: Springer Proceedings in Mathematics, vol. 12. Springer, Berlin (2012)
9. Ben Zineb, T., Gobet, E.: Preliminary control variates to improve empirical regression methods. Monte Carlo Methods Appl. **19**(4), 331–354 (2013)
10. Bossy, M., Gobet, E., Talay, D.: Symmetrized Euler scheme for an efficient approximation of reflected diffusions. J. Appl. Probab. **41**(3), 877–889 (2004)
11. Bouchard, B., Touzi, N.: Discrete time approximation and Monte Carlo simulation of backward stochastic differential equations. Stoch. Process. Their Appl. **111**, 175–206 (2004)
12. Breiman, L.: Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1992) [Corrected reprint of the 1968 original]
13. Cessenat, M., Dautray, R., Ledanois, G., Lions, P.L., Pardoux, E., Sentis, R.: Méthodes probabilistes pour les équations de la physique. Collection CEA, Eyrolles (1989)
14. Crisan, D., Manolarakis, K.: Solving backward stochastic differential equations using the cubature method. Preprint (2010)

15. Durrett, R.: Brownian Motion and Martingales in Analysis. Wadsworth Mathematics Series. Wadsworth International Group, Belmont (1984)
16. El Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S., Quenez, M.C.: Reflected solutions of backward SDE's and related obstacle problems for PDE's. Ann. Probab. **25**(2), 702–737 (1997)
17. El Karoui, N., Peng, S.G., Quenez, M.C.: Backward stochastic differential equations in finance. Math. Financ. **7**(1), 1–71 (1997)
18. El Karoui, N., Hamadène, S., Matoussi, A.: Backward stochastic differential equations and applications. In: Carmona, R. (ed.) Indifference Pricing: Theory and Applications, Chap. 8, pp. 267–320. Princeton University Press, Princeton (2008)
19. Föllmer, H.: Calcul d'Itô sans probabilités. In: Seminar on Probability, XV (University of Strasbourg, Strasbourg, 1979/1980) (French), pp. 143–150. Springer, Berlin (1981)
20. Freidlin, M.I.: Functional Integration and Partial Differential Equations. Annals of Mathematics Studies. Princeton University Press, Princeton (1985)
21. Friedman, A.: Partial Differential Equations of Parabolic Type. Prentice-Hall, Eglewood Cliffs (1964)
22. Friedman, A.: Stochastic Differential Equations and Applications, vol. 1. Academic, New York (1975) [A Subsidiary of Harcourt Brace Jovanovich, Publishers, XIII]
23. Friedman, A.: Stochastic Differential Equations and Applications, vol. 2. Academic, New York (1976) [A Subsidiary of Harcourt Brace Jovanovich, Publishers, XIII]
24. Gavalas, G.R.: Nonlinear Differential Equations of Chemically Reacting Systems. Springer Tracts in Natural Philosophy, vol. 17. Springer, New York (1968)
25. Geiss, C., Geiss, S., Gobet, E.: Generalized fractional smoothness and L_p-variation of BSDEs with non-Lipschitz terminal condition. Stoch. Process. Their Appli. **122**(5), 2078–2116 (2012)
26. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2003)
27. Gobet, E.: Euler schemes for the weak approximation of killed diffusion. Stoch. Process. Their Appl. **87**, 167–197 (2000)
28. Gobet, E.: Euler schemes and half-space approximation for the simulation of diffusions in a domain. ESAIM Probab. Stat. **5**, 261–297 (2001)
29. Gobet, E.: Advanced Monte Carlo methods for barrier and related exotic options. In: Ciarlet, P.G., Bensoussan, A., Zhang, Q. (eds.) Handbook of Numerical Analysis, vol. XV. Special Volume: Mathematical Modeling and Numerical Methods in Finance, pp. 497–528. Elsevier/North-Holland, Netherlands (2009)
30. Gobet, E.: Méthodes de Monte-Carlo et processus stochastiques: du linéaire au non-linéaire. Editions de l'École Polytechnique, Palaiseau (2013)
31. Gobet, E., Labart, C.: Error expansion for the discretization of backward stochastic differential equations. Stoch. Process. Their Appl. **117**(7), 803–829 (2007)
32. Gobet, E., Labart, C.: Solving BSDE with adaptive control variate. SIAM Numer. Anal. **48**(1), 257–277 (2010)
33. Gobet, E., Lemor, J.P.: Numerical simulation of BSDEs using empirical regression methods: Theory and practice. In: Proceedings of the Fifth Colloquium on BSDEs, 29th May–1st June 2005, Shangai. Available at http://hal.archives-ouvertes.fr/hal-00291199/fr/ (2006)
34. Gobet, E., Maire, S.: Sequential control variates for functionals of Markov processes. SIAM J. Numer. Anal. **43**(3), 1256–1275 (2005)
35. Gobet, E., Maire, S.: Sequential Monte Carlo domain decomposition for the Poisson equation. In: Proceedings of the 17th IMACS World Congress, Scientific Computation, Applied Mathematics and Simulation, Paris, 11–15 July 2005
36. Gobet, E., Makhlouf, A.: L_2-time regularity of BSDEs with irregular terminal functions. Stoch. Proces. Their Appli. **120**, 1105–1132 (2010)
37. Gobet, E., Menozzi, S.: Stopped diffusion processes: Boundary corrections and overshoot. Stoch. Process. Their Appl. **120**, 130–162 (2010)
38. Gobet, E., Munos, R.: Sensitivity analysis using Itô-Malliavin calculus and martingales. Application to stochastic control problem. SIAM J. Control Optim. 43(5), 1676–1713 (2005)

39. Gobet, E., Turkedjiev, P.: Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions (2013) [In Revision for Mathematics of Computation]. Available at http://hal.archives-ouvertes.fr/hal-00642685
40. Gobet, E., Lemor, J.P., Warin, X.: A regression-based Monte Carlo method to solve backward stochastic differential equations. Ann. Appl. Probab. **15**(3), 2172–2202 (2005)
41. Golub, G., Van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Balimore (1996)
42. Gyorfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics. Springer, New York (2002)
43. Hodgkin, A.L., Huxley, A.F., Katz, B.: Measurement of current-voltage relations in the membrane of the giant axon of Loligo. J. Physiol. **116**, 424–448 (1952). Available at http://www.sfn.org/skins/main/pdf/HistoryofNeuroscience/hodgkin1.pdf
44. Ito, K.: On stochastic differential equations. Mem. Am. Math. Soc. (4), 1–51 (1951)
45. Itô, K., McKean, H.P.: Diffusion Processes and Their Sample Paths. Springer, Berlin (1965)
46. Jacod, J., Protter, P.: Probability Essentials, 2nd edn. Springer, Berlin (2003)
47. Jourdain, B., Lelong, J.: Robust adaptive importance sampling for normal random vectors. Ann. Appl. Probab. **19**(5), 1687–1718 (2009)
48. Karatzas, I., Shreve, S.E.: Brownian Motion and Stochastic Calculus, 2nd edn. Springer, New York (1991)
49. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1995)
50. Kobylanski, M.: Backward stochastic differential equations and partial differential equations with quadratic growth. Ann. Probab. **28**(2), 558–602 (2000)
51. Kolmogorov, A.N., Petrovsky, I.G., Piskunov, N.S.: Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. Bulletin Université d'État à Moscou, Série internationale A 1, pp. 1–26 (1937)
52. Kunita, H.: Stochastic differential equations and stochastic flows of diffeomorphisms. In: Ecole d'Eté de Probabilités de St-Flour XII, 1982. Lecture Notes in Mathematics, vol. 1097, pp. 144–305. Springer, Berlin (1984)
53. Lapeyre, B., Pardoux, E., Sentis, R.: Methodes de Monte Carlo pour les processus de transport et de diffusion. Collection Mathématiques et Applications, vol. 29. Springer, Berlin (1998)
54. Lemor, J.P., Gobet, E., Warin, X.: Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations. Bernoulli **12**(5), 889–916 (2006)
55. Lévy, P.: Sur certains processus stochastiques homogènes. Compos. Math. **7**, 283–339 (1939)
56. Ma, J., Yong, J.: Forward-Backward Stochastic Differential Equations. Lecture Notes in Mathematics, vol. 1702. Springer, Berlin (1999)
57. Nelson, E.: Dynamical Theories of Brownian Motion. Princeton University Press, Princeton (1967)
58. Newton, N.J.: Variance reduction for simulated diffusions. SIAM J. Appl. Math. **54**(6), 1780–1805 (1994)
59. Pardoux, E.: Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order. In: Stochastic Analysis and Related Topics, VI (Geilo, 1996). Progress in Probability, vol. 42, pp. 79–127. Birkhäuser, Boston (1998)
60. Pardoux, E., Peng, S.G.: Adapted solution of a backward stochastic differential equation. Syst. Control Lett. **14**(1), 55–61 (1990)
61. Pardoux, E., Peng, S.: Backward stochastic differential equations and quasilinear parabolic partial differential equations. In: Stochastic Partial Differential Equations and Their Applications. Proceedings of the IFIP International Conference, Charlotte/NC, 1991. Lecture Notes in Control and Information Sciences, vol. 176, pp. 200–217. Springer, Berlin (1992)
62. Peirano, É., Talay, D.: Domain decomposition by stochastic methods. In: Domain Decomposition Methods in Science and Engineering, pp. 131–147 (electronic). National Autonomous University of Mexico, México (2003)

63. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion, 3rd edn. Comprehensive Studies in Mathematics. Springer, Berlin (1999)
64. Sabelfeld, K.K.: Monte Carlo Methods in Boundary Value Problems. Springer Series in Computational Physics. Springer, Berlin (1991) [translated from Russian]
65. Samuelson, P.A.: Proof that properly anticipated prices fluctuate randomly. Ind. Manag. Rev. **6**, 42–49 (1965)
66. Shigesada, N., Kawasaki, K. (eds.): Biological Invasions: Theory and Practice. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford (1997)
67. Stroock, D.W., Varadhan, S.R.S.: Multidimensional Diffusion Processes, 2nd edn. Comprehensive Studies in Mathematics. Springer, Berlin (1997)
68. Talay, D., Tubaro, L.: Expansion of the global error for numerical schemes solving stochastic differential equations. Stoch. Anal. Appl. **8**(4), 94–120 (1990)
69. Zhang, J.: A numerical scheme for BSDEs. Ann. Appl. Probab. **14**(1), 459–488 (2004)

Structure-Preserving Shock-Capturing Methods: Late-Time Asymptotics, Curved Geometry, Small-Scale Dissipation, and Nonconservative Products

Philippe G. LeFloch

Abstract We consider weak solutions to nonlinear hyperbolic systems of conservation laws arising in compressible fluid dynamics and we describe recent work on the design of structure-preserving numerical methods. We focus on preserving, on one hand, the late-time asymptotics of solutions and, on the other hand, the geometrical effects that arise in certain applications involving curved space. First, we study here nonlinear hyperbolic systems with stiff relaxation in the late time regime. By performing a singular analysis based on a Chapman–Enskog expansion, we derive an effective system of parabolic type and we introduce a broad class of finite volume schemes which are consistent and accurate even for asymptotically late times. Second, for nonlinear hyperbolic conservation laws posed on a curved manifold, we formulate geometrically consistent finite volume schemes and, by generalizing the Cockburn–Coquel–LeFloch theorem, we establish the strong convergence of the approximate solutions toward entropy solutions.

1 Introduction

1.1 Objective

We present some recent developments on shock capturing methods for nonlinear hyperbolic systems of balance laws, whose prototype is the Euler system of compressible fluid flows, and especially discuss structure-preserving techniques. The problems under consideration arise with complex fluids in realistic applications when friction terms, geometrical terms, viscosity and capillarity effects, etc., need

P.G. LeFloch (✉)
Laboratoire Jacques-Louis Lions, Centre National de la Recherche Scientifique and Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Paris, France
e-mail: contact@philippelefloch.org

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_4,

to be taken into account in order to achieve a proper description of the physical phenomena. For these problems, it is necessary to design numerical methods that are not only consistent with the given partial differential equations, but remain accurate and robust in certain asymptotic regimes of physical interest. That is, certain structural properties of these hyperbolic problems (conservation or balance law, equilibrium state, monotonicity properties, etc.) are essential in many applications, and one seeks that the numerical solutions preserve these properties, which is often a very challenging task.

To be able to design structure-preserving methods, a theoretical analysis of the hyperbolic problems under consideration must be performed first by investigating certain singular limits as well as certain classes of solutions of physical relevance. The mathematical analysis allows one to exhibit the key properties of solutions and derive effective equations that describe the limiting behavior of solutions, etc. This step requires a deep understanding of the initial value problem, as is for instance the case of small-scale dissipation sensitive, viscosity-capillarity driven shock waves which, as it turns out, do not satisfy standard entropy criteria; see LeFloch [45] for a review. Such a study is in many physical applications involving hyperbolic systems in nonconservative form, in order to avoid the appearance of spurious solutions with wrong speed; see Hou and LeFloch [38].

The design of structure-preserving schemes forces us to go beyond the basic property of consistency with the conservative form of the equations, and requires to revisit the standard strategies, based on finite volumes, finite differences, Runge–Kutta techniques, etc. By mimicking the theoretical analysis at the discrete numerical level, we can arrive at structure-preserving schemes, which preserve the relevant structure of the systems and the asymptotic behavior of solutions.

The techniques developed for model problems provide us with the proper tools to tackle the full problems of physical interest. A variety of nonlinear hyperbolic problems arising in the applications do involve small scales or enjoy important structural or asymptotic properties. By going beyond the consistency with the conservation form of the equations, one can now develop a variety of numerical methods that preserve these properties at the discrete level. By avoid physically wrong solutions, one can understand first the physical phenomena in simplified situations, and next contribute to validate the "full" physical models.

We will only review here two techniques which allows one to preserve late-asymptotics and geometrical terms and, for further reading on this broad topic, we refer to the textbooks [12, 45, 56], as well as the lecture notes [44, 47, 49]. Another challenging application arises in continuum physics in the regime of (small) viscosity and capillarity, which may still drive the propagation of certain (nonclassical undercompressive) shock waves. This is relevant in material science for the modeling of smart (martensite) materials, as well as in fluid dynamics for the modeling of multiphase flows (for instance in the context of nuclear plants) and for the coupling of physical models across interfaces.

1.2 Preserving Late-Time Asymptotics with Stiff Relaxation

In Sect. 2, this strategy is developed for a class of hyperbolic systems with stiff relaxation in the regime of late times. Such systems arise in the modeling of a complex multi-fluid flow when two (or more) scales drive the behavior of the flow. Many examples from continuum physics fall into the proposed framework, for instance the Euler equations with (possibly nonlinear) friction. In performing a singular analysis of these hyperbolic systems, we keep in mind the analogy with the passage from Boltzmann equation (microscopic description) to the Navier–Stokes equations (macroscopic description). Our aim here is, first, to derive via a formal Chapman–Enskog expansion an effective system of parabolic type and, second, to design a scheme which provides consistent and accurate discretizations for all times, including asymptotically late times.

Indeed, we propose and analyze a class of asymptotic-preserving finite volume methods, which are consistent with, both, the given nonlinear hyperbolic system and the effective parabolic system. It thus preserves the late-time asymptotic regime and, importantly, requires only a classical CFL (Courant, Friedrichs, Lewy) condition of hyperbolic type, rather than a more restrictive, parabolic-type stability condition. This section is based on the joint work [9, 11].

1.3 Geometry-Preserving Finite Volume Methods

The second topic of interest here is provided by the class of hyperbolic conservation laws posed on a curved space. Such equations are relevant in geophysical applications, for which the prototype is given by shallow water equations on the sphere with topography. Computations of large-scale atmospheric flows and oceanic motions (involving the Coriolis force, Rosby waves, etc.) requires robust numerical methods. Another motivation is provided conservation laws on moving surfaces describing combustion phenomena. We should astrophysical applications, involving fluids or plasmas, and the study of the propagation of linear waves (wave operator, Dirac equations, etc.) on curved backgrounds of general relativity (such as Schwarschild or, more generally, Kerr spacetime). These applications provide important examples where the partial differential equations of interest are naturally posed on a curved manifold.

Scalar conservation laws yield a drastically simplified, yet very challenging, mathematical model for understanding nonlinear aspects of shock wave propagation on manifolds. In Sect. 3, based on the work [52], we introduce the geometry-preserving finite volume method for hyperbolic balance laws formulated on surfaces or, more generally, manifolds. First, we present some theoretical tools to handle the interplay between the nonlinear waves propagating on solutions and the underlying geometry of the problem. A generalization of the standard Kruzkov theory is obtained on a manifold, by formulating the hyperbolic equation under

consideration from a field of differential forms. The proposed finite volume method is geometry-consistent and relies on a coordinate-independent formulation. The actual implementation of this finite volume scheme on the sphere is realized in [3, 5].

2 Late-Time Asymptotics with Stiff Relaxation

2.1 A Class of Nonlinear Hyperbolic Systems of Balance Laws

Consider the following system of partial differential equations

$$\epsilon\, \partial_t U + \partial_x F(U) = -\frac{R(U)}{\epsilon}, \qquad U = U(t,x) \in \Omega \subset \mathbb{R}^N, \tag{1}$$

in which $t > 0$, $x \in \mathbb{R}$ denote the time and space variables and the flux $F : \Omega \to \mathbb{R}^N$ is defined on the convex and open subset Ω. The first-order part of (1) is assumed to be hyperbolic in the sense that the matrix-valued map $A(U) := D_U F(U)$ admits real eigenvalues and a full basis of eigenvectors.

In order to analyze the singular limit $\epsilon \to 0$ of late-time and stiff relaxation, we distinguish between two distinct regimes. In the hyperbolic-to-hyperbolic regime, one replaces $\epsilon \partial_t U$ by $\partial_t U$ and establishes that solutions to

$$\partial_t U + \partial_x F(U) = -\frac{R(U)}{\epsilon}, \qquad U = U(t,x),$$

are driven by an effective system of hyperbolic type. Such a study was pioneered by Chen, Levermore, and Liu [21]. On the other hand, in the hyperbolic-to-parabolic regime which is under consideration in the present work, we obtain effective equations of parabolic type. In the earlier papers [31, 58], Marcati et al. established rigorous convergence theorems for several classes of models. Our objective here is to introduce a general framework to design numerical methods for such problems.

We make the following assumptions.

Condition 1. There exists an $n \times N$ matrix Q with (maximal) rank $n < N$ such that

$$QR(U) = 0, \qquad U \in \Omega, \tag{2}$$

hence, $QU \in Q\Omega =: \omega$ satisfies

$$\epsilon\, \partial_t \big(QU\big) + \partial_x \big(QF(U)\big) = 0. \tag{3}$$

Condition 2. There exists a map $\mathscr{E} : \omega \subset \mathbb{R}^N \to \Omega$ describing the equilibria $u \in \omega$, with

$$R(\mathscr{E}(u)) = 0, \qquad u = Q\,\mathscr{E}(u). \tag{4}$$

We introduce the equilibrium submanifold $\mathscr{M} := \{U = \mathscr{E}(u)\}$.

Condition 3. It is assumed that

$$QF(\mathscr{E}(u)) = 0, \qquad u \in \omega. \tag{5}$$

Observe that the term $\partial_x\big(QF(\mathscr{E}(u))\big)$ must vanish identically, so that $QF(\mathscr{E}(u))$ must be a constant, which we normalize to be 0.

Condition 4. For all $u \in \omega$, we impose

$$\begin{aligned} \dim\Big(\ker(B(\mathscr{E}(u)))\Big) &= n, \\ \ker\big(B(\mathscr{E}(u))\big) \cap \mathrm{Im}\,\big(B(\mathscr{E}(u))\big) &= \{0\}, \end{aligned} \tag{6}$$

hence, the $N \times N$ matrix $B := DR_U$ has "maximal" kernel on the equilibrium manifold.

2.2 Models Arising in Compressible Fluid Dynamics

2.2.1 Stiff Friction

We begin with the Euler system for compressible fluids with friction:

$$\begin{aligned} &\epsilon\,\partial_t\rho + \partial_x(\rho v) = 0, \\ &\epsilon\,\partial_t(\rho v) + \partial_x\big(\rho v^2 + p(\rho)\big) = -\frac{\rho v}{\epsilon}. \end{aligned} \tag{7}$$

The density $\rho \geq 0$ and the velocity v are the main unknowns, while the pressure $p : \mathbb{R}^+ \to \mathbb{R}^+$ is a prescribed function satisfying the hyperbolicity condition $p'(\rho) > 0$ (for $\rho > 0$). The first-order homogeneous system is strictly hyperbolic and (7) fits into our late-time/stiff relaxation framework in Sect. 2.1 if we set

$$U = \begin{pmatrix} \rho \\ \rho v \end{pmatrix}, \quad F(U) = \begin{pmatrix} \rho v \\ \rho v^2 + p(\rho) \end{pmatrix}, \quad R(U) = \begin{pmatrix} 0 \\ \rho v \end{pmatrix}$$

and $Q = (1\ 0)$. The local equilibria $u = \rho$ are found to be scalar-valued with $\mathscr{E}(u) = (\rho, 0)^T$ and we immediately check that $QF(\mathscr{E}(u)) = 0$.

2.2.2 Stiff Radiative Transfer

The following model arises in the theory of radiative transfer:

$$\begin{aligned} &\epsilon\, \partial_t e + \partial_x f = \frac{\tau^4 - e}{\epsilon}, \\ &\epsilon\, \partial_t f + \partial_x \Big(\chi\,(f/e)\, e \Big) = -\frac{f}{\epsilon}, \\ &\epsilon\, \partial_t \tau = \frac{e - \tau^4}{\epsilon}. \end{aligned} \tag{8}$$

The radiative energy $e > 0$ and the radiative flux f are the main unknowns, restricted so that $|f/e| \leq 1$, while $\tau > 0$ is the temperature. The so-called Eddington factor $\chi : [-1, 1] \to \mathbb{R}^+$ is, typically, taken to be $\chi(\xi) = \frac{3+4\xi^2}{5+2\sqrt{4-3\xi^2}}$. Again, this system fits within our general framework.

2.2.3 Coupling Stiff Friction and Stiff Radiative Transfert

By combining the previous two examples together, one can consider to the following coupled Euler/$M1$ model

$$\begin{aligned} &\epsilon \partial_t \rho + \partial_x \big(\rho v\big) = 0, \\ &\epsilon \partial_t \rho v + \partial_x \big(\rho v^2 + p(\rho)\big) = -\frac{\kappa}{\epsilon} \rho v + \frac{\sigma}{\epsilon} f, \\ &\epsilon \partial_t e + \partial_x f = 0, \\ &\epsilon \partial_t f + \partial_x \Big(\chi \left(\frac{f}{e} \right) e \Big) = -\frac{\sigma}{\epsilon} f. \end{aligned} \tag{9}$$

Here, κ and σ are positive constants and, in the applications, a typical choice for the pressure is $p(\rho) = C_p \rho^\eta$ with $C_p \ll 1$ and $\eta > 1$. Now, we should set

$$U = \begin{pmatrix} \rho \\ \rho v \\ e \\ f \end{pmatrix}, \qquad F(U) = \begin{pmatrix} \rho v \\ \rho v^2 + p(\rho) \\ f \\ \chi(\frac{f}{e}) e \end{pmatrix}, \qquad R(U) = \begin{pmatrix} 0 \\ \kappa \rho v - \sigma f \\ 0 \\ \sigma f \end{pmatrix},$$

and the local equilibria read

$$\mathscr{E}(u) = \begin{pmatrix} \rho \\ 0 \\ e \\ 0 \end{pmatrix}, \qquad u = QU = \begin{pmatrix} \rho \\ e \end{pmatrix}, \qquad Q = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix},$$

so that, once again, $QF(\mathscr{E}(u)) = 0$.

2.3 An Expansion Near Equilibria

Our singular analysis proceeds with a Chapman–Engskog expansion around a local equilibria $u = u(t, x) \in \omega$. We set

$$U^\epsilon = \mathscr{E}(u) + \epsilon\, U_1 + \epsilon^2\, U_2 + \ldots, \qquad u := QU^\epsilon,$$

and requires that $\epsilon\, \partial_t U^\epsilon + \partial_x F(U^\epsilon) = -R(U^\epsilon)/\epsilon$. We thus obtain $QU_1 = QU_2 = \ldots = 0$ and then

$$\begin{aligned} F(U^\epsilon) &= F(\mathscr{E}(u)) + \epsilon\, A(\mathscr{E}(u))\, U_1 + \mathscr{O}(\epsilon^2), \\ \frac{R(U^\epsilon)}{\epsilon} &= B(\mathscr{E}(u))\, U_1 + \frac{\epsilon}{2} D_U^2 R(\mathscr{E}(u)).(U_1, U_1) + \epsilon B(\mathscr{E}(u))\, U_2 + \mathscr{O}(\epsilon^2). \end{aligned}$$

In turn, we deduce that

$$\begin{aligned} &\epsilon\, \partial_t \big(\mathscr{E}(u)\big) + \partial_x \big(F(\mathscr{E}(u))\big) + \epsilon\, \partial_x \Big(A(\mathscr{E}(u))\, U_1\Big) \\ &= -B(\mathscr{E}(u))\, U_1 - \frac{\epsilon}{2} D_U^2 R(\mathscr{E}(u)).(U_1, U_1) - \epsilon B(\mathscr{E}(u))\, U_2 + \mathscr{O}(\epsilon^2). \end{aligned}$$

The zero-order terms imply that $U_1 \in \mathbb{R}^N$ satisfies the algebraic system

$$B(\mathscr{E}(u))\, U_1 = -\partial_x \big(F(\mathscr{E}(u))\big) \in \mathbb{R}^N,$$

which we can solve in U_1. At this juncture, we rely on the condition $QU_1 = 0$ and the following lemma.

Lemma 1 (Technical Lemma). *If C is an $N \times N$ matrix satisfying* $\dim \ker C = n$ *and* $\ker C \cap Im\, C = \{0\}$, *and if Q is an $n \times N$ matrix of rank n, then for all $J \in \mathbb{R}^N$, there exists a unique solution $V \in \mathbb{R}^N$ to $C\, V = J$ and $QV = 0$ $QJ = 0$.*

Proposition 1 (First-Order Corrector Problem). *The first-order term U_1 is characterized by $B(\mathscr{E}(u))\, U_1 = -\partial_x \big(F(\mathscr{E}(u))\big)$ and $QU_1 = 0$.*

Considering next the first-order terms, we arrive at

$$\partial_t\big(\mathscr{E}(u)\big) + \partial_x\Big(A(\mathscr{E}(u))\,U_1\Big) = -\frac{1}{2}D_U^2\,R(\mathscr{E}(u)).(U_1,U_1) - B(\mathscr{E}(u))\,U_2$$

and, after multiplication by Q and using $Q\mathscr{E}(u) = u$,

$$\partial_t u + \partial_x\Big(Q\;A(\mathscr{E}(u))\,U_1\Big) = -\frac{1}{2}Q\;D_U^2\,R(\mathscr{E}(u)).(U_1,U_1) - Q\;B(\mathscr{E}(u))\,U_2.$$

On the other hand, by differentiating $QR(U) = 0$, we get $Q\;D_U^2\,R.(U_1,U_1) \equiv 0$ and $Q\;B\;U_2 \equiv 0$. This leads us to the following conclusion.

Theorem 1 (Late Time/Stiff Relaxation Effective Equations). *The effective system reads*

$$\partial_t u = -\partial_x\Big(QA(\mathscr{E}(u))\,U_1\Big) =: \partial_x\,(\mathscr{M}(u)\partial_x u)$$

for some $n \times n$ matrix $\mathscr{M}(u)$ and with U_1 being the unique solution to

$$B(\mathscr{E}(u))\,U_1 = -A(\mathscr{E}(u))\partial_x\big(\mathscr{E}(u)\big), \qquad QU_1 = 0.$$

2.4 Mathematical Entropy Pair for Stiff Balance Laws

We now assuming now that a mathematical entropy $\Phi : \Omega \to \mathbb{R}$ exists and satisfies the following two additional conditions:

Condition 5. There exists an entropy-flux $\Psi : \Omega \to \mathbb{R}$ such that $D_U\Phi\,A = D_U\Psi$ in Ω. So, all smooth solutions satisfy

$$\epsilon\partial_t\big(\Phi(U^\epsilon)\big) + \partial_x\big(\Psi(U^\epsilon)\big) = -D_U\Phi(U^\epsilon)\,\frac{R(U^\epsilon)}{\epsilon}$$

and, consequently, the matrix $D_U^2\Phi\,A$ is symmetric in Ω. Moreover, the map Φ is convex, i.e. the $N \times N$ matrix $D_U^2\Phi$ is positive definite on $\mathscr{M}$.

Condition 6. The entropy is compatible withe the relaxation in the sense that

$$D_U\Phi\;R \geq 0 \qquad \text{in } \Omega,$$
$$D_U\Phi(U) = v(U)Q \in \mathbb{R}^N, \qquad v(U) \in \mathbb{R}^d.$$

Next, we return to the effective equations $\partial_t u = \partial_x\mathscr{D}$ sand $\mathscr{D} := -QA(\mathscr{E}(u))\,U_1$ and, multiplying it by the Hessian of the entropy, we see that $U_1 \in \mathbb{R}^N$ is characterized by

$$\mathscr{L}(u)U_1 = -\big(D_U^2\Phi\big)(\mathscr{E}(u))\partial_x\big(F(\mathscr{E}(u))\big),$$
$$QU_1 = 0,$$

with $\mathscr{L}(u) = D_U^2\Phi(\mathscr{E}(u))B(\mathscr{E}(u))$.

Denoting by $\mathscr{L}(u)^{-1}$ the generalized inverse with constraint and setting $S(u) := QA(\mathscr{E}(u))$, we obtain

$$\mathscr{D} = S\mathscr{L}^{-1}\big(D_U^2\Phi\big)(\mathscr{E})\partial_x\big(F(\mathscr{E})\big).$$

Finally, one can check that, with $v := \partial_x\,(D_u\Phi(\mathscr{E}))^T$,

$$\big(D_U^2\Phi\big)(\mathscr{E})\partial_x\big(F(\mathscr{E})\big) = S^T v.$$

Theorem 2 (Entropy Structure of the Effective System). *When a mathematical entropy is available, the effective equations take the form*

$$\partial_t u = \partial_x\Big(L(u)\,\partial_x\big(D_u\Phi(\mathscr{E}(u))\big)^T\Big),$$

with

$$L(u) := S(u)\mathscr{L}(u)^{-1}S(u)^T, \qquad S(u) := QA(\mathscr{E}(u)),$$
$$\mathscr{L}(u) := \big(D_U^2\Phi\big)(\mathscr{E}(u))B(\mathscr{E}(u)),$$

where, for all b satisfying $Qb = 0$, the unique solution to $\mathscr{L}(u)V = b$, $QV = 0$ is denoted by $\mathscr{L}(u)^{-1}b$ (generalized inverse).

This result can be formulated in the so-called entropy variable $\big(D_u\Phi(\mathscr{E}(u))\big)^T$. Furthermore, a dissipation property follows from our assumptions and, specifically, from the entropy and equilibrium properties (see $R(\mathscr{E}(u)) = 0$), we obtain

$$D_U\Phi R \geq 0 \qquad \text{in } \Omega,$$
$$(D_U\Phi R)\,|_{U=\mathscr{E}(u)} = 0 \qquad \text{in } \omega.$$

Thus, the matrix $D_U^2\Big(D_U\Phi R\Big)|_{U=\mathscr{E}(u)}$ is non-negative definite. It follows that

$$D_U^2\Big(D_U\Phi R\Big) = D_U^2\Phi B + \big(D_U^2\Phi B\big)^T \qquad \text{when } U = \mathscr{E}(u),$$

so that $D_U^2\Phi\ B|_{U=\mathscr{E}(u)} \geq 0$ in ω.

For the equilibrium entropy $\Phi(\mathscr{E}(u))$, the associated (entropy) flux $u \mapsto \Psi(\mathscr{E}(u))$ is constant on the equilibrium manifold ω. For the map $\Psi(\mathscr{E})$, we have

$$D_u\big(\Psi(\mathscr{E})\big) = D_U\Psi(\mathscr{E})D_u\mathscr{E} = D_U\Phi(\mathscr{E})A(\mathscr{E})D_u\mathscr{E}.$$

Observe that $\big(D_U \Phi\big)(\mathscr{E}) = D_u\big(\Phi(\mathscr{E})\big) Q$, so that

$$D_u\Big(\Psi(\mathscr{E}(u))\Big) = D_u\Phi(\mathscr{E}(u))QA(\mathscr{E}(u))D_u\mathscr{E}(u)$$
$$= D_u\Big(\Phi(\mathscr{E}(u))\Big) D_u QF(\mathscr{E}(u)).$$

Since $QF(\mathscr{E}) = 0$, then $D_u QF(\mathscr{E}) = 0$ and the proof is completed.

Therefore, $D_u\big(\Psi(\mathscr{E}(u))\big) = 0$ for all $u \in \omega$. From the expansion $U^\epsilon = \mathscr{E}(u) + \epsilon U_1 + \ldots$, where U_1 is given by the first-order corrector problem, we deduce

$$\Psi(U^\epsilon) = \Psi(\mathscr{E}(u)) + \epsilon\, D_U\Psi(\mathscr{E}(u))\, U_1 + \mathscr{O}(\epsilon^2),$$

and then $\partial_x\Psi(U^\epsilon) = \epsilon\, \partial_x D_U\Psi(\mathscr{E}(u))\, U_1 + \mathscr{O}(\epsilon^2)$. Similarly, for the relaxation source, we have

$$D_U\Phi(U^\epsilon)R(U^\epsilon) = \epsilon^2 D_U^2\Phi(\mathscr{E}(u))D_U R(\mathscr{E}(u))U_1 + \mathscr{O}(\epsilon^3).$$

We thus get

$$\partial_t\big(\Phi(\mathscr{E}(u))\big) + \partial_x\Big(D_U\Psi(\mathscr{E}(u))\, U_1\Big)$$
$$= -U_1^T\ \big(D_U^2\Phi(\mathscr{E}(u))B(\mathscr{E}(u))\big)\, U_1.$$

At this juncture, recall that $X\ \big(D_U^2\Phi\big)(\mathscr{E})B(\mathscr{E})\ X \geq 0$ for $X \in \mathbb{R}^N$.

Proposition 2 (Monotonicity of the Entropy). *The entropy is non-increasing, i.e.*

$$\partial_t\big(\Phi(\mathscr{E}(u))\big) + \partial_x\,(D_U\Psi(\mathscr{E}(u))\, U_1) \leq 0$$

and

$$\partial_t\big(\Phi(\mathscr{E}(u))\big) = \partial_x\Big(\big(D_u\big(\Phi(\mathscr{E}(u))\big)L(u)\,\partial_x\big(D_u\big(\Phi(\mathscr{E}(u))\big)^T\Big).$$

2.5 *Effective Models*

2.5.1 Effective Model for Stiff Friction

We now analyze the diffusive regime for the Euler equations with friction. According to the general theory, the equilibria satisfy $\partial_t\rho = -\partial_x\Big(QA(\mathscr{E}(u))\, U_1\Big)$ with

$$D_U F(\mathscr{E}(u)) = \begin{pmatrix} 0 & 1 \\ p'(\rho) & 0 \end{pmatrix}.$$

Here, U_1 is the unique solution to $B(\mathscr{E}(u))U_1 = -\partial_x\big(F(\mathscr{E}(u))\big)$ and $QU_1 = 0$ with

$$B(\mathscr{E}(u)) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \partial_x\Big(F(\mathscr{E}(u))\Big) = \begin{pmatrix} 0 \\ \partial_x\big(p(\rho)\big) \end{pmatrix}.$$

The effective diffusion equation for the Euler equations with friction thus read:

$$\partial_t \rho = \partial_x^2\big(p(\rho)\big), \tag{10}$$

which is a nonlinear parabolic equation (away from vacuum) since $p'(\rho) > 0$. Near the vacuum, this equation is often degenerate since $p'(\rho)$ typically vanishes at $\rho = 0$. For instance for polytropic gases $p(\rho) = \kappa\rho^\gamma$ with $\kappa > 0$ and $\gamma \in (1, \gamma)$ we get

$$\partial_t \rho = \kappa\gamma\, \partial_x\big(\rho^{\gamma-1}\partial\rho\big). \tag{11}$$

Defining the internal energy $e(\rho) > 0$ by $e'(\rho) = p(\rho)/\rho^2$ we see that, for all smooth solutions to (7),

$$\epsilon\, \partial_t\Big(\rho\frac{v^2}{2} + \rho e(\rho)\Big) + \partial_x\Big(\rho\frac{v^3}{2} + (\rho e(\rho) + p(\rho))v\Big) = -\frac{\rho v^2}{\epsilon}, \tag{12}$$

so that $\Phi(U) = \rho\frac{v^2}{2} + \rho e(\rho)$ is a convex entropy and is compatible with the relaxation. All the conditions of the general framework are therefore satisfied.

2.5.2 Effective Model for Stiff Radiative Transfer

This system is compatible with our late-time/stiff relaxation framework with now

$$U = \begin{pmatrix} e \\ f \\ \tau \end{pmatrix}, \qquad F(U) = \begin{pmatrix} f \\ \chi(\frac{f}{e})e \\ 0 \end{pmatrix}, \qquad R(U) = \begin{pmatrix} e - \tau^4 \\ f \\ \tau^4 - e \end{pmatrix}.$$

The equilibria read $u = \tau + \tau^4$ and

$$\mathscr{E}(u) = \begin{pmatrix} \tau^4 \\ 0 \\ \tau \end{pmatrix}, \qquad Q := (1\ 0\ 1)$$

and we have $QF(\mathscr{E}(u)) = 0$.

We determine the diffusive regime for the $M1$ model from

$$\big(D_U F\big)(\mathscr{E}(u)) = \begin{pmatrix} 0 & 1 & 0 \\ \chi(0) & \chi'(0) & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where U_1 is the solution to

$$\begin{pmatrix} 1 & 0 & -4\tau^3 \\ 0 & 1 & 0 \\ -1 & 0 & 4\tau^3 \end{pmatrix} U_1 = \begin{pmatrix} 0 \\ \partial_x\big(\tau^4/3\big) \\ 0 \end{pmatrix},$$

$$(1\,0\,1)U_1 = 0.$$

Therefore, we have $U_1 = \begin{pmatrix} 0 \\ \frac{4}{3}\tau^3\partial_x\tau \\ 0 \end{pmatrix}$ and the effective diffusion equation reads

$$\partial_t(\tau + \tau^4) = \partial_x\left(\frac{4}{3}\tau^3\partial_x\tau\right), \tag{13}$$

which admits an entropy.

2.5.3 Effective Model for Stiff Friction and Stiff Radiative Transfert

Here, we have

$$D_U F(\mathscr{E}(u)) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ p'(\rho) & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 \end{pmatrix}, \qquad U_1 = \begin{pmatrix} 0 \\ \frac{1}{\kappa}\Big(-\partial_x p(\rho) - \frac{1}{3}\partial_x e\Big) \\ 0 \\ -\frac{1}{3\sigma}\partial_x e \end{pmatrix},$$

and the effective diffusion system for the coupled Euler/$M1$ model reads

$$\begin{aligned} &\partial_t\rho - \frac{1}{\kappa}\partial_x^2 p(\rho) - \frac{1}{3\kappa}\partial_x^2 e = 0, \\ &\partial_t e - \frac{1}{3\sigma}\partial_x^2 e = 0. \end{aligned} \tag{14}$$

The second equation is a heat equation, and its solution appears as a source-term in the first one.

2.5.4 Effective Model for Stiff Nonlinear Friction

Our framework encompass handle certain nonlinear diffusion regime under the scaling

$$\epsilon\, \partial_t U + \partial_x F(U) = -\frac{R(U)}{\epsilon^q}.$$

The parameter $q \geq 1$ introduces a new scale and is necessary when the relaxation is nonlinear. We assume that

$$R\Big(\mathscr{E}(u) + \epsilon\, U\Big) = \epsilon^q R\Big(\mathscr{E}(u) + M(\epsilon)\, U\Big), \qquad U \in \Omega, \quad u \in \omega$$

for some matrix $M(\epsilon)$. In that regime, the effective equations are now nonlinear parabolic.

Our final example requires this more general theory and reads

$$\begin{aligned} &\epsilon \partial_t h + \partial_x \big(hv\big) = 0, \\ &\epsilon \partial_t \big(hv\big) + \partial_x \Big(h\, v^2 + p(h)\Big) = -\frac{\kappa^2(h)}{\epsilon^2}\, g\, hv|hv|, \end{aligned} \tag{15}$$

where h is the fluid height and v the fluid velocity v. The pressure reads $p(h) = g\, h^2/2$ while $g > 0$ is the gravity constant. The friction $\kappa : \mathbb{R}^+ \to \mathbb{R}^+$ is a positive function, and for instance one can take $\kappa(h) = \frac{\kappa_0}{h}$ with $\kappa_0 > 0$.

The nonlinear version of the late-time/stiff relaxation framework applies by introducing

$$U = \begin{pmatrix} h \\ hv \end{pmatrix}, \quad F(U) = \begin{pmatrix} hv \\ hv^2 + p(h) \end{pmatrix}, \quad R(U) = \begin{pmatrix} 0 \\ \kappa^2(h) g hv|hv| \end{pmatrix}.$$

The equilibria $u = h$ are associated with

$$\mathscr{E}(u) = \begin{pmatrix} h \\ 0 \end{pmatrix}, \qquad\qquad Q = (1\ 0).$$

The relaxation is nonlinear and

$$R(\mathscr{E}(u) + \epsilon U) = \epsilon^2 R\big(\mathscr{E}(U) + M(\epsilon) U\big),$$

with

$$M(\epsilon) := \begin{pmatrix} \epsilon & 0 \\ 0 & 1 \end{pmatrix}.$$

in turn, we obtain a nonlinear effective equation for the Euler equations with nonlinear friction, i.e.

$$\partial_t h = \partial_x \left(\frac{\sqrt{h}}{\kappa(h)} \frac{\partial_x h}{\sqrt{|\partial_x h|}} \right), \tag{16}$$

which is a parabolic and fully nonlinear.

Introducing the internal energy $e(h) := gh/2$, we see that all smooth solutions to (15) satisfy the entropy inequality

$$\epsilon \partial_t \left(h \frac{v^2}{2} + g \frac{h^2}{2} \right) + \partial_x \left(h \frac{v^2}{2} + gh^2 \right) v = -\frac{\kappa^2(h)}{\epsilon^2} ghv^2 |hv|. \tag{17}$$

The entropy $\Phi(U) := h\frac{v^2}{2} + g\frac{h^2}{2}$ satisfies the compatibility properties for the nonlinear late-time/stiff relaxation theory, with

$$R(\mathscr{E}(u) + M(0)\bar{U}_1) = \begin{pmatrix} 0 \\ \partial_x p(h) \end{pmatrix},$$

where $\bar{U}_1 = (0\ \beta)$. We obtain $R(\mathscr{E}(u) + M(0)\bar{U}_1) = c(u)\bar{U}_1$ with

$$c(u) = g\kappa(h)\sqrt{h|\partial_x h|} \geq 0.$$

2.6 A Class of Asymptotic-Preserving Finite Volume Method

2.6.1 The General Strategy

We now will design a class of finite volume schemes which are consistent with the asymptotic regime $\epsilon \to 0$ and allow us to recover the effective diffusion equation (independently of the mesh-size) for the limiting solutions. Hence, we develop here a rather general framework adapted to the hyperbolic-to-parabolic relaxation regime.

Step 1. We rely on a arbitrary finite volume scheme for the homogeneous system

$$\partial_t U + \partial_x F(U) = 0,$$

as described below.

Step 2. Next, we modify this scheme and include a matrix-valued free parameter in order to consistently approximate the non-homogeneous system (for any $\gamma > 0$)

$$\partial_t U + \partial_x F(U) = -\gamma\, R(U).$$

Step 3. By performing an asymptotic analysis of this scheme after replacing the discretization parameter Δt by $\epsilon \Delta t$, and γ by $1/\epsilon$, we then determine the free parameters and ensure the desired asymptotic-preserving property.

For definiteness, the so-called HLL discretization of the homogeneous system (Harten, Lax, and van Leer [36]) are now discussed. We present the solver based on a single intermediate state and on a uniform mesh with cells of length Δx, that is,

$$[x_{i-1/2}, x_{i+1/2}], \qquad x_{i+1/2} = x_i + \frac{\Delta x}{2}$$

for all $i = \ldots, -1, 0, 1, \ldots$. The time discretization is based on some Δt restricted by the CFL condition [28] with $t^{m+1} = t^m + \Delta t$.

Given any initial data (lying in Ω):

$$U^0(x) = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} U(x,0)dx, \qquad x \in [x_{i-1/2}, x_{i+1/2}).$$

we design approximations that are piecewise constant at each t^m, that is,

$$U^m(x) = U_i^m, \qquad x \in [x_{i-1/2}, x_{i+1/2}), \quad i \in \mathbb{Z}.$$

At each cell interface we use the approximate Riemann solver

$$\tilde{U}_{\mathscr{R}}(\frac{x}{t}; U_L, U_R) = \begin{cases} U_L, & \frac{x}{t} < -b, \\ \tilde{U}^\star, & -b < \frac{x}{t} < b, \\ U_R, & \frac{x}{t} > b, \end{cases}$$

where $b > 0$ is (sufficiently) large. The "numerical cone" (and numerical diffusion) is determined by some $b > 0$ and, for simplicity in the presentation, we assume a single constant b. More generally, one can introduce distinct speeds $b^-_{i+1/2} < b^+_{i+1/2}$ at each interface.

We introduce the intermediate state

$$\tilde{U}^\star = \frac{1}{2}(U_L + U_R) - \frac{1}{2b}\big(F(U_R) - F(U_L)\big)$$

and, under the CFL condition $b\frac{\Delta t}{\Delta x} \leq 1/2$, the underlying Riemann solutions are non-interacting. Our global approximations

$$\tilde{U}^m_{\Delta x}(x, t^m + t), \qquad t \in [0, \Delta t), \quad x \in \mathbb{R},$$

are defined as follows.

At the time t^{m+1}, we set

$$\tilde{U}_i^{m+1} = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{U}_{\Delta x}^m(x, t^m + \Delta t)\, dx$$

and, recalling $\tilde{U}_{i+1/2}^\star = \frac{1}{2}(U_i^m + U_{i+1}^m) - \frac{1}{2b}(F(U_{i+1}^m) - F(U_i^m))$, and integrating out the expression given by the Riemann solutions, we arrive at the scheme adapted to our homogeneous system

$$\tilde{U}_i^{m+1} = U_i^m - \frac{\Delta t}{\Delta x}\Big(F_{i+1/2}^{HLL} - F_{i-1/2}^{HLL}\Big),$$

where

$$F_{i+1/2}^{HLL} = \frac{1}{2}\Big(F(U_i^m) + F(U_{i+1}^m)\Big) - \frac{b}{2}(U_{i+1}^m - U_i^m).$$

More generally one can include here two speeds $b_{i+1/2}^- < b_{i+1/2}^+$.

This scheme enjoys an invariant domain property, as follows. The intermediate states $\tilde{U}_{i+1/2}^\star$ can be written in the form of a convex combination

$$\tilde{U}_{i+1/2}^\star = \frac{1}{2}\left(U_i^m + \frac{1}{b}F(U_i^m)\right) + \frac{1}{2}\left(U_{i+1}^m - \frac{1}{b}F(U_{i+1}^m)\right) \in \Omega,$$

provided b is large enough. An alternative decomposition is

$$\tilde{U}_{i+1/2}^\star = \frac{1}{2}\Big(I + \frac{1}{b}\overline{A}(U_i^m, U_{i+1}^m)\Big)\, U_i^m + \frac{1}{2}\Big(I - \frac{1}{b}\overline{A}(U_i^m, U_{i+1}^m)\Big)\, U_{i+1}^m,$$

where $\overline{A}$ is an "average" of $D_U F$. By induction, we conclude that $\tilde{U}_i^m$ in Ω for all m, i.

2.6.2 Handling the Stiff Relaxation

Consider the modified Riemann solver:

$$U_{\mathscr{R}}(\frac{x}{t}; U_L, U_R) = \begin{cases} U_L, & \frac{x}{t} < -b, \\ U^{\star L}, & -b < \frac{x}{t} < 0, \\ U^{\star R}, & 0 < \frac{x}{t} < b, \\ U_R, & \frac{x}{t} > b, \end{cases}$$

with, at the interface,

$$U^{\star L} = \underline{\alpha}\tilde{U}^{\star} + (I - \underline{\alpha})\big(U_L - \bar{R}(U_L)\big),$$
$$U^{\star R} = \underline{\alpha}\tilde{U}^{\star} + (I - \underline{\alpha})\big(U_R - \bar{R}(U_R)\big).$$

We have introduced an arbitrary $N \times N$-matrix and an N-vector by

$$\underline{\alpha} = \left(I + \frac{\gamma \Delta x}{2b}(I + \underline{\sigma})\right)^{-1}, \qquad \bar{R}(U) = (I + \underline{\sigma})^{-1} R(U).$$

The term $\underline{\sigma}$ is a parameter matrix and we require that all inverse matrices are well-defined and, importantly, the correct asymptotic regime arises at the discrete level (see below).

At each $x_{i+1/2}$, we use the Riemann solver $U_{\mathscr{R}}(\frac{x - x_{i+1/2}}{t - t^m}; U_i^m, U_{i+1}^m)$ and superimpose non-interacting Riemann solutions

$$U_{\Delta x}^m(x, t^m + t), \qquad t \in [0, \Delta t), \quad x \in \mathbb{R}.$$

The approximation at the time t^{m+1} reads $U_i^{m+1} = \int_{x_{i-1/2}}^{x_{i+1/2}} U_{\Delta x}^m(x, t^m + \Delta t)\, dx$. By integration of the Riemann solutions, we arrive at the following discrete form of the balance law

$$\begin{aligned} &\frac{1}{\Delta t}(U_i^{m+1} - U_i^m) + \frac{1}{\Delta x}\Big(\underline{\alpha}_{i+1/2} F_{i+1/2}^{HLL} - \underline{\alpha}_{i-1/2} F_{i-1/2}^{HLL}\Big) \\ &= \frac{1}{\Delta x}(\underline{\alpha}_{i+1/2} - \underline{\alpha}_{i-1/2}) F(U_i^m) - \frac{b}{\Delta x}(I - \underline{\alpha}_{i-1/2})\bar{R}_{i-1/2}(U_i^m) \\ &\quad - \frac{b}{\Delta x}(I - \underline{\alpha}_{i+1/2})\bar{R}_{i+1/2}(U_i^m). \end{aligned} \tag{18}$$

The source can rewritten as

$$\begin{aligned} \frac{b}{\Delta x}(I - \underline{\alpha}_{i+1/2})\bar{R}_{i+1/2}(U_i^m) &= \frac{b}{\Delta x}\underline{\alpha}_{i+1/2}(\underline{\alpha}_{i+1/2}^{-1} - I)\bar{R}_{i+1/2}(U_i^m) \\ &= \frac{\gamma}{2}\underline{\alpha}_{i+1/2} R(U_i^m) \end{aligned}$$

and

$$\frac{b}{\Delta x}(I - \underline{\alpha}_{i-1/2})\bar{R}_{i-1/2}(U_i^m) = \frac{\gamma}{2}\underline{\alpha}_{i-1/2} R(U_i^m).$$

Our finite volume scheme for late-time/stiff-relaxation problems finally read

$$\begin{aligned}&\frac{1}{\Delta t}(U_i^{m+1}-U_i^m)+\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2}F_{i+1/2}^{HLL}-\underline{\alpha}_{i-1/2}F_{i-1/2}^{HLL})\\&=\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2}-\underline{\alpha}_{i-1/2})F(U_i^m)-\frac{\gamma}{2}(\underline{\alpha}_{i+1/2}+\underline{\alpha}_{i-1/2})R(U_i^m).\end{aligned}\tag{19}$$

Theorem 3 (A Class of Finite Volume Schemes for Relation Problems). *When*

$$\underline{\sigma}_{i+1/2}-\underline{\sigma}_{i-1/2}=\mathcal{O}(\Delta x)$$

and the matrix-valued map $\underline{\sigma}$ *is smooth, the finite volume scheme above is* consistent *with the hyperbolic system with relaxation and satisfies the following invariant domain property: provided all states*

$$\begin{aligned}U_{i+1/2}^{\star L}&=\underline{\alpha}_{i+1/2}\tilde{U}_{i+1/2}^{\star}+(I-\underline{\alpha}_{i+1/2})(U_i^m-\bar{R}(U_i^m)),\\U_{i+1/2}^{\star R}&=\underline{\alpha}_{i+1/2}\tilde{U}_{i+1/2}^{\star}+(I-\underline{\alpha}_{i+1/2})(U_{i+1}^m-\bar{R}(U_{i+1}^m))\end{aligned}$$

belong to Ω*, then all of the states* U_i^m *belong to* Ω.

2.7 Effective Equation for the Discrete Asymptotics

We replace Δt by $\Delta t/\epsilon$ and γ by $1/\epsilon$ and consider the expression

$$\begin{aligned}&\frac{\epsilon}{\Delta t}(U_i^{m+1}-U_i^m)+\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2}F_{i+1/2}^{HLL}-\underline{\alpha}_{i-1/2}F_{i-1/2}^{HLL})\\&=\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2}-\underline{\alpha}_{i-1/2})F(U_i^m)-\frac{1}{2\epsilon}(\underline{\alpha}_{i+1/2}+\underline{\alpha}_{i-1/2})R(U_i^m),\end{aligned}$$

in which

$$\underline{\alpha}_{i+1/2}=\left(I+\frac{\Delta x}{2\epsilon b}(I+\underline{\sigma}_{i+1/2})\right)^{-1}.$$

We expand near an equilibrium state $U_i^m=\mathcal{E}(u_i^m)+\epsilon(U_1)_i^m+\mathcal{O}(\epsilon^2)$ and find

$$\begin{aligned}F_{i+1/2}^{HLL}&=\frac{1}{2}F\big(\mathcal{E}(u_i^m)\big)+\frac{1}{2}F\big(\mathcal{E}(u_{i+1}^m)\big)-\frac{b}{2}\big(\mathcal{E}(u_{i+1}^m)-\mathcal{E}(u_i^m)\big)+\mathcal{O}(\epsilon),\\\frac{1}{\epsilon}R(U_i^m)&=B(\mathcal{E}(u_i^m))(U_1)_i^m+\mathcal{O}(\epsilon),\\\underline{\alpha}_{i+1/2}&=\frac{2b\epsilon}{\Delta x}\big(I+\underline{\sigma}_{i+1/2}\big)^{-1}+\mathcal{O}(1).\end{aligned}$$

The first-order terms yield us

$$
\begin{aligned}
&\frac{1}{\Delta t}(\mathscr{E}(u_i^{m+1}) - \mathscr{E}(u_i^m)) \\
&= -\frac{2b}{\Delta x^2}\left((I + \underline{\sigma}_{i+1/2})^{-1} F^{HLL}_{i+1/2}|_{\mathscr{E}(u)} - (I + \underline{\sigma}_{i-1/2})^{-1} F^{HLL}_{i-1/2}|_{\mathscr{E}(u)}\right) \\
&\quad + \frac{2b}{\Delta x^2}\left((I + \underline{\sigma}_{i+1/2})^{-1} - (I + \underline{\sigma}_{i-1/2})^{-1}\right) F(\mathscr{E}(u_i^m)) \\
&\quad - \frac{b}{\Delta x}\left((I + \underline{\sigma}_{i+1/2})^{-1} + (I + \underline{\sigma}_{i-1/2})^{-1}\right) B(\mathscr{E}(u_i^m))(U_1)_i^m.
\end{aligned}
$$

Assuming here the existence of an $n \times n$ matrix $\mathscr{M}_{i+1/2}$ satisfying

$$
Q\left(I + \underline{\sigma}_{i+1/2}\right)^{-1} = \frac{1}{b^2}\mathscr{M}_{i+1/2} Q
$$

and multiplying the equation above by Q, we get

$$
\frac{1}{\Delta t}(u_i^{m+1} - u_i^m) = -\frac{2}{b\Delta x^2}\left(\mathscr{M}_{i+1/2} Q F^{HLL}_{i+1/2}|_{\mathscr{E}(u)} - \mathscr{M}_{i-1/2} Q F^{HLL}_{i-1/2}|_{\mathscr{E}(u)}\right),
$$

where

$$
\begin{aligned}
Q F^{HLL}_{i+1/2}|_{\mathscr{E}(u)} &= \frac{Q}{2} F(\mathscr{E}(u_i^m)) + \frac{Q}{2} F(\mathscr{E}(u_{i+1}^m)) - \frac{b}{2} Q\left(\mathscr{E}(u_{i+1}^m) - \mathscr{E}(u_i^m)\right) \\
&= -\frac{b}{2}(u_{i+1}^m - u_i^m).
\end{aligned}
$$

The asymptotic system for the scheme thus reads

$$
\frac{1}{\Delta t}(u_i^{m+1} - u_i^m) = \frac{1}{\Delta x^2}\left(\mathscr{M}_{i+1/2}(u_{i+1}^m - u_i^m) + \mathscr{M}_{i-1/2}(u_{i-1}^m - u_i^m)\right). \tag{20}
$$

Recall that for some matrix $\mathscr{M}(u)$, the effective equation reads $\partial_t u = \partial_x\left(\mathscr{M}(u)\partial_x u\right)$.

Theorem 4 (Discrete Late-Time Asymptotic-Preserving Property). *Assume that the matrix-valued coefficients satisfy the following conditions:*

- *The matrices $I + \underline{\sigma}_{i+1/2}$ and $\left(1 + \frac{\Delta x}{2\epsilon b}\right) I + \underline{\sigma}_{i+1/2}$ are invertible for all $\epsilon \in [0, 1]$. There exists a matrix $\mathscr{M}_{i+1/2}$ satisfying the commutation condition*

$$
Q(I + \underline{\sigma}_{i+1/2})^{-1} = \frac{1}{b^2}\mathscr{M}_{i+1/2} Q.
$$

- *The discrete formulation of $\mathscr{M}(u)$ at each interface $x_{i+1/2}$ satisfies*

$$
\mathscr{M}_{i+1/2} = \mathscr{M}(u) + \mathscr{O}(\Delta x).
$$

Then the effective system associated with the proposed finite volume scheme coincides with the effective system determined in the late-time/stiff relaxation framework.

Finally, wee refer to [9] for various numerical experiments demonstrating the relevance of the proposed scheme and its efficiency in order to compute late-time behaviors of solutions. Asymptotic solutions may have large gradients but are in fact regular. Note that our CFL stability condition is based on the homogeneous hyperbolic system and therefore imposes a restriction on $\Delta t/\Delta x$ only. In our test, for simplicity, the initial data were taken in the image of Q, while the reference solutions (needed for the purpose of comparison) were computed separately by solving the associated parabolic equations, of course under a (much more restrictive) restriction on $\Delta t/(\Delta x)^2$.

The proposed theoretical framework for late-time/stiff relaxation problems thus led us to the development of a good strategy to design asymptotic-preserving schemes involving matrix-valued parameter. The convergence analysis ($\epsilon \to 0$) and the numerical analysis ($\Delta x \to 0$) for the problems under consideration are important and challenging open problems. It would be very interesting to apply our technique to plasma mixtures in a multi-dimensional setting.

Furthermore, high-order accurate Runge–Kutta methods have been recently developed for these stiff relaxation problems by Boscarino and Russo [10] and by Boscarino, LeFloch, and Russo [11].

3 Geometry-Preserving Finite Volume Methods

3.1 Objective and Background Material

On a smooth $(n+1)$-dimensional manifold M referred to as a spacetime, we consider the class of nonlinear conservation laws

$$d(\omega(u)) = 0, \qquad u = u(x),\ x \in M. \tag{21}$$

For all $\overline{u} \in \mathbb{R}$, $\omega = \omega(\overline{u})$ is a smooth field of n-forms, referred to as the flux field of the conservation law under consideration.

Two examples are of particular interest. When $M = \mathbb{R}_+ \times N$ and the n-manifold N is endowed with a Riemannian metric h, (21) reads

$$\partial_t u + \mathrm{div}_h(b(u)) = 0, \qquad u = u(t, y),\ t \geq 0,\ y \in N,$$

where div_h denotes the divergence operator for the metric h. The flux field is then considered as a flux vector field $b = b(\overline{u})$ on the n-manifold N and is independent of the time variable.

More generally, when M is endowed with a Lorentzian metric g, (21) reads

$$\mathrm{div}_g(a(u)) = 0, \qquad u = u(x),\ x \in M,$$

in which the flux $a = a(\overline{u})$ is now a vector field on M. In this Riemannian or Lorentzian settings, the theory of weak solutions on manifolds was initiated by Ben-Artzi and LeFloch [4] and developed in [1,2,46,51].

In the present paper, we discuss the novel approach in which the conservation law is written in the form (21), that is, the flux $\omega = \omega(\overline{u})$ is defined as a field of differential forms of degree n. No geometric structure is assumed on M and the sole flux field structure is assumed. The Eq. (21) is a "conservation law" for the unknown quantity u, as follows from Stokes theorem for sufficiently smooth solutions u: the total flux

$$\int_{\partial \mathscr{U}} \omega(u) = 0, \qquad \mathscr{U} \subset M, \tag{22}$$

vanishes for every smooth open subset $\mathscr{U}$. By relying on (21) rather than the equivalent expressions in the special cases of Riemannian or Lorentzian manifolds, we develop a theory of entropy solutions which is technically and conceptually simpler and provides a generalization of earlier works. From a numerical perspective, relying o (21) leads us to a geometry-consistent class of finite volume schemes, as we will now present it. So, our main objective i this presentation will be a generalization of the formulation and convergence of the finite volume method for general conservation law (21). In turn, this will also establish the existence of a contracting semi-group of entropy solutions.

We will proceed as follows:

- First we will formulate the initial and boundary problem for (21) by taking into account the nonlinearity and hyperbolicity of the equation. We need to impose that the manifold satisfies a global hyperbolicity condition, which provides a global time-orientation and allow us to distinguish between "future" and "past" directions in the time-evolution and we suppose that the manifold is foliated by compact slices.
- Second, we introduce a geometry-consistent version of the finite volume method which provides a natural discretization of the conservation law (21), which solely uses the n-volume form structure associated with the flux field ω.
- Third, we derive stability estimates, especially certain discrete versions of the entropy inequalities. We obtain a uniform control of the entropy dissipation measure, which, however, is not sufficient by itself to establish the compactness of the sequence of solutions. Yet, these stability estimates imply that the sequence of approximate solutions generated by the finite volume scheme converges to an entropy measure-valued solution in the sense of DiPerna.
- Fourth, to conclude we rely on DiPerna's uniqueness theorem [30] and establish the existence of entropy solutions to the corresponding initial value problem.

In the course of our analysis, we will derive the following contraction property: for any entropy solutions u, v and any hypersurfaces H, H' such that H' lies in the

future of H, one has

$$\int_{H'} \boldsymbol{\Omega}(u_{H'}, v_{H'}) \leq \int_{H} \boldsymbol{\Omega}(u_H, v_H). \tag{23}$$

Here, for all reals $\overline{u}, \overline{v}$, the n-form field $\boldsymbol{\Omega}(\overline{u}, \overline{v})$ is determined from the flux field $\omega(\overline{u})$ and is a generalization (to the spacetime setting) of the notion (introduced in [42]) of Kruzkov entropy $|\overline{u} - \overline{v}|$.

DiPerna's measure-valued solutions were first used to establish the convergence of schemes by Szepessy [64], Coquel and LeFloch [25–27], and Cockburn, Coquel, and LeFloch [22, 23]. Further hyperbolic models including a coupling with elliptic equations and many applications were investigated by Kröner [40], and Eymard, Gallouet, and Herbin [34]. For higher-order schemes, see Kröner, Noelle, and Rokyta [41]. See also Westdickenberg and Noelle [66].

3.2 *Entropy Solutions to Conservation Laws Posed on a Spacetime*

We assume that M is an oriented, compact, differentiable $(n+1)$-manifold with boundary. Given an $(n+1)$-form α, its modulus is defined as the $(n+1)$-form $|\alpha| := |\overline{\alpha}|\, dx^0 \wedge \cdots \wedge dx^n$, where $\alpha = \overline{\alpha}\, dx^1 \wedge \cdots \wedge dx^n$ is written in an oriented frame determined in coordinates $x = (x^\alpha) = (x^0, \ldots, x^n)$. If H is a hypersurface, we denote by $i = i_H : H \to M$ the canonical injection map, and by $i^* = i_H^*$ is the pull-back operator acting on differential forms defined on M.

We introduce the following notion:

- A flux field ω on the $(n+1)$-manifold M is a parametrized family $\omega(\overline{u}) \in \Lambda^n(M)$ of smooth fields of differential forms of degree n, that depends smoothly upon the real parameter $\overline{u}$.
- The conservation law associated with a flux field ω and with unknown $u : M \to \mathbb{R}$ is

$$d\big(\omega(u)\big) = 0, \tag{24}$$

  where d is the exterior derivative operator and, therefore, $d\big(\omega(u)\big)$ is a field of differential forms of degree $(n+1)$.
- A flux field ω is said to grow at most linearly if for every 1-form ρ on M

$$\sup_{\overline{u} \in \mathbb{R}} \int_M |\rho \wedge \partial_u \omega(\overline{u})| < +\infty. \tag{25}$$

In local coordinates $x = (x^\alpha)$ we write (for all $\overline{u} \in \mathbb{R}$) $\omega(\overline{u}) = \omega^\alpha(\overline{u})\, (\widehat{dx})_\alpha$ and $(\widehat{dx})_\alpha := dx^0 \wedge \ldots \wedge dx^{\alpha-1} \wedge dx^{\alpha+1} \wedge \ldots \wedge dx^n$. Here, the coefficients $\omega^\alpha = \omega^\alpha(\overline{u})$

are smooth. The operator d acts on differential forms and that, given a p-form ρ and a p'-form ρ', one has $d(d\rho) = 0$ and $d(\rho \wedge \rho') = d\rho \wedge \rho' + (-1)^p \rho \wedge d\rho'$. The Eq. (24) makes sense for unknowns that are Lipschitz continuous. However, solutions to nonlinear hyperbolic equations need not be continuous and we need to recast (24) in a weak form.

Given a smooth solution u of (24) we apply Stokes theorem on any open subset $\mathscr{U}$ (compactly included in M and with smooth boundary $\partial\mathscr{U}$) and find

$$0 = \int_{\mathscr{U}} d(\omega(u)) = \int_{\partial\mathscr{U}} i^*(\omega(u)). \tag{26}$$

Similarly, given any smooth function $\psi : M \to \mathbb{R}$ we write $d(\psi\,\omega(u)) = d\psi \wedge \omega(u) + \psi\, d(\omega(u))$, where $d\psi$ is a 1-form field. Provided u satisfies (24), we deduce that

$$\int_M d(\psi\,\omega(u)) = \int_M d\psi \wedge \omega(u)$$

and, by Stokes theorem,

$$\int_M d\psi \wedge \omega(u) = \int_{\partial M} i^*(\psi\omega(u)). \tag{27}$$

A suitable orientation of the boundary ∂M is required for this formula to hold.

Definition 1 (Weak Solutions on a Spacetime). Given a flux field (with at most linear growth) ω, a function $u \in L^1(M)$ is a weak solution to (24) on the spacetime M if $\int_M d\psi \wedge \omega(u) = 0$ for every $\psi : M \to \mathbb{R}$ that is compactly supported in the interior $\overset{\circ}{M}$.

Observe that the function u is integrable and $\omega(\overline{u})$ has at most linear growth in $\overline{u}$, so that the $(n+1)$-form $d\psi \wedge \omega(u)$ is integrable on the compact manifold M.

Definition 2. A (smooth) field of n-forms $\Omega = \Omega(\overline{u})$ is a (convex) entropy flux field for (24) if there exists a (convex) function $U : \mathbb{R} \to \mathbb{R}$ such that

$$\Omega(\overline{u}) = \int_0^{\overline{u}} \partial_u U(\overline{v})\, \partial_u \omega(\overline{v})\, d\overline{v}, \qquad \overline{u} \in \mathbb{R}.$$

It is admissible if, moreover, $\sup |\partial_u U| < \infty$.

If we choose the function $U(\overline{u}, \overline{v}) := |\overline{u} - \overline{v}|$, where $\overline{v}$ is a real parameter, the entropy flux field reads

$$\boldsymbol{\Omega}(\overline{u}, \overline{v}) := \operatorname{sgn}(\overline{u} - \overline{v})\, (\omega(\overline{u}) - \omega(\overline{v})). \tag{28}$$

This is a generalization to spacetimes of the so-called Kruzkov's entropy pairs.

Next, given any smooth solution u to (24), we multiply (24) by $\partial_u U(u)$ and obtain the conservation law

$$d(\Omega(u)) - (d\Omega)(u) + \partial_u U(u)(d\omega)(u) = 0.$$

For discontinuous solutions, we impose the entropy inequalities

$$d(\Omega(u)) - (d\Omega)(u) + \partial_u U(u)(d\omega)(u) \leq 0 \tag{29}$$

in the sense of distributions for all admissible entropy pair (U, Ω). This is justified, for instance, via the vanishing viscosity method, i.e. by searching for weak solutions realizable as limits of smooth solutions to a parabolic regularization.

It remains to prescribe initial and boundary conditions. We emphasize that, without further assumption on the flux field (to be imposed shortly below), points along the boundary ∂M can not be distinguished and it is natural to prescribe the trace of the solution along the whole of the boundary ∂M. This is possible provided the boundary data, $u_B : \partial M \to \mathbb{R}$, is assumed by the solution in a suitably weak sense. Following Dubois and LeFloch [32], we use the notation

$$u\big|_{\partial M} \in \mathscr{E}_{U,\Omega}(u_B) \tag{30}$$

for all convex entropy pair (U, Ω), where for all reals $\overline{u}$

$$\mathscr{E}_{U,\Omega}(\overline{u}) := \Big\{ \overline{v} \in \mathbb{R} \;\Big|\; E(\overline{u}, \overline{v}) := \Omega(\overline{u}) + \partial_u U(\overline{u})(\omega(\overline{v}) - \omega(\overline{u})) \leq \Omega(\overline{v}) \Big\}.$$

Definition 3 (Entropy Solutions on a Spacetime with Boundary). Let $\omega = \omega(\overline{u})$ be a flux field (with at most linear growth) and let $u_B \in L^1(\partial M)$ be a boundary function. A function $u \in L^1(M)$ is an entropy solution to the boundary value problem (24) and (30) if there exists a bounded and measurable field of n-forms $\gamma \in L^1\Lambda^n(\partial M)$ such that, for every admissible convex entropy pair (U, Ω) and every smooth function $\psi : M \to \mathbb{R}_+$,

$$\int_M \Big(d\psi \wedge \Omega(u) + \psi\, (d\Omega)(u) - \psi\, \partial_u U(u)(d\omega)(u) \Big)$$
$$+ \int_{\partial M} \psi_{|\partial M} \left(i^* \Omega(u_B) + \partial_u U(u_B) \big(\gamma - i^* \omega(u_B) \big) \right) \geq 0.$$

This definition makes sense since each of the terms $d\psi \wedge \Omega(u)$, $(d\Omega)(u)$, $(d\omega)(u)$ belong to $L^1(M)$. Following DiPerna [30], we can also consider solutions that are no longer functions but Young measures, i.e, weakly measurable maps $\nu : M \to \mathrm{Prob}(\mathbb{R})$ taking values within is the set of probability measures $\mathrm{Prob}(\mathbb{R})$.

Definition 4. Let $\omega = \omega(\overline{u})$ be a flux field with at most linear growth and let $u_B \in L^\infty(\partial M)$ be a boundary function. A compactly supported Young measure

$\nu : M \to \mathrm{Prob}(\mathbb{R})$ is an entropy measure-valued solution to the boundary value problem (24), (30) if there exists a bounded and measurable field of n-forms $\gamma \in L^\infty \Lambda^n(\partial M)$ such that, for all convex entropy pair (U, Ω) and all smooth functions $\psi \geq 0$,

$$\int_M \Big\langle \nu, d\psi \wedge \Omega(\cdot) + \psi \left(d(\Omega(\cdot)) - \partial_u U(\cdot)(d\omega)(\cdot) \right) \Big\rangle + \int_{\partial M} \psi_{|\partial M} \Big\langle \nu, \Big(i^* \Omega(u_B) + \partial_u U(u_B) \big(\gamma - i^* \omega(u_B) \big) \Big) \Big\rangle \geq 0.$$

3.3 Global Hyperbolicity and Geometric Compatibility

The manifold M is now assumed to be foliated by hypersurfaces, say

$$M = \bigcup_{0 \leq t \leq T} H_t, \tag{31}$$

where each slice has the topology of a (smooth) n-manifold N with boundary. Topologically we have $M \simeq [0, T] \times N$, and

$$\begin{aligned} \partial M &= H_0 \cup H_T \cup B, \\ B = (0, T) \times N &:= \bigcup_{0 < t < T} \partial H_t. \end{aligned} \tag{32}$$

We impose a non-degeneracy condition on the averaged flux on the hypersurfaces.

Definition 5. Let M be a manifold endowed with a foliation (31)–(32) and let $\omega = \omega(\overline{u})$ be a flux field. Then, the conservation law (24) on M satisfies the global hyperbolicity condition if there exist constants $0 < \underline{c} < \overline{c}$ such that, for every non-empty hypersurface $e \subset H_t$, the integral $\int_e i^* \partial_u \omega(0)$ is positive and the function $\varphi_e : \mathbb{R} \to \mathbb{R}$,

$$\varphi_e(\overline{u}) := \fint_e i^* \omega(\overline{u}) = \frac{\int_e i^* \omega(\overline{u})}{\int_e i^* \partial_u \omega(0)}, \qquad \overline{u} \in \mathbb{R}$$

satisfies

$$\underline{c} \leq \partial_u \varphi_e(\overline{u}) \leq \overline{c}, \qquad \overline{u} \in \mathbb{R}. \tag{33}$$

The function φ_e represents the averaged flux along e. From now, we assume that the conditions above are satisfied and we refer to H_0 as an initial hypersurface and we prescribe an initial data $u_0 : H_0 \to \mathbb{R}$ on this hypersurface. We impose a

boundary data u_B on the submanifold B. We sometimes refer to H_t as spacelike hypersurfaces.

Under the global hyperbolicity condition (31)–(33), the initial and boundary value problem takes the following form. The boundary condition (30) decomposes into an initial data

$$u_{H_0} = u_0 \tag{34}$$

and a boundary condition

$$u\big|_B \in \mathscr{E}_{U,\Omega}(u_B). \tag{35}$$

Correspondingly, the condition in Definition 3 reads

$$\begin{aligned}&\int_M \Big(d\psi \wedge \Omega(u) + \psi\,(d\Omega)(u) - \psi\,\partial_u U(u)(d\omega)(u)\Big)\\ &+ \int_B \psi_{|\partial M}\left(i^*\Omega(u_B) + \partial_u U(u_B)\big(\gamma - i^*\omega(u_B)\big)\right) + \int_{H_T} i^*\Omega(u_{H_T})\\ &- \int_{H_0} i^*\Omega(u_0) \geq 0.\end{aligned}$$

Definition 6. A flux field ω is geometry-compatible if it is closed for each value of the parameter,

$$(d\omega)(\overline{u}) = 0, \qquad \overline{u} \in \mathbb{R}. \tag{36}$$

This condition ensures that constants are trivial solutions, a property shared by many models of fluid dynamics (such as the shallow water model). When (36) holds, it follows from Definition 2 that every entropy flux field Ω satisfies $(d\Omega)(\overline{u}) = 0$ (for all $\overline{u} \in \mathbb{R}$) and the entropy inequalities (29) for a solution $u : M \to \mathbb{R}$ take the simpler form

$$d(\Omega(u)) \leq 0. \tag{37}$$

3.4 The Spacetime Finite Volume Method

We now assume that $M = [0,T]\times N$ is foliated by slices with *compact* topology N, and the initial data u_0 is bounded. We assume that the global hyperbolicity condition holds and the flux field ω is geometry-compatible. Let $\mathscr{T}^h = \bigcup_{K\in\mathscr{T}^h} K$ be a triangulation of M, that is, a collection of cells (or elements), determined as the images of polyhedra of $\mathbb{R}^{n+1}$, satisfying:

- The boundary ∂K of an element K is a piecewise smooth, n-manifold, $\partial K = \bigcup_{e \subset \partial K} e$ and contains exactly two spacelike faces e_K^+ and e_K^- and "vertical" elements

$$e^0 \in \partial^0 K := \partial K \setminus \{e_K^+, e_K^-\}.$$

- The intersection $K \cap K'$ of two distinct elements $K, K' \in \mathscr{T}^h$ is either a common face of K, K' or else a submanifold with dimension at most $(n-1)$.
- The triangulation is compatible with the foliation in the sense that there exist times $t_0 = 0 < t_1 < \ldots < t_N = T$ such that all spacelike faces are submanifolds of $H_n := H_{t_n}$ for some $n = 0, \ldots, N$, and determine a triangulation of the slices. We denote by $\mathscr{T}_0^h$ the set of all K which admit one face belonging to the initial hypersurface H_0.

We define the measure $|e|$ of a hypersurface $e \subset M$ by

$$|e| := \int_e i^* \partial_u \omega(0). \tag{38}$$

This quantity is positive if e is sufficiently "close" to one of the hypersurfaces along which we have the hyperbolicity condition (33). Provided $|e| > 0$ which is the case if e is included in one of the slices of the foliation, we associate to e the function $\varphi_e : \mathbb{R} \to \mathbb{R}$. The following hyperbolicity condition holds along the triangulation since the spacelike elements are included in the spacelike slices:

$$\underline{c} \leq \partial_u \varphi_{e_K^\pm}(\overline{u}) \leq \overline{c}, \qquad K \in \mathscr{T}^h. \tag{39}$$

Next, we introduce the finite volume method by averaging (24) over each element $K \in \mathscr{T}^h$. Applying Stokes theorem with a smooth solution u to (24), we get

$$0 = \int_K d(\omega(u)) = \int_{\partial K} i^* \omega(u).$$

Decomposing the boundary ∂K into its parts e_K^+, e_K^-, and $\partial^0 K$, we obtain

$$\int_{e_K^+} i^* \omega(u) - \int_{e_K^-} i^* \omega(u) + \sum_{e^0 \in \partial^0 K} \int_{e^0} i^* \omega(u) = 0. \tag{40}$$

Given the averaged values u_K^- along e_K^- and $u_{K_{e^0}}^-$ along $e^0 \in \partial^0 K$, we need an approximation u_K^+ of the solution u along e_K^+. The second term in (40) can be approximated by

$$\int_{e_K^-} i^* \omega(u) \approx \int_{e_K^-} i^* \omega(u_K^-) = |e_K^-| \varphi_{e_K^-}(u_K^-)$$

and the last term by $\int_{e^0} i^*\omega(u) \approx q_{K,e^0}(u_K^-, u_{K_{e^0}}^-)$, where the *total discrete flux* $q_{K,e^0} : \mathbb{R}^2 \to \mathbb{R}$ (i.e., a scalar-valued function) must be prescribed.

Finally, the proposed version of the finite volume method for the conservation law (24) takes the form

$$\int_{e_K^+} i^*\omega(u_K^+) = \int_{e_K^-} i^*\omega(u_K^-) - \sum_{e^0 \in \partial^0 K} q_{K,e^0}(u_K^-, u_{K_{e^0}}^-) \tag{41}$$

or, equivalently,

$$|e_K^+|\varphi_{e_K^+}(u_K^+) = |e_K^-|\varphi_{e_K^-}(u_K^-) - \sum_{e^0 \in \partial^0 K} q_{K,e^0}(u_K^-, u_{K_{e^0}}^-). \tag{42}$$

We assume that the functions q_{K,e^0} satisfy the following properties for all $\overline{u}, \overline{v} \in \mathbb{R}$:

- *Consistency:*

$$q_{K,e^0}(\overline{u}, \overline{u}) = \int_{e^0} i^*\omega(\overline{u}). \tag{43}$$

- *Conservation:*

$$q_{K,e^0}(\overline{v}, \overline{u}) = -q_{K_{e^0},e^0}(\overline{u}, \overline{v}). \tag{44}$$

- *Monotonicity:*

$$\partial_{\overline{u}} q_{K,e^0}(\overline{u}, \overline{v}) \geq 0, \qquad \partial_{\overline{v}} q_{K,e^0}(\overline{u}, \overline{v}) \leq 0. \tag{45}$$

We need to specify the discretization of the initial data and define constant initial values $u_{K,0} = u_K^-$ (for $K \in \mathscr{T}_0^h$) associated with H_0, by setting

$$\int_{e_K^-} i^*\omega(u_K^-) := \int_{e_K^-} i^*\omega(u_0), \qquad e_K^- \subset H_0. \tag{46}$$

We also define a piecewise constant function $u^h : M \to \mathbb{R}$ by, for every element $K \in \mathscr{T}^h$,

$$u^h(x) = u_K^-, \qquad x \in K. \tag{47}$$

We introduce $N_K := \#\partial^0 K$, the total number of "vertical" neighbors of an element $K \in \mathscr{T}^h$, supposed to be uniformly bounded. We fix a finite family of local charts covering the manifold M, and assume that the parameter h coincides with the largest diameter of faces $e_K^\pm$ of elements $K \in \mathscr{T}^h$, where the diameter is computed with the Euclidian metric in chosen local coordinates.

We also impose the Courant–Friedrich–Levy condition (for all $K \in \mathscr{T}^h$)

$$\frac{N_K}{|e_K^+|} \max_{e^0 \in \partial^0 K} \sup_u \left| \int_{e^0} \partial_u \omega(u) \right| < \inf_u \partial_u \varphi_{e_K^+}, \tag{48}$$

in which the supremum and infimum in u are taken over the range of the initial data. Finally, we assume that the family of triangulations satisfy

$$\lim_{h \to 0} \frac{\tau_{\max}^2 + h^2}{\tau_{\min}} = \lim_{h \to 0} \frac{\tau_{\max}^2}{h} = 0 \tag{49}$$

where $\tau_{\max} := \max_i (t_{i+1} - t_i)$ and $\tau_{\min} := \min_i (t_{i+1} - t_i)$. For instance, these conditions are satisfied if $\tau_{\max}$, $\tau_{\min}$, and h vanish at the same order.

Our main objective in this presentation is establishing the convergence of the proposed finite volume schemes towards an entropy solution. Our analysis of the finite volume method will rely on a decomposition of (42) into (essentially) one-dimensional schemes, a technique that goes back to Tadmor [65], Coquel and LeFloch [25], and Cockburn, Coquel, and LeFloch [24].

By applying Stokes theorem to (36) with some $\overline{u} \in \mathbb{R}$, we obtain

$$\begin{aligned} 0 &= \int_K d(\omega(\overline{u})) = \int_{\partial K} i^* \omega(\overline{u}) \\ &= \int_{e_K^+} i^* \omega(\overline{u}) - \int_{e_K^-} i^* \omega(\overline{u}) + \sum_{e^0 \in \partial^0 K} q_{K,e^0}(\overline{u}, \overline{u}). \end{aligned}$$

Choosing $\overline{u} = u_K^-$, we deduce

$$|e_K^+| \varphi_{e_K^+}(u_K^-) = |e_K^-| \varphi_{e_K^-}(u_K^-) - \sum_{e^0 \in \partial^0 K} q_{K,e^0}(u_K^-, u_K^-), \tag{50}$$

which can be combined with (42):

$$\begin{aligned} \varphi_{e_K^+}(u_K^+) &= \varphi_{e_K^+}(u_K^-) - \sum_{e^0 \in \partial^0 K} \frac{1}{|e_K^+|} \Big(q_{K,e^0}(u_K^-, u_{K_{e^0}}^-) - q_{K,e^0}(u_K^-, u_K^-) \Big) \\ &= \sum_{e^0 \in \partial^0 K} \left(\frac{1}{N_K} \varphi_{e_K^+}(u_K^-) - \frac{1}{|e_K^+|} \Big(q_{K,e^0}(u_K^-, u_{K_{e^0}}^-) - q_{K,e^0}(u_K^-, u_K^-) \Big) \right). \end{aligned}$$

We introduce the intermediate values $\tilde{u}_{K,e^0}^+$:

$$\varphi_{e_K^+}(\tilde{u}_{K,e^0}^+) := \varphi_{e_K^+}(u_K^-) - \frac{N_K}{|e_K^+|} \Big(q_{K,e^0}(u_K^-, u_{K_{e^0}}^-) - q_{K,e^0}(u_K^-, u_K^-) \Big), \tag{51}$$

and thus arrive at the convex decomposition

$$\varphi_{e_K^+}(u_K^+) = \frac{1}{N_K} \sum_{e^0 \in \partial^0 K} \varphi_{e_K^+}(\tilde{u}_{K,e^0}^+). \tag{52}$$

Given any entropy pair (U, Ω) and hypersurface $e \subset M$ satisfying $|e| > 0$ we introduce the averaged entropy flux along e: $\varphi_e^\Omega(u) := \fint_e i^* \Omega(u)$.

Lemma 2. *For every convex entropy flux Ω one has*

$$\varphi_{e_K^+}^\Omega(u_K^+) \leq \frac{1}{N_K} \sum_{e^0 \in \partial^0 K} \varphi_{e_K^+}^\Omega(\tilde{u}_{K,e^0}^+). \tag{53}$$

In fact, the function $\varphi_{e_K^+}^\Omega \circ (\varphi_{e_K^+}^\omega)^{-1}$ is convex.

Proof. It suffices to show the inequality for the entropy flux, and then average this inequality over e. We need to check

$$\Omega(u_K^+) \leq \frac{1}{N_K} \sum_{e^0 \in \partial^0 K} \Omega(\tilde{u}_{K,e^0}^+), \tag{54}$$

namely

$$\begin{aligned} &\frac{1}{N_K} \sum_{e^0 \in \partial^0 K} \big(\Omega(\tilde{u}_{K,e^0}^+) - \Omega(u_K^+)\big) \\ &= \frac{1}{N_K} \sum_{e^0 \in \partial^0 K} \big(\omega(u_K^+) - \omega(\tilde{u}_{K,e^0}^+)\big) \partial_u U(u_K^+) + \frac{1}{N_K} \sum_{e^0 \in \partial^0 K} D_{K,e^0}, \end{aligned}$$

with

$$D_{K,e^0} := \int_0^1 \partial_{uu} U(u_K^+) \Big(\omega(\tilde{u}_{K,e^0}^+ + a(u_K^+ - \tilde{u}_{K,e^0}^+)) - \omega(\tilde{u}_{K,e^0}^+)\Big) (u_K^+ - \tilde{u}_{K,e^0}^+)\, da.$$

In the right-hand side, the former term vanishes identically (see (51)) and the latter term is non-negative, since $U(u)$ is convex and $\partial_u \omega$ is positive. □

3.5 Discrete Entropy Estimates

From the decomposition (52), we derive the discrete entropy inequalities of interest.

Lemma 3 (Entropy Inequalities for the Faces). *For all convex entropy pair (U, Ω) and all $K \in \mathscr{T}^h$ and $e^0 \in \partial^0 K$, there exists numerical entropy flux functions $Q_{K,e^0} : \mathbb{R}^2 \to \mathbb{R}$ satisfying (for all $u, v \in \mathbb{R}$):*

- *Q_{K,e^0} is consistent with the entropy flux Ω:*

$$Q_{K,e^0}(u,u) = \int_{e^0} i^* \Omega(u). \tag{55}$$

- *Conservation property:*

$$Q_{K,e^0}(u,v) = -Q_{K_{e^0},e^0}(v,u). \tag{56}$$

- *Discrete entropy inequality:*

$$\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}) - \varphi^{\Omega}_{e_K^+}(u_K^-) + \frac{N_K}{|e_K^+|}\Big(Q_{K,e^0}(u_K^-, u^-_{K_{e^0}}) - Q_{K,e^0}(u_K^-, u_K^-)\Big) \le 0. \tag{57}$$

Proof. *Step 1.* For $u, v \in \mathbb{R}$ and $e^0 \in \partial^0 K$, let us set

$$H_{K,e^0}(u,v) := \varphi_{e_K^+}(u) - \frac{N_K}{|e_K^+|}\Big(q_{K,e^0}(u,v) - q_{K,e^0}(u,u)\Big)$$

and note that $H_{K,e^0}(u,u) = \varphi_{e_K^+}(u)$. We now check that H_{K,e^0} satisfies

$$\frac{\partial}{\partial u} H_{K,e^0}(u,v) \ge 0, \qquad \frac{\partial}{\partial v} H_{K,e^0}(u,v) \ge 0. \tag{58}$$

The second property is immediate by the monotonicity (45). For the first one, we recall the CFL condition (48) and the monotonicity (45). From the definition of $H_{K,e^0}(u,v)$, we have

$$H_{K,e^0}(u, u_{K_{e^0}}) = \big(1 - \sum_{e^0 \in \partial^0 K} \alpha_{K,e^0}\big)\varphi_{e_K^+}(u) + \sum_{e^0 \in \partial^0 K} \alpha_{K,e^0} \varphi_{e_K^+}(u_{K_{e^0}}),$$

and

$$\alpha_{K,e^0} := \frac{1}{|e_K^+|} \frac{q_{K,e^0}(u, u_{K_{e^0}}) - q_{K,e^0}(u,u)}{\varphi_{e_K^+}(u) - \varphi_{e_K^+}(u_{K_{e^0}})}.$$

This gives a convex combination of $\varphi_{e_K^+}(u)$ and $\varphi_{e_K^+}(u_{K_{e^0}})$. By (45) we have $\sum_{e^0 \in \partial^0 K} \alpha_{K,e^0} \ge 0$ and, with (48),

$$\sum_{e^0 \in \partial^0 K} \alpha_{K,e^0} \le \sum_{e^0 \in \partial^0 K} \frac{1}{|e_K^+|} \Big| \frac{q_{K,e^0}(u, u_{K_{e^0}}) - q_{K,e^0}(u,u)}{\varphi_{e_K^+}(u) - \varphi_{e_K^+}(u_{K_{e^0}})} \Big| \le 1.$$

Step 2. We will establish the entropy inequalities for Kruzkov's entropies $\boldsymbol{\Omega}$. Introduce the discrete version of Kruzkov's entropy flux

$$\mathbf{Q}(u,v,c) := q_{K,e^0}(u\vee c, v\vee c) - q_{K,e^0}(u\wedge c, v\wedge c),$$

where $a\vee b = \max(a,b)$ and $a\wedge b = \min(a,b)$. Note that $Q_{K,e^0}(u,v)$ satisfies the first two properties of the lemma with the entropy flux replaced by the Kruzkov's family $\Omega = \boldsymbol{\Omega}$ in (28).
First, we observe:

$$\begin{aligned}
&H_{K,e^0}(u\vee c, v\vee c) - H_{K,e^0}(u\wedge c, v\wedge c)\\
&= \varphi_{e_K^+}(u\vee c) - \frac{N_K}{|e_K^+|}\big(q_{K,e^0}(u\vee c, v\vee c) - q_{K,e^0}(u\vee c, u\vee c)\big)\\
&\quad - \Big(\varphi_{e_K^+}(u\wedge c) - \frac{N_K}{|e_K^+|}\big(q_{K,e^0}(u\wedge c, v\wedge c) - q_{K,e^0}(u\wedge c, u\wedge c)\big)\Big)\\
&= \varphi^{\Omega}_{e_K^+}(u,c) - \frac{N_K}{|e_K^+|}\Big(\mathbf{Q}(u,v,c) - \mathbf{Q}(u,u,c)\Big),
\end{aligned} \tag{59}$$

where $\varphi_{e_K^+}(u\vee c) - \varphi_{e_K^+}(u\wedge c) = \int_{e_K^+} i^*\boldsymbol{\Omega}(u,c) = \varphi^{\Omega}_{e_K^+}(u,c)$.
Second, we prove that for $u = u^-_K$, $v = u^-_{K_{e^0}}$ and for any $c\in\mathbb{R}$

$$H_{K,e^0}(u^-_K\vee c, u^-_{K_{e^0}}\vee c) - H_{K,e^0}(u^-_K\wedge c, u^-_{K_{e^0}}\wedge c) \geq \varphi^{\Omega}_{e_K^+}(\tilde u^+_{K,e^0}, c). \tag{60}$$

Indeed, we have

$$\begin{aligned}
H_{K,e^0}(u,v)\vee H_{K,e^0}(\lambda,\lambda) &\leq H_{K,e^0}(u\vee\lambda, v\vee\lambda),\\
H_{K,e^0}(u,v)\wedge H_{K,e^0}(\lambda,\lambda) &\geq H_{K,e^0}(u\wedge\lambda, v\wedge\lambda),
\end{aligned}$$

where H_{K,e^0} is monotone in both variables. Since $\varphi_{e_K^+}$ is monotone, we have

$$\begin{aligned}
&H_{K,e^0}(u^-_K\vee c, u^-_{K_{e^0}}\vee c) - H_{K,e^0}(u^-_K\wedge c, u^-_{K_{e^0}}\wedge c)\\
&\geq \Big|H_{K,e^0}(u^-_K, u^-_{K_{e^0}}) - H_{K,e^0}(c,c)\Big| = \Big|\varphi_{e_K^+}(\tilde u^+_{K,e^0}) - \varphi_{e_K^+}(c)\Big|\\
&= \operatorname{sgn}\big(\varphi_{e_K^+}(\tilde u^+_{K,e^0}) - \varphi_{e_K^+}(c)\big)\big(\varphi_{e_K^+}(\tilde u^+_{K,e^0}) - \varphi_{e_K^+}(c)\big)\\
&= \operatorname{sgn}\big(\tilde u^+_{K,e^0} - c\big)\big(\varphi_{e_K^+}(\tilde u^+_{K,e^0}) - \varphi_{e_K^+}(c)\big) = \varphi^{\Omega}_{e_K^+}(\tilde u^+_{K,e^0}, c).
\end{aligned}$$

Combining this with (59) (with $u = u^-_K$, $v = u^-_{K_{e^0}}$), we obtain the inequality

$$\varphi^{\Omega}_{e^+_K}(\tilde{u}^+_{K,e^0}, c) - \varphi^{\Omega}_{e^+_K}(u^-_K, c) + \frac{N_K}{|e^+_K|}\Big(\mathbf{Q}(u, v, c) - \mathbf{Q}(u, u, c)\Big) \leq 0,$$

which implies a similar inequality for all convex entropy flux fields. □

We now combine Lemma 2 with Lemma 3.

Lemma 4 (Entropy Inequalities for the Elements). *For each $K \in \mathscr{T}^h$, one has*

$$|e^+_K|\left(\varphi^{\Omega}_{e^+_K}(u^+_K) - \varphi^{\Omega}_{e^+_K}(u^-_K)\right) + \sum_{e^0 \in \partial^0 K}\left(Q(u^-_K, u^-_{K_{e^0}}) - Q(u^-_K, u^-_K)\right) \leq 0. \tag{61}$$

If V is convex, then a *modulus of convexity* for V is a positive real $\beta < \inf V''$ (where the infimum is taken over the range of the data and solutions). In view of the proof of Lemma 2, $\varphi^{\Omega}_e \circ (\varphi^{\omega}_e)^{-1}$ is convex for every spacelike hypersurface e and every convex function U. (Note that the discrete entropy flux terms do not appear in (62) below.)

Lemma 5 (Entropy Balance Inequality Between Two Hypersurfaces). *For $K \in \mathscr{T}^h$, denote by $\beta_{e^+_K}$ a modulus of convexity for $\varphi^{\Omega}_{e^+_K} \circ \left(\varphi^{\omega}_{e^+_K}\right)^{-1}$ and set $\beta = \min_{K \in \mathscr{T}^h} \beta_{e^+_K}$. Then, for $i \leq j$ one has*

$$\sum_{K \in \mathscr{T}^h_{t_j}} |e^+_K| \varphi^{\Omega}_{e^+_K}(u^+_K) + \sum_{\substack{K \in \mathscr{T}^h_{[t_i, t_j)} \\ e^0 \in \partial^0 K}} \frac{\beta}{2N_K} |e^+_K| \left|\tilde{u}^+_{K,e^0} - u^+_K\right|^2 \leq \sum_{K \in \mathscr{T}^h_{t_i}} |e^-_K| \varphi^{\Omega}_{e^-_K}(u^-_K), \tag{62}$$

where $\mathscr{T}^h_{t_i}$ is the subset of all K satisfying $e^-_K \in H_{t_i}$, and one sets $\mathscr{T}^h_{[t_i, t_j)} := \bigcup_{i \leq k < j} \mathscr{T}^h_{t_k}$.

Proof. Multiplying (57) by $|e^+_K|/N_K$ and summing in $K \in \mathscr{T}^h$, $e^0 \in \partial^0 K$ yield

$$\sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \frac{|e^+_K|}{N_K} \varphi^{\Omega}_{e^+_K}(\tilde{u}^+_{K,e^0}) - \sum_{K \in \mathscr{T}^h} |e^+_K| \varphi^{\Omega}_{e^+_K}(u^-_K) + \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \left(Q_{K,e^0}(u^-_K, u^-_{K_{e^0}}) - Q_{K,e^0}(u^-_K, u^-_K)\right) \leq 0.$$

The conservation property (56) gives

$$\sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} Q_{K,e^0}(u^-_K, u^-_{K_{e^0}}) = 0 \tag{63}$$

and so

$$\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\varphi^{\Omega}_{e_K^+}(\tilde u^+_{K,e^0}) - \sum_{K\in\mathscr{T}^h} |e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^-) - \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} Q_{K,e^0}(u_K^-,u_K^-) \le 0. \tag{64}$$

If V is convex and if $v=\sum_j \alpha_j v_j$ is a convex combination of v_j, then

$$V(v) + \frac{\beta}{2}\sum_j \alpha_j |v_j - v|^2 \le \sum_j \alpha_j V(v_j),$$

where $\beta = \inf V''$, the infimum being taken over all v_j. We apply this with $v = \varphi_{e_K^+}(u_K^+)$ and $V = \varphi^{\Omega}_{e_K^+}\circ(\varphi^{\omega}_{e_K^+})^{-1}$, which is convex.

In view of (52) and by multiplying the above inequality by $|e_K^+|$ and summing in $K\in\mathscr{T}^h$, we obtain

$$\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} |e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^+) + \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{\beta}{2}\frac{|e_K^+|}{N_K}|\tilde u^+_{K,e^0} - u_K^+|^2 \le \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\varphi^{\Omega}_{e_K^+}(\tilde u^+_{K,e^0}).$$

Combining the result with (64), we conclude that

$$\begin{aligned}&\sum_{K\in\mathscr{T}^h} |e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^+) - \sum_{K\in\mathscr{T}^h} |e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^-) + \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{\beta}{2}\frac{|e_K^+|}{N_K}|\tilde u^+_{K,e^0} - u_K^+|^2\\ &\le \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} Q_{K,e^0}(u_K^-,u_K^-).\end{aligned} \tag{65}$$

Finally, using

$$\begin{aligned}0 &= \int_K d(\Omega(u_K^-)) = \int_{\partial K} i^*\Omega(u_K^-)\\ &= |e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^-) - |e_K^-|\varphi^{\Omega}_{e_K^-}(u_K^-) + \sum_{e^0\in\partial^0 K} Q_{K,e^0}(u_K^-,u_K^-),\end{aligned}$$

we obtain the desired inequality, after further summation over all of K within two arbitrary hypersurfaces. □

We apply Lemma 5 and obtain an important uniform estimate.

Lemma 6 (Global Entropy Dissipation Estimate). *The entropy dissipation is globally bounded, as follows:*

$$\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\left|\tilde{u}^+_{K,e^0}-u_K^+\right|^2 \lesssim C\int_{H_0} i^*\Omega(u_0) \tag{66}$$

for some constant $C>0$ depending upon the flux field and the sup-norm of the initial data. Here, Ω is the n-form entropy flux field associated with $U(u)=u^2/2$.

Proof. We apply (62) with the choice $U(u)=u^2$

$$0\geq \sum_{K\in\mathscr{T}^h}\left(|e_K^+|\varphi^{\Omega}_{e_K^+}(u_K^+)-|e_K^-|\varphi^{\Omega}_{e_K^-}(u_K^-)\right)+\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}}\frac{\beta}{2}\frac{|e_K^+|}{N_K}\left|\tilde{u}^+_{K,e^0}-u_K^+\right|^2.$$

After summing up in the "vertical" direction and keeping the contribution of all $K\in\mathscr{T}_0^h$ on H_0, we deduce that

$$\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\beta\left|\tilde{u}^+_{K,e^0}-u_K^+\right|^2 \leq \frac{2}{\beta}\sum_{K\in\mathscr{T}_0^h}|e_K^-|\varphi^{\Omega}_{e_K^-}(u_{K,0}).$$

For some constant $C>0$, we have $\sum_{K\in\mathscr{T}_0^h}|e_K^-|\varphi^{\Omega}_{e_K^-}(u_{K,0})\leq C\int_{H_0} i^*\Omega(u_0)$. These are essentially L^2 norm of the initial data, and this inequality is checked by fixing a reference volume form on H_0 and using the discretization (46) of the initial data u_0. □

3.6 Global Form of the Discrete Entropy Inequalities

One additional notation now is needed in order to handle "vertical face" of the triangulation: we fix a reference field of non-degenerate n-forms $\tilde{\omega}$ on M (to measure the "area" of the faces $e^0\in\partial K^0$). This is used in the convergence proof only, but not in the formulation of the finite volume schemes. For every $K\in\mathscr{T}^h$ we define

$$|e^0|_{\tilde{\omega}}:=\int_{e^0} i^*\tilde{\omega} \quad \text{for faces } e^0\in\partial^0 K \tag{67}$$

and the non-degeneracy condition is equivalent to $|e^0|_{\tilde{\omega}}>0$. Given a smooth function ψ defined on M and given a face $e^0\in\partial^0 K$ of some element, we introduce

$$\psi_{e^0}:=\frac{\int_{e^0}\psi\, i^*\tilde{\omega}}{\int_{e^0} i^*\tilde{\omega}}, \qquad \psi_{\partial^0 K}:=\frac{1}{N_K}\sum_{e^0\in\partial^0 K}\psi_{e^0}.$$

Lemma 7 (Global Form of the Discrete Entropy Inequalities). *Let Ω be a convex entropy flux field and let $\psi \geq 0$ be a smooth function supported away from the hypersurface $t = T$. Then, the finite volume scheme satisfies the entropy inequality*

$$- \sum_{K \in \mathscr{T}^h} \int_K d(\psi \Omega)(u_K^-) - \sum_{K \in \mathscr{T}_0^h} \int_{e_K^-} \psi \, i^* \Omega(u_{K,0}) \leq A^h(\psi) + B^h(\psi) + C^h(\psi), \tag{68}$$

with

$$A^h(\psi) := \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \frac{|e_K^+|}{N_K} \big(\psi_{\partial^0 K} - \psi_{e^0}\big) \big(\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}) - \varphi^{\Omega}_{e_K^+}(u_K^+)\big),$$

$$B^h(\psi) := \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \int_{e^0} \big(\psi_{e^0} - \psi\big) \, i^* \Omega(u_K^-),$$

$$C^h(\psi) := - \sum_{K \in \mathscr{T}^h} \int_{e_K^+} \big(\psi_{\partial^0 K} - \psi\big) \big(i^* \Omega(u_K^+) - i^* \Omega(u_K^-)\big).$$

Proof. From the discrete entropy inequalities (57), we get

$$\begin{aligned} &\sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \frac{|e_K^+|}{N_K} \psi_{e^0} \Big(\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}) - \varphi^{\Omega}_{e_K^+}(u_K^-)\Big) \\ &\quad + \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \psi_{e^0} \left(Q_{K,e^0}(u_K^-, u^-_{K_{e^0}}) - Q_{K,e^0}(u_K^-, u_K^-)\right) \leq 0. \end{aligned} \tag{69}$$

Thanks (56), we have $\sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \psi_{e^0} Q_{K,e^0}(u_K^-, u^-_{K_{e^0}}) = 0$ and, from (55),

$$\begin{aligned} \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \psi_{e^0} Q_{K,e^0}(u_K^-, u_K^-) &= \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \psi_{e^0} \int_{e^0} i^* \Omega(u_K^-) \\ &= \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \int_{e^0} \psi \, i^* \Omega(u_K^-) + \sum_{\substack{K \in \mathscr{T}^h \\ e^0 \in \partial^0 K}} \int_{e^0} (\psi_{e^0} - \psi) i^* \Omega(u_K^-). \end{aligned}$$

Next, we observe

$$
\begin{aligned}
&\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\psi_{e^0}\,\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0})\\
&= \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\psi_{\partial^0 K}\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}) + \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}(\psi_{e^0}-\psi_{\partial^0 K})\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0})\\
&\geq \sum_{K\in\mathscr{T}^h} |e_K^+|\psi_{\partial^0 K}\varphi^{\Omega}_{e_K^+}(u_K^+) + \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}(\psi_{e^0}-\psi_{\partial^0 K})\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}),
\end{aligned}
$$

where, we recalled (53) and the convex combination (52). From

$$
\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}\psi_{e^0}\,\varphi^{\Omega}_{e_K^+}(u_K^-) = \sum_{K\in\mathscr{T}^h} |e_K^+|\psi_{\partial^0 K}\,\varphi^{\Omega}_{e_K^+}(u_K^-),
$$

the inequality (69) reads

$$
\begin{aligned}
&\sum_{K\in\mathscr{T}^h} |e_K^+|\psi_{\partial^0 K}\Big(\varphi^{\Omega}_{e_K^+}(u_K^+)-\varphi^{\Omega}_{e_K^+}(u_K^-)\Big) - \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}}\int_{e^0}\psi\; i^*\Omega(u_K^-)\\
&\leq -\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}(\psi_{e^0}-\psi_{\partial^0 K})\varphi^{\Omega}_{e_K^+}(\tilde{u}^+_{K,e^0}) + \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}}\int_{e^0}(\psi_{e^0}-\psi)\, i^*\Omega(u_K^-).
\end{aligned}
\tag{70}
$$

The first term in (70) reads

$$
\begin{aligned}
&\sum_{K\in\mathscr{T}^h} |e_K^+|\psi_{\partial^0 K}\Big(\varphi^{\Omega}_{e_K^+}(u_K^+)-\varphi^{\Omega}_{e_K^+}(u_K^-)\Big)\\
&= \sum_{K\in\mathscr{T}^h}\int_{e_K^+}\psi(i^*\Omega(u_K^+)-i^*\Omega(u_K^-))\\
&\quad + \sum_{K\in\mathscr{T}^h}\int_{e_K^+}(\psi_{\partial^0 K}-\psi)\,(i^*\Omega(u_K^+)-i^*\Omega(u_K^-)).
\end{aligned}
$$

We sum up with respect to K the identities

$$
\begin{aligned}
\int_K d(\psi\Omega)(u_K^-) &= \int_{\partial K}\psi\; i^*\Omega(u_K^-)\\
&= \int_{e_K^+}\psi\; i^*\Omega(u_K^-) - \int_{e_K^-}\psi\; i^*\Omega(u_K^-) + \sum_{e^0\in\partial^0 K}\int_{e^0}\psi\; i^*\Omega(u_K^-)
\end{aligned}
$$

and we combine them with (70). We arrive at the desired conclusion by observing that

$$\sum_{K\in\mathscr{T}^h}\Big(\int_{e_K^+}\psi\, i^*\Omega(u_K^+)-\int_{e_K^-}\psi\, i^*\Omega(u_K^-)\Big)=-\sum_{K\in\mathscr{T}_0^h}\int_{e_K^-}\psi\, i^*\Omega(u_{K,0}).$$

□

3.7 Convergence and Well-Posedness Results

This is the final step of our analysis.

Theorem 5 (Convergence Theory). *Under the assumptions in Sect. 3.4, the family of approximate solutions u^h generated by the finite volume scheme converges (as $h\to 0$) to an entropy solution to the initial value problem (24), (34).*

This theorem generalizes to spacetimes the technique originally introduced by Cockburn, Coquel and LeFloch [22,23] for the (flat) Euclidean setting and extended to Riemannian manifolds by Amorim et al. [1] and to Lorentzian manifolds by Amorim et al. [2].

Corollary 1 (Well-Posedness Theory on a Spacetime). *Fix $M=[0,T]\times N$ a $(n+1)$-dimensional spacetime foliated by n-dimensional hypersurfaces H_t ($t\in[0,T]$) with compact topology N (cf. (24)). Consider also a geometry-compatible flux field ω on M satisfying the global hyperbolicity condition (33). Given any initial data u_0 on H_0, the initial value problem (24), (34) admits a unique entropy solution $u\in L^\infty(M)$ which has well-defined L^1 traces on spacelike hypersurface of M. These solutions determines a (Lipschitz continuous) contracting semi-group:*

$$\int_{H'} i_{H'}^*\boldsymbol{\Omega}\left(u_{H'},v_{H'}\right)\le\int_H i_H^*\boldsymbol{\Omega}\left(u_H,v_H\right) \tag{71}$$

for any two hypersurfaces H, H' such that H' lies in the future of H, and the initial condition is assumed in the sense

$$\lim_{\substack{t\to 0\\ t>0}}\int_{H_t} i_{H_t}^*\boldsymbol{\Omega}\left(u(t),v(t)\right)=\int_{H_0} i_{H_0}^*\boldsymbol{\Omega}\left(u_0,v_0\right). \tag{72}$$

The following conclusion was originally established by DiPerna [30] for conservation laws posed on the Euclidian space.

Theorem 6. *Fix ω a geometry-compatible flux field on M satisfying the global hyperbolicity condition (33). Then, any entropy measure-valued solution ν to the initial value problem (24), (34) reduces to a Dirac mass and, more precisely,*

$$\nu = \delta_u, \tag{73}$$

where $u \in L^\infty(M)$ is the entropy solution to the problem.

We now give a proof of Theorem 5. By definition, a Young measure ν represents all weak-$*$ limits of composite functions $a(u^h)$ for all continuous functions a (as $h \to 0$):

$$a(u^h) \overset{*}{\rightharpoonup} \langle \nu, a \rangle = \int_{\mathbb{R}} a(\lambda)\, d\nu(\lambda). \tag{74}$$

Lemma 8 (Entropy Inequalities for Young Measures). *Given any Young measure ν associated with the approximations u^h. and for all convex entropy flux field Ω and smooth functions $\psi \geq 0$ supported away from the hypersurface $t = T$, one has*

$$\int_M \langle \nu, d\psi \wedge \Omega(\cdot) \rangle - \int_{H_0} i^* \Omega(u_0) \leq 0. \tag{75}$$

Thanks to (75), for all convex entropy pairs (U, Ω) we have $d\langle \nu, \Omega(\cdot) \rangle \leq 0$ on M. On the initial hypersurface H_0 the Young measure ν coincides with the Dirac mass δ_{u_0}. By Theorem 6 there exists a unique function $u \in L^\infty(M)$ such that the measure ν coincides with the Dirac mass δ_u. This implies that u^h converge strongly to u, and this concludes our proof of convergence.

Proof. We pass to the limit in (68), by using the property (74) of the Young measure. Observe that the left-hand side of (68) converges to the left-hand side of (75). Indeed, since ω is geometry-compatible, the first term

$$\sum_{K \in \mathscr{T}^h} \int_K d(\psi \Omega)(u_K^-) = \sum_{K \in \mathscr{T}^h} \int_K d\psi \wedge \Omega(u_K^-) = \int_M d\psi \wedge \Omega(u^h)$$

converges to $\int_M \langle \nu, d\psi \wedge \Omega(\cdot) \rangle$. On the other hand, one has

$$\sum_{K \in \mathscr{T}_0^h} \int_{e_K^-} \psi\, i^* \Omega(u_{K,0}) = \int_{H_0} \psi\, i^* \Omega(u_0^h) \to \int_{H_0} \psi\, i^* \Omega(u_0),$$

in which u_0^h is the initial discretization of the data u_0 converges strongly to u_0 since the maximal diameter h tends to zero.

The terms on the right-hand side of (68) also vanish in the limit $h \to 0$. We begin with the first term $A^h(\psi)$. Taking the modulus, applying Cauchy–Schwarz inequality, and using (66), we obtain

$$
\begin{aligned}
|A^h(\psi)| &\leq \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}|\psi_{\partial^0 K}-\psi||\tilde{u}^+_{K,e^0}-u_K^-| \\
&\leq \Big(\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}|\psi_{\partial^0 K}-\psi|^2\Big)^{1/2}\Big(\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}|\tilde{u}^+_{K,e^0}-u_K^-|^2\Big)^{1/2} \\
&\leq \Big(\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K}(C\,(\tau_{\max}+h))^2\Big)^{1/2}\Big(\int_{H_0} i^*\Omega(u_0)\Big)^{1/2},
\end{aligned}
$$

hence

$$
|A^h(\psi)| \leq C'\,(\tau_{\max}+h)\Big(\sum_{K\in\mathscr{T}^h}|e_K^+|\Big)^{1/2} \leq C''\,\frac{\tau_{\max}+h}{(\tau_{\min})^{1/2}}.
$$

Here, Ω is associated with the quadratic entropy and we used that $|\psi_{\partial^0 K}-\psi| \leq C\,(\tau_{\max}+h)$. Our conditions (49) imply that the upper bound for $A^h(\psi)$ tends to zero with h.

Next, we rely on the regularity of ψ and Ω and we estimate the second term in the right-hand side of (68). By setting $C_{e^0} := \frac{\int_{e^0} i^*\Omega(u_K^-)}{\int_{e^0} i^*\tilde{\omega}}$, we obtain

$$
\begin{aligned}
|B^h(\psi)| &= \Big|\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}}\int_{e^0}(\psi_{e^0}-\psi)\left(i^*\Omega(u_K^-)-C_{e^0}i^*\tilde{\omega}\right)\Big| \\
&\leq \sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}} \sup_K|\psi_{e^0}-\psi|\int_{e^0}\left|i^*\Omega(u_K^-)-C_{e^0}i^*\tilde{\omega}\right| \\
&\leq C\sum_{\substack{K\in\mathscr{T}^h\\ e^0\in\partial^0 K}}(\tau_{\max}+h)^2\,|e^0|_{\tilde{\omega}},
\end{aligned}
$$

hence $|B^h(\psi)| \leq C\,\frac{(\tau_{\max}+h)^2}{h}$. This implies the upper bound for $B^h(\psi)$ tends to zero with h.

Finally, we treat the last term in the right-hand side of (68)

$$
|C^h(\psi)| \leq \sum_{K\in\mathscr{T}^h}|e_K^+|\sup_K|\psi_{\partial^0 K}-\psi|\int_{e_K^+}|i^*\Omega(u_K^+)-i^*\Omega(u_K^-)|,
$$

using the modulus defined earlier. In view of (54), we obtain

$$|C^h(\psi)| \leq C \sum_{\substack{K\in\mathscr{T}^h \\ e^0\in\partial^0 K}} \frac{|e_K^+|}{N_K} |\psi_{\partial^0 K} - \psi| \left|\tilde{u}^+_{K,e^0} - u_K^-\right|,$$

and it is now clear that $C^h(\psi)$ satisfies the same estimate as the one we derived for $A^h(\psi)$. □

Acknowledgements The author was partially supported by the Agence Nationale de la Recherche (ANR) through the grant ANR SIMI-1-003-01, and by the Centre National de la Recherche Scientifique (CNRS). These notes were written at the occasion of a short course given by the authors at the University of Malaga for the XIV Spanish-French School Jacques-Louis Lions. The author is particularly grateful to C. Vázquez-Cendón and C. Parés for their invitation, warm welcome, and efficient organization during his stay in Malaga.

References

1. Amorim, P., Ben-Artzi, M., LeFloch, P.G.: Hyperbolic conservation laws on manifolds: total variation estimates and the finite volume method. Methods Appl. Anal. **12**, 291–324 (2005)
2. Amorim, P., LeFloch, P.G., Okutmustur, B.: Finite volume schemes on Lorentzian manifolds. Commun. Math. Sci. **6**, 1059–1086 (2008)
3. Beljadid, A., LeFloch, P.G., Mohamadian, M.: A geometry-preserving finite volume method for conservation laws in curved geometries (2003, preprint HAL-00922214)
4. Ben-Artzi, M., LeFloch, P.G.: The well-posedness theory for geometry compatible hyperbolic conservation laws on manifolds. Ann. Inst. H. Poincaré Nonlinear Anal. **24**, 989–1008 (2007)
5. Ben-Artzi, M., Falcovitz, J., LeFloch, P.G.: Hyperbolic conservation laws on the sphere: a geometry-compatible finite volume scheme. J. Comput. Phys. **228**, 5650–5668 (2009)
6. Berthon, C., Turpault, R.: Asymptotic preserving HLL schemes. Numer. Meth. Partial Differ. Equ. **27**, 1396–1422 (2011)
7. Berthon, C., Charrier, P., Dubroca, B.: An HLLC scheme to solve the M1 model of radiative transfer in two space dimensions. J. Sci. Comput. **31**, 347–389 (2007)
8. Berthon, C., Coquel, F., LeFloch, P.G.: Why many theories of shock waves are necessary: kinetic relations for nonconservative systems. Proc. R. Soc. Edinb. **137**, 1–37 (2012)
9. Berthon, C., LeFloch, P.G., Turpault, R.: Late-time/stiff-relaxation asymptotic-preserving approximations of hyperbolic equations. Math. Comput. **82**, 831–860 (2013)
10. Boscarino, S., Russo, G.: On a class of uniformly accurate IMEX Runge-Kutta schemes and application to hyperbolic systems with relaxation. SIAM J. Sci. Comput. **31**, 1926–1945 (2009)
11. Boscarino, S., LeFloch, P.G., Russo, G.: Highorder asymptotic preserving methods for fully nonlinear relaxation problems. SIAM J. Sci. Comput. (2014). See also ArXiv :1210.4761
12. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources. Birkhäuser, Zurich (2004)
13. Bouchut, F., Ounaissa, H., Perthame, B.: Upwinding of the source term at interfaces for Euler equations with high friction. J. Comput. Math. Appl. **53**, 361–375 (2007)
14. Boutin, B., Coquel, F., LeFloch, P.G.: Coupling techniques for nonlinear hyperbolic equations. I: self-similar diffusion for thin interfaces. Proc. R. Soc. Edinb. **141A**, 921–956 (2011)
15. Boutin, B., Coquel, F., LeFloch, P.G.: Coupling techniques for nonlinear hyperbolic equations. III: the well–balanced approximation of thick interfaces. SIAM J. Numer. Anal. **51**, 1108–1133 (2013)
16. Boutin, B., Coquel, F., LeFloch, P.G.: Coupling techniques for nonlinear hyperbolic equations. IV: multicomponent coupling and multidimensional wellbalanced schemes. Math. Comput. (2014). See ArXiv: 1206.0248

17. Buet, C., Cordier, S.: An asymptotic preserving scheme for hydrodynamics radiative transfer models: numerics for radiative transfer. Numer. Math. **108**, 199–221 (2007)
18. Buet, C., Després, B.: Asymptotic preserving and positive schemes for radiation hydrodynamics. J. Comput. Phys. **215**, 717–740 (2006)
19. Castro, M.J., LeFloch, P.G., Munoz-Ruiz, M.L., Pares, C.: Why many theories of shock waves are necessary: convergence error in formally path-consistent schemes. J. Comput. Phys. **227**, 8107–8129 (2008)
20. Chalons, C., LeFloch, P.G.: Computing undercompressive waves with the random choice scheme: nonclassical shock waves. Interfaces Free Boundaries **5**, 129–158 (2003)
21. Chen, G.Q., Levermore, C.D., Liu, T.P.: Hyperbolic conservation laws with stiff relaxation terms and entropy. Comm. Pure Appl. Math. **47**, 787–830 (1995)
22. Cockburn, B., Coquel, F., LeFloch, P.G.: An error estimate for high-order accurate finite volume methods for scalar conservation laws. Preprint 91-20, AHCRC Institute, Minneapolis, 1991
23. Cockburn, B., Coquel, F., LeFloch, P.G.: Error estimates for finite volume methods for multidimensional conservation laws. Math. Comput. **63**, 77–103 (1994)
24. Cockburn, B., Coquel, F., LeFloch, P.G.: Convergence of finite volume methods for multidimensional conservation laws. SIAM J. Numer. Anal. **32**, 687–705 (1995)
25. Coquel, F., LeFloch, P.G.: Convergence of finite difference schemes for conservation laws in several space dimensions. C. R. Acad. Sci. Paris Ser. I **310**, 455–460 (1990)
26. Coquel, F., LeFloch, P.G.: Convergence of finite difference schemes for conservation laws in several space dimensions: the corrected antidiffusive flux approach. Math. Comp. **57**, 169–210 (1991)
27. Coquel, F., LeFloch, P.G.: Convergence of finite difference schemes for conservation laws in several space dimensions: a general theory. SIAM J. Numer. Anal. **30**, 675–700 (1993)
28. Courant, R., Friedrichs, K., Lewy, H.: Über die partiellen Differenzengleichungen der mathematischen Physik. Math. Ann. **100**, 32–74 (1928)
29. Dal Maso, G., LeFloch, P.G., Murat, F.: Definition and weak stability of nonconservative products. J. Math. Pures Appl. **74**, 483–548 (1995)
30. DiPerna, R.J.: Measure-valued solutions to conservation laws. Arch. Ration. Mech. Anal. **88**, 223–270 (1985)
31. Donatelli, D., Marcati, P.: Convergence of singular limits for multi-D semilinear hyperbolic systems to parabolic systems. Trans. Am. Math. Soc. **356**, 2093–2121 (2004)
32. Dubois, F., LeFloch, P.G.: Boundary conditions for nonlinear hyperbolic systems of conservation laws. J. Differ. Equ. **31**, 93–122 (1988)
33. Ernest, J., LeFloch, P.G., Mishra, S.: Schemes with well-controlled dissipation (WCD). I. SIAM J. Numer. Anal. (2014)
34. Eymard, R., Gallouët, T., Herbin, R.: The finite volume method. In: Handbook of Numerical Analysis, vol. VII, pp. 713–1020. North-Holland, Amsterdam (2000)
35. Greenberg, J.M., Leroux, A.Y.: A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. SIAM J. Numer. Anal. **33**, 1–16 (1996)
36. Harten, A., Lax, P.D., van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM Rev. **25**, 35–61 (1983)
37. Hayes, B.T., LeFloch, P.G.: Nonclassical shocks and kinetic relations: finite difference schemes. SIAM J. Numer. Anal. **35**, 2169–2194 (1998)
38. Hou, T.Y., LeFloch, P.G.: Why nonconservative schemes converge to wrong solutions: error analysis. Math. Comput. **62**, 497–530 (1994)
39. Jin, S., Xin, Z.: The relaxation scheme for systems of conservation laws in arbitrary space dimension. Commun. Pure Appl. Math. **45**, 235–276 (1995)
40. Kröner, D.: Finite volume schemes in multidimensions. In: Numerical Analysis 1997 (Dundee). Pitman Research Notes in Mathematics Series, vol. 380, pp. 179–192. Longman, Harlow (1998)
41. Kröner, D., Noelle, S., Rokyta, M.: Convergence of higher-order upwind finite volume schemes on unstructured grids for scalar conservation laws with several space dimensions. Numer. Math. **71**, 527–560 (1995)

42. Kruzkov, S.: First-order quasilinear equations with several space variables. Math. USSR Sb. **10**, 217–243 (1970)
43. Lax, P.D.: Hyperbolic systems of conservation laws and the mathematical theory of shock waves. In: Regional Conference Series in Applied Mathematics, vol. 11. SIAM, Philadelphia (1973)
44. LeFloch, P.G.: An introduction to nonclassical shocks of systems of conservation laws. In: Kroner, D., Ohlberger, M., Rohde, C. (eds.) International School on Hyperbolic Problems, Freiburg, Germany, Oct. 1997. Lecture Notes on Computer Engineering, vol. 5, pp. 28–72. Springer, Berlin (1999)
45. LeFloch, P.G.: Hyperbolic systems of conservation laws: the theory of classical and nonclassical shock waves. In: Lectures in Mathematics. ETH Zürich/Birkhäuser, Basel (2002)
46. LeFloch, P.G.: Hyperbolic conservation laws and spacetimes with limited regularity. In: Benzoni, S., Serre, D. (eds.) Proceedings of 11th International Conference on Hyperbolic Problems: Theory, Numerics, and Applications, pp. 679–686. ENS Lyon, 17–21 July 2006. Springer, Berlin (See arXiv:0711.0403)
47. LeFloch, P.G.: Kinetic relations for undercompressive shock waves: physical, mathematical, and numerical issues. Contemp. Math. **526**, 237–272 (2010)
48. LeFloch, P.G., Makhlof, H.: A geometry-preserving finite volume method for compressible fluids on Schwarzschild spacetime. Commun. Comput. Phys. (2014). See also ArXiv :1212.6622
49. LeFloch, P.G., Mishra, S.: Numerical methods with controled dissipation for small-scale dependent shocks. Acta Numer. (2014). See Preprint ArXiv 1312.1280
50. LeFloch, P.G., Mohamadian, M.: Why many shock wave theories are necessary: fourth-order models, kinetic functions, and equivalent equations. J. Comput. Phys. **227**, 4162–4189 (2008)
51. LeFloch, P.G., Okutmustur, B.: Hyperbolic conservation laws on manifolds with limited regularity. C. R. Math. Acad. Sci. Paris **346**, 539–543 (2008)
52. LeFloch, P.G., Okutmustur, B.: Hyperbolic conservation laws on spacetimes: a finite volume scheme based on differential forms. Far East J. Math. Sci. **31**, 49–83 (2008)
53. LeFloch, P.G., Rohde, C.: High-order schemes, entropy inequalities, and nonclassical shocks. SIAM J. Numer. Anal. **37**, 2023–2060 (2000)
54. LeFloch, P.G., Makhlof, H., Okutmustur, B.: Relativistic Burgers equations on curved spacetimes: derivation and finite volume approximation. SIAM J. Numer. Anal. (2012, preprint). ArXiv:1206.3018
55. LeVeque, R.J.: Balancing source terms and flux gradients in high-resolution Godunov methods: the quasi-steady wave-propagation algorithm. J. Comput. Phys. **146**, 346–365 (1998)
56. LeVeque, R.J.: Finite volume methods for hyperbolic problems. In: Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (2002)
57. Marcati, P.: Approximate solutions to conservation laws via convective parabolic equations. Commun. Partial Differ. Equ. **13**, 321–344 (1988)
58. Marcati, P., Milani, A.: The one-dimensional Darcy's law as the limit of a compressible Euler flow. J. Differ. Equ. **84**, 129–146 (1990)
59. Marcati, P., Rubino, B.: Hyperbolic to parabolic relaxation theory for quasilinear first order systems. J. Differ. Equ. **162**, 359–399 (2000)
60. Nessyahu, H., Tadmor, E.: Non-oscillatory central differencing for hyperbolic conservation laws. J. Comput. Phys. **87**, 408–463 (1990)
61. Russo, G.: Central schemes for conservation laws with application to shallow water equations. In: Rionero, S., Romano, G. (eds.) Trends Applied Mathematics and Mechanics, STAMM 2002, pp. 225–246. Springer, Italia SRL (2005)
62. Russo, G.: High-order shock-capturing schemes for balance laws. In: Numerical Solutions of Partial Differential Equations. Advanced Courses in Mathematics, CRM Barcelona, pp. 59–147. Birkhäuser, Basel (2009)
63. Russo, G., Khe, A.: High-order well-balanced schemes for systems of balance laws. In: Hyperbolic Problems: Theory, Numerics and Applications. Proceedings of Symposia in Applied Mathematics, Part 2, vol. 67, pp. 919–928. American Mathematical Society, Providence (2009)

64. Szepessy, S.: Convergence of a shock-capturing streamline diffusion finite element method for a scalar conservation law in two space dimensions. Math. Comput. **53**, 527–545 (1989)
65. Tadmor, E.: Approximate solutions of nonlinear conservation laws. In: Advanced Numerical Approximation of Nonlinear Hyperbolic Equations (Cetraro, 1997). Lecture Notes in Mathematics, vol. 1697, pp. 1–149. Springer, Berlin (1998)
66. Westdickenberg, M., Noelle, S.: A new convergence proof for finite volume schemes using the kinetic formulation of conservation laws. SIAM J. Numer. Anal. **37**, 742–757 (2000)

Part II
Talks

Gradient Calculus for a Class of Optimal Design Problems in Engineering

Carlos Castro

Abstract This chapter reviews some recent works in which the analysis and control of partial differential equations are applied to optimal design in some problems appearing in aerodynamics and elasticity. From a mathematical point of view, the idea is to apply a descent algorithm to a cost functional defined on a part of the boundary. More specifically, we focus here on problems where the cost functional is defined on the part of the boundary to be optimized. This is the case, for instance, when the goal is to improve the lift or the drag in aerodynamic problems or to uniformize the tangential stresses along the boundary of a elastic material.

1 Introduction

This work contains a series of applications of control problems to aerodynamics and elasticity problems with the aim of improving the industrial software in simulation. We focus mainly on aerodynamic applications since they have been more extensively studied in the last years. However the methodology considered here is general and can be easily adapted to structural optimization, as we show in Sect. 6.

In the last years, advanced software for automatic aerodynamic design optimization has been extensively used by engineers to avoid expensive experimental proofs in wind tunnels (see the early works by A. Jameson [15] and O. Pironneau [22] or the more recent book [20] and the references therein). This optimization software is based on gradient methods to minimize a suitable cost or objective function

C. Castro (✉)
Dep. Matemática e Informática, ETSI Caminos Canales y Puertos,
Universidad Politécnica de Madrid, Madrid, Spain
e-mail: carlos.castro@upm.es

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*,
SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_5,

(drag coefficient, deviation from a prescribed surface pressure distribution, etc.) with respect to a set of design variables (defining, for example, an airfoil profile). This is a complex problem where several difficulties arise: parametrization of complex geometries, suitable choice of the correct systems of equations to model the fluid according to the underlying physics (Euler equations, Navier–Stokes, RANS, turbulence models, etc.), numerical methods to solve the differential equations, mesh generation, mesh adaptivity to small changes in the geometry, cost function approximation, gradient approximation, etc. These and other industrial constraints make any practical application of such a technology a very complex task. Mathematical analysis can be useful to improve some of the factors involved in this process. Here we focus on the computation of the gradient of cost functionals associated to optimal design.

To fix the problem we consider a fluid domain Ω bounded by a typically disconnected boundary $\partial\Omega$ which is divided into a far-field component Γ_∞ and a wall boundary S (Fig. 1). Aeronautic optimization problems seek the minimization of a certain cost function, such as the deviation of the pressure on S from a prescribed pressure distribution in the so-called inverse design problems, or integrated force coefficients (drag or lift) in force optimization problems. In these examples the cost function J can be defined as an integral over the wall boundary S of a suitable function $f(U, S)$ of the flow variables, referred to as a vector U, and the geometry S

$$J(S) = \int_S f(U, S)\, ds. \tag{1}$$

The flow variables U satisfy a suitable flow model (Euler, Navier–Stokes, RANS, etc.), that we write as

$$R(U) = 0, \qquad x \in \Omega, \tag{2}$$

including initial and boundary conditions.

Note that the cost functional depends on a part S of the boundary of the domain, which will be referred to as the control variable. The set of admissible controls is therefore a set of different geometries for S that we refer as S_{ad}. We are interested in the following problem: Find $S_{min} \in S_{ad}$ such that,

$$J(S_{min}) = \min_{S \in S_{min}} J(S). \tag{3}$$

To prove the existence of solution for the above minimization problem is, in general, a difficult problem which strongly depends on the flow equations, the restrictions included in S_{ad}, and the functional itself.

However, since these aerodynamical problems are very sensitive to perturbations of the domain, rather than looking for an optimal S, in the applications one tries to improve a given "natural" design by performing small perturbations. Therefore,

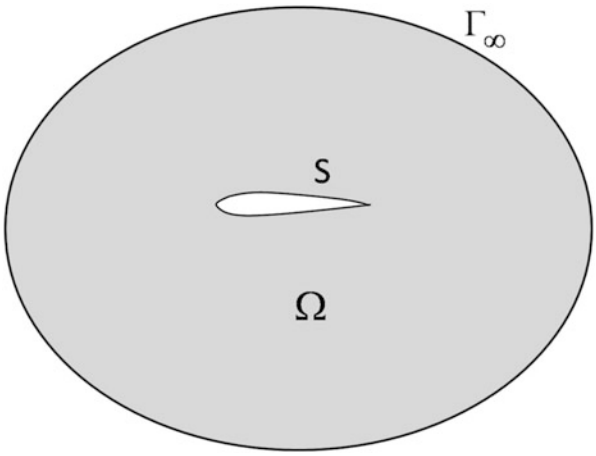


Fig. 1 Exterior domain with boundary S

the main interest is to make a sensitivity analysis of J with respect to small perturbations of the boundary S. Once this is done, the deformation for which the functional decreases with highest rate is chosen: the best descent direction. In other words, the main objective is to compute the shape derivative of J.

Another important point is that, in the engineering practice, instead of computing the exact continuous objective function, one computes a discrete approximation in which the time and physical domain are discretized. The objective function is evaluated by means of a discrete integration rule and the variables U are estimated by means of a numerical approximation scheme for solving the flow equations. Therefore our real optimization problem is in fact a discrete version of (1)–(3), and the sensitivity analysis should be done for such discrete version. This is usually referred to as a *discrete approach* to obtain sensitivity (see for example [19, 20]). Note that this sensitivity analysis will depend on the discretization aspects, such as the numerical scheme used to approximate the flow variables, the mesh, the numerical approximation of the cost functional, and even on the implementation issues such as multigrid techniques and, possibly, parallel computation.

In contrast with this *discrete approach* there is the alternative *continuous approach* where the sensitivity analysis is obtained for the continuous system and then discretized to obtain the optimal descent direction for the discrete model (see [14]). The validity of this *continuous approach* to obtain an accurate sensitivity analysis of the discrete model is not obvious. It is usually based on strong convergence results of the chosen discretization for (1)–(3) and the smoothness of solutions. On the other hand, the continuous approach makes easier the analysis and reduce the dependence on the numerical scheme chosen to obtain the flow variables. We refer to [21] for a comparison between both approaches, the discrete and continuous.

In this work we focus on the continuous approach of the sensitivity analysis, that we will briefly describe.

In order to define the shape deformation of the control boundary S we introduce a suitable parametrization of S given by $\mathbf{x} : [0, 1] \to \mathbb{R}^2$. A generic deformation of the boundary can be described as a vector field $\delta\mathbf{x}(s)$ such that the new geometry S' is parametrized by $\mathbf{x}'(s) = \mathbf{x}(s) + \delta\mathbf{x}(s)$. For sufficiently small perturbations, $\delta\mathbf{x}(s)$ can be described by normal displacements as follows:

$$\delta\mathbf{x}(s) = \alpha(s)\mathbf{n}, \quad \mathbf{n} \text{ normal vector to } S, \tag{4}$$

since tangent deformations are equivalent to reparameterizations of the boundary. The function α represents a perturbation profile which describes the amount of displacement, in the normal direction, at each point of S. This α is usually taken in a finite dimensional subspace generated by some basis functions (polynomial, trigonometric, etc.)

$$\tilde{U}_{ad} = span(\alpha_1, \alpha_2, \ldots, \alpha_n).$$

The sensitivity analysis for the continuous model consists in finding the shape derivative of J, i.e. the derivative of J with respect to any deformation profile $\alpha \in \tilde{U}_{ad}$, and then the best decreasing rate is chosen. This will constitute the descent direction for J. There are two main approaches that have been tried in industrial applications: *finite differences* and *adjoint methodology*.

In the finite difference approach, shape derivatives are calculated by computing the finite difference

$$\frac{J(S_{\alpha_k,\varepsilon}) - J(S)}{\epsilon}, \quad \varepsilon << 1,$$

where $S_{\alpha_k,\varepsilon}$ is the new geometry obtained from S with the parametrization $\mathbf{x}(s) + \varepsilon\alpha_k\mathbf{n}(s)$. This is done for each $k = 1, .., n$. The parameter ε should be chosen small enough to recover the linear behavior but not too small to avoid round errors. In this way, partial derivatives with respect to each α_k are computed. The one with the highest decreasing rate is chosen as the descent direction for computing the new geometry. The main drawback of this approach is that it is computationally too expensive. Note that each finite difference of J requires an evaluation of the cost functional and therefore a new solution of the flow equations. On the other hand, the choice of the value of ε is difficult and an adequate strategy to estimate it has to be used.

A more efficient way to compute a descent direction for J is the adjoint method, in which one seeks for the following representation of the Gateaux derivative of J with respect to α,

$$\delta J = \int_S G(s)\alpha(s)\, ds,$$

for some function $G(s)$, usually known as gradient of J. Once this is known, an optimal descent direction is chosen by projecting $-G(s)$ in the subspace of admissible deformations $\tilde{U}_{ad}$.

The computation of G involves shape derivatives, in the sense given by Hadamard (see [12]), and classical control theory which reduces the computation of the gradient to the resolution of a suitable adjoint system. In contrast with the finite difference approach, only one system has to be solved to obtain the descent direction. However, this adjoint system does not issue from a physical fluid problem but from an algebraic calculation. Therefore the usual numerical methods for fluids are not well-adapted to solve it, in general, and a particular numerical analysis is needed to find efficient methods.

The adjoint method is in fact a particular application of the classical control theory for partial differential equations. This theory was significantly developed due, in particular, to the works of J.-L. Lions [18]. Later on O. Pironneau investigated the application of the control theory to the optimal shape design for elliptic equations [22]. In the late eighties A. Jameson [15] was the first to apply these techniques to the Euler and Navier–Stokes equations in the field of aeronautical applications. From these pioneering works a lot of new results and applications have made of this topic an essential tool in optimal design.

In this work we review the continuous adjoint, when considering different models to approximate the flow variables, namely the Euler equation (Sect. 3), Navier–Stokes equations (Sect. 4), and Euler equations in presence of shock waves (Sect. 5). The analysis has been validated with two-dimensional and three-dimensional examples. At this moment, the Navier–Stokes sensitivity analysis is implemented in experimental versions of high performance codes as SU2 (Standford University) and TAU (developed in Germany by DLR). It is worth mentioning that the extension of the continuous approach to the sensitivity analysis of RANS equations with Spalart–Allmaras model for turbulence has been studied in [6], where gradient formulas are derived. In Sect. 6 we show an application of this technique in the context of elasticity problems.

2 Gradient Computation

In this section we describe the methodology to obtain gradient formulas for the cost functional in a systematic way. It is worth mentioning that this calculus is formal since it assumes that solutions of the underlying differential equations are smooth. This is not true in general. As it is well known, Euler equations may produce discontinuities even for a smooth initial data. For simplicity, we focus on dimension $n = 2$ but the case $n = 3$ can be treated similarly.

Let us consider $U_{ad} \subset L^2(0, 1)$ and the functional $J : U_{ad} \to \mathbb{R}$

$$J(\alpha) = \int_S j(U)\, ds \tag{5}$$

where S is described as a normal perturbation of a reference geometry S_0 in such a way that $S = S_0 + \alpha(s)\mathbf{n}$ and $\alpha \in \tilde{U}_{ad} \subset L^2(0,1)$. The vector function U is the solution of system (2). All the functionals considered below can be written in this form.

Classical shape derivatives allow us to write the Gateaux derivative of J, δJ, in the generic direction α (see [23]), as follows:

$$\delta J = \int_S \left(\frac{\partial j}{\partial U} \delta U + (\partial_n j + \kappa j)\alpha \right) ds \tag{6}$$

where κ represents the curvature of S (in 3-d the boundary S will be a surface and κ should be replaced by $2H$ with H the mean curvature of S). The vector function δU represents the Gateaux derivative of U in the direction given by α and it is obtained by linearization of system (2),

$$\frac{\partial R}{\partial U} \delta U = 0, \qquad x \in \Omega. \tag{7}$$

Now we introduce an adjoint state Ψ for which

$$\int_S \frac{\partial j}{\partial U} \delta U \, ds = \int_S \mathscr{B}\Psi \delta\alpha \, ds, \tag{8}$$

where $\mathscr{B}$ is a certain operator and Ψ satisfies the so-called adjoint system

$$\mathscr{A}\Psi = 0, \qquad x \in \Omega. \tag{9}$$

The operators $\mathscr{A}$ and $\mathscr{B}$ strongly depend on the flow equations and boundary conditions included in $R(U)$, and they must be computed specifically for each problem. We show an example in the appendix below.

Once obtained the adjoint state we can replace (8) into (6),

$$\delta J = \int_S (\mathscr{B}\Psi + \partial_n j + \kappa j) \, \delta\alpha \, ds, \tag{10}$$

and therefore

$$G(s) = \mathscr{B}\Psi(s) + \partial_n(j(U(s))) + \kappa j(U(s)).$$

Remark 1. In general, the operator $\mathscr{A}$ is closely related to the linearized system (7) and its numerical approximation should take into account this fact. There are several ways to deduce numerical schemes for (7) but the more stable ones are usually obtained by a suitable adjoint of the linearization of the numerical methods for $R(U)$. This can be done at several levels, from a specific code based on the linearized numerical scheme to automatic differentiation tools that provides a linearization of the whole numerical code used to solve $R(U)$, including parallelization, multigrid, preconditioners, etc.

3 Continuous Adjoint Formulation for Euler System

We first consider the case of steady inviscid two dimensional flows. We present a brief description of the continuous adjoint formulas. We refer to [1,7] for a complete analysis and full formulation.

The governing equations in this case are

$$\nabla \cdot F = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} = 0, \quad \text{in } \Omega, \quad U = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{pmatrix}, \tag{11}$$

$$F_x = \begin{pmatrix} \rho u \\ \rho u^2 + P \\ \rho uv \\ \rho u H \end{pmatrix}, \qquad F_y = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + P \\ \rho v H \end{pmatrix}. \tag{12}$$

Here, ρ is the density, u and v are the Cartesian velocity components, E is the total energy, and P and H are the pressure and enthalpy, given by the following relations:

$$P = (\gamma - 1)\rho \left[E - \frac{1}{2}(u^2 + v^2) \right], \quad H = E + \frac{P}{\rho}, \tag{13}$$

where γ is the ratio of specific heats. The above system must be completed with suitable boundary conditions. We consider characteristic-type boundary conditions [16] on the far-field boundary Γ_∞, and non-penetrating boundary conditions on solid wall boundaries,

$$\mathbf{v} \cdot \mathbf{n} = 0, \quad \mathbf{v} = (u, v) \quad \mathbf{n} = (n_x, n_y), \quad \text{normal vector on } S. \tag{14}$$

$$\text{Far field boundary conditions on } \Gamma_\infty. \tag{15}$$

The operator $R(U)$ in this case contains the whole system of equations and boundary conditions (11)–(15).

Concerning the cost function, there are several possibilities according to different interests. Conventional cost functions include specified pressure distributions (inverse design), force (drag or lift) or moment coefficients, efficiency (i.e., lift over drag), etc. All these cost functionals can be written in the general form:

$$J(S) = \int_S g(P, \mathbf{n})\, ds \tag{16}$$

for some function $g(P, \mathbf{n})$. For example, in the particular case of lift-drag coefficients, the cost functional take the form

$$J(S) = \int_S C_p(\mathbf{n}\cdot\mathbf{d})\,ds, \qquad \mathbf{d} = \begin{cases} (\cos\beta, \sin\beta), & (drag), \\ (\sin\beta, \cos\beta), & (lift), \end{cases} \tag{17}$$

where β is a constant parameter (angle of attack), $C_p = (P - P_\infty)/C_\infty$, $C_\infty = \gamma M_\infty^2 P_\infty/2$, and P_∞ and M_∞ are freestream pressure and Mach number respectively.

Following the general framework in Sect. 2 above $j(U) = g(P(U), \mathbf{n})$.

The final expression for G in the case of (16) is given by

$$G = \frac{\partial g}{\partial P}\partial_n P + \mathbf{t}\cdot\partial_{tg}\frac{\partial g}{\partial \mathbf{n}} - \kappa\left(g - \frac{\partial g}{\partial \mathbf{n}}\cdot\mathbf{n}\right)$$
$$-\nabla\cdot\mathbf{v}(\rho\psi_1 + \rho\mathbf{v}\cdot\boldsymbol{\varphi} + \rho H\psi_4) + \mathbf{t}\cdot\mathbf{v}\partial_{tg}(\rho\psi_1 + \rho\mathbf{v}\cdot\boldsymbol{\varphi} + \rho H\psi_4)$$

where κ is the curvature of S (for 3D flows the mean curvature appears), $\mathbf{t}$ is the unitary tangent vector to S, ∂_n the normal derivative and ∂_{tg} the tangential derivative. The adjoint variables

$$\Psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{pmatrix}, \qquad \boldsymbol{\varphi} = (\psi_2, \psi_3),$$

satisfy the adjoint system

$$A^T\nabla\Psi = 0, \qquad A = \left(\frac{\partial F_x}{\partial U}, \frac{\partial F_y}{\partial U}\right),$$

and the boundary conditions

$$\varphi\cdot\mathbf{n} = \frac{\partial g}{\partial P}, \qquad \text{on } S,$$
$$\text{adjoint far field conditions on } \Gamma_\infty.$$

We refer to [13] for details on how this adjoint boundary conditions are obtained and implemented numerically.

4 Continuous Adjoint Formulation for Navier–Stokes System

In this section we consider the Navier–Stokes system. We refer to [7] for the full expression of the adjoint system, the derivation of the gradient formula, and some numerical experiments. The gradient formula for 3D flows is also given in [7].

The governing equations, for steady viscous laminar flows in two dimensions, are

$$\nabla \cdot F - \nabla \cdot F^v = 0, \quad \text{in } \Omega, \tag{18}$$

where $F = (F_x, F_y)$ has been defined in (12) and

$$F_x^v = \begin{pmatrix} 0 \\ \sigma_{xx} \\ \sigma_{xy} \\ u\sigma_{xx} + v\sigma_{xy} + k\frac{\partial T}{\partial x} \end{pmatrix}, \qquad F_y^v = \begin{pmatrix} 0 \\ \sigma_{xy} \\ \sigma_{yy} \\ u\sigma_{yx} + v\sigma_{yy} + k\frac{\partial T}{\partial y} \end{pmatrix}. \tag{19}$$

The viscous stresses may be written as

$$\sigma_{xx} = \frac{2}{3}\mu\left(2u_x - v_y\right), \qquad \sigma_{yx} = \sigma_{xy} = \mu\left(u_y + v_x\right),$$
$$\sigma_{yy} = \frac{2}{3}\mu\left(2v_y - u_x\right),$$

where μ is the laminar viscosity coefficient. The coefficient of thermal conductivity and the temperature are computed as follows:

$$k = \frac{c_p}{Pr}\mu, \qquad T = \frac{P}{R\rho},$$

where c_p is the specific heat at constant pressure, Pr is the Prandtl number, and R is the gas constant.

Equation (18) is complemented with characteristic-type boundary conditions on the far field, and nonslip conditions on solid walls

$$u = v = 0, \quad \text{on } S.$$

An additional boundary condition has to be imposed to the temperature on the solid walls, which can be either adiabatic or isothermal (constant temperature)

$$\partial_n T = 0, \quad \text{adiabatic},$$
$$T = T_0, \quad \text{constant temperature}.$$

In the adiabatic case, the expression for G is given by

$$G = \frac{\partial g}{\partial P}\partial_n P + \mathbf{t} \cdot \partial_{tg}\frac{\partial g}{\partial \mathbf{n}} - \kappa\left(g - \frac{\partial g}{\partial \mathbf{n}} \cdot \mathbf{n}\right)$$
$$-(\mathbf{n} \cdot \partial_n \mathbf{v})(\rho\psi_1 + \rho H\psi_4) + \mathbf{n} \cdot \Sigma \cdot \partial_n \mathbf{v} - \psi_4(\mathbf{n} \cdot \sigma \cdot \partial_n \mathbf{v})$$
$$+\psi_4(\sigma : \nabla \mathbf{v}) - k(\partial_{tg}\psi_4)(\partial_{tg} T),$$

where ' :' denotes the double dot contraction of second order tensor fields. The adjoint variables satisfy the adjoint system

$$(A + A^v)^T \nabla \Psi = 0, \qquad A^v = \left(\frac{\partial F_x^v}{\partial U}, \frac{\partial F_y^v}{\partial U} \right),$$

with boundary conditions

$$\varphi = \frac{\partial g}{\partial P} \mathbf{n}, \qquad \text{on } S$$

and adjoint farfield boundary conditions on Γ_∞.

The second order tensor Σ is defined as follows

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \qquad \Sigma_{xx} = \frac{2}{3}\mu(2\partial_x \psi_2 - \partial_y \psi_3),$$

$$\Sigma_{xy} = \Sigma_{yx} = \mu(\partial_y \psi_2 + \partial_x \psi_3), \qquad \Sigma_{yy} = \frac{2}{3}\mu(2\partial_y \psi_3 - \partial_x \psi_2).$$

5 Continuous Adjoint Formulation for Euler System in the Presence of Shock Waves

So far, we have considered smooth solutions of flow equations. In this case, the perturbation of the flow field variables with respect to shape changes can be approximated by linearizing the governing equations. However, inviscid flows described by the Euler equations can develop discontinuities (shocks or contact discontinuities) due to the intersection of characteristics. In this case, the smooth analysis in Sect. 3 is no longer valid. We refer to [2] for the complete formulation and analysis of this section.

In this section we restrict ourselves to the particular case where there is a single discontinuity located on a smooth curve Σ (Fig. 2). When this occurs, Euler system (11) must be completed with the Rankine–Hugoniot conditions that relate the flow variables on both sides of the discontinuity. Thus, we replace (11) by

$$\begin{cases} \nabla \cdot F = 0, & \text{in } \Omega \backslash \Sigma, \\ [F \cdot \mathbf{n}]_\Sigma = 0, & \text{on } \Sigma, \end{cases} \tag{20}$$

where $[A]_\Sigma$ represents the jump of A across the discontinuity curve Σ, i.e. $[A]_\Sigma = A^+ - A^-$.

The sensitivity analysis in this case is much more complex since a perturbation of the boundary S may affect to the position of the discontinuity Σ. Thus, the

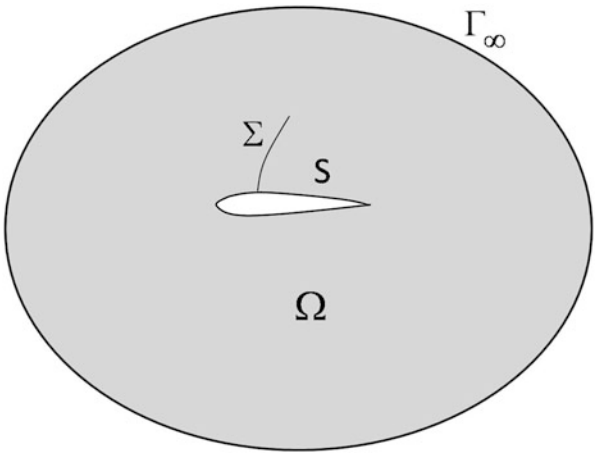


Fig. 2 Exterior domain with boundary S and shockwave

variational calculus must be modified to take into account the position of the discontinuity as a new variable. This analysis has been done in simpler models by different authors (see for instance [3, 4, 10, 11]). Moreover, in [8, 17] the position of the discontinuity is used to improve gradient algorithms in the case of the inviscid Burgers equation in 1-d.

A formal calculus based on this approach allows us to obtain a formula for G in this particular case.

We must distinguish two different situations: either the shock wave Σ meets the boundary S at a point $x_b \in S$, or it does not. We focus on the first case since the second one is simpler. We have the following

$$G = \frac{\partial g}{\partial P}\partial_n P + \mathbf{t}\cdot\partial_{tg}\frac{\partial g}{\partial \mathbf{n}} - \kappa\left(g - \frac{\partial g}{\partial \mathbf{n}}\cdot\mathbf{n}\right)$$
$$-\nabla\cdot\mathbf{v}(\rho\psi_1 + \rho\mathbf{v}\cdot\boldsymbol{\varphi} + \rho H\psi_4) + \mathbf{t}\cdot\mathbf{v}\partial_{tg}(\rho\psi_1 + \rho\mathbf{v}\cdot\boldsymbol{\varphi} + \rho H\psi_4),$$

for $x \in S$ but $x \neq x_b$. Note that this formula is analogous to the gradient formula for smooth solutions. The only difference is that we do not have to compute it at the discontinuity point x_b where the flow variables may have discontinuities and their derivatives may produce singularities. The adjoint system is given by

$$\begin{cases} A^T\nabla\Psi = 0, & \text{in } \Omega\backslash\Sigma \\ \mathbf{t}\cdot\partial_{tg}\boldsymbol{\varphi} = 0, & \text{on } \Sigma, \\ [\Psi]_\Sigma = 0, \end{cases} \tag{21}$$

together with the adjoint boundary conditions for the far field and

$$\varphi \cdot \mathbf{n} = \frac{\partial g}{\partial P}, \qquad \text{on } S.$$

The second and third equations in (21) are transmission conditions that comes from the Rankine–Hugoniot conditions by duality. They are usually referred as adjoint Rankine–Hugoniot conditions. They establish, in particular, that the adjoint vector variables Ψ must be continuous at Σ.

It is worth mentioning that well-posedness of the adjoint system (21) is a difficult task due to the discontinuity of the matrix coefficients A at Σ. This is a challenging problem even for simpler scalar conservation laws in one dimension [5].

6 An Example in Elasticity

In this section we apply the same strategy in the context of elasticity problems. In particular, we consider optimal design problems whose cost functions depend on the stresses at the boundary to be optimized. An example described in [9] considers the shape optimization of the cross-sectional vault of a tunnel in order to have uniform stresses along the profile (see also [24]). In this way, we avoid regions with larger compression stresses at the boundary that could produce more fatigue. For this specific problem, a two-dimensional elastic problem is solved for the cross-section of the tunnel with the following objective function

$$J(\alpha) = \frac{1}{2}\int_S (\sigma_t - \sigma_{tm})^2 \, ds, \tag{22}$$

where σ_t represents the tangential stresses along S ($\sigma_t = \mathbf{t}\cdot\sigma\cdot\mathbf{t}$ where σ is the stress tensor and $\mathbf{t}$ the tangent vector to S) and σ_{tm} a reference value that can be either a given constant or the average of the tangential stresses along S,

$$\sigma_{tm} = \frac{\int_S \sigma_t ds}{\int_S ds}.$$

Of course, other functionals are also possible according to the interest of the application.

Let us state the problem: consider the elasticity system defined on a domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega = \Gamma \cup S$ and $\Gamma \bigcap S = \emptyset$, (Fig. 3) and the objective function

$$J(\alpha) = \int_S j(\mathbf{t}\cdot\sigma\cdot\mathbf{t}) \, ds, \tag{23}$$

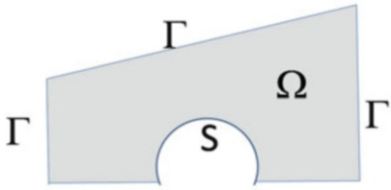


Fig. 3 Cross section of a tunnel vault

for some function j, where $\sigma = \sigma_{\alpha\beta}$ is the second order stress tensor and $\mathbf{t}$ the tangent vector to S that we obtain rotating $\pi/2$ the outward normal clockwise. The stress tensor is obtained by solving the elasticity system

$$\sigma_{\alpha\beta,\beta} + f_\alpha = 0, \quad \mathbf{x} \in \Omega, \quad \alpha, \beta = 1, 2, \tag{24}$$

$$\sigma_{33} = \nu(\sigma_{11} + \sigma_{22}), \quad \mathbf{x} \in \Omega, \tag{25}$$

$$\varepsilon_{\alpha\beta} = \frac{1+\nu}{E}\sigma_{\alpha\beta} - \frac{\nu}{E}\sigma_{kk}\delta_{\alpha\beta}, \tag{26}$$

$$\varepsilon_{13} = \varepsilon_{23} = \varepsilon_{33} = 0, \tag{27}$$

where $\mathbf{x} = (x_1, x_2) \in \Omega$ is a generic point of the elastic body, (f_1, f_2) the components of the external forces, $\delta_{\alpha\beta}$ the Kronecker delta and $\varepsilon_{\alpha\beta}$ are the components of the strain tensors respectively. The elastic constants of the isotropic material are the Young modulus E and Poisson ratio ν. Partial derivative is denoted by a comma (,). The expression of the strain tensor components are given as a function of the displacements as follows:

$$\varepsilon_{\alpha\beta} = \frac{1}{2}\left(u_{\alpha,\beta} + u_{\beta,\alpha}\right). \tag{28}$$

To fix ideas, the following boundary conditions are assumed

$$u_\alpha = \bar{u}_\alpha, \quad \mathbf{x} \in \Gamma, \quad \alpha = 1, 2 \tag{29}$$

$$\sigma_{\alpha\beta} n_\beta = 0, \quad \mathbf{x} \in S, \quad \alpha, \beta = 1, 2 \tag{30}$$

where $\bar{u}_\alpha$ are specified displacement, $\mathbf{n} = (n_1, n_2)$ is the outward normal unit vector to the boundary Γ. Other boundary conditions are also possible.

The gradient in this case is given by

$$G = -\partial_{tg}(j'(\sigma_t)(\mathbf{n}\cdot\sigma\cdot\mathbf{t} + \mathbf{t}\cdot\sigma\cdot\mathbf{n})) - \Psi\cdot\partial_n\sigma\cdot\mathbf{n} + \partial_{tg}(\Psi\cdot\sigma\cdot\mathbf{t})$$
$$-\frac{\nu}{1-\nu}j'(\sigma_t)\mathbf{n}\cdot\partial_n(\sigma)\cdot\mathbf{n} + [\kappa j(\sigma_t) + \partial_n\left(j(\sigma_t)\right)]$$

where ∂_{tg} and ∂_n represent the tangential and normal derivatives respectively, and $\Psi = (\psi_1, \psi_2)$ satisfies the adjoint problem,

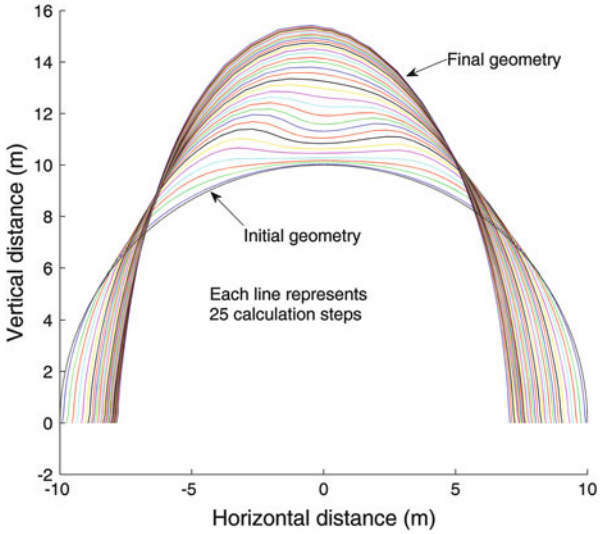


Fig. 4 Cross-sectional designs obtained with a gradient type method

$$\frac{\partial \sigma^*_{\alpha\beta}}{\partial \beta} = 0, \quad x \in \Omega, \quad \alpha, \beta = 1, 2,$$
$$\sigma^*_{33} = \nu(\sigma^*_{11} + \sigma^*_{22}), \quad x \in \Omega,$$
$$\epsilon^*_{\alpha\beta} = \frac{1+\nu}{E}\sigma^*_{\alpha\beta} - \frac{\nu}{E}\sigma^*_{kk}\delta_{\alpha\beta},$$
$$\epsilon^*_{\alpha\beta} = \frac{1}{2}\left(\frac{\partial \psi_\beta}{\partial \alpha} + \frac{\partial \psi_\alpha}{\partial \beta}\right),$$
$$\epsilon^*_{13} = \epsilon^*_{23} = \epsilon^*_{33} = 0,$$

with the following boundary conditions:

$$\psi_\alpha = 0, \quad x \in \Gamma, \quad \alpha = 1, 2,$$
$$\sigma^*_{\alpha\beta} n_\beta = \frac{-E}{1-\nu^2}\partial_{tg}\left(j'(\sigma_t)\right) t_\alpha, \quad x \in S, \quad \alpha = 1, 2.$$

The practical implementation of these formulas and some numerical experiments are given in [9]. As an example Figs. 4 and 5 show the different profiles obtained by a gradient-type algorithm applied to the functional (22). In this example, the initial geometry is a semicircle that is transformed into a parabolic profile after the optimization procedure.

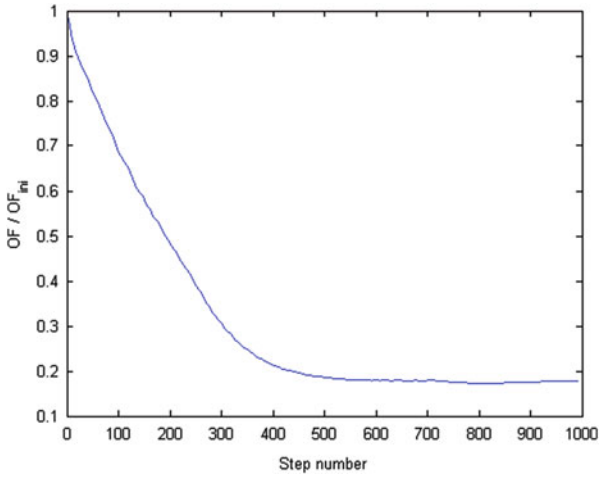


Fig. 5 Evolution of the cost functional

Appendix

In this section we show how to compute the adjoint operators $\mathscr{A}$ and $\mathscr{B}$ in (8)–(9) for one of the examples above. As it has been said, this computation strongly depends on the specific problem, but nevertheless the methodology is straightforward as it will be shown here.

We focus on the 2D elasticity problem described in Sect. 6. The objective function is given by (23) so that, according to (8) we look for $\mathscr{B}$, $\mathscr{A}$ and an adjoint state Ψ such that

$$\int_S \delta(j(\mathbf{t}\cdot\sigma\cdot\mathbf{t}))\,ds = \int_S \mathscr{B}\Psi\delta\alpha\,ds, \tag{31}$$

with Ψ satisfying $\mathscr{A}\Psi = 0$.

First of all, observe that, as $\delta\mathbf{t} = \delta\alpha'\mathbf{n}$, we have

$$\begin{aligned}\delta(j(\mathbf{t}\cdot\sigma\cdot\mathbf{t})) = {} & j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})(\mathbf{n}\cdot\sigma\cdot\mathbf{t}+\mathbf{t}\cdot\sigma\cdot\mathbf{n})\delta\alpha' \\ & + j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})\mathbf{t}\cdot\delta\sigma\cdot\mathbf{t},\end{aligned} \tag{32}$$

where the tensor $\delta\sigma$ is the solution of the linearized elasticity system

$$\delta\sigma_{\alpha\beta,\beta} = 0, \quad \mathbf{x}\in\Omega, \quad \alpha,\beta = 1,2, \tag{33}$$

$$\delta\sigma_{33} = \nu(\delta\sigma_{11}+\delta\sigma_{22}), \quad \mathbf{x}\in\Omega, \tag{34}$$

$$\delta\varepsilon_{\alpha\beta} = \frac{1+\nu}{E}\delta\sigma_{\alpha\beta} - \frac{\nu}{E}\delta\sigma_{kk}\delta_{\alpha\beta}, \quad \mathbf{x}\in\Omega, \tag{35}$$

$$\delta\varepsilon_{\alpha\beta} = \frac{1}{2}\left(\delta u_{\alpha,\beta} + \delta u_{\beta,\alpha}\right), \quad \mathbf{x} \in \Omega, \tag{36}$$

$$\delta\varepsilon_{13} = \delta\varepsilon_{23} = \delta\varepsilon_{33} = 0, \quad \mathbf{x} \in \Omega, \tag{37}$$

and the linearized boundary conditions

$$\delta\mathbf{u} = 0, \quad \mathbf{x} \in \Gamma, \tag{38}$$

$$\delta\sigma \cdot \mathbf{n} + \partial_n(\sigma) \cdot \mathbf{n}\delta\alpha - \sigma \cdot \mathbf{t}\alpha'(s) = 0, \quad \mathbf{x} \in S. \tag{39}$$

The only term that requires further analysis in (32) is the last one, i.e.

$$\int_S j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\mathbf{t} \cdot \delta\sigma \cdot \mathbf{t}\, ds. \tag{40}$$

To simplify it we write the linearized stress-strain tensor given in (34)–(35) with respect to the local system of coordinates associated to S, $\{\mathbf{t}, \mathbf{n}\}$. The following expression for $\delta\sigma_{\alpha\beta}$ is obtained:

$$\begin{aligned}\mathbf{t} \cdot \delta\sigma \cdot \mathbf{t} &= \frac{E}{1-\nu^2}\mathbf{t} \cdot \delta\varepsilon \cdot \mathbf{t} + \frac{\nu}{1-\nu}\mathbf{n} \cdot \delta\sigma \cdot \mathbf{n} \\ &= \frac{E}{1-\nu^2}\partial_{tg}(\delta\mathbf{u} \cdot \mathbf{t}) - \frac{\nu}{1-\nu}\mathbf{n} \cdot \partial_n(\sigma) \cdot \mathbf{n}\delta\alpha. \end{aligned} \tag{41}$$

In the last equality, we have used

$$\mathbf{t} \cdot \delta\varepsilon \cdot \mathbf{t} = \partial_{tg}(\delta\mathbf{u} \cdot \mathbf{t}),$$

and the boundary conditions to be satisfied for $\delta\mathbf{u}$ and $\mathbf{u}$ on S.

Therefore (40) can be written as

$$\begin{aligned}\int_S j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\mathbf{t} \cdot \delta\sigma \cdot \mathbf{t}\, ds &= \frac{E}{1-\nu^2}\int_S j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\partial_{tg}(\delta\mathbf{u} \cdot \mathbf{t})\, ds \\ &\quad - \frac{\nu}{1-\nu}\int_S j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\mathbf{n} \cdot \partial_n(\sigma_{\alpha\beta}) \cdot \mathbf{n}\delta\alpha\, ds \\ &= -\frac{E}{1-\nu^2}\int_S \partial_{tg} j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})(\delta\mathbf{u} \cdot t)\, ds \\ &\quad - \frac{\nu}{1-\nu}\int_S j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\mathbf{n} \cdot \partial_n(\sigma_{\alpha\beta}) \cdot \mathbf{n}\delta\alpha\, ds. \end{aligned} \tag{42}$$

In order to eliminate the term $\delta\mathbf{u}$ the adjoint problem to the linearized system is introduced

$$\frac{\partial \sigma^*_{\alpha\beta}}{\partial \beta} = 0, \quad x \in \Omega, \quad \alpha, \beta = 1, 2, \tag{43}$$

$$\sigma^*_{33} = \nu(\sigma^*_{11} + \sigma^*_{22}), \quad x \in \Omega, \tag{44}$$

$$\epsilon^*_{\alpha\beta} = \frac{1+\nu}{E}\sigma^*_{\alpha\beta} - \frac{\nu}{E}\sigma^*_{kk}\delta_{\alpha\beta}, \tag{45}$$

$$\epsilon^*_{\alpha\beta} = \frac{1}{2}\left(\frac{\partial \psi_\beta}{\partial \alpha} + \frac{\partial \psi_\alpha}{\partial \beta}\right) \tag{46}$$

$$\delta\epsilon^*_{13} = \delta\epsilon^*_{23} = \delta\epsilon^*_{33} = 0, \tag{47}$$

with the following boundary conditions

$$\psi_\alpha = 0, \quad x \in \Gamma, \quad \alpha = 1, 2, \tag{48}$$

$$\sigma^* \cdot \mathbf{n} = \frac{-E}{1-\nu^2}\partial_{tg} j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\mathbf{t}, \quad x \in S. \tag{49}$$

Multiplying the equations of the linearized system by $\Psi = (\psi_1, \psi_2)$ and integrating by parts it is easily obtained

$$0 = -\int_\Omega \delta\sigma : \epsilon^* \, dx - \int_S \Psi \cdot \left(\partial_n \sigma \cdot \mathbf{n}\delta\alpha - \sigma \cdot \mathbf{t}\delta\alpha'\right) \, ds, \tag{50}$$

where : represents the double dot product of second order tensors.

A straightforward computation allows us to write the first term in this formula as follows,

$$\int_\Omega \delta\sigma : \epsilon^* \, dx = \int_\Omega \sigma^* : \delta\epsilon \, dx. \tag{51}$$

Now we integrate by parts in the right hand side of (51), taking into account the boundary conditions for $\mathbf{u}$ and Ψ,

$$\begin{aligned} \int_\Omega \sigma^* : \delta\epsilon \, dx &= -\int_\Omega \delta u_\alpha \frac{\partial \sigma^*_{\alpha\beta}}{\partial \beta} dx + \int_S \delta\mathbf{u} \cdot \sigma^* \cdot \mathbf{n} \, ds \\ &= -\frac{E}{1-\nu^2}\int_S \partial_{tg} j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\boldsymbol{\delta}\mathbf{u} \cdot \mathbf{t} \, ds. \end{aligned} \tag{52}$$

Therefore, combining (50)–(52) the following equation is obtained

$$\frac{E}{1-\nu^2}\int_S \partial_{tg} j'(\mathbf{t} \cdot \sigma \cdot \mathbf{t})\boldsymbol{\delta}\mathbf{u} \cdot \mathbf{t} \, ds = \int_S \Psi \cdot \left(\partial_n \sigma \cdot \mathbf{n}\delta\alpha - \sigma \cdot \mathbf{t}\delta\alpha'\right) \, ds. \tag{53}$$

Substituting (53) into (42) we obtain the final expression for (40),

$$\int_S j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})\mathbf{t}\cdot\delta\sigma\cdot\mathbf{t}\,ds = -\int_S \Psi\cdot\left(\partial_n\sigma\cdot\mathbf{n}\delta\alpha - \sigma\cdot\mathbf{t}\delta\alpha'\right)\,ds \\ -\frac{\nu}{1-\nu}\int_S j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})\mathbf{n}\cdot\partial_n(\sigma_{\alpha\beta})\cdot\mathbf{n}\delta\alpha\,ds. \tag{54}$$

From this formula together with (32) we obtain in the left hand side of (31) an expression where all the terms contain a factor with $\delta\alpha$ or its derivative. Integrating by parts on S and assuming that either S has no boundary or $\delta\alpha = 0$ at the boundary of S we easily obtain the expression for $\mathscr{B}$ in (31),

$$\mathscr{B} = -\partial_{tg}(j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})(\mathbf{n}\cdot\sigma\cdot\mathbf{t}+\mathbf{t}\cdot\sigma\cdot\mathbf{n})) \\ -\Psi\cdot\partial_n\sigma\cdot\mathbf{n} + \partial_{tg}(\Psi\cdot\sigma\cdot\mathbf{t}) - \frac{\nu}{1-\nu}j'(\mathbf{t}\cdot\sigma\cdot\mathbf{t})\mathbf{n}\cdot\partial_n(\sigma)\cdot\mathbf{n}.$$

The operator $\mathscr{A}\Psi = 0$ contain all the adjoint equations and boundary conditions (43)–(49).

Acknowledgements Supported by project MTM2011-29306-C02-02 from MICINN (Spain). The author is grateful to J. García-Palacios and A. Samartín for their advice and help on the elasticity application and the numerical experiment presented in the paper.

References

1. Anderson, W., Venkatakrishnan, V.: Aerodynamic Design Optimization on Unstructured Grids with a Continuous Adjoint Formulation. AIAA Paper, **97-0643** (1997)
2. Baeza, A., Castro, C., Palacios, F., Zuazua, E.: 2-D Euler shape design on nonregular flows using adjoint Rankine-Hugoniot relations. AIAA J. **47**(3), 552–562 (2009)
3. Bardos, C., Pironneau, O.: A formalism for the differentiation of conservation laws. C. R. Acad. Sci. Paris Ser. I **335**, 839–845 (2002)
4. Bardos, C., Pironneau, O.: Derivatives and control in presence of shocks. Comput. Fluid Dyn. J. **11**(4), 383–392 (2003)
5. Bouchut, F., James, F.: One-dimensional transport equations with discontinuous coefficients. Nonlinear Anal. Theory Appl. **32**(7), 891–933 (1998)
6. Bueno, A., Castro, C., Palacios, F., Zuazua, E.: Continuous adjoint approach for the Spalart-Allmaras model in aerodynamic optimization. AIAA J. **50**(3), 631–646 (2012)
7. Castro, C., Lozano, C., Palacios, F., Zuazua, E.: A systematic continuous adjoint approach to viscous aerodynamic design on unstructured grids. AIAA J. **45**(9), 2125–2139 (2007)
8. Castro, C., Palacios, F., Zuazua, E.: An alternating descent method for the optimal control of the inviscid Burgers equation in the presence of shocks. Math. Models Methods Appl. Sci. **18**(3), 369–416 (2008)
9. Garcia-Palacios, J., Castro, C., Samartin, A.: Optimal design in elasticity: A systematic adjoint approach to boundary cost functionals. Preprint (2013)

10. Giles, M.B., Pierce, N.A.: Analytic adjoint solutions for the quasi one-dimensional Euler equations. J. Fluid Mech. **426**, 327–345 (2001)
11. Godlewski, E., Raviart, P.A.: The linearized stability of solutions of nonlinear hyperbolic systems of conservation laws. A general numerical approach. Math. Comput. Simul. **50**, 77–95 (1999)
12. Hadamard, J.: Leçons sur le Calcul des Variations. Gauthier-Villars, Paris (1910)
13. Hirsch, C.: Numerical Computation of Internal and External Flows, vol. 2. Wiley, Chichester (1990)
14. Jameson, A.: Aerodynamic design via control theory. J. Sci. Comput. **3**, 233–260 (1988)
15. Jameson, A.: Optimum Aerodynamic Design Using CFD and Control Theory. AIAA Paper, **95** (1995)
16. Jameson, A., Schmidt, W., Turkel, E.: Numerical Solution of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes. AIAA Paper **81-1259** (1981)
17. Lecaros, R., Zuazua, E.: Tracking control of the 1D scalar conservation laws in the presence of shocks. Preprint (2013)
18. Lions, J.-L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, New York (1971)
19. Mavriplis, D., Discrete adjoint-based approach for optimization problems on three-dimensional unstructured meshes. AIAA J. **45**(4), 740–750 (2007)
20. Mohammadi, B., Pironneau, O.: Shape optimization in fluid mechanics. Annu. Rev. Fluids Mech. **36**, 255–279 (2004)
21. Nadarajah, S., Jameson, A.: A Comparison of the Continuous and Discrete Adjoint Approach to Automatic Aerodynamic Optimization. AIAA Paper 2000-0667. 38th Aerospace Sciences Meeting and Exhibit, Reno, NV (January 2000)
22. Pironneau, O.: On optimum design in fluid mechanics. J. Fluid Mech. **64**, 97–110 (1974)
23. Simon, J.: Differenciation with respect to the domain in boundary value problems. Numer. Funct. Optim. **2**(7–8), 649–687 (1980)
24. García-Palacios, J., Castro, C., Samartín, A.: Optimal boundary geometry in an elasticity problem: a systematic adjoint approach. In: Domingo A., Lázaro C. (eds.) Proc. of the Int. Assoc. for Shell and Spatial Struct. (IASS), pp. 509–524. Symposium 2009, Valencia (2009)

Medical Image Processing: Mathematical Modelling and Numerical Resolution

Emanuele Schiavi, Juan Francisco Garamendi, and Adrián Martín

Abstract Medical image processing is an interdisciplinary research field attracting expertise from applied mathematics, computer sciences, engineering, statistics, physics, biology and medicine. In this context we shall present an introduction to basic techniques and concepts as well as more advanced methods to promote interests for further study and research in the field.

1 Introduction

Medical image processing is a growing field in medicine and mathematics which aims to improve the diagnostic power of some acquisition data modalities such as MRI, fMRI, PET, MEG, CT, etc. This leads to improved treatment control and therapies. In this work we shall consider some digital image processing related mathematical problems such as filtering, denoising and segmentation of digital images with a particular view to medical image processing and restoration. Our research is based at Fundación CIEN-Fundación Reina Sofía, Madrid, Spain

`http://www.fundacionreinasofia.es/ES/Paginas/home.aspx`

E. Schiavi (✉)
Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, Madrid, Spain
e-mail: emanuele.schiavi@urjc.es

J.F. Garamendi
Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain

A. Martín
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_6,

where a interdisciplinary group of scientists coming from different areas and institutions is working on biomarkers for neurological diseases such as Alzheimer and Parkinson.

These lecture notes cover some basic aspects of the mathematical modelling but they also aim to introduce the reader to the most recent techniques and numerical algorithms. A very straightforward and applied introduction to the field can be found in [11] where basic, routinary algorithms are implemented. A more advanced introduction to the theoretical material we shall consider in this work is described in the book by Chan and Shen [5] where the mathematical foundations of modern image processing and low-level computer vision are presented, bridging contemporary mathematics with state-of-the-art methodologies in modern image processing. An interesting medical images processing overview can be found in http://www.math.wisc.edu/~angenent/preprints/medicalBAMS.pdf and a more general, geometric approach to PDE image processing is in the book by Osher [15].

2 Digital Image Processing

Digital image processing is a recent and challenging branch of applied mathematics which developes models and numerical algorithms for Filtering, Denoising, Deblurring, Edge-enhancing, Segmentation, Registration, Tracking, Impainting, Smoothing, Compression, Features Extraction and Pattern Recognition. The great improvement in computational power as well as the design of specific patient tailored acquisition modalities which took place in the last decade have motivated the implementation of advanced mathematical theories to the pre-processing analysis and the statistical post-processing interpretation of huge amounts of possibly multimodal patient data. In short, fast and accurate mathematical analysis are both possible and necessary paradigms, contrary to the past view where fast but very approximated results where looked for. In this section we shall briefly introduce the reader to the key steps of the mathematical analysis focusing on the models and results which made possible the evolution and implementation of numerical algorithms for the resolution of the PDE that appear in image processing and enhancement. We shall consider two basic approaches. In the first one we shall see how it is possible to filter an image using directly an evolution diffusion equation. Then we shall move to the variational approach in order to solve the energy minimization problems which arise when we try to solve the associated inverse problems and their Tikhonov regularization. This introduces the need for nonlinear operators which include the very famous Total Variation Model by Rudin, Osher and Fatemi (1987), see [17].

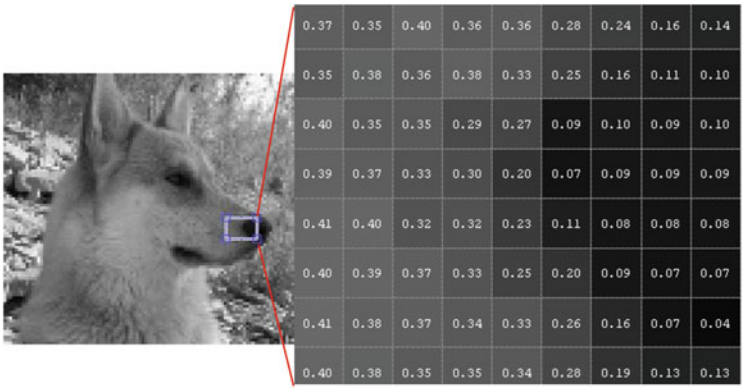


Fig. 1 A detail of the image showing the matricial coding

2.1 Linear Filtering and Convolution

The basic material of this section covers linear diffusion filtering and its relations with gaussian smoothing. This introduces the use of partial differential equations into image processing. More advanced properties of this linear approach like scale-space properties and its applications, generalizations and limitations can be found in http://www.lpi.tel.uva.es/muitic/pim/docus/anisotropic_diffusion.pdf. Here we briefly introduce some concepts. Digital images are commonly defined as matrices of scalars for grayscale images or as vectors for multimodality and/or multichannel images as well as simple multichannel colour RGB images. In a discrete setting images are then $u = (u_{i,j}), 1 \leq i, j \leq N, u_{i,j} \in [0, 1]$ or $u_{i,j} \in [0, 255]$ 2D discrete, bounded signals (Fig. 1). In the variational framework we shall adopt a continuous world view so that a grayscale image is a real valued function $u : \Omega \to \mathbb{R}$ on an open set $\Omega \subset \mathbb{R}^2$. A color image is a vector-valued function $u : \Omega \to \mathbb{R}^3$ on an open set $\Omega \subset \mathbb{R}^2$ which maps into RGB color space.
In fact the digital images can also be organized into functional and algebraic structures such as multichannel images where different data acquisition modalities can be grouped to form a unique vectorial description of the image.

These matrices can be seen as the values of a distribution (generalized function) $u_0(x)$ defined on an open and bounded 2D or 3D domain Ω x being a pixel (2D) or a voxel (3D). This allows a functional analytic setting for image-processing problems and in particular, for the design of digital processing algorithms through partial differential equations (PDEs) models. More recent and advanced acquisition techniques in Magnetic Resonance, such as scalar Diffusion Weighted Images (DWI) or Diffusion Tensor Images (DTI) provide 3D volumes of tensorial data, a sort of matrices of matrices which inform about the (anisotropic) movement of water molecules through the fibers of the white matter of the brain (Figs. 2 and 3).

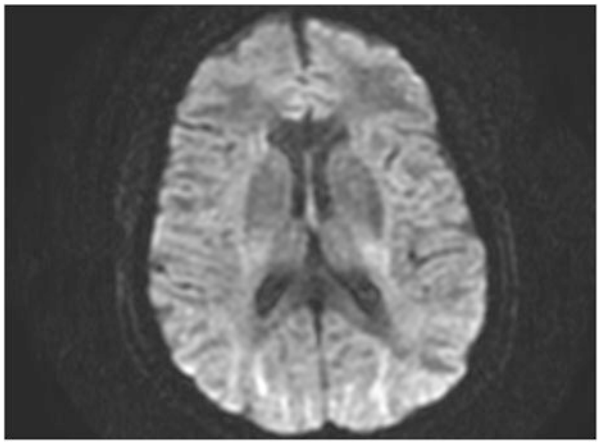


Fig. 2 A DW-MR image courtesy of Fundación Reina Sofia, Centro de Alzheimer, Madrid

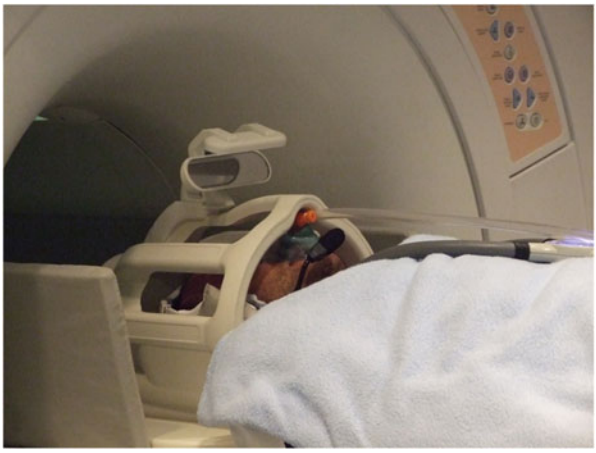


Fig. 3 The MR scanner of General Electric 3 T Signa at Fundación Reina Sofia, Centro de Alzheimer, Madrid (Research agreement with General Electric). Image courtesy of CIEN Foundation

Filtering is a technique for modifying or enhancing an image. For example, you can filter an image to emphasize certain features or remove other features. Image processing operations implemented with filtering include smoothing, sharpening, and edge enhancement. In the continuous case we can understand the analogy between filtering and convolution by means of the heat equation which is a linear diffusion equation (Figs. 4 and 5).

Let $\mathbf{J}$ be a flux of any scalar magnitude such as intensity of the signal, temperature or concentration of a chemical substance. The flux is generated by local differences in the intensity and we have $\mathbf{J} = -D\nabla u$ where D is a tensor characterizing the

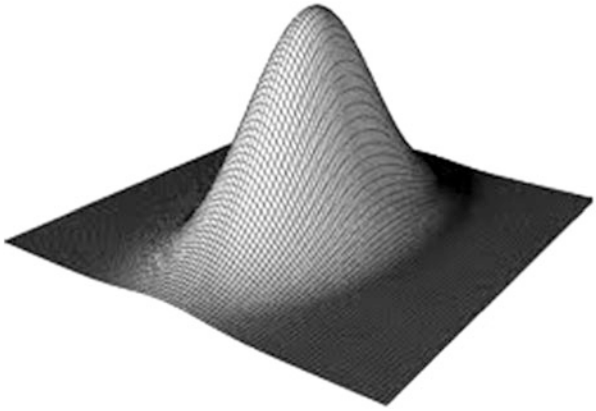


Fig. 4 A plot of a 2D gaussian function

possible anisotropy of the diffusion. In the isotropic case $D = I_d$, the identity matrix. The mass conservation equation states that, without sources or sinks, the local variation of the magnitude of u is caused by the divergence of the flux, $\partial_t u = -\text{div}\mathbf{J}$ which is

$$\partial_t u = \text{div}(\nabla u) = \Delta u.$$

Let n denote the spatial dimension and consider the Cauchy problem

$$\begin{cases} \partial_t u = \Delta u, & \mathbb{R}^n \times (0, +\infty) > 0, \\ u(x, 0) = u_0(x), \mathbb{R}^n \end{cases}$$

associated to the initial data $u_0(x)$.

If we assume that $u_0(x) = \delta_0$, the Delta function located at $x = 0$ the explicit solution (or Gauss kernel) of the Cauchy problem is:

$$G(x, t) = \frac{e^{-|x|^2/t}}{(4\pi t)^{n/2}}$$

where the gaussian is represented in Fig. 4.

The solution of the original problem can be expressed in terms of the convolution:

$$u = G * u_0 = \int_{\mathbb{R}^2} G(x - y)u_0(y)dy.$$

Defining $\sigma = \sqrt{2t}$ we see that the solution of our problem is given by the convolution of the initial data with a gaussian function with *standard deviation* σ

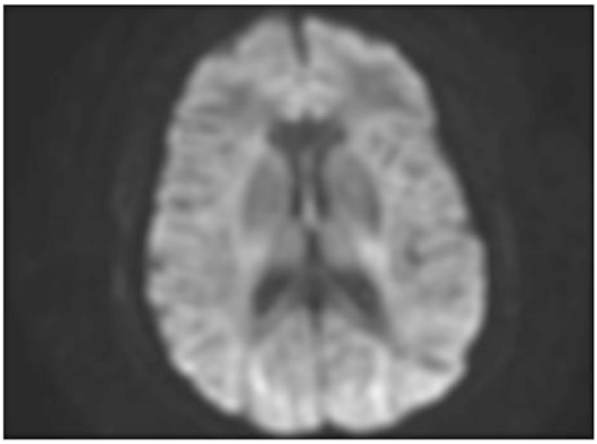


Fig. 5 The oversmoothed, blurred image obtained by convolution

(the width of the gaussian kernel) which corresponds to a linear diffusion process during exactly $T = \sigma^2/2$ where σ^2 is the estimated *variance* of the noise affecting the data. In the discrete case filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. Linear filtering of an image is accomplished through an operation called convolution. Convolution is a neighborhood operation in which each output pixel is the weighted sum of neighboring input pixels (Fig. 5).

A fundamental property of the convolution operation is that it regularizes the data and, even with $u_0 \in L^1(\mathbb{R})$ we have $G * u_0 \in C^\infty(\mathbb{R})$ for any $t > 0$. This clearly is a poor result in image processing because this low pass filter smooths out all the high frequencies of the image, where noise and details are involved. The need for nonlinear filtering became readily evident.

2.2 Nonlinear Filtering

It has been introduced into the digital imaging community through the *intriguing* model proposed by Perona and Malik, in [16]. Details about the theoretical difficulties associated to this *forward-backward* nonlinear diffusion model can be found in http://www.lpi.tel.uva.es/muitic/pim/docus/anisotropic_diffusion.pdf. The associated PDE is

$$\partial_t u = \operatorname{div}(g(\nabla u)\nabla u)$$

with $u(x,0) = u_0(x)$ and

$$g(s^2) = \frac{1}{1 + s^2/\lambda^2}, \qquad \lambda > 0.$$

The consideration of the 1D case

$$\partial_t u = \partial_x (g(u_x) u_x)$$

with flux function

$$\Phi(s) = s g(s^2) = \frac{s}{1 + s^2/\lambda^2}$$

reveals that

$$\Phi'(s) \geq 0 \quad |s| \leq \lambda, \qquad \Phi'(s) < 0 \quad |s| > \lambda$$

and the equation

$$\partial_t u = \partial_x (\Phi(u_x)) = \Phi'(u_x) u_{xx}$$

has *negative* diffusion when the gradient is big, e.g. near the edges of the image. Despite of this, the numerical resolution of this equation introduces *numerical* diffusion which stabilizes the solution and the model provide quite good results. A simple and straightforward introduction to nonlinear diffusion and related algorithms in MATLAB can be found in http://staff.science.uva.nl/~rein/nldiffusionweb/nldiffusioncode.pdf. The original and detailed analysis of nonlinear diffusion and anisotropy is in the excellent book by Weickert http://www.lpi.tel.uva.es/muitic/pim/docus/anisotropic_diffusion.pdf.

2.3 *Modelling Medical Images Processing and Restoration*

Digital image denoising and segmentation are basic problems in image processing and computer vision which can be dealt with in the variational framework. Roughly speaking this amounts to the minimization of an energy functional defined in a suitable functional space. The minima of the functional can be characterized as the weak solutions of the associated Euler–Lagrange equations which are, typically, nonlinear second order elliptic partial differential equations. These nonlinearities are necessary in order to avoid oversmoothing as predicted by the general linear elliptic regularity theory. This introduces both, mathematical and numerical difficulties in the analysis of such models and makes the implementation of efficient numerical methods challenging.

We shall review some aspects of what is called the Tikhonov Regularization for ill-posed inverse problems. This introduces a General Regularization Model which can be justified by means of a Bayesian formulation. In fact many of the tasks encountered in image processing can be considered as problems in statistical inference. In particular, they fit naturally into a Bayesian framework:

$$\log p(u|f) \propto \log p(f|u) + \log p(u)$$

and a MAP (Maximum A Posteriori) estimation of u is:

$$\max_u \{\log p(f|u) + \log p(u)\}$$

where $p(f|u) = \exp(-H(u, f))$ is the likelihood term and $p(u) = (1/\lambda)\exp(-J(u))$ is the *prior*. Following this Bayesian modelling approach we consider the minimization problem

$$\min_{u \in BV(\Omega)} J(u) + \lambda H(u, f) \tag{1}$$

where $J(u)$ is the convex nonnegative Total Variation regularization functional

$$J(u) = |u|_{BV} = \int_\Omega |Du| \tag{2}$$

and the data fidelity term (modelling gaussian noise) is

$$H(u, f) = \int_\Omega |f - u|^2 dx.$$

The term $\int_\Omega |Du|$ denotes the Total Variation of u with Du being its generalized gradient (a vector bounded Radon measure). When $u \in W^{1,1}(\Omega)$ we have $\int_\Omega |Du| = \int_\Omega |\nabla u| dx$. The λ parameter in (1) is a scale parameter tuning the model. In this (weak) setting it is a very common and useful approach to describe images as distributions.

One popular model for image denoising is the Rudin, Osher and Fatemi's (ROF) model, where we seek for a distribution u in the space of the Bounded Variation ($BV(\Omega)$) distributions, which is the solution to the following nonlinear minimization problem.

Given $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ which represents the data, minimize the (strictly convex) energy

$$E(u) = \int_\Omega |Du| + \frac{1}{2\lambda} \int_\Omega |f - u|^2 dx \tag{3}$$

where Ω is a Lipschitz domain (the unit square or a cube for the sake of simplicity) and $f \in L^\infty(\Omega)$ is the image affected by Gaussian white noise. Due to the fact that

the functional in (2) is not differentiable at the origin we introduce the notion of the subdifferential of $J(u)$ at a point u by

$$\partial J(u) = \{p \in BV(\Omega)^* | \, J(v) \geq J(u) + < p, v - u >\}$$

for all $v \in BV(\Omega)$, to give a (weak and multivalued) meaning to the Euler–Lagrange equation associated to the minimization problem. Using variational calculus and convex analysis the associated Euler–Lagrange Equation is then

$$\lambda \partial J(u) + (u - f) \ni 0$$

which is a multivalued equation which reflects the non differentiability of the TV operator. The proper setting for such multivalued equations is in terms of variational inequalities which can be deduced from the so called *Complementary Formulation*. Typically this difficulty is avoided using the approximating minimization problems

$$J(u_\epsilon) = \int_\Omega \sqrt{|\nabla u_\epsilon|^2 + \epsilon} dx + \frac{1}{2\lambda} \int_\Omega |f - u_\epsilon|^2 dx \tag{4}$$

with Euler–Lagrange Equation

$$-\lambda \text{div} \left(\frac{\nabla u_\epsilon}{\sqrt{|\nabla u_\epsilon|^2 + \epsilon}} \right) + (u_\epsilon - f) = 0.$$

It is standard to look for a solution to (3) [and (4)] by solving a related nonlinear parabolic equation using a pseudo-time-stepping algorithm in order to approximate the steady-state configuration $u(x)$. This approach, known as (primal) gradient descent, has two serious drawbacks: the approximating problems have *continuous* solutions u_ϵ which are unfeasible in medical imaging because different organs and subcortical structures are characterized by discontinuities; moreover, the numerical method is slowly convergent. An elegant and brilliant solution to these problems can be found in Chambolle [3]. A deep theoretical study of this kind of *linear* energy functionals is considered in [1].

In what follows we shall describe some advanced models that our group has proposed and applied in the last years.

3 Advanced Models

In this section we shall present some advanced models for image segmentation and denoising. Notice that image denoising can be considered as a pre-processing step previous to the segmentation task. A PDE approach to image segmentation is based on the celebrated Mumford and Shah model, [14]. When piecewise constant

solutions of the Mumford-Shah model are considered we have a *minimal partition problem* and a huge literature is concerned with the analysis of such a model [6]. Here we shall consider an anisotropic version of the Mumford and Shah functional which has been proposed in [8,9] for multichannel and multiphase image segmentation.

Let $\bar{f}$ be a vector valued function such as $\bar{f} \in L^\infty(\Omega;\mathbb{R}^M)$ defined on a bounded open domain $\Omega \subset \mathbb{R}^D$, where each scalar component $f_i(x) : \Omega \to \mathbb{R}$ is a channel. Let $\bar{u}$ be a vector valued piecewise constant function such as $\bar{u} = \sum_1^N \bar{c}_i \chi_i$ with $\bar{c}_i \in \mathbb{R}^M$ and χ_i the characteristic functions of the domain partition. Then we can perform multiclass (N classes) and multichannel (M channels) image segmentation minimizing

$$J(\mathbf{C},\Gamma) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left|\left|\bar{c}_j - \bar{c}_i\right|\right|_{L^p(\Omega;\mathbb{R}^M)} \int_{\Gamma_{ij}} d\mathcal{H}^{D-1} + \frac{1}{2\lambda} \sum_{i=1}^{N} \int_{\Omega_i} |\bar{c}_i - \bar{f}|^2 dx \tag{5}$$

where

$$\left|\left|\bar{c}_j - \bar{c}_i\right|\right|_{L^p(\Omega;\mathbb{R}^M)} = \left[\sum_{m=1}^{M} \left|c_{j,m} - c_{i,m}\right|^p\right]^{1/p}.$$

In our current numerical implementation we choose $p = 2$. Notice that the functional is expressed in terms of a matrix $\mathbf{C}$ with components $c_{i,j}$ which reflect the different values of the piecewise solution as well as in terms of a curve Γ along which the solution is discontinuous. The key idea of our method relies on the strong analogy between this anisotropic Mumford and Shah functional (AMS) and the ROF model we introduced before. In fact, in the class of piecewise constant functions both energies coincides. This suggests that the minima of the AMS model can be obtained thresholding the ROF minima. To show an application of these ideas we consider a four classes segmentation problem, as in MRI brain segmentation where white and gray matter, together with liquid and background are the relevant classes. Let $u_{rof} \in BV(\Omega) \cap [0,1]$ be the minimum of the ROF functional (3). If we threshold this solution by mean of a vector $\bar{t} \in \mathbb{R}^3$ we generate a piecewise constant approximation of u_{rof} for every $\bar{t}$ in form $\bar{c}(\bar{t}) \cdot \bar{\chi}(\bar{t})$. The problem is then to minimize the Anisotropic Mumford Shah energy (5) finding the best threshold $\bar{t} \in \mathbb{R}^3$ for the solution of the ROF model (3). This can be accomplished by using a genetic algorithm (notice that the problem is not convex) where the search is restricted to simple functions $u_{\bar{t}}(x) \in SBV(\Omega)$ taking, for a.e. $x \in \Omega$, the (possibly re-ordered) values $\bar{c} = (c_1, c_2, c_3, c_4)$ as defined by formula (6) which we shall deduce below. Let $\bar{\chi} = \{\chi_i\}_{i=1}^4$ be a given partition. Then

$$\frac{\partial J}{\partial c_j} = \left(\sum_{i=1, i\neq j}^{4} \frac{(j-i)}{|j-i|} |\Gamma_{i,j}| \right) - \frac{1}{\lambda} \int_{\Omega_j} f\, dx + \frac{|\Omega_j|}{\lambda} c_j, \qquad j = 1, \ldots, 4$$

and the functional $J(\bar{c}, \bar{\chi})$ is optimized by the choice

$$c_j = \bar{f}^j + \frac{\lambda}{|\Omega_j|} \left(\sum_{i=1, i\neq j}^{4} \frac{(i-j)}{|i-j|} |\Gamma_{ij}| \right), \qquad j = 1, \ldots, 4 \tag{6}$$

where $\bar{f}^j$ are the local averaged data values as predicted by the partition:

$$\bar{f}^j = \int_{\Omega_j} f\, dx / |\Omega_j|, \quad j = 1, \ldots, 4$$

Moreover we have:

$$\sum_{j=1}^{4} c_j |\Omega_j| = \sum_{j=1}^{4} \int_{\Omega_j} f\, dx + \lambda \sum_{j=1}^{4} \left(\sum_{i=1, i\neq j}^{4} \frac{(i-j)}{|i-j|} |\Gamma_{ij} \right) = \int_{\Omega} f\, dx. \tag{7}$$

This implies that, if we calculate u as the (unique) minimum of $J(u)$ in (3) and we threshold u by means of a threshold vector $\bar{t} = (t_1, t_2, t_3) \in \mathbb{R}^3$, then we generate a partition $\bar{\chi} = (\chi_1, \chi_2, \chi_3, \chi_4)$ (defining $\Omega_i = \{x \in \Omega \,/\, t_{i-1} \leq u(x) < t_i\}$) and, using formula (6) for the best constants, an optimal representation of u for the given partition in form $u = \bar{c} \cdot \bar{\chi}$. Notice that a relabeling is performed to ensure the ordering of the optimal constants once the threshold $\bar{t}$ is applied. More details in this procedure can be found in [8].

The numerics are performed using the dual formulation of the problem. This provides a convenient framework to solve the multiphase systems associated with the minimization of the AMS functional. More advanced staggered schemes are proposed in [10]. Segmentation results with different values of the λ parameter are presented below. Figure 6 shows the segmentation of a brain phantom slice with three levels of added noise with different values of the parameter.

Finally, we segmented real MRI images acquired at Fundación Reina Sofía in Madrid. Figure 7 shows the result of the automatic segmentation with two different values of the parameter. Both results are visually correct, while the λ parameter allows to obtain segmentations at different scales of detail.

More sophisticated results can be obtained when segmenting FA color code DT-MR images as we show in figures below. A brief introduction to this kind of scalar MR images which are obtained from tensorial data is presented in the next section (Figs. 8 and 9).

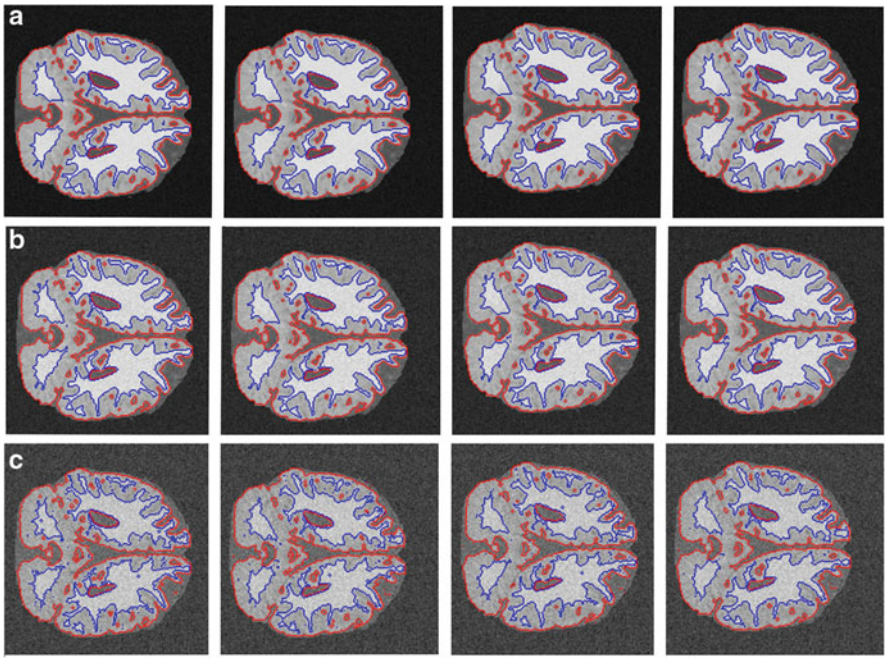


Fig. 6 Segmentation of the same slice of a phantom with different noise levels and different values of λ. *From left to right*, results with λ values 0.08, 0.09, 0.1 and 0.11. (**a**) Phantom with 5 % noise; (**b**) phantom with 10 % noise; (**c**) phantom with 20 % noise

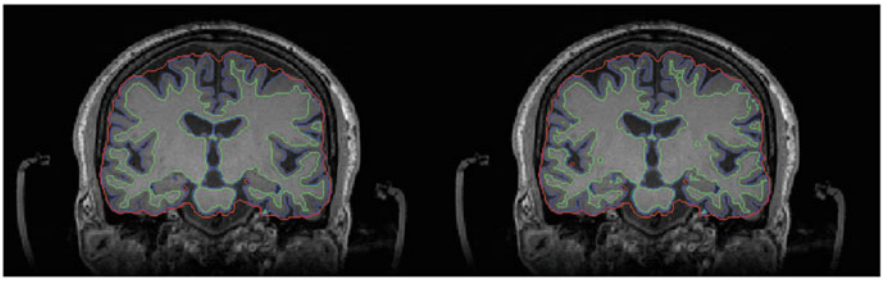


Fig. 7 Segmentation of real MRI brain data with two vales of the λ parameter. *Left*: $\lambda = 0.12$; *Right*: $\lambda = 0.08$

3.1 MRI Denoising

We now step forward in the modelling exercise. In fact, accurate MRI noise modelling is a fundamental issue in medical image processing which leads naturally to the assumption that MR magnitude images are corrupted by Rician noise which is a signal dependent noise. Indeed this noise is originated in the computation of the

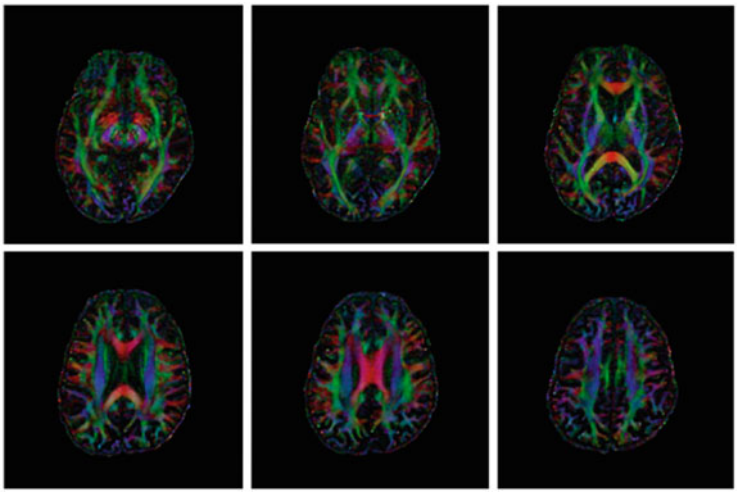


Fig. 8 A color code Fractional Anisotropy (FA) image which is obtained computing the eigenvalues of the Diffusion Tensor Image (DTI) reconstructed from the Diffusion Weighted Images (DWI) acquired at the Hospital Reina Sofía

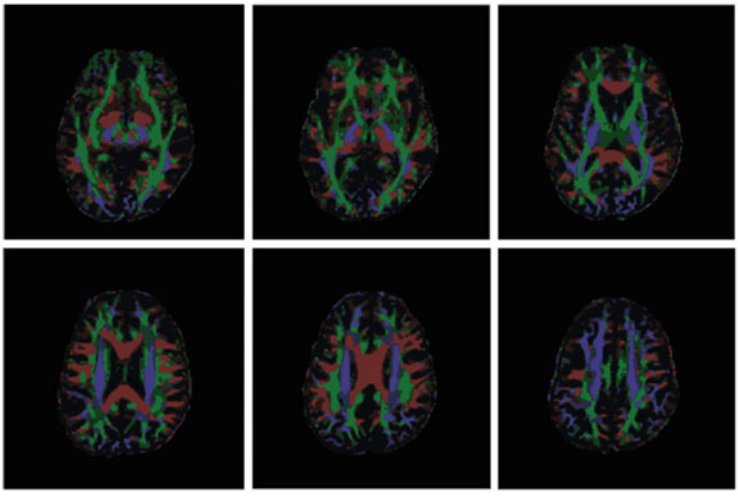


Fig. 9 The obtained segmentation. Notice that we segment the directions along which the fibers propagate in the brain

magnitude image from the real and imaginary images, that are obtained from the inverse Fourier Transform applied to the original raw data. This process involves a non-linear operation which maps the original Gaussian distribution of the noise to a Rician distribution. Nevertheless it is usually argued that this bias does not affect seriously the processing and subsequent analysis of MR images so that a

(identically distributed and signal independent) Gaussian noise is modelled. This assumptions fails when low signal-to-noise ratio are considered. With this purpose we consider, in a variational framework, a denoising model for MR Rician noise contaminated images proposed in [12] which combines the Total Variation semi-norm with a Rician data fitting term.

The data term $H(u, f)$ is a fitting functional which is nonnegative with respect to u for fixed f. To model Rician noise $H(u, f)$ has been deduced previously in [2] in the context of diffusion tensor MR images. The Rician likelihood term is of the form:

$$H(u, f) = \frac{1}{2\sigma^2} \int_\Omega u^2 dx - \int_\Omega \log I_0 \left(\frac{uf}{\sigma^2} \right) dx \tag{8}$$

where σ is the standard deviation of the Rician noise of the data and I_0 is the modified zeroth-order Bessel function of the first kind. It can be shown that functional (8) is possibly non-convex depending on the data f, λ and σ. Using (1), (2) and (8) the minimization problem is formulated as: Fixed λ and σ and given a noisy image $f \in L^\infty(\Omega)$ recover $u \in BV(\Omega) \cap L^\infty(\Omega)$ minimizing the energy:

$$J(u) + \lambda H(u, f) = \int_\Omega |Du| + \frac{\lambda}{2\sigma^2} \int_\Omega u^2 dx - \lambda \int_\Omega \log I_0 \left(\frac{uf}{\sigma^2} \right) dx. \tag{9}$$

When the functional in (9) is considered for minimization, the variational approach leads to the resolution of a nonlinear multivalued PDE elliptic equation which is the Euler Lagrange equation for optimization. In fact the first order optimality condition reads

$$\partial J(u) + \lambda \partial_u H(u, f) \ni 0 \tag{10}$$

with (Gâteaux) differential

$$\partial_u H(u, f) = \frac{u}{\sigma^2} - \frac{I_1 \left(uf/\sigma^2 \right)}{I_0 \left(uf/\sigma^2 \right)} \frac{f}{\sigma^2} \tag{11}$$

where I_1 is the modified first-order Bessel function of the first kind and verifies $0 \leq I_1(\xi)/I_0(\xi) < 1$, $\forall \xi > 0$. As we introduced before, this gives rise to a number of interesting theoretical problems when the Total Variation operator is considered as a *prior*, because the energy functional is not differentiable at the origin (i.e. $\nabla u = \bar{0}$) and regular approximated problems must be solved. A number of mathematical difficulties is associated with the multivalued formulation (10) and a regularization of the diffusion term $\operatorname{div}(\nabla u/|\nabla u|)$ in form $\operatorname{div}(\nabla u/|\nabla u|_\epsilon)$, with $|\nabla u|_\epsilon = \sqrt{|\nabla u|^2 + \epsilon^2}$ and $0 < \epsilon \ll 1$ is implemented to avoid degeneration of the equation where $\nabla u = \bar{0}$. Using this approximation it is possible to give a (weak) meaning to the following formulation: Fixed λ, σ and (small) ϵ and given

$f \in L^\infty(\Omega) \cap [0,1]$ find $u_\epsilon \in W^{1,1}(\Omega) \cap [0,1]$ solving

$$-\text{div}\left(\frac{\nabla u}{|\nabla u|_\epsilon}\right) + \frac{\lambda}{\sigma^2}\left(u - \left[I_1\left(\frac{uf}{\sigma^2}\right)/I_0\left(\frac{uf}{\sigma^2}\right)\right] f\right) = 0$$

which we write in form

$$-\text{div}\left(\frac{\nabla u_\epsilon}{|\nabla u_\epsilon|_\epsilon}\right) + \frac{\lambda}{\sigma^2}\left[u_\epsilon - r_\epsilon(u_\epsilon, f) f\right] = 0 \tag{12}$$

complemented with Neumann homogeneous boundary conditions $\partial u_\epsilon / \partial n = 0$ and where, for notational simplicity, we introduced the nonlinear function

$$r_\epsilon(u_\epsilon, f) = I_1(u_\epsilon f/\sigma^2)/I_0(u_\epsilon f/\sigma^2).$$

This is a nonlinear (in fact quasilinear) elliptic problem that we solve with a gradient descent scheme until stabilization (when $t \to +\infty$) of the evolutionary solution to steady state, i.e. a solution of the elliptic problem (12) which is a minimum of the approximating energy functionals

$$\begin{aligned} E_\epsilon(u_\epsilon) &= J_\epsilon(u_\epsilon) + \lambda H(u_\epsilon, f) = \\ &= \int_\Omega j_\epsilon(u_\epsilon)dx + \lambda \int_\Omega h(u_\epsilon)dx = \\ &= \int_\Omega \sqrt{|\nabla u_\epsilon|^2 + \epsilon^2}dx + \frac{\lambda}{2\sigma^2}\int_\Omega u_\epsilon^2 dx - \lambda \int_\Omega \log I_0\left(\frac{u_\epsilon f}{\sigma^2}\right)dx. \end{aligned} \tag{13}$$

When $\epsilon \to 0$ we have $u_\epsilon \to u$, $J_\epsilon(u_\epsilon) \to J(u)$ and the energies in (9) and (13) coincide.

The gradient descent approach amounts to solve the associated nonlinear parabolic problem:

$$\frac{\partial u_\epsilon}{\partial t} = \text{div}\left(\frac{\nabla u_\epsilon}{|\nabla u_\epsilon|_\epsilon}\right) - \frac{\lambda}{\sigma^2}\left[u_\epsilon - r_\epsilon(u_\epsilon, f) f\right] \tag{14}$$

complemented with Neumann homogeneous boundary conditions $\partial u_\epsilon / \partial n = 0$ and initial condition $u_\epsilon(0, x) = u_0^\epsilon(x)$ whose (weak) solution stabilizes (when $t \to +\infty$) to the steady state of (12), i.e. a minimum of (13) which approximates, for ϵ sufficiently small, a minimum of the energy functional (9). A direct gradient descent method has been used in [12] in order to validate the model assumption of Rician noise. This approach is found to be inherently slow because a stabilization at the steady state is needed. Also, that scheme is finally explicit and very small time steps have to be used to avoid numerical oscillations. Here we present a framework to solve numerically and efficiently the gradient descent scheme (gradient flow)

associated to the Rician energy minimization problem introducing a semi-implicit formulation. Details can be found in [13].

Using a simple Euler discretization of the time derivative, stationary problems of the ROF type [17] are deduced. This allows to use the well known dual formulation of the ROF model proposed in [3] to speed up the computations. As a by-product of this approach the exact Total Variation operator can be computed and this provides accuracy of the solution in so far truly (discontinuous) bounded variation solutions are numerically approximated. In fact we considered the approximated Euler–Lagrange equation (12) associated to the minimization of the energy (9). This is a modelling approximation and we can get rid of it. We argue as follows. Considering the original Euler–Lagrange equation associated to the energy (9) we have (with abuse of notation for the diffusive term)

$$-\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)+\frac{\lambda}{\sigma^2}\left[u-r(u,f)f\right]=0 \tag{15}$$

with $r(u,f)=I_1(uf/\sigma^2)/I_0(uf/\sigma^2)$. A rigorous treatment of Eq. (15) should follow the multivalued formalism of (10).

Using again a gradient descent scheme we have to solve the parabolic problem:

$$\frac{\partial u}{\partial t}=\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)-\frac{\lambda}{\sigma^2}\left[u-r(u,f)f\right] \tag{16}$$

together with Neumann homogeneous boundary conditions $\partial u/\partial n=0$ and initial condition $u(0,x)=u_0(x)$. For comparison purposes we used $u_0(x)=u_0^{\epsilon}(x)$ in all numerical tests.

Using forward finite differences for the temporal derivative in (16) and a semi-implicit scheme where only the term depending on the ratio of the Bessel's functions is delayed, results in the numerical scheme:

$$\left(1+\tau\frac{\lambda}{\sigma^2}\right)u^{n+1}=u^n+\tau\left(\operatorname{div}\left(\frac{\nabla u^{n+1}}{|\nabla u^{n+1}|}\right)+\frac{\lambda}{\sigma^2}r(u^n,f)f\right) \tag{17}$$

where the diffusive term is (formally) exact and implicitly considered. Defining $\beta=(\tau\lambda)/\sigma^2$, $\gamma=(1+\beta)/\tau$ and

$$g^n=\left(\frac{1}{1+\beta}\right)u^n+\left(\frac{\beta}{1+\beta}\right)r(u^n,f)f \tag{18}$$

we can write:

$$-\operatorname{div}\left(\frac{\nabla u^{n+1}}{|\nabla u^{n+1}|}\right)+\left(\frac{1}{\gamma}\right)\left(u^{n+1}-g^n\right)=0 \tag{19}$$

which is the Euler–Lagrange equation of a ROF energy functional.

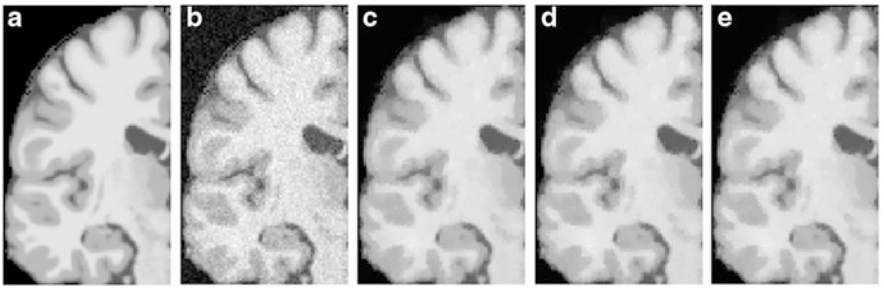


Fig. 10 The original image and the contaminated phantom are shown in (**a**) original phantom and (**b**) noisy for $\sigma = 0.05$. The denoised images obtained with the R-ROF-Dual, R-ROF-Primal-Dual, R-Primal-Dual algorithms and for the parametric values $\sigma = 0.05$, $\lambda = 0.075$ are presented in the sub-plots (**c**) R-ROF-D denoised, (**d**) R-ROF-PD denoised and (**e**) R-PD denoised, respectively

$$E_n(u) = \int_\Omega |Du| + \left(\frac{1}{2\gamma}\right) \int_\Omega (u - g^n)^2 dx \tag{20}$$

for any positive integer $n > 0$, with (artificial) time $t_n = n\tau$. Hence, at each gradient descent step τ, we can solve a ROF problem associated to the minimization of the energy (20) in the space $BV(\Omega) \cap [0, 1]$. This problem is mathematically well-posed and it can be numerically solved by very efficient methods, when formulated using well known duality arguments in [3] or primal-dual algorithms in [4, 18].

In our study we first compared different algorithms using synthetic brain images from the BrainWeb Simulated Brain Database[1] at the Montreal Neurological Institute. The original phantoms were artificially contaminated with Rician noise considering the data as a complex image with zero imaginary part and adding random Gaussian perturbations to both the real and imaginary part, before computing the magnitude image (Fig. 10).

Apart from the modelling exercise and the implementation details of the algorithms presented above, our main interest relies in the application to real brain images. In the following we present some preliminary results we obtained in [13] for Diffusion Weighted Magnetic Resonance Images (DW-MRI) denoising. The DW-MRI are images acquired in order to obtain a Diffusion Tensor Image (DTI). Accurate denoising of the DW-MRI is crucial for a good DTI reconstruction because of their characteristic very low SNR, [2].

Diffusion Tensor Imaging is becoming one of the most popular methods for the analysis of the white matter (WM) structure of the brain, where some alterations can be found from early stages in some degenerative diseases. This technique measures Brownian motion (random motion) of the water molecules in the brain, which is assumed to be isotropic when it is not restricted by the surrounding structure. In the WM regions, which contain densely packed fibre bundles, they cause an anisotropic diffusion of water molecules along the perpendicular directions to them. At each

[1] Available at http://www.bic.mni.mcgill.ca/brainweb.

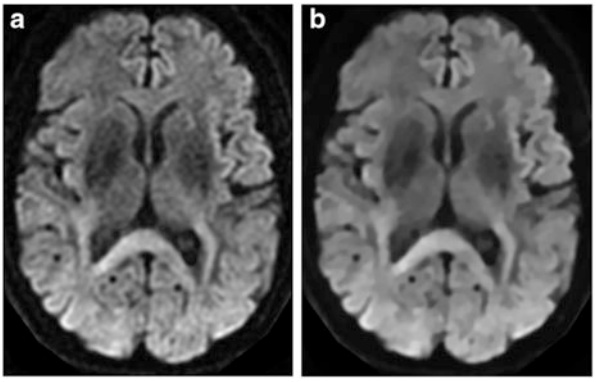


Fig. 11 A slice of the original Diffusion Weighted Image corresponding to the (1, 0, 0) gradient direction and the corresponding denoised image. (**a**) Original; (**b**) denoised with $\lambda = \sigma/2$

voxel of a DTI the water diffusion is represented by a symmetric 3×3 tensor, where the information of the preferred directions of the motion and the relevance of these directions is found in the eigenvectors and the eigenvalues of the tensor. These tensorial data can be represented as different scalar measurements, one of them is the Fractional Anistropy (FA) of the tissue, which is defined as

$$\mathrm{FA} = \sqrt{\frac{3\left((\hat{\lambda}-\lambda_1)^2 + (\hat{\lambda}-\lambda_2)^2 + (\hat{\lambda}-\lambda_3)^2\right)}{2\left(\lambda_1^2+\lambda_2^2+\lambda_3^2\right)}}$$

where the λ_i are the eigenvalues of the tensor and $\hat{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$. The FA values vary from 0, (when the motion in the voxel is completely isotropic) to 1 (totally anisotropic). For the reconstruction of the DTI a set of DWI has to be acquired, scanning the tissue in different directions of the space. At least six DWI volumes are needed in order to be able to calculate the DTI, which is a positive defined matrix. The noise present into the DWI scalar images can generate small, negative eigenvalues. Increasing the number of directions along which the brain is scanned improves the image quality but at the expenses of a longer acquisition time. The importance of pre-processing the DW Images previously to the DTI reconstruction is then two-fold: to improve the DT image quality through accurate Rician denoising so allowing shorter scanning time.

The data we used consist of a DW-MR brain volume provided by Fundación CIEN-Fundación Reina Sofía which was acquired with a 3 T General Electric scanner equipped with an 8-channel coil. The DW images have been obtained with a single-shot spin-eco EPI sequence (FOV = 24 cm, TR = 9,100, TE = 88.9, slice thickness = 3 mm, spacing = 0.3, matrix size = 128×128, NEX = 2). The DW-MRI data consists on a volume obtained with $b = 0\,/\mathrm{mm}^2$ and 15 volumes with $b = 1{,}000\,\mathrm{s/mm}^2$.

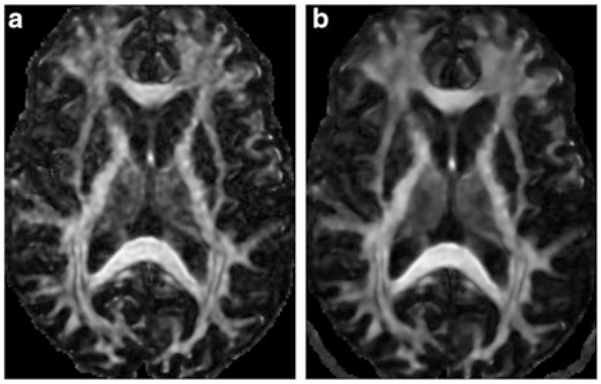


Fig. 12 A slice of the Fractional Anisotropy estimated from the Tensor Image. *Dark colour* corresponds to values near zero (isotropic regions) and *bright color* corresponds to values near one (anisotropic regions). (**a**) From original DWI data; (**b**) from denoised DWI data with $\lambda = \sigma/2$

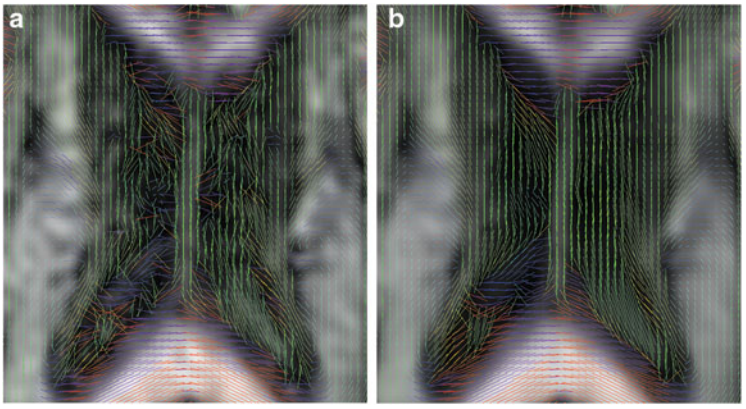


Fig. 13 A detail of the first eigenvectors of the DTI over the FA image. The color is based on the main orientation of the tensorial data. *Red* means right-left direction, *green* anterior-posterior and *blue* inferior-superior. Fibres with an oblique angle have a color that is a mixture of the principal colors and dark color is used for the isotropic regions. (**a**) From original DWI data; (**b**) from denoised DWI data with $\lambda = \sigma/2$

These DW-MR images, which represent diffusion measurements along multiples directions, are denoised with the proposed method previously to the Diffusion Tensorial Image reconstruction, which was done with the 3d Slicer tools.[2] In Fig. 11a we show a slice of the original DWI data corresponding to the (1, 0, 0) gradient direction where the affecting noise is clearly visible. The complete DW-MRI data volume is denoised using the proposed method. The Rician noise

[2]Free available in http://www.slicer.org/.

standard deviation (σ) has been estimated for each slice of each gradient direction while we used a value of $\lambda = \sigma/2$ for the denoising. The slice resulting from the denoising process is shown in Fig. 11b. It can be observed how noise has been removed in the denoised images but the details and the edges have been fully preserved, as we should expect when the exact TV model is solved. The effect of this denoising process over the reconstructed tensor and their derived scalar measurements (obtained with the 3d Slicer tools) is presented in Figs. 12 and 13. Figure 12 shows a Fractional Anisotropy image where the structures and details are clearly enhanced if the DW-MRI volume is denoised previously. When finer details are considered the denoising step is yet more crucial. For instance in Fig. 13 the main eigenvector of the tensor is represented, where the noise on the original DWI data cause inhomogeneities (see Fig. 13a) in the eigenvectors field which are product of the noise (Fig. 13b).

Acknowledgements The authors wish to thank to all the Research Institutions involved into the Alzheimer Project (Fundación CIEN-Fundación Reina Sofía, Lab. de Neuroimagen. Centro de Tecnología Biomédica (UPM-URJC) Universidad Rey Juan Carlos) and very specially to the MICINN Spanish Minister for Science and Innovation for supporting Project TEC2012-39095-C03-02. Finally the first author wish to thank to the organizers of the XV Escuela Hispano-Francesa Jacques-Louis Lions for inviting him as a lecturer.

References

1. Andreu-Vaillo, F., Caselles, V., Mazón, J.M.: Parabolic quasilinear equations minimizing linear growth functionals. In: Progress in Mathematics. Birkhauser, Basel (2004). ISBN 3-7643-6619-2
2. Basu, S., Fletcher, T., Whitaker, R.: Rician noise removal in diffusion tensor MRI. In: MICCAI 2006. Lecture Notes in Computer Science, vol. 4190, pp. 117–125. Springer, Heidelberg (2006)
3. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**(1–2), 89–97 (2004)
4. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
5. Chan, T., Shen, J.: Image processing and analysis: variational, PDE, wavelet, and stochastic methods. In: Society for Industrial and Applied Mathematics (2005). ISBN: 089871589X
6. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
7. Garamendi, J.F., Malpica, N., Martel, J., Schiavi, E.: Automatic segmentation of the liver in CT using level sets without edges. In: Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, vol. 4477, pp. 161–168. Springer, Berlin, Heidelberg, New York (2007). ISBN-10 3-540-72846-5
8. Garamendi, J.F., Malpica, N., Schiavi, E.: A fast anisotropic mumford-shah functional based segmentation. In: Lecture Notes in Computer Science, vol. 5524, pp. 322–329. Springer, Berlin, Heidelberg (2009). http://www.springerlink.com/content/v2u2nq61502hw50x/
9. Garamendi, J.F., Malpica, N., Schiavi, E. (2009). Multiphase systems for medical image region classification. In: Mathematical Models in Engineering, Biology and Medicine. Conference on Boundary Value Problems, Santiago de Compostela, Spain, 16–19 September 2008, p. 104
10. Garamendi, J.F., Gaspar, F.J., Malpica, N., Schiavi, E.: Box relaxation schemes in staggered discretizations for the dual formulation of total variation minimization. IEEE Trans. Image Process. **22**, 2030–2043 (2013)

11. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing using MATLAB, 2nd edn. Gatesmark Publishing (2009). ISBN: 978-0-9820854-0-0
12. Martín, A., Garamendi, J.F., Schiavi, E.: Iterated rician denoising. In: IPCV'11, pp. 959–963. CSREA Press (2011)
13. Martin, A., Garamendi, J.-F., Schiavi, E.: MRI TV-rician denoising. Commun. Comput. Inf. Sci. **357**, 255–268 (2013)
14. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**(5), 577–685 (1989)
15. Osher, S., Paragios, N.: Geometric Level Set Methods in Imaging, Vision, and Graphics. Springer, New York (2003). ISBN: 0387954880
16. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**(7), 629–639 (1990)
17. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D Nonlinear Phenomena **60**(1), 259–268 (1992)
18. Zhu, M., Chan, T.: An Efficient Primal-Dual Hybrid Gradient Algorithm for Total Variation Image Restoration. UCLA CAM Report 08–34 (2008)

On Probabilistic Analytical and Numerical Approaches for Divergence Form Operators with Discontinuous Coefficients

Denis Talay

Abstract In this paper we review some recent results on stochastic analytical and numerical approaches to parabolic and elliptic partial differential equations involving a divergence form operator with a discontinuous coefficient and a compatibility transmission condition.

In the one-dimensional case existence and uniqueness results for such PDEs can be obtained by stochastic methods. The probabilistic interpretation of the solutions allows one to develop and analyze a low complexity Monte Carlo numerical resolution method. In addition, it allows to get accurate pointwise estimates for the derivatives of the solutions from which sharp convergence rate estimates are deduced for the stochastic numerical method.

A stochastic approach is also developed for the linearized Poisson–Boltzmann equation in Molecular Dynamics. As in the one-dimensional case, the probabilistic interpretation of the solution involves the solution of a SDE including a non standard local time term related to the discontinuity interface. We present an extended Feynman–Kac formula for the Poisson–Boltzmann equation. This formula justifies various probabilistic numerical methods to approximate the free energy of a molecule and bases error analyzes.

We finally present probabilistic interpretations of the non-linearized Poisson–Boltzmann equation in terms of backward stochastic differential equations.

1 Introduction

Let us consider a positive matrix-valued function a which is smooth except at the interface surfaces between subdomains of $\mathbb{R}^d$ and the parabolic diffraction problem

D. Talay (✉)
2002 Route des Lucioles, BP 93, 06901 Sophia Antipolis Cedex, France
e-mail: denis.talay@inria.fr

C. Parés et al. (eds.), *Advances in Numerical Simulation in Physics and Engineering*, SEMA SIMAI Springer Series 3, DOI 10.1007/978-3-319-02839-2_7,

$$\begin{cases} \partial_t u(t,x) - \dfrac{1}{2}\mathrm{div}(a(x)\nabla)u(t,x) = 0 \text{ for all } (t,x) \in (0,T] \times \mathbb{R}^d, \\ u(0,x) = f(x) \text{ for all } x \in \mathbb{R}^d, \\ \text{Compatibility transmission conditions along the interfaces surfaces.} \end{cases} \tag{1}$$

Suppose that $L := \frac{1}{2}\mathrm{div}(a\nabla)$ is a strongly elliptic operator. One can find in, e.g., Ladyzenskaya et al. [12, Chap. III, Sect. 13] the proof of existence and uniqueness of continuous solutions with possibly discontinuous derivatives along the interface surfaces.

Various probabilistic interpretations of the operator L have been developed by many authors: for example, Fukushima et al. [10] and Rozkosz [22] use the theory of Dirichlet forms to construct an abstract Markov process whose generator is L. However, these constructions are neither favorable to derive stochastic numerical resolution methods for (1), nor to get the accurate pointwise estimates for partial derivatives of the function u which are necessary to analyze the convergence rate of the numerical methods.

In the one dimensional case $d = 1$, the differential operator $\frac{1}{2}\partial_x(a\partial_x)$ is the generator of the solution to a stochastic differential equation (SDE) involving its own local time: see, e.g., Bass and Chen [2], Étoré [7], Martinez and Talay [15]. This new description is the starting point for recent numerical studies: Lejay and Martinez [14] and Étoré [7, 8] proposed simulation methods for this solution based on approximations of $a(x)$ and random walks simulations, and they analyzed the convergence rates of these methods.

We here focus on a numerical method based on the Euler discretization scheme for stochastic differential equations with weighted local times. We obtain sharp convergence rate estimates owing to our probabilistic interpretation of the strong solutions to (1) in terms of exact solutions to such SDEs.

The extension of this new analytical and numerical approach to general multi-dimensional cases is still in progress: see [17]. However recent advances concern the Poisson–Boltzmann equation in Molecular Dynamics in its linear and semi-linear forms, which we summarize in the last sections.

In all the paper we emphasize the tools from stochastic numerics and stochastic analysis which allow us to deal with the transmission boundary conditions.

1.1 Notation

For a left continuous function g we denote by $g_-(x)$ and $g(x-)$ the left limit of g at point x, respectively. When g is right continuous, we denote either by $g_+(x)$ or by $g(x+)$ the right limit of g at point x.

We denote by $\mathcal{C}_b^\ell(\mathbb{R})$ the set of all bounded continuous functions with bounded continuous derivatives up to order ℓ, and by $\partial_x^i g$ the i-th derivative of g.

For all integers $0 \leq \ell < \infty$ and $1 \leq p \leq \infty$ we denote the $L^p(\mathbb{R})$ norm of the function g by $\|g\|_p$ and we set

$$\|g\|_{\ell,p} := \sum_{i=0}^{\ell} \|\partial_x^i g\|_p. \tag{2}$$

2 One-Dimensional Diffraction Problems

All the results of this section come from Martinez and Talay [15].

We consider the case $d = 1$ and $a(x) = (\sigma(x))^2$ is a real function on $\mathbb{R}$ which is right continuous at point 0 and differentiable on $\mathbb{R} - \{0\}$ with a bounded derivative. All the contents of the section hold true when a has a finite number of discontinuities.

We rewrite the partial differential equation (PDE) (1) and its transmission condition as

$$\begin{cases} \partial_t u(t,x) - \frac{1}{2}\partial_x(a(x)\partial_x u(t,x)) = 0, \ (t,x) \in (0,T] \times (\mathbb{R} - \{0\}), \\ u(t,0+) = u(t,0-), \ t \in [0,T], \\ u(0,x) = f(x), \ x \in \mathbb{R}, \\ a(0+)\partial_x u(t,0+) = a(0-)\partial_x u(t,0-), \ t \in [0,T]. \quad (\star) \end{cases} \tag{3}$$

We assume

$$\exists \lambda > 0, \ \Lambda > 0, \ 0 < \lambda \leq a(x) = (\sigma(x))^2 \leq \Lambda < +\infty \text{ for all } x \in \mathbb{R}. \tag{4}$$

We also assume that σ is of class $\mathcal{C}_b^3(\mathbb{R} - \{0\})$ and is left and right continuous at point 0, and that the first derivative of σ has finite left and right limits at 0.

Our first result shows that, for a wide class of functions f, the solution of the PDE (1) with $d = 1$ can be represented as

$$u(t,x_0) := \mathbb{E}^{x_0} f(X_t), \tag{5}$$

where the process (X_t) is the unique weak solution to the one-dimensional stochastic differential equation (SDE) with weighted local time

$$dX_t = \sigma(X_t)dB_t + \sigma(X_t)\sigma'_-(X_t)\,dt + \frac{a(0+) - a(0-)}{2a(0+)} dL_t^0(X), \quad X_0 = x_0. \tag{6}$$

Here σ'_- is the left derivative of σ, $(B_t, t \geq 0)$ is a one-dimensional standard Brownian motion on a filtered probability space, and $L_t^0(X)$ is the right-sided local

time of $X := (X_t)$ corresponding to the sign function defined as $\mathrm{sgn}(x) := 1$ for $x > 0$ and $\mathrm{sgn}(x) := -1$ for $x \leq 0$ (see, e.g., Revuz an Yor [21]).

Notice that Eq. (6) involves the local time of the solution. Under conditions weaker than those ones of Theorem 2.2 below, the unique weak solution exists, and this solution is a strong Markov process: see Le Gall [13]. For all real number x_0 we denote by $\mathbb{P}^{x_0}$ the probability distribution of the solution such that $X_0 = x_0$, $\mathbb{P}^{x_0} - \text{a.s.}$

From a numerical point of view, the stochastic representation (5) is unsatisfying because of the difficulty to numerically approximate the local time process $(L_t^0(X))$ with good accuracy and weak computational cost. We thus apply a transformation which removes the local time of X already used by Le Gall [13]. We thus get a new stochastic differential equation without local time which can be discretized by the standard Euler scheme. As the transformation is one-to-one and its inverse is explicit, one then readily deduces an approximation $\overline{X}$ of X. Choosing $\overline{X}_0 = X_0$ we then approximate $u(t, x_0)$ by $\mathbb{E}^{x_0} f(\overline{X}_t)$, the latter being computed by Monte Carlo simulations of $\overline{X}$.

We below state sharp convergence rate estimates for $\mathbb{E}^{x_0} f(\overline{X}_t)$ to $u(t, x_0)$ according to different hypotheses on f. These convergence rates are new in the literature because the SDE obtained by removing the local time has discontinuous coefficients: see [11,24] for a review when the coefficients are smooth, and Yan [25] for a weak convergence of the Euler scheme for general SDEs with discontinuous coefficients (without precise convergence rates).

2.1 A Probabilistic Interpretation of the One-Dimensional PDE (1)

The SDE (6) allows us to construct a stochastic interpretation of Eq. (3).

Theorem 2.1. *Let us assume condition (4) and that the function σ is of class $\mathcal{C}_b^3(\mathbb{R} - \{0\})$ and is left and right continuous at point 0. Moreover, we assume that the first derivative of the function σ has finite left and right limits at 0. Let (X_t) be the solution to (6). Let the bounded function f be in the set*

$$\mathcal{W}^2 = \left\{ g \in \mathcal{C}_b^2(\mathbb{R} - \{0\}), g^{(i)} \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}) \text{ for } i = 1, 2, \right.$$
$$\left. a(0+)g'(0+) = a(0-)g'(0-) \right\}. \tag{7}$$

Then the function

$$u(t, x) := \mathbb{E}^x f(X_t), \ (t, x) \in [0, T] \times \mathbb{R},$$

is the unique one in $\mathcal{C}_b^{1,2}([0,T]\times(\mathbb{R}-\{0\}))$ *and continuous on* $[0,T]\times\mathbb{R}$ *which satisfies (3).*

Using the preceding stochastic representation of $u(t,x)$ one can get the following accurate pointwise estimates for its derivatives.

Theorem 2.2. *(i) Under the hypotheses on the function* σ *made in Theorem 2.1, the probability distribution of* X_t *under* $\mathbb{P}^x$ *has a density* $q^X(x,t,y)$ *which satisfies:*

$$\exists C>0,\ \forall x\in\mathbb{R},\ \forall t>0,\ \textit{Leb-a-e. } y\in\mathbb{R}-\{0\},\ q^X(x,t,y)\le\frac{C}{\sqrt{t}} \tag{8}$$

and

$$\exists C>0,\ \forall x\in\mathbb{R},\ \forall t\in(0,T],\ \forall f\in L^1(\mathbb{R}),\ |u(t,x)|=|\mathbb{E}^x f(X_t)|\le\frac{C}{\sqrt{t}}\|f\|_1. \tag{9}$$

(ii) Suppose in addition that the function σ *is of class* $\mathcal{C}_b^4(\mathbb{R}-\{0\})$ *and that its three first derivatives have finite left and right limits at* 0*. Set*

$$\begin{aligned}\mathcal{W}^4:=\big\{&g\in\mathcal{C}_b^4(\mathbb{R}-\{0\}),\ g^{(i)}\in L^2(\mathbb{R})\cap L^1(\mathbb{R})\ \textit{for } i=1,\dots,4\\ &a(0+)g'(0+)=a(0-)g'(0-)\ \textit{and}\\ &a(0+)(\mathcal{L}g)'(0+)=a(0-)(\mathcal{L}g)'(0-)\big\},\end{aligned} \tag{10}$$

where

$$\mathcal{L}g(x):=\sigma(x)\sigma'_-(x)\partial_x g_-(x)+\frac{1}{2}a(x)\partial^2_{xx}g(x)\mathbb{I}_{x\neq0}. \tag{11}$$

Then, for all $j=0,1,2$ *and* $i=1,\dots,4$ *such that* $2j+i\le4$,

$$\exists C>0,\ \forall x\in\mathbb{R},\ \forall t\in(0,T],\ \forall f\in\mathcal{W}^4,\ |\partial_t^j\partial_x^i u(t,x)|\le\frac{C}{\sqrt{t}}\|f'\|_{\gamma,1}, \tag{12}$$

where $\gamma=1$ *if* $2j+i=1$ *or 2, and* $\gamma=3$ *if* $2j+i=3$ *or 4, and* $\|\cdot\|_{\gamma,1}$ *is defined as in (2).*

The proofs of the two preceding theorems are quite long. We here summarize their key steps.

2.2 *Proof of Theorem 2.1*

The two following observations are key ingredients in the proof.

First, for all function g of class $\mathcal{C}_b^2(\mathbb{R}-\{0\})$ having a second derivative in the sense of the distributions which is a Radon measure and satisfying the transmission condition

$$a(0+)g'(0+) = a(0-)g'(0-),$$

the Itô–Tanaka formula applied to $g(X_t)$ and the definition (11) of $\mathcal{L}$ lead to

$$\forall x \in \mathbb{R},\ \forall t > 0,\ \mathbb{E}^x g(X_t) = g(x) + \int_0^t \mathbb{E}^x \mathcal{L}g(X_s)\, ds. \tag{13}$$

Second, let $\sigma^+(x)$ be an arbitrary $\mathcal{C}_b^3(\mathbb{R})$ extension of the function $\sigma(x)\mathbb{I}_{x>0}$ which satisfies, for $a^+(x) := (\sigma^+(x))^2$,

$$0 < \lambda \le a^+(x) \le \Lambda < +\infty \text{ for all } x \in \mathbb{R}.$$

Denote by (X_t^+) the unique strong solution to

$$dX_t^+ = \sigma^+(X_t^+)dB_t + \sigma^+(X_t^+)(\sigma^+)'(X_t^+)\, dt.$$

Let $\tau_0(X)$ be the first passage time of the process (X_t) at point 0:

$$\tau_0(X) := \inf\{s > 0 \ :\ X_s = 0\}.$$

Given $T > 0$, let $r_0^x(s)$ be the density under $\mathbb{P}^x$ of $\tau_0(X) \wedge T$. Notice that $\tau_0(X) = \tau_0(X^+)$. For all function ϕ such that $\mathbb{E}|\phi(X_t)|$ is finite we have, for all $x > 0$ and $0 \le t \le T$,

$$\begin{aligned}
\mathbb{E}^x\phi(X_t) &= \mathbb{E}^x\left[\phi(X_t)\mathbb{I}_{\{\tau_0 \ge t\}}\right] + \mathbb{E}^x\left[\phi(X_t)\mathbb{I}_{\{\tau_0 < t\}}\right] \\
&= \mathbb{E}^x\left[\phi(X_t^+)\mathbb{I}_{\{\tau_0 \ge t\}}\right] + \int_0^t \mathbb{E}^0\phi(X_{t-s})r_0^x(s)\, ds \\
&= \mathbb{E}^x\phi(X_t^+) - \mathbb{E}^x\left[\phi(X_t^+)\mathbb{I}_{\{\tau_0 < t\}}\right] + \int_0^t \mathbb{E}^0\phi(X_s)r_0^x(t-s)\, ds \\
&= \mathbb{E}^x\phi(X_t^+) - \int_0^t \mathbb{E}^0\phi(X_s^+)r_0^x(t-s)ds + \int_0^t \mathbb{E}^0\phi(X_s)r_0^x(t-s)\, ds.
\end{aligned} \tag{14}$$

Of course, a similar representation holds true for all $x < 0$ provided the introduction of a diffusion process X^- obtained by smoothly extending $\sigma(x)\mathbb{I}_{x<0}$.

We need the following lemma.

Lemma 2.3. *There exists $\tilde{C} > 0$ such that, for all $0 \leq \alpha < 1$, and all function H bounded on $[0,T]$, continuously differentiable on $(0,T]$, satisfying $H(0) = 0$ and*

$$|H'(s)| \leq \frac{C_H}{s^\alpha} \text{ for all } s \in (0,T],$$

it holds

$$\forall t \in (0,T],\ \forall x \neq 0,\ \left|\partial_x \int_0^t r_0^x(t-s)H(s)ds\right| \leq C_H\ \tilde{C},$$

and

$$\forall t \in (0,T],\ \forall x \neq 0,\ \left|\partial_{xx}^2 \int_0^t r_0^x(t-s)H(s)ds\right| \leq C_H\ \tilde{C}\left(1 + \frac{1}{t^\alpha}\right).$$

First Step: Smoothness and Boundedness. In this paragraph we prove that the function $u(t,x) := \mathbb{E}^x f(X_t)$ is in $C_b^{1,2}([0,T]\times(\mathbb{R}-\{0\})$. Without loss of generality, we limit ourselves to the case $x > 0$. From the representation (14) with $\phi \equiv f$ and a representation of $r_0^x(s)$ in terms of the joint distribution of Brownian motion and Bessel bridges (see, e.g., Pauwels [19]), it is easy to deduce the continuity of $u(t,x)$ w.r.t. t and x. In particular, the second and third equalities in (3) are satisfied. Next, to study the boundedness of the function $\partial_x u(t,x)$, we differentiate the flow of (X_t^+):

$$\begin{aligned}
&\partial_x \mathbb{E}^x f(X_t^+) \\
&= \mathbb{E}^x\left[f'(X_t^+)\exp\left(\int_0^t (\sigma^+)'(X_s^+)dB_s\right.\right. \\
&\left.\left. + \frac{1}{2}\int_0^t \{((\sigma^+)'(X_s^+))^2 + \sigma^+(X_s^+)(\sigma^+)''(X_s^+)\}ds\right)\right].
\end{aligned}$$

If we integrate by parts the stochastic integral in the right-hand side of the previous expression, then there exists a bounded continuous function G such that

$$\partial_x \mathbb{E}^x f(X_t^+) = \mathbb{E}^x\left[f'(X_t^+)\exp\left(\sigma^+(X_t^+) - \sigma^+(x) + \int_0^t G(X_s^+)ds\right)\right]. \tag{15}$$

Therefore

$$\exists C > 0,\ \forall 0 < t \leq T,\ \forall x \in \mathbb{R},\ |\partial_x \mathbb{E}^x f(X_t^+)| \leq C\,\|f'\|_\infty.$$

We then consider the two last terms of the right-hand side of (14). We now use (13) and Lemma 2.3 with

$$H(s) = \mathbb{E}^0 f(X_s^+) - \mathbb{E}^0 f(X_s)$$

and $C_H = C(||\mathcal{L}^+ f||_\infty + ||\mathcal{L} f||_\infty)$, where $\mathcal{L}^+$ is the infinitesimal generator of the process (X_t^+), that is,

$$\mathcal{L}^+ f(x) := \frac{1}{2} a^+(x) f''(x) + \frac{1}{2} (a^+)'(x) f'(x).$$

Then, we have

$$\exists C > 0,\ \forall 0 < t \le T,\ \forall x \ne 0,\ |\partial_x u(t,x)| \le C \|f'\|_\infty + C \|f''\|_\infty.$$

We proceed similarly to prove that

$$\exists C > 0,\ \forall 0 < t \le T,\ \forall x \ne 0,\ |\partial^2_{xx} u(t,x)| \le C \|f'\|_\infty + C \|f''\|_\infty, \qquad (16)$$

noticing that, from (15),

$$\exists C > 0,\ \forall 0 < t \le T,\ \forall x \in \mathbb{R},\ |\partial^2_{xx} \mathbb{E}^x f(X_t^+)| \le C \|f'\|_\infty + C \|f''\|_\infty.$$

Second Step: $u(t,x)$ Satisfies the First Equality in (3). In view of (13) we have, for all $0 < t < T$, $0 < \epsilon < T - t$ and x in $\mathbb{R}$,

$$u(t+\epsilon, x) - u(t,x) = \mathbb{E}^x f(X_{t+\epsilon}) - \mathbb{E}^x f(X_t) = \int_t^{t+\epsilon} \mathbb{E}^x \mathcal{L} f(X_s) ds. \qquad (17)$$

Changing ϕ into $\mathcal{L} f$ in (14) shows that $\mathbb{E}^x \mathcal{L} f(X_t)$ is a continuous function w.r.t. t. Therefore $\partial_t u(t,x)$ is well defined for all $0 < t \le T$ and all x in $\mathbb{R}$.

In addition, as (X_t) is strong Markov,

$$u(t+\epsilon, x) - u(t,x) = \mathbb{E}^x u(t, X_\epsilon) - u(t,x). \qquad (18)$$

Itô's formula leads to

$$\begin{aligned}
\mathbb{E}^x u(t, X_\epsilon) - u(t,x) &= \mathbb{E}^x u(t, X_\epsilon) \mathbb{I}_{\tau_0 \ge \epsilon} + \mathbb{E}^x u(t, X_\epsilon) \mathbb{I}_{\tau_0 < \epsilon} - u(t,x) \\
&= \int_0^\epsilon \mathbb{E}^x \mathcal{L} u(t, X_s) ds\, \mathbb{I}_{\tau_0 \ge \epsilon} - u(t,x)\, \mathbb{P}^x(\tau_0 \le \epsilon) \\
&\quad + \int_0^\epsilon \mathbb{E}^0 u(t, X_s) r_0^x(\epsilon - s) ds.
\end{aligned}$$

Divide by ϵ the left and right-hand sides and observe that, for all $x \ne 0$,

$$\mathbb{P}^x - \text{a.s.},\ \lim_{\epsilon \searrow 0} \frac{1}{\epsilon} \int_0^\epsilon \mathcal{L} u(t, X_s) ds = \mathcal{L} u(t,x).$$

Applying Lebesgue's Dominated Convergence theorem we deduce

$$\lim_{\epsilon \searrow 0} \frac{\mathbb{E}^x u(t, X_\epsilon) - u(t,x)}{\epsilon} = \mathcal{L}u(t,x) - u(t,x) r_0^x(0) + \lim_{\epsilon \searrow 0} \frac{\int_0^\epsilon \mathbb{E}^0 u(t, X_s) r_0^x(\epsilon - s) ds}{\epsilon}.$$

As $r_0^x(0) = 0$, Lebesgue's Dominated Convergence theorem implies that, for all $x \neq 0$,

$$\partial_t u(t,x) = \mathcal{L}u(t,x). \tag{19}$$

Third Step: $u(t,x)$ Satisfies the Transmission Condition ($\star$). In view of the preceding first step, for all fixed t the second partial derivative w.r.t. x of $u(t,x)$ is a Radon measure. Thus we may apply the Itô–Tanaka formula to $u(t, X_s)$ for $0 \leq s \leq \epsilon$ and fixed time t. Our first step also ensures that the resulting Brownian integrals are martingales. Therefore

$$\begin{aligned} \mathbb{E}^0 u(t, X_\epsilon) - u(t,0) = \mathbb{E}^0 \int_0^\epsilon \mathcal{L}u(t, X_s) ds \\ + \frac{1}{2a(0+)} (a(0+)\partial_x u(t, 0+) \\ - a(0-)\partial_x u(t, 0-)) \mathbb{E}^0 L_\epsilon^0(X). \end{aligned} \tag{20}$$

Observe that the equality (18) holds true for $x = 0$ since it only results from the Markov property of (X_t) and that, combined with (17) it leads to

$$\mathbb{E}^0 u(t, X_\epsilon) - u(t,0) = \int_t^{t+\epsilon} \mathbb{E}^0 \mathcal{L}f(X_s) ds.$$

Therefore we deduce from (20) that

$$\begin{aligned} &(a(0+)\partial_x u(t,0+) - a(0-)\partial_x u(t,0-)) \, \mathbb{E}^0 L_\epsilon^0(X) \\ &\qquad = 2a(0+) \left(\int_t^{t+\epsilon} \mathbb{E}^0 \mathcal{L}f(X_s) ds - \int_0^\epsilon \mathbb{E}^0 \mathcal{L}u(t, X_s) ds \right). \end{aligned}$$

Since $\mathcal{L}f$ and $\mathcal{L}u(t,\cdot)$ are bounded functions, the compatibility transmission condition ($\star$) will be proved if we show that

$$\liminf_{\epsilon \searrow 0} \frac{\mathbb{E}^0 L_\epsilon^0(X)}{\epsilon} = +\infty. \tag{21}$$

This is achieved by reducing the question to Brownian local times.

Last Step: Uniqueness. We finally prove that $u(t,x) := \mathbb{E}^x f(X_t)$ is the unique solution to (3) in the sense of Theorem 2.1. The standard method to prove stochastic representations of solutions $v(t,x)$ of parabolic equations with smooth coefficients (see, e.g., Friedman [9]) relies on Itô's formula applied to $v(t, X_t)$. Here, as the first space derivative of $u(t,x)$ is discontinuous at $x = 0$ for all t, one would rather need to apply a formula of Itô–Tanaka type. However the classical Itô–Tanaka's formula cannot be extended to functions which depend on time and space. In order to circumvent this difficulty we use a trick taken from Peskir [20, Sect. 3] which, according to the author, is due to Kurtz.

As, for all real number x, $x \vee 0 = \frac{1}{2}(x + |x|)$ and $x \wedge 0 = \frac{1}{2}(x - |x|)$, Itô–Tanaka's formula implies

$$\begin{aligned} d(X_t \vee 0) &= \frac{1}{2} dX_t + \frac{1}{2} \mathrm{sgn}(X_t)\, dX_t + \frac{1}{2}\, dL_t^0(X) \\ &= \mathbb{I}_{X_t>0}\, dX_t + \frac{1}{2}\, dL_t^0(X), \\ d(X_t \wedge 0) &= \frac{1}{2}\, dX_t - \frac{1}{2}\, \mathrm{sgn}(X_t)\, dX_t - \frac{1}{2}\, dL_t^0(X) \\ &= \mathbb{I}_{X_t<0}\, dX_t - \frac{a(0-)}{2a(0+)}\, dL_t^0(X). \end{aligned}$$

Now, let $U(t,x)$ be an arbitrary solution to (3). For all fixed t in $[0,T]$ the function $U(t-s,x)$ is of class $\mathcal{C}_b^{1,2}([0,t] \times \mathbb{R} - \{0\})$ and its partial derivatives have left and right limits when x tends to 0. Thus we may apply the classical Itô's formula to this function and the semi-martingales $(X_s \vee 0)$ and $(X_s \wedge 0)$. As the resulting Brownian integrals are martingales we obtain:

$$\begin{aligned} \mathbb{E}^x U(0, X_t \vee 0) = U(t, x \vee 0) &- \mathbb{E}^x \int_0^t \partial_t U(t-s, X_s \vee 0)\, ds \\ &+ \mathbb{E}^x \int_0^t \partial_x U(t-s, X_s \vee 0)\, \mathbb{I}_{X_s>0}\, \sigma(X_s)\sigma'(X_s)\, ds \\ &+ \frac{1}{2}\mathbb{E}^x \int_0^t \partial_{xx}^2 U(t-s, X_s)\, \mathbb{I}_{X_s>0}\, a(X_s)\, ds \\ &+ \frac{1}{2}\mathbb{E}^x \int_0^t \partial_x U(t-s, 0+) dL_s^0(X). \end{aligned}$$

Similarly, we get

$$\begin{aligned} \mathbb{E}^x U(0, X_t \wedge 0) = U(t, x \wedge 0) &- \mathbb{E}^x \int_0^t \partial_t U(t-s, X_s \wedge 0)\, ds \\ &+ \mathbb{E}^x \int_0^t \partial_x U(t-s, X_s \wedge 0)\, \mathbb{I}_{X_s<0}\, \sigma(X_s)\sigma'(X_s)\, ds \end{aligned}$$

$$+\frac{1}{2}\mathbb{E}^x\int_0^t \partial^2_{xx}U(t-s,X_s)\,\mathbb{I}_{X_s<0}\,a(X_s)\,ds$$
$$-\frac{a(0-)}{2a(0+)}\mathbb{E}^x\int_0^t \partial_x U(t-s,0-)dL^0_s(X).$$

We finally use that $U(t,x)=U(t,x\vee 0)+U(t,x\wedge 0)-U(t,0)$ and $U(0,x)=f(x)$. In view of the first equality in (3), it follows that

$$\mathbb{E}^x f(X_t)=U(t,x)+\frac{1}{2a(0+)}\mathbb{E}^x\int_0^t (a(0+)\partial_x U(t-s,0+)$$
$$-a(0-)\partial_x U(t-s,0-))dL^0_s(X).$$

It now remains to use that, by hypothesis, $U(t,x)$ satisfies the transmission condition $(\star)$. That ends the proof.

2.3 *Proof of Theorem 2.2*

One proves (9) by closely following a part of the proof of Aronson's estimate (see, e.g., Bass [1, Chap. 7, Sect. 4] and Stroock [23]).

Proposition 2.4. *There exists $C>0$ such that, for all $t\in(0,T]$,*

$$\sup_{x\neq 0}|\partial_t u(t,x)|\le\frac{C}{\sqrt{t}}\|f'\|_{1,1}. \tag{22}$$

Proof. As above, w.l.g. we may and do assume $x>0$. We start from (14) and write

$$u(t,x)=\mathbb{E}^x f(X_t^+)+v(t,x), \tag{23}$$

where

$$v(t,x):=-\int_0^t \mathbb{E}^0 f(X_s^+)r_0^x(t-s)\,ds+\int_0^t \mathbb{E}^0 f(X_s)r_0^x(t-s)\,ds. \tag{24}$$

We have

$$v(t,x)=\int_0^t\int_0^{t-s}\left(\mathbb{E}^0\mathcal{L}f(X_\xi)-\mathbb{E}^0\mathcal{L}^+f(X_\xi^+)\right)d\xi\,r_0^x(s)\,ds, \tag{25}$$

and thus

$$\partial_t v(t,x)=\int_0^t\left(\mathbb{E}^0\mathcal{L}f(X_s)-\mathbb{E}^0\mathcal{L}^+f(X_s^+)\right)r_0^x(t-s)\,ds. \tag{26}$$

One can prove the following estimate: for all $0 \leq \alpha < 1$ there exists $C > 0$ such that

$$\forall 0 \leq t \leq T,\ \forall x \neq 0,\ \int_0^t \frac{1}{s^\alpha} r_0^x(t-s,x)\,ds \leq \frac{C}{t^\alpha}. \tag{27}$$

Successively using Inequalities (9) and (27) we obtain

$$\begin{aligned} |\partial_t v(t,x)| &\leq C\left(\|\mathcal{L}^+ f\|_1 + \|\mathcal{L} f\|_1\right) \int_0^t \frac{1}{\sqrt{s}} r_0^x(t-s)\,ds \\ &\leq \frac{C}{\sqrt{t}}\left(\|\mathcal{L}^+ f\|_1 + \|\mathcal{L} f\|_1\right). \end{aligned}$$

We now use the following well known estimate (see, e.g., Friedman [9]): for all $t > 0$, the probability density $q^{X^+}(x,t,y)$ of X_t^+ under $\mathbb{P}^x$ satisfies

$$\exists C > 0,\ \exists \nu > 0,\ \forall 0 < t \leq T,\ q^{X^+}(x,t,y) \leq \frac{C}{\sqrt{t}} \exp\left(-\tfrac{(y-x)^2}{\nu t}\right). \tag{28}$$

From Itô's formula and the preceding inequality we have

$$\sup_{x\in\mathbb{R}} |\partial_t \mathbb{E}^x f(X_t^+)| \leq \frac{C}{\sqrt{t}} \|\mathcal{L}^+ f\|_1.$$

In view of (23) we thus are in a position to obtain (22). □

Similar calculations lead to: There exists $C > 0$ such that, for all $t \in (0,T]$,

$$\sup_{x\neq 0} |\partial_{tt}^2 u(t,x)| \leq \frac{C}{\sqrt{t}} \|f'\|_{3,1}.$$

Proposition 2.5. *There exists $C > 0$ such that, for all $t \in (0,T]$,*

$$\sup_{x\neq 0} |\partial_x u(t,x)| \leq \frac{C}{\sqrt{t}} \|f'\|_{1,1}. \tag{29}$$

Proof. In view of (15) and the Gaussian estimate (28) we have

$$\|\partial_x \mathbb{E}^x f(X_t^+)\|_\infty \leq \frac{C}{\sqrt{t}} \|f'\|_1. \tag{30}$$

Therefore it suffices to prove

$$\sup_{x\neq 0} |\partial_x v(t,x)| \leq C\left(\|\mathcal{L}^+ f\|_1 + \|\mathcal{L} f\|_1\right). \tag{31}$$

In view of (25) this inequality results from Lemma 2.3 applied to the function

$$H(s) := \int_0^s \left(\mathbb{E}^0 \mathcal{L}^+ f(X_\xi^+) - \mathbb{E}^0 \mathcal{L} f(X_\xi) \right) d\xi,$$

noticing that, in view of (9), we may choose

$$C_H := C \left(\|\mathcal{L}^+ f\|_1 + \|\mathcal{L} f\|_1 \right).$$

□

We proceed similarly to get: For $\ell = 2, \cdots, 4$, there exists $C > 0$ such that, for all t in $(0, T]$,

$$\sup_{x \neq 0} |\partial_{x\ell}^{\ell} u(t, x)| \leq \frac{C}{\sqrt{t}} \|f'\|_{1,1}.$$

2.4 A Transformed Euler Scheme

Without loss of generality, we assume that $a(0+) - a(0-)$ is strictly positive. Using the symmetric local time $\tilde{L}$ as in [13], Eq. (6) writes

$$dX_t = \sigma(X_t)\, dB_t + \sigma(X_t)\sigma'_-(X_t)\, dt + \frac{a(0+) - a(0-)}{a(0+) + a(0-)} d\tilde{L}_t^0(X),$$

so that the hypotheses of Theorem 2.3 in [13] are well satisfied since

$$-1 < \tfrac{a(0+)-a(0-)}{a(0+)+a(0-)} < 1.$$

Therefore Girsanov's theorem implies that the stochastic differential equation (6) has a unique weak solution.

Set

$$\beta_+ := \tfrac{2a(0-)}{a(0+)+a(0-)} \text{ and } \beta_- := \tfrac{2a(0+)}{a(0+)+a(0-)}, \tag{32}$$

and

$$\begin{cases} \beta(x) := x \left(\beta_- \mathbb{I}_{x \leq 0} + \beta_+ \mathbb{I}_{x>0} \right), \\ \beta^{-1}(x) := \frac{x}{\beta_-} \mathbb{I}_{x \leq 0} + \frac{x}{\beta_+} \mathbb{I}_{x>0}. \end{cases} \tag{33}$$

Set also

$$\begin{cases} \tilde{\sigma}(x) & := \sigma \circ \beta^{-1}(x)\,(\beta_- \mathbb{I}_{x\le 0} + \beta_+ \mathbb{I}_{x>0})\,, \\ \tilde{b}(x) & := \sigma \circ \beta^{-1}(x)\sigma'_- \circ \beta^{-1}(x)\,(\beta_- \mathbb{I}_{x\le 0} + \beta_+ \mathbb{I}_{x>0})\,. \end{cases} \tag{34}$$

From Itô–Tanaka's formula (see, e.g., Revuz and Yor [21, Chap. VI]) to $\beta(X_t)$ we get that the process $Y := \beta(X)$ satisfies the SDE with discontinuous coefficients:

$$Y_t = \beta(X_0) + \int_0^t \tilde{\sigma}(Y_s)\, dB_s + \int_0^t \tilde{b}(Y_s)\, ds. \tag{35}$$

Now denote by h_n the step-size of the discretization, that is, $h_n := \frac{T}{n}$. For all $0 \le k \le n$ set $t_k^n := k\, h_n$. Let $(\overline{Y}_t^n)$ be the Euler approximation of (Y_t) defined by $\overline{Y}_0^n = \beta(X_0)$ and, for all $t_k^n \le t \le t_{k+1}^n$,

$$\overline{Y}_t^n = \overline{Y}_{t_k^n}^n + \tilde{\sigma}(\overline{Y}_{t_k^n}^n)\mathbb{I}_{\overline{Y}_{t_k^n}^n \neq 0}(B_t - B_{t_k^n}) + \tilde{b}(\overline{Y}_{t_k^n}^n)\mathbb{I}_{\overline{Y}_{t_k^n}^n \neq 0}(t - t_k^n). \tag{36}$$

The transformed Euler scheme for(X_t) is then defined by

$$\overline{X}_t^n = \beta^{-1}\left(\overline{Y}_t^n\right),\ 0 \le t \le T. \tag{37}$$

When the coefficient $a(x)$ is smooth, the weak convergence rate of the classical Euler scheme is of order $1/n$ and the discretization error can even be expanded in terms of powers of $1/n$: for a survey, see, e.g., Talay [24]. Our next theorem states that the discretization error of the transformed Euler scheme is of order $1/n^{1/2-\epsilon}$ for all $0 < \epsilon < \frac{1}{2}$ when the function f belongs to $\mathcal{W}^4$.

Theorem 2.6. *Under the hypotheses made on the function σ in Theorem 2.2-(ii), there exists a positive number C such that, for all initial condition f in $\mathcal{W}^4$, all parameter $0 < \epsilon < \frac{1}{2}$, all n large enough, and all x_0 in $\mathbb{R}$,*

$$\left|\mathbb{E}^{x_0} f(X_T) - \mathbb{E}^{x_0} f(\overline{X}_T^n)\right| \le C\,\|f'\|_{1,1} h_n^{(1-\epsilon)/2} + C\,\|f'\|_{1,1}\sqrt{h_n} + C\,\|f'\|_{3,1} h_n^{1-\epsilon}. \tag{38}$$

2.5 Convergence Rate Analysis: Proof of Theorem 2.6

For all $k \le n$ set

$$\theta_k^n := T - t_k^n.$$

The proof of Theorem 2.6 proceeds as follows. Since $u(0,x) = f(x)$ and $u(T,x) = \mathbb{E}^x f(X_T)$ for all x, the discretization error at time T can be decomposed as follows:

$$\begin{aligned}\epsilon_T^{x_0} &= \left|\mathbb{E}^{x_0} f \circ \beta^{-1}(Y_T) - \mathbb{E}^{x_0} f \circ \beta^{-1}\left(\overline{Y}_T^n\right)\right| \\ &= \left|\sum_{k=0}^{n-1}(\mathbb{E}^{x_0} u(T - t_k^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) - \mathbb{E}^{x_0} u(T - t_{k+1}^n, \beta^{-1}(\overline{Y}_{t_{k+1}^n}^n)))\right|,\end{aligned} \tag{39}$$

and thus

$$\begin{aligned}\epsilon_T^{x_0} \leq &\left|\sum_{k=0}^{n-2} \mathbb{E}^{x_0}\left\{u(\theta_k^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) - u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\right.\right. \\ &\left.\left. + u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) - u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_{k+1}^n}^n))\right\}\right| \\ &+ \left|\mathbb{E}^{x_0} u(\theta_1^n, \beta^{-1}(\overline{Y}_{t_{n-1}^n}^n)) - \mathbb{E}^{x_0} u(0, \beta^{-1}(\overline{Y}_T^n))\right|.\end{aligned} \tag{40}$$

One readily proves that

$$\left|\mathbb{E}^{x_0} u(\theta_1^n, \beta^{-1}(\overline{Y}_{t_{n-1}^n}^n)) - \mathbb{E}^{x_0} u(0, \beta^{-1}(\overline{Y}_T^n))\right| \leq C\|f'\|_{1,1}\sqrt{h_n}. \tag{41}$$

The rest of this section is devoted to the analysis of

$$\left|\sum_{k=0}^{n-2} \mathbb{E}^{x_0}(T_k - S_k)\right|,$$

where the time increment T_k is defined as

$$T_k := u(\theta_k^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) - u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) \tag{42}$$

and the space increment is defined as

$$S_k := u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_{k+1}^n}^n)) - u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)). \tag{43}$$

In all the calculation below, we use the following notation: given some real number $r(n)$ depending on n, and two positive numbers μ and ν,

$$r(n) = \mathcal{Q}_3\left(\frac{h_n^\nu}{(t_k^n)^\mu}\right) \text{ means } \exists C > 0,\ \forall n \geq 1,\ \forall 0 \leq k \leq n,\ |r(n)| \leq C\frac{h_n^\nu}{(t_k^n)^\mu}\|f'\|_{3,1}. \tag{44}$$

We distinguish two cases. On the one hand, when $\overline{Y}^n_{t^n_k}$ and $\overline{Y}^n_{t^n_{k+1}}$ are simultaneously positive or negative, we use a Taylor expansion of $u(t^n_{k+1},\cdot)$ around $(t^n_k, \overline{Y}^n_{t^n_k})$ and then apply accurate estimates of the derivatives of $u(t,x)$ for t in $(0,T]$ and $x \neq 0$. On the other hand, we combine two tricks: first, we prove that $\overline{Y}^n_{t^n_k}$ and $\overline{Y}^n_{t^n_{k+1}}$ have opposite signs with small probability when $\overline{Y}^n_{t^n_k}$ is large enough; second, when $\overline{Y}^n_{t^n_k}$ is small, we explicit the expansion of $u(t^n_{k+1},\cdot)$ around 0 and use Theorem 2.1; these two calculations allow us to cancel the lower order term in the expansion. We emphasize that using the transmission condition $(\star)$ is natural: it results from the construction of the approximation scheme by means of the function β^{-1} whose derivatives are discontinuous at 0.

In view of (12) one easily gets

$$\mathbb{E}^{x_0} T_k = \mathbb{E}^{x_0} \partial_t u(\theta^n_{k+1}, \beta^{-1}(\overline{Y}^n_{t^n_k}))h_n + \mathcal{Q}_3\left(\frac{h_n^2}{\sqrt{\theta^n_{k+1}}}\right). \tag{45}$$

Let S_k be defined as in (43). Set

$$\begin{aligned}
\triangle_{k+1} B &:= B_{t^n_{k+1}} - B_{t^n_k},\\
\triangle_{k+1} \overline{Y}^n &:= \tilde{\sigma}(\overline{Y}^n_{t^n_k})\triangle_{k+1} B + \tilde{b}(\overline{Y}^n_{t^n_k})h_n,\\
\triangle^\sharp_{k+1} \overline{X}^n &:= \sigma(\overline{X}^n_{t^n_k})\triangle_{k+1} B + \sigma\sigma'_-(\overline{X}^n_{t^n_k})h_n.
\end{aligned}$$

We emphasize that, due to the asymmetry of the definition β^{-1}, $\triangle^\sharp_{k+1}\overline{X}^n$ does not coincide with $\overline{X}^n_{t^n_{k+1}} - \overline{X}^n_{t^n_k}$ when $\overline{X}^n_{t^n_{k+1}}$ and $\overline{X}^n_{t^n_k}$ have opposite signs, which explains the two notations $\triangle$ and $\triangle^\sharp$. However the definitions (34) and (36) imply

$$\frac{\triangle_{k+1}\overline{Y}^n}{\beta_+}\mathbb{I}_{[\overline{Y}^n_{t^n_k}>0]} + \frac{\triangle_{k+1}\overline{Y}^n}{\beta_-}\mathbb{I}_{[\overline{Y}^n_{t^n_k}\leq 0]} = \triangle^\sharp_{k+1}\overline{X}^n. \tag{46}$$

We need to introduce the four following events:

$$\begin{cases}
\Omega^{++}_k &:= [\overline{Y}^n_{t^n_k} > 0 \text{ and } \overline{Y}^n_{t^n_{k+1}} > 0],\\
\Omega^{--}_k &:= [\overline{Y}^n_{t^n_k} \leq 0 \text{ and } \overline{Y}^n_{t^n_{k+1}} \leq 0],\\
\Omega^{+-}_k &:= [\overline{Y}^n_{t^n_k} > 0 \text{ and } \overline{Y}^n_{t^n_{k+1}} \leq 0],\\
\Omega^{-+}_k &:= [\overline{Y}^n_{t^n_k} \leq 0 \text{ and } \overline{Y}^n_{t^n_{k+1}} > 0].
\end{cases} \tag{47}$$

In view of the definition of the function β^{-1} in Sect. 2.4 we have

$$\text{On } \Omega^{++}_k,\ \beta^{-1}(\overline{Y}^n_{t^n_{k+1}}) = \frac{1}{\beta_+}\overline{Y}^n_{t^n_{k+1}} = \beta^{-1}(\overline{Y}^n_{t^n_k}) + \frac{1}{\beta_+}\triangle_{k+1}\overline{Y}^n.$$

Therefore

$$
\begin{aligned}
S_k \mathbb{I}_{\Omega_k^{++}} = & \frac{\triangle_{k+1}\overline{Y}^n}{\beta_+}\partial_x u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{++}} \\
& + \frac{1}{2}\frac{(\triangle_{k+1}\overline{Y}^n)^2}{(\beta_+)^2}\partial_{xx}^2 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{++}} \\
& + \frac{1}{6}\frac{(\triangle_{k+1}\overline{Y}^n)^3}{(\beta_+)^3}\partial_{x^3}^3 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{++}} \\
& + \frac{(\triangle_{k+1}\overline{Y}^n)^4}{(\beta_+)^4}\int_{[0,1]^4} \partial_{x^4}^4 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n) + \alpha_1\alpha_2\alpha_3\alpha_4 \triangle_{k+1}\overline{Y}^n) \\
& \qquad \alpha_1\alpha_2\alpha_3\, d\alpha_1 \dots d\alpha_4\, \mathbb{I}_{\Omega_k^{++}} \\
=: & \, S_k^{++1} + S_k^{++2} + S_k^{++3} + S_k^{++4}.
\end{aligned}
$$

A similar decomposition holds for $S_k \mathbb{I}_{\Omega_k^{--}}$.

We now use that $\Omega_k^{++} \cup \Omega_k^{--} = \Omega - (\Omega_k^{+-} \cup \Omega_k^{-+})$ and notice that $\Omega_k^{+-} \cup \Omega_k^{-+}$ belongs to the σ-field generated by (B_t) up to time t_{k+1}^n. In view of (46) we get

$$
\begin{aligned}
\mathbb{E}^{x_0}(S_k^{++1} + S_k^{--1}) = & \, \mathbb{E}^{x_0}\left[\sigma\sigma' \circ \beta^{-1}(\overline{Y}_{t_k^n}^n)\partial_x u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\right] h_n \\
& - \mathbb{E}^{x_0}\left[\triangle_{k+1}^{\sharp}\overline{X}^n\, \partial_x u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{+-} \cup \Omega_k^{-+}}\right].
\end{aligned}
$$

Proceeding similarly and making it explicit the conditional expectation of $(\triangle_{k+1}^{\sharp}\overline{X}^n)^2$ w.r.t. the past of (B_t) up to time t_k^n, we obtain

$$
\begin{aligned}
\mathbb{E}^{x_0}(S_k^{++2} + S_k^{--2}) = & \frac{1}{2}\mathbb{E}^{x_0}\left[a \circ \beta^{-1}(\overline{Y}_{t_k^n}^n)\partial_{xx}^2 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\right] h_n \\
& - \frac{1}{2}\mathbb{E}^{x_0}\left[(\triangle_{k+1}^{\sharp}\overline{X}^n)^2 \partial_{xx}^2 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{+-} \cup \Omega_k^{-+}}\right],
\end{aligned}
$$

and, since $\mathbb{E}^{x_0}(\triangle_{k+1}B)^3 = 0$,

$$
\begin{aligned}
\mathbb{E}^{x_0}(S_k^{++3} + S_k^{--3}) = & \frac{1}{2}\mathbb{E}^{x_0}\left[a \circ \beta^{-1}(\overline{Y}_{t_k^n}^n)\sigma\sigma' \circ \beta^{-1}(\overline{Y}_{t_k^n}^n)\partial_{x^3}^3 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\right] h_n^2 \\
& - \frac{1}{6}\mathbb{E}^{x_0}\left[(\triangle_{k+1}^{\sharp}\overline{X}^n)^3 \partial_{x^3}^3 u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n))\, \mathbb{I}_{\Omega_k^{+-} \cup \Omega_k^{-+}}\right] \\
& + \mathcal{Q}_3\left(\frac{h_n^2}{\sqrt{\theta_{k+1}^n}}\right).
\end{aligned}
$$

In addition, in view of Theorem 2.2 we have

$$\mathbb{E}^{x_0}|S_k^{++4} + S_k^{--4}| \leq \frac{C h_n^2}{\sqrt{\theta_{k+1}^n}} \|f'\|_{3,1}.$$

To summarize the calculations of this subsection, we have obtained

$$\mathbb{E}^{x_0} S_k =: \mathbb{E}^{x_0} \mathcal{L}u(\theta_{k+1}^n, \beta^{-1}(\overline{Y}_{t_k^n}^n)) h_n + \mathbb{E}^{x_0} \mathcal{R}_k + \mathcal{Q}_3\left(\frac{h_n^2}{\sqrt{\theta_{k+1}^n}}\right). \tag{48}$$

We now estimate the remaining term $\mathbb{E}^{x_0}\mathcal{R}_k$.

2.6 *Estimate for $\mathbb{E}^{x_0}\mathcal{R}_k$: Localization Around 0*

Arbitrarily fix $0 < \epsilon < \frac{1}{2}$. We aim to show

$$\begin{aligned} |\mathbb{E}^{x_0}\mathcal{R}_k| \leq & \frac{C h_n^{1-2\epsilon}}{\sqrt{\theta_k^n}} \|f'\|_{1,1} \mathbb{P}^{x_0}\left[|\overline{Y}_{t_{k+1}^n}^n| \leq h_n^{1/2-\epsilon}\right] \\ & + \frac{C h_n^{3/2(1-\epsilon)}}{\sqrt{\theta_{k+1}^n}} \|f'\|_{3,1} \mathbb{P}^{x_0}\left[|\overline{Y}_{t_k^n}^n| \leq h_n^{1/2-\epsilon}\right]. \end{aligned} \tag{49}$$

To get this precise estimate we need to use the transmission condition ($\star$) in Eq. (3). This explains that we localize on the event where $\overline{Y}_{t_k^n}^n$ is close to 0. We start with checking that we may neglect the complementary event.

Define $\Gamma(y)$ by

$$\Gamma(y) := \frac{-y - \tilde{b}(y)h_n}{\tilde{\sigma}(y)}.$$

Observe that

$$\begin{aligned} \Omega_k^{+-} = & \left[0 < \overline{Y}_{t_k^n}^n \leq h_n^{1/2-\epsilon} \text{ and } \overline{Y}_{t_{k+1}^n}^n \leq -h_n^{1/2-\epsilon}\right] \cup \left[0 < \overline{Y}_{t_k^n}^n \leq h_n^{1/2-\epsilon} \text{ and } \right. \\ & \left. -h_n^{1/2-\epsilon} \leq \overline{Y}_{t_{k+1}^n}^n \leq 0\right] \cup \left[\overline{Y}_{t_k^n}^n \geq h_n^{1/2-\epsilon} \text{ and } \triangle_{k+1}B \leq \Gamma(\overline{Y}_{t_k^n}^n)\right]. \end{aligned}$$

Notice that

$$\mathbb{P}^{x_0}\left[\overline{Y}_{t_k^n}^n \geq h_n^{1/2-\epsilon} \text{ and } \triangle_{k+1}B \leq \Gamma(\overline{Y}_{t_k^n}^n)\right] \leq C \exp(-\tfrac{1}{C} n^{-\epsilon}),$$

and, similarly,

$$\mathbb{P}^{x_0}\left[0 \le \overline{Y}^n_{t^n_k} \le h_n^{1/2-\epsilon} \text{ and } \overline{Y}^n_{t^n_{k+1}} \le -h_n^{1/2-\epsilon}\right] \le C\,\exp(-\tfrac{1}{C}n^{-\epsilon}).$$

We proceed analogously on the event Ω_k^{-+}. This leads us to limit ourselves to consider the events

$$\Omega_k^{+-*} := \left[0 < \overline{Y}^n_{t^n_k} \le h_n^{1/2-\epsilon} \text{ and } -h_n^{1/2-\epsilon} \le \overline{Y}^n_{t^n_{k+1}} \le 0\right]$$

and

$$\Omega_k^{-+*} := \left[-h_n^{1/2-\epsilon} \le \overline{Y}^n_{t^n_k} \le 0 \text{ and } 0 \le \overline{Y}^n_{t^n_{k+1}} \le h_n^{1/2-\epsilon}\right].$$

Notice that, on these events, equality (46) implies that $|\triangle^\sharp_{k+1}\overline{X}^n| \le Ch_n^{1/2-\epsilon}$. Therefore, in view of the estimates (12) one has

$$\begin{aligned}&\left|\mathbb{E}^{x_0}\left[(\triangle^\sharp_{k+1}\overline{X}^n)^2\partial^2_{xx}u(\theta^n_{k+1},\beta^{-1}(\overline{Y}^n_{t^n_k}))\,\mathbb{I}_{\Omega_k^{+-*}\cup\Omega_k^{-+*}}\right]\right| \\ &\quad\le \frac{Ch_n^{1-2\epsilon}}{\sqrt{\theta^n_{k+1}}}\|f'\|_{1,1}\mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_k}| \le h_n^{1/2-\epsilon}\right],\end{aligned}$$

and

$$\begin{aligned}&\left|\mathbb{E}^{x_0}\left[(\triangle^\sharp_{k+1}\overline{X}^n)^3\partial^3_{x^3}u(\theta^n_{k+1},\beta^{-1}(\overline{Y}^n_{t^n_k}))\,\mathbb{I}_{\Omega_k^{+-*}\cup\Omega_k^{-+*}}\right]\right| \\ &\quad\le \frac{Ch_n^{3/2(1-2\epsilon)}}{\sqrt{\theta^n_{k+1}}}\|f'\|_{3,1}\mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_k}| \le h_n^{1/2-\epsilon}\right].\end{aligned}$$

Therefore, to show (49) it suffices to show

$$\begin{aligned}&\left|\mathbb{E}^{x_0}\left[(S_k - \triangle^\sharp_{k+1}\overline{X}^n\partial_x u(\theta^n_{k+1},\beta^{-1}(\overline{Y}^n_{t^n_k})))\,\mathbb{I}_{\Omega_k^{+-*}\cup\Omega_k^{-+*}}\right]\right| \\ &\quad\le \frac{Ch_n^{1-2\epsilon}}{\sqrt{\theta^n_{k+1}}}\|f'\|_{1,1}\mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_k}| \le h_n^{1/2-\epsilon}\right].\end{aligned} \tag{50}$$

2.7 *Proof of (50): Expansion Around 0*

On the event Ω_k^{+-*} we have that $\overline{Y}^n_{t^n_{k+1}}$ and $\overline{Y}^n_{t^n_k}$ are close to 0. On this event, we also have that $\overline{Y}^n_{t^n_{k+1}}$ is negative and $\overline{Y}^n_{t^n_k}$ is positive, so that $\beta^{-1}(\overline{Y}^n_{t^n_{k+1}}) = \frac{1}{\beta_-}\overline{Y}^n_{t^n_{k+1}}$ and $\beta^{-1}(\overline{Y}^n_{t^n_k}) = \frac{1}{\beta_+}\overline{Y}^n_{t^n_k}$. As $u(t,x)$ is continuous at point 0, we get

$$
\begin{aligned}
&\mathbb{E}^{x_0}\left[(S_k - \triangle^{\sharp}_{k+1}\overline{X}^n \partial_x u(\theta^n_{k+1}, \beta^{-1}(\overline{Y}^n_{t^n_k}))) \, \mathbb{I}_{\Omega_k^{+-*}}\right] \\
&= \frac{1}{\beta_-}\mathbb{E}^{x_0}\left[\overline{Y}^n_{t^n_{k+1}} \partial_x u(\theta^n_{k+1}, 0-) \, \mathbb{I}_{\Omega_k^{+-*}}\right] - \frac{1}{\beta_+}\mathbb{E}^{x_0}\left[\overline{Y}^n_{t^n_k} \partial_x u(\theta^n_{k+1}, 0+) \, \mathbb{I}_{\Omega_k^{+-*}}\right] \\
&\quad - \mathbb{E}^{x_0}\left[\triangle^{\sharp}_{k+1}\overline{X}^n \partial_x u(\theta^n_{k+1}, 0+) \, \mathbb{I}_{\Omega_k^{+-*}}\right] \\
&\quad + \mathbb{E}^{x_0}\left[\left((\beta^{-1}(\overline{Y}^n_{t^n_{k+1}}))^2 \int_{[0,1]^2} \partial^2_{xx} u(\theta^n_{k+1}, \alpha_1\alpha_2\beta^{-1}(\overline{Y}^n_{t^n_{k+1}}))\alpha_1 \, d\alpha_1 d\alpha_2 \right.\right. \\
&\qquad -(\beta^{-1}(\overline{Y}^n_{t^n_k}))^2 \int_{[0,1]^2} \partial^2_{xx} u(\theta^n_{k+1}, \alpha_1\alpha_2\beta^{-1}(\overline{Y}^n_{t^n_k}))\alpha_1 \, d\alpha_1 d\alpha_2 \\
&\qquad \left.\left. -\triangle^{\sharp}_{k+1}\overline{X}^n \, \beta^{-1}(\overline{Y}^n_{t^n_k}) \int_0^1 \partial^2_{xx} u(\theta^n_{k+1}, \alpha_1\beta^{-1}(\overline{Y}^n_{t^n_k})) \, d\alpha_1\right) \mathbb{I}_{\Omega_k^{+-*}}\right].
\end{aligned}
$$

The absolute value of the last expectation in the right-hand side can be bounded from above by

$$
\frac{Ch_n^{1-2\epsilon}}{\sqrt{\theta^n_{k+1}}} \| f' \|_{1,1} \mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_k}| \le h_n^{1/2-\epsilon}\right]
$$

since

$$
\text{on } \Omega_k^{+-*}, \ |\beta^{-1}(\overline{Y}^n_{t^n_{k+1}})| + |\beta^{-1}(\overline{Y}^n_{t^n_k})| \le Ch_n^{1/2-\epsilon}.
$$

In addition, in view of (46) the sum of the three first terms in the right-hand side reduces to

$$
\mathbb{E}^{x_0}\left[\overline{Y}^n_{t^n_{k+1}} \, \mathbb{I}_{\Omega_k^{+-*}}\left(\frac{1}{\beta_-}\partial_x u(\theta^n_{k+1}, 0-) - \frac{1}{\beta_+}\partial_x u(\theta^n_{k+1}, 0+)\right)\right],
$$

so that now are in a position to use the transmission condition ($\star$) in Eq. (3). Remembering the definition (32) of β_+ and β_- we deduce that the preceding expression is null. We may proceed similarly as above on the event Ω_k^{-+*}. We thus have proven (50), which ends the proof of (49).

2.8 Summing Up

Gather the expansions (45) and (48). One can easily prove that the law of $\overline{Y}^n_{t_k}$ has a density for all k, from which

$$
\mathbb{E}^{x_0}\partial_t u\left(T - t^n_k, \beta^{-1}\left(\overline{Y}^n_{t_k}\right)\right) - \mathbb{E}^{x_0}\mathcal{L}u\left(T - t^n_k, \beta^{-1}\left(\overline{Y}^n_{t_k}\right)\right) = 0. \qquad (51)
$$

Use the equality (51) and the inequalities (49), (41). It follows:

$$\begin{aligned}\epsilon_T^{x_0} \le & \sum_{k=0}^{n-2} \frac{Ch_n^{1-2\epsilon}}{\sqrt{\theta_k^n}} \|f'\|_{1,1} \mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_{k+1}}| \le h_n^{1/2-\epsilon}\right] \\ & + \sum_{k=0}^{n-2} \frac{Ch_n^{3/2(1-2\epsilon)}}{\sqrt{\theta^n_{k+1}}} \|f'\|_{3,1} \mathbb{P}^{x_0}\left[|\overline{Y}^n_{t^n_k}| \le h_n^{1/2-\epsilon}\right] \\ & + C\|f'\|_{1,1}\sqrt{h_n} + C\|f'\|_{3,1} h_n.\end{aligned}$$

To deduce (38) it now remains to apply Theorem 2.7 below to the Itô process $(\overline{Y}^n_t)$.

2.9 *Estimate for the Number of Visits of Small Balls by the Euler Scheme*

In this subsection we recall a result from Bernardin et al. [3] which was essential to estimate the remaining terms in the above error expansion. This estimate is useful to analyze convergence rates of discretization schemes for SDEs with irregular coefficients.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ be a filtered probability space satisfying the usual conditions. Let (W_t) be a m-dimensional standard Brownian motion on this space. Given two progressively measurable processes (b_t) and (σ_t) taking values respectively in $\mathbb{R}^d$ and in the space of real $d \times m$ matrices, Z_t is the $\mathbb{R}^d$ valued Itô process

$$Z_t = Z_0 + \int_0^t b_s\, ds + \int_0^t \sigma_s\, dW_s. \tag{52}$$

Suppose: There exists a positive number $K \ge 1$ such that, $\mathbb{P}$-a.s.,

$$\forall t \ge 0,\ \|b_t\| \le K,$$

and

$$\forall 0 \le s \le t,\ \frac{1}{K^2}\int_s^t \psi(s)\, ds \le \int_s^t \psi(s)\|\sigma_s \sigma_s^*\|\, ds \le K^2 \int_s^t \psi(s)\, ds$$

for all positive locally integrable map $\psi : \mathbb{R}_+ \to \mathbb{R}_+$.

Theorem 2.7. *Let (Z_t) be as in (52). Let f be a positive and increasing function in $\mathcal{C}^1([0,T);\mathbb{R}_+)$ such that f^α is integrable on $[0,T)$ for all $1 \le \alpha < 2$. Assume also that there exists $1 < \nu < 1 + \eta$, where $\eta := \frac{1}{4K^4}$, such that*

$$\int_0^T f^{2\nu-1}(s)\, f'(s) \frac{(T-s)^{1+\eta}}{s^\eta}\, ds < +\infty.$$

Then there exists $C > 0$, depending only on ν, K and T, such that, for all $\xi \in \mathbb{R}^d$ and $0 < \varepsilon < 1/2$, there exists $h_0 > 0$ satisfying

$$\forall\, h \le h_0,\ h \sum_{k=0}^{N_h} f(kh)\mathbb{P}(\|Z_{ph} - \xi\| \le h^{1/2-\varepsilon}) \le C h^{1/2-\varepsilon}, \tag{53}$$

where $N_h := \lfloor T/h \rfloor - 1$.

2.10 Extensions

One can relax the condition that the functions f and $\mathcal{L}f$ satisfy the transmission conditions in the definition (10) of $\mathcal{W}^4$.

Theorem 2.8. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be in the space*

$$\mathcal{W} := \left\{ g \in \mathcal{C}_b^4(\mathbb{R} - \{0\}),\ g^{(i)} \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}) \text{ for } i = 1, \ldots, 4, \right\}. \tag{54}$$

Under the hypotheses on the function σ made in Theorem 2.2-(ii), there exists a positive number C (depending on f) such that, for all $0 < \epsilon < \frac{1}{2}$, all n large enough, and all x_0 in $\mathbb{R}$,

$$\left| \mathbb{E}^{x_0} f(X_T) - \mathbb{E}^{x_0} f(\overline{X}_T^n) \right| \le C h_n^{1/2-\epsilon}. \tag{55}$$

In the case where $a(x)$ has a finite number of discontinuities, one can split the real line into intervals whose boundary points are the discontinuity points of $a(x)$ and introduce transmission conditions at each of these points. One can also construct an explicit transformation β removing the local time of (X_t) at these discontinuity points. Thus one can readily extend our transformed Euler Scheme. The above convergence rate estimates still hold true.

Now consider the equation

$$\begin{cases} \partial_t v(t,x) - \mathcal{L}v(t,x) - b(x)\dfrac{\partial}{\partial x} v(t,x) = 0 \text{ for all } (t,x) \in (0,T] \times \mathbb{R}, \\ v(0,x) = f(x) \text{ for all } x \in \mathbb{R}, \\ \text{Compatibility transmission conditions at the discontinuity points of } a(x). \end{cases} \tag{56}$$

If the bounded function b is smooth enough (e.g. b is in $\mathcal{C}_b^6(\mathbb{R})$), one can represent the solution of (56) by means of a SDE similar to (6) except that the drift term involves $b(X_t)$, a new modified Euler scheme can easily be constructed, and all our results remain true.

The multi-dimensional setting requires non trivial additional works to define the suitable stochastic differential equation with weighted local time which provides a representation similar to (5), to derive accurate pointwise estimates on the derivatives of the solution to the PDE, to construct a discretization scheme which may easily be simulated, and to estimate its convergence rate. See [17].

3 The Linear 3D Poisson–Boltzmann PDE in Molecular Dynamics

The results in this section come from Bossy et al. [4].

The Poisson–Boltzmann (PB) PDE in Molecular Dynamics describes the electrostatic potential around a biomolecular assembly, and is used to compute global characteristics of the system such as the solvatation free energy and the electrostatic forces exerted by the solvent on the molecule.

The implicit solvent equation, which means that the solvent is considered as a continuum, reads

$$\mathcal{L}u(x) := -\nabla \cdot (\varepsilon(x)\nabla u(x)) + \kappa^2(x)u(x) = f(x), \qquad x \in \mathbb{R}^3,$$

where $\varepsilon(x)$ is the permittivity of the medium and $\kappa^2(x)$ is called the ion accessibility parameter. The singular source term f is defined as

$$f := \sum_{i=1}^{N} q_i \delta_{x_i},$$

the atomic structure of the molecule being modelled as N atoms at positions $x_1, \ldots, x_N$ with charges q_i.

Difficulties arising from the singularity of f can be removed as follows. Let χ be a $\mathcal{C}^\infty$ function with compact support in Ω_{int} such that $\chi(x) = 1$ in the neighborhood of the points $\{x_1, \ldots, x_N\}$. Consider the function $u_0 := \sum_i G_i$ where

$$G_i(x) := \frac{1}{4\pi} \frac{q_i}{\epsilon_{\text{int}}} \frac{1}{|x - x_i|}, \qquad x \in \mathbb{R}^3.$$

Notice that the functions G_i satisfy

$$-\nabla \cdot (\epsilon_{\text{int}} \nabla G_i) = q_i \delta_i, \qquad x \in \mathbb{R}^3.$$

Define the $\mathcal{C}^\infty$ function g with compact support in Ω_{int} as

$$g(x) = \epsilon_{\text{int}} \left(u_0(x) \Delta \chi(x) + \nabla u_0(x) \cdot \nabla \chi(x) \right), \qquad x \in \mathbb{R}^3. \tag{57}$$

Then, the function $v := u - \chi G$ solves the PB equation with regularized source term g' (RPBE in the terminology of Chen et al. [6] where this transformation appears)

$$-\nabla \cdot (\epsilon(x) \nabla v(x)) + \kappa^2(x) v(x) = g(x), \qquad x \in \mathbb{R}^3. \tag{58}$$

Therefore the singularity of the source term is not an issue. However it remains to face the discontinuities of the function κ and the fact that the Poisson–Boltzmann PDE involves a divergence form operator with discontinuous coefficient ε.

Assume that Γ is a smooth (C^3) manifold in $\mathbb{R}^d$. We denote by $\pi(x)$ the orthogonal projection of x on Γ, $n(y)$ the outward normal to Γ for $y \in \Gamma$, and $\rho(x)$ the signed distance between x and Γ, that is, $\rho(x) := (x - \pi(x)) \cdot n(\pi(x))$.

We say that $(\mathbb{P}^x)_{x \in \mathbb{R}^d}$ on $(\mathcal{C}, \mathcal{B})$ solves the martingale problem (MP) for $\mathcal{L}$ if

$$\mathbb{P}^x \{ w \in \mathcal{C} : w(0) = x \} = 1,$$

for all $x \in \mathbb{R}^d$, and

$$M_t^\varphi(w) := \varphi(w(t)) - \varphi(w(0)) - \int_0^t \mathcal{L}\varphi(w(s))\, ds \quad \text{is a } \mathbb{P}^x \text{ martingale,}$$

for all φ satisfying

$$\varphi \in C_b^0(\mathbb{R}^d) \cap C_b^2(\mathbb{R}^d \setminus \Gamma), \quad \varepsilon \nabla \varphi \cdot (n \circ \pi) \in C_b^0(\mathcal{N}).$$

Notice that the test functions satisfy the transmission condition

$$\varepsilon_{\text{int}} \nabla^{\text{int}} \varphi(x) \cdot n(x) = \varepsilon_{\text{ext}} \nabla^{\text{ext}} \varphi(x) \cdot n(x).$$

The following theorem bases the probabilistic interpretation of the linear and non-linear Poisson–Boltzmann equations. The technical difficulties of its proof come from the fact that the dynamics of the unknown process (X_t) depends on the local time of the auxiliary process $(\rho(X_t))$.

Theorem 3.1. *For all x the above martingale problem has a unique solution $\mathbb{P}$ which is the unique weak solution to the following SDE with weighted local time:*

$$\begin{cases} X_t &= x + \int_0^t \sqrt{2\epsilon(X_s)}\, dB_s + \dfrac{\epsilon_{ext} - \epsilon_{int}}{2\epsilon_{ext}} \int_0^t \mathbf{n}(X_s)\, dL_s^0(D), \\ D_t &= \rho(X_t), \end{cases} \tag{59}$$

where B is a d-dimensional Brownian motion and $(L_t^0(D))_{t\geqslant 0}$ is the right-sided local time at 0 of the continuous semi-martingale (D_t).

Now, to prove that the solution to the martingale problem (MP) provides a stochastic representation of the solution to the Poisson–Boltzmann equation, a key step consists in proving the next lemma.

Lemma 3.2 (Generalized Itô–Meyer Formula). *If X is a continuous semi-martingale, $Y := \rho(X)$, and if ϕ is a test function for the martingale problem for $\mathcal{L}$, then*

$$\phi(X_t) = \phi(X_0) + \int_0^t \nabla^{int}\phi(X_s)\cdot dX_s + \frac{1}{2}\sum_{i,j=1}^{3}\int_0^t \frac{\partial^2 u}{\partial x_i\,\partial x_j}(X_s) d\langle X^i, X^j\rangle_s + \frac{1}{2}\int_0^t h(X_s)\, dL_s^0(Y), \quad \forall t \geq 0 \ a.s.,$$

where $h(x) := \left(\dfrac{\varepsilon_{int}}{\varepsilon_{ext}} - 1\right) \nabla^{int}\phi(\pi(x))\cdot n(\pi(x))$.

The preceding formula would be easily obtained from Itô's and Itô–Tanaka's formulas if the functions $\phi(x) - g(x)[\rho(x)]_+$ and $g(x)$ were $\mathcal{C}^2$.

Theorem 3.3 (First Feynman–Kac Representation). *Let v be the solution of $-\nabla\cdot(\varepsilon\nabla v) + \kappa^2 v = g$, where g is a smooth function. Then, for all $x \in \mathbb{R}^3$,*

$$v(x) = \mathbb{E}^x\left[\int_0^{+\infty} g(X_t)\exp\left(-\int_0^t \kappa^2(X_s)\, ds\right) dt\right].$$

This representation does not allow to develop an efficient numerical scheme: first, one needs to precisely discretize X everywhere where g is nonzero; second, generally the computation of g is costly. Thus the next representation is more favorable to the derivation of a simulation algorithm because it only involves the entrance time and entrance position in small neighborhoods of Γ of (scaled) Brownian paths.

Fix $h > 0$ and define the following sequence of stopping times

$$\tau_k = \inf\{t \geq \tau'_{k-1} : \rho(X_t) = -h\}$$
$$\tau'_k = \inf\{t \geq \tau_k : X_t \in \Gamma\}.$$

Since $\Delta(u-G)=0$ in Ω_i, for all x such that $\rho(x)\le -h$,

$$u(x)=\mathbb{E}^x[u(X_{\tau_1'})-G(X_{\tau_1'})]+G(x).$$

For all $x\in\Omega_{\rm ext}$,

$$u(x)=\mathbb{E}^x\left[u(X_{\tau_1})\exp\left(-\int_0^{\tau_1}\kappa^2(X_t)\,dt\right)\right].$$

Recursively applying the two preceding formulas leads to the following result.

Theorem 3.4. *One has*

$$u(x)=\mathbb{E}^x\left[\sum_{k=1}^{+\infty}\left(G(X_{\tau_k})-G(X_{\tau_k'})\right)\exp\Big(-\int_0^{\tau_k}\kappa^2(X_t)\,dt\Big)\right].$$

This new representation allows one to justify Walk on Spheres algorithms introduced in this context by Mascagni and Simonov [16], construct improved versions of these algorithms, and analyze the convergence rates of all these methods.

4 The Semi-linear 3D Poisson–Boltzmann PDE in Molecular Dynamics

The results of this section come from Champagnat et al. [5].

The semi-linear Poisson–Boltzmann equation reads

$$-\nabla\cdot(\varepsilon(x)\nabla v(x))+\kappa^2(x)\sinh(v(x))=f(x),\qquad x\in\mathbb{R}^3. \tag{60}$$

The semi-linear structure of this PDE leads to interpret it in terms of Backward Stochastic Differential Equations. When the differential operator in the PDE has smooth coefficients and the zero order term satisfies suitable strict monotonicity conditions, the theory is well developed: see e.g. Pardoux [18] for a survey. Here the context requires new arguments to face the discontinuity of ε and the fact that κ is null in $\Omega_{\rm ext}$.

As in the preceding section, consider the Poisson–Boltzmann equation with regularized source term

$$-\nabla\cdot(\epsilon(x)\nabla v(x))+\kappa^2(x)\sinh(v(x))=g(x),\qquad x\in\mathbb{R}^3. \tag{61}$$

Consider the Backward Stochastic Differential Equation

$$\forall T>0,\ \forall 0\le t\le T,\qquad Y_t^x=Y_T^x+\int_t^T(g(X_s^x)-\kappa^2(X_s^x)\ \sinh(Y_s^x))\,ds-\int_t^T Z_s^x\,dB_s.$$

Consider the following subspace M of the Sobolev space $H^1(\mathbb{R}^3)$:

$$M := \left\{ v \in H^1(\mathbb{R}^3) | \cosh^2 v - 1 \in L^1(\mathbb{R}^3) \right\}.$$

Observe that

$$\left\{ v \in H^1(\mathbb{R}^3) | \cosh^2(v) - 1 \in L^1(\mathbb{R}^3) \right\} = \left\{ v \in H^1(\mathbb{R}^3) | \sinh(v) \in L^2(\mathbb{R}^3) \right\}$$
$$\subset \left\{ v \in H^1(\mathbb{R}^3) | \sinh(v/2) \in L^2(\mathbb{R}^3) \right\} = \left\{ v \in H^1(\mathbb{R}^3) | \cosh(v) - 1 \in L^1(\mathbb{R}^3) \right\}.$$

Definition 4.1. Suppose that Ω_{int} is a bounded connex subdomain of $\mathbb{R}^d$ with boundary Γ of class $\mathcal{C}^\infty$.A weak solution to the RPBE (61) is a map v belonging to M such that

$$\mathcal{E}_0(v, \phi) + \int \kappa^2(y) \sinh(v(y))\phi(y)\, dy - \int g(y)\phi(y)\, dy = 0, \qquad \phi \in H^1(\mathbb{R}^3), \tag{62}$$

where $\mathcal{E}_0(v, \phi) := (\epsilon \nabla v, \nabla \phi) = \int \epsilon(y) \nabla v(y) \nabla \phi(y)\, dy$ and where g is defined as in (57).

Theorem 4.2. *The RPBE (61) has a unique weak solution v in the sense of Definition 4.1. This solution belongs to $\mathcal{C}_b^0(\mathbb{R}^3) \cap \mathcal{C}^2(\mathbb{R}^3 \backslash \Gamma)$ and its trace $v_{|\Gamma}$ belongs to $\mathcal{C}^3(\Gamma)$.*

In addition, there exists a function $r(x)$ on $\mathcal{C}^2(\mathbb{R}^3)$ such that

$$r(x) = \left(\frac{\epsilon_{\text{int}}}{\epsilon_{\text{ext}}} - 1 \right) \nabla^{\text{int}} v(\pi(x)) \cdot \mathbf{n}(\pi(x)), \qquad x \in \mathcal{N}, \tag{63}$$

and such that the map

$$\hat{v}(x) := v(x) - r(x)[\rho(x)]_+, \qquad x \in \mathbb{R}^3 \tag{64}$$

is in $\mathcal{C}^2(\mathbb{R}^3 \backslash \Gamma) \cap W^{2,\infty}_{\text{loc}}(\mathbb{R}^3)$. Finally, the gradient ∇v belongs to $L^\infty(\mathbb{R}^3)$.

On a filtered probability space equipped with a Brownian motion (B_t) where a weak solution (X_t) to the SDE (59) has been constructed, let τ be a $(\mathcal{F}_t)_{t \geqslant 0}$ stopping time. We allow τ to take infinite values. Let ξ be a $\mathcal{F}_\tau$ measurable random variable and f a progressively measurable map from $\Omega \times \mathbb{R}^+ \times \mathbb{R}^k \times \mathbb{R}^{k \times d}$ to $\mathbb{R}^k$.

The map $\bar{f}(s, y, z) := f(X_s, y, z)$ is assumed to satisfy the following conditions. Almost surely, for all (t, z) in $[0, T] \times \mathbb{R}^{k \times d}$ the map $y \in \mathbb{R}^k \to \bar{f}(t, y, z)$ is continuous. There exists a progressively measurable bounded process (K_t) such that, a.s., for all (t, z, z') in $[0, T] \times \mathbb{R}^{k \times d} \times \mathbb{R}^{k \times d}$, for all y in $\mathbb{R}^d$,

$$|\bar{f}(t, y, z) - \bar{f}(t, y, z')| \leqslant K(t) \|z - z'\|.$$

There exists a progressively measurable bounded process (μ_t) such that, a.s., for all (t, y, y', z) in $[0, T] \times \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^{k\times d}$,

$$\langle y - y', \bar{f}(t, y, z) - \bar{f}(t, y', z)\rangle \leqslant \mu(t)|y - y'|^2.$$

There exists a progressively measurable process λ satisfying

$$\forall t > 0,\ \lambda(t) - 2\mu(t) - K^2(t) > \bar{\lambda} > 0$$

such that

$$\mathbb{E}\int_0^\tau e^{\int_0^t \lambda(s)ds}|\bar{f}(t, 0, 0)|^2\, dt < \infty.$$

For all real number $r > 0$ and all integer $n > 0$, one has

$$\sup_{|y|\leqslant r} |\bar{f}(t, y, 0) - \bar{f}(t, 0, 0)| \in L^1(]0, n[\times\Omega, dt \otimes \mathbb{P}).$$

The random variable ξ is supposed to satisfy

$$\mathbb{E}e^{\int_0^\tau \lambda(s)ds}|\xi|^2 < \infty$$

and

$$\mathbb{E}\int_0^\tau e^{\int_0^t \lambda(s)ds}|\bar{f}(t, e^{-1/2\int_0^t \tilde{\lambda}(s)ds}\bar{\xi}_t, e^{-1/2\int_0^t \tilde{\lambda}(s)ds}\bar{\eta}_t)|^2\, dt < \infty,$$

where we have set $\tilde{\lambda}(t) := 2\mu(t) - K^2(t)$, $\bar{\xi} := e^{\int_0^\tau \tilde{\lambda(s)}\, ds}\xi$, $\bar{\xi}_t := \mathbb{E}[\bar{\xi}|\mathcal{F}_t]$ and $\bar{\eta}$ is a predictable process satisfying

$$\bar{\xi} = \mathbb{E}\bar{\xi} + \int_0^\infty \bar{\eta}_t\, dB_t,$$
$$\mathbb{E}\int_0^\infty |\bar{\eta}_t|^2\, dt < \infty.$$

We now are in a position to exhibit our stochastic interpretation of the non-linear Poisson–Boltzmann equation.

Theorem 4.3. *There exists a unique progressively measurable process (Y_t, Z_t) such that*

$$\forall T > 0,\ \ Y_{t\wedge\tau} = Y_{T\wedge\tau} + \int_{t\wedge\tau}^{T\wedge\tau} f(X_s, Y_s, Z_s)\, ds - \int_{t\wedge\tau}^{T\wedge\tau} Z_s dB_s,\ 0 \leqslant t \leqslant T, \tag{65}$$

and satisfying

$$\mathbb{E}\left[\sup_{0\leqslant t} e^{\int_0^t \lambda(X_s)ds}|Y_t|^2 + \int_0^\infty e^{\int_0^t \lambda(X_s)ds}\left(|Y_t|^2 + |Z_t|^2\right)\,dt\right] \leqslant C\,\mathbb{E}\left[\int_0^\infty e^{\int_0^\lambda (X_s)ds}|f(X_t,0,0)|^2 dt\right].$$

Let $u(x)$ be the unique weak solution u in the sense of Definition 4.1 of the Poisson–Boltzmann equation (60). It admits the following probabilistic representation:

$$u(x) = \chi(x)u_0(x) + Y_0, \quad x \in \mathbb{R}^3,$$

where (Y, Z) solves (65).

References

1. Bass, R.F.: Diffusions and Elliptic Operators. Springer, New York (1998)
2. Bass, R.F., Chen, Z.Q.: Stochastic differential equations for Dirichlet processes. Probab. Theory Relat. Fields **121**(3), 422–446 (2001)
3. Bernardin, F., Bossy, M., Martinez, M., Talay, D.: On mean discounted numbers of passage times in small balls of Itô processes observed at discrete times. Electron. Commun. Probab. **14**, 302–316 (2009)
4. Bossy, M., Champagnat, N., Maire, S., Talay, D.: Probabilistic interpretation and random walk on spheres algorithms for the Poisson-Boltzmann equation in molecular dynamics. ESAIM:M2AN Math. Model. Numer. Anal. **44**(5), 997–1048 (2010)
5. Champagnat, N., Perrin, N., Talay, D.: (in preparation)
6. Chen, L., Holst, M.J., Xu, J.: The finite element approximation of the nonlinear Poisson-Boltzmann equation. SIAM J. Numer. Anal. **45**(6), 2298–2320 (2007)
7. Étoré, P.: On random walk simulation of one-dimensional diffusion processes with discontinuous coefficients. Electron. J. Probab. **11**(9), 249–275 (2006)
8. Étoré, P., Lejay, A.: A Donsker theorem to simulate one-dimensional processes with measurable coefficients. ESAIM Probab. Stat. **11**(9), 301–326 (2007)
9. Friedman, A.: Stochastic Differential Equations and Applications. Dover, Mineola (2006)
10. Fukushima, M., Oshima, Y., Takeda, M.: Dirichlet forms and symmetric Markov processes. In: de Gruyter Studies in Mathematics, vol. 19. (Walter) de Gruyter, Berlin (2011)
11. Graham, C., Talay, D.: Stochastic simulation and Monte Carlo methods, mathematical foundations of stochastic simulation. In: Stochastic Modelling and Applied Probability, vol. 68. Springer, Heildeberg (2013)
12. Ladyzenskaya, O.A., Solonnikov, V.A., Uralćeva, N.N.: Linear and quasi-linear equations of parabolic type. In: Translations of Mathematical Monographs, vol. 23. American Mathematical Society, Providence (1967)
13. Le Gall, J.-F.: One-dimensional stochastic differential equations involving the local times of the unknown process. In: Proceedings stochastic analysis and applications (Swansea, 1983). Lecture Notes in Mathematics, vol. 1095, pp. 51–82. Springer, Berlin (1984)
14. Lejay, A., Martinez, M.: A scheme for simulating one-dimensional diffusions with discontinuous coefficients. Ann. Appl. Probab. **16**(1), 107–139 (2006)

15. Martinez, M., Talay, D.: One-dimensional parabolic diffraction equations: pointwise estimates and discretization of related stochastic differential equations with weighted local times. Electron. J. Probab. **17**(27), 1–30 (2012)
16. Mascagni, M., Simonov, N.A.: Monte Carlo method for calculating the electrostatic energy of a molecule. In: Computational science—ICCS 2003, Part I. Lecture Notes in Computur Science, vol. 2657, pp. 63–72. Springer, Berlin (2003)
17. Niklitschek-Soto, S., Talay, D.: (in preparation)
18. Pardoux, É.: Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order. In: Decreusefond, L., Gjerde, J., Øksendal, B., Üstünel, A.S. (eds.) Stochastic Analysis and Related Topics: The Geilo Workshop, (1996). Birkhäuser, Boston (1998)
19. Pauwels, E.J.: Smooth first-passage densities for one-dimensional diffusions. J. Appl. Probab. **24**(2), 370–377 (1987)
20. Peskir,G.: A change-of-variable formula with local time on curves. J. Theoret. Probab. **18**(3), 499–535 (2005)
21. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Springer, Berlin (1999)
22. Rozkosz, A.: Weak convergence of diffusions corresponding to divergence form operators. Stoch. Stoch. Rep. **57**(1–2), 129–157 (1996)
23. Stroock, D.W.: Diffusion semi-groups corresponding to uniformly elliptic divergence form operators (I): Aronson's estimate for elliptic operators in divergence form. In: Séminaire de probabilités XXII. Lecture Notes in Mathematics, vol. 1321, pp. 316–347. Springer, Berlin (1988)
24. Talay, D.: Probabilistic numerical methods for partial differential equations: elements of analysis. In: Talay, D., Tubaro, L. (eds.) Probabilistic Models for Nonlinear Partial Differential Equations and Numerical Applications. Lecture Notes in Mathematics, vol. 1627, pp. 148–196. Springer, Berlin (1996)
25. Yan, L.: The Euler scheme with irregular coefficients. Ann. Probab. **30**(3), 1172–1194 (2002)