Master Project

Coupled Mathematical Models for Heart Integration
A Stability Study

Author:
Andrea DI BLASIO

Supervisors:
Prof. Alfio QUARTERONI
Dr. Toni LASSILA

A thesis submitted in fulfilment of the requirements
for the degree of Master in Computational Science and Engineering

in the

Chair of Modelling and Scientific Computing
Mathematics Section

20 June, 2014

Abstract

In this thesis we consider a fully coupled model which aims at reproducing some qualitative features of the electro-mechanical activity of the heart. The models used to describe both the electrical and mechanical activities are relatively simple. However, coupling them together can give rise to numerical instabilities or incorrect predictions. After having introduced each of the sub-models of the fully coupled system we perform some numerical experiments to draw some insights on the numerical approximation of this problem. Firstly we focus on the numerical approximation of the Aliev-Panfilov model, which controls the electrical activation of the muscle. We verify that different approaches can be followed to solve such a problem by the finite element method reducing the computational effort. However each approach can lead to inaccurate predictions of the front velocity. Then we suggest also two numerical schemes for time integration particularly suited for PDEs such as the Aliev-Panfilov model: the operator splitting method and the Runge-Kutta-Chebyshev method. When considering the fully coupled problem, we examine two ways of reducing the computational cost: treating some of the coupling terms explicitly or solving the linearised system iteratively. We verify that with the first choice we can experience numerical instabilities depending on the numerical scheme used for time integration. On the other hand, when solving the linearised system iteratively, key points to solve the problem efficiently are the choice of an adaptive stopping criterion and a good preconditioner. From the numerical experiments performed we conclude that the coupling between the active stress and the mechanics is very influential on the stability of the system and on the convergence of the residuals.

Acknowledgements

I would like to sincerely thank Prof. Alfio Quarteroni for the opportunity he gave me to work on this project within the Chair of Modelling and Scientific Computing.

Special thanks go to Dr. Toni Lassila for his precious advices and the availability he has always shown to me during the whole semester.

I would like to take this opportunity to thank Dr. Luca Dedé and Dr. Simone Deparis as well. They have been always available to discuss and exchange opinions, not only on scientific related topics. Their comprehension and advices were very helpful throughout the whole duration of my studies.

Lastly but not least, my family. Their continuous support and confidence in me made all this possible.

Contents

List of Figures

List of Tables

Introduction

Mathematical modelling is a useful technique to investigate electro-mechanical activity of the heart. Over the past 40 years many complex and detailed models of electro-mechanical activity have been developed to reproduce various experimental observations. However, qualitative predictions can be made by using simpler models, which give rise to easier numerical problems to be solved. In this thesis we consider a fully coupled model describing the cardiac activity. We use simple models to describe both the electrical and the mechanical activities. Furthermore only one-dimensional models in space are considered. In particular the two-variables model developed by R. R. Aliev and A. V. Panfilov [1] is used to describe the electrical activation of the muscle. To model the muscle contraction mechanism, we adopt another two-variables model which was proposed by J. Bestel, F. Clément and M. Sorine [2]. To model the mechanical response of the cardiac tissue, we consider the latter as an neo-Hookean incompressible solid. Several ways of coupling the different state variables can be considered. Of course the electrical signal has a direct effect on the dynamics of the active stress, which in turn generates contraction. However, other less well-known ways of coupling can be considered, such as, for example, the effect of the mechanical deformation on the propagation of the electrical signal through the medium (mechanoelectrical feedback). Even if the models we choose are simple when solved individually, having them coupled together can give rise to numerical issues, depending on how we actually decide to solve the whole problem. In particular, we investigate if numerical instabilities can occur if we decide to treat some of the coupling terms explicitly, or, alternatively, if solving the numerical problem by using iterative solvers represents a suitable possibility to reduce the computational cost. If so, choosing a good preconditioner is a key point, which can deserve particular attention. These kind of numerical observations are the subject of the last chapter of this thesis.

The outline of the work is the following. In Chapter 1 we review some of the most historically important membrane models presented in literature. After having discussed some of the properties of the Hodgkin-Huxley model (1952) [13] and the FitzHugh-Nagumo model (1961) [6, 7, 8], we conclude the first section by presenting the Aliev-Panfilov equations (AP model) (1996) [1], which are used in the fully coupled cardiac model to describe excitation in cardiac cells. The second part of the chapter is devoted to illustrating the numerical approximation of the AP model, and several numerical aspects that should be considered when solving these equations. We solve the problem by means of the finite element method (FEM). We underline how different approaches can be considered to approximate elementwise integrals, and we show the numerical effects that each approach can have on the approximated solution. Moreover the AP model, as many diffusion-reaction problems, suffers from severe time step restriction when schemes as explicit Euler's method are used for time integration. We therefore present two valid alternatives to perform time integration, which are particularly suited for such problems: the operator splitting method (OS) and the family of Runge-Kutta-Chebyshev methods (RKC).
Chapter 2 starts by illustrating how muscle cells are organized. The main reference for this part is [19]. Then, similarly as done in Chapter 1, we review some historically important models describing muscle contraction: the Hill model (1938) [12] and the Huxley model (1957) [17, 18]. We conclude the section by presenting the derivation of the model developed by Bestel, Clément and Sorine (BCS model) [2], which is obtained through statistical analysis of the Huxley model.

Chapter 3 gives an overview on basic concepts in continuum mechanics. We refer to [29] for this part. We build the theoretical framework to finally introduce the constitutive equations for hyperelastic materials, which are used to describe the response to stress of the cardiac tissue. In particular we focus on the description of some kinematics concepts and on the derivation of some of the governing equations of the mechanics, such as the conservation of mass and the balance of linear momentum.

In Chapter 4 we give the equations characterizing the coupled model we choose to solve. The coupling terms between the different state variables are reported explicitly. We conclude the chapter by reporting explicitly the linearisation of the non-liner system we aim at solving.

In Chapter 5 are shown the numerical results. After a brief introduction on the parameter set we decide to adopt, we discuss two different ways of solving the whole problem more efficiently, reducing the computational effort. The first choice is given by treating some of the coupling terms explicitly. We point out that when some of the coupling terms are solved explicitly we can experience numerical instabilities, depending on the time step chosen and/or on the numerical scheme used for time integration. The second way to reduce the cost is represented by avoiding the exact solution of the linearised system, and therefore by deciding to use an iterative solver. Key points of this approach are the choice of an adaptive stopping criterion and a suitable preconditioner to ensure fast convergence. In particular we find that the coupling between active stress and mechanics is very influential on the convergence of the residuals. We conclude the chapter by presenting a simple numerical experiment which aims at showing how stretch-activated currents [24, 25] can induce termination of re-entrant waves.

Chapter 1

Model of Cardiac Excitation

1.1 Review of Membrane Models

In 1952 Hodgkin and Huxley [13, 19] proposed the first quantitative mathematical model for the study of generation and propagation of signals in excitable systems. Although their work was devoted to the study of the action potential in the long giant axon of a squid nerve cell, their ideas have been extended and applied to a wide range of excitable cells during the past years. In particular FitzHugh [7, 19] managed to transfer the essential behaviour of the excitable process into a simpler model suitable for mathematical analysis. This simplified model turned out to be of great theoretical interest and contributed enormously to the study of excitable systems. Indeed, despite detailed ionic models are able to accurately reproduce most of the basic features of cardiac tissue, they are suitable to simulate only limited spatial regions, due to their numerical complexity. The modern ionic models consists of dozens of ODEs and it is a great numerical challenge to solve them for relatively large spatial domains, especially if high resolution is required. Then many important problems, such as re-entrant cardiac arrhythmias, which involve only large areas of cardiac tissue, can not be solved numerically by means of these models. The two-variable model of cardiac excitation presented by Aliev and Panfilov in 1995 [1] was built upon the FitzHugh-Nagumo model, retaining its simplicity while adequately representing the shape of the action potential and the pulse propagation in patches of cardiac cells. Then it can be used effectively in computer simulations which can involve even the whole heart.

1.1.1 The Hodgkin-Huxley Model

The functioning of many cells, such as neurons and muscle cells, is dependent on the generation and propagation of electrical signals. The membrane of each cell contains ion channels that allow specific ions to pass through the cell membrane. The flow of these currents is completely driven by the ionic concentration gradient and the cell membrane potential, used by many cells as signal. In particular there are cells for which, if the applied current is strong enough, the membrane potential goes through a large excursion, called action potential, before then returning to rest. Such cells are called excitable. It is the case for cardiac cells, smooth and skeletal muscle cells and most neurons. Then excitable cells have the ability to respond fully to a stimulus or not at all, being able to distinguish between a stimulus of sufficient amplitude and background noise.
Cell membrane can be modelled as a capacitor in parallel with an ionic current, giving result to the equation

$$C_m \frac{\mathrm{d}V}{\mathrm{d}t} + I_{ion}(V,t) = 0, \tag{1.1.1}$$

being C_m the membrane capacitance, while V measures the difference between the internal and the external potential. In many neural cells the principal ionic currents taking part to the excitation process are the Na^+ current and the K^+ current. Moreover Hodgkin and Huxley considered the contribution of

all the other ionic species, by lumping them together into one current called leakage current. The ionic currents were assumed to be a linear function of the membrane potential, with a driving force given by their respective Nernst potential. Thus their model reads as

$$C_m \frac{\mathrm{d}V}{\mathrm{d}t} = -g_{Na}(V - V_{Na}) - g_K(V - V_K) - g_L(V - V_L) + I_{app}, \tag{1.1.2}$$

where I_{app} denotes the applied current. Defining

$$g_{eff} = g_{Na} + g_K + g_L \quad \text{and} \quad V_{eq} = \left(g_{Na}V_{Na} + g_K V_k + g_L V_L\right)/g_{eff},$$

it is easy to realize that with a steady applied current the membrane voltage equilibrate to

$$V_{eq} = V_{eff} + I_{app}/g_{eff}.$$

Indeed this is what happens for a sufficiently small current, while for larger inputs the response observed is quite different. Therefore if we assume the model (1.1.2) correct, the only explanation for the differences observed is that the conductances are in some way dependent on the voltage. Moreover Hodgkin and Huxley observed in their experiments that, even with voltage fixed, the conductances show time-dependent behaviour. In particular they observed that, when the voltage is stepped up and held fixed at a higher level, the potassium conductance increases over time in a sigmoidal fashion to finally reach a steady level. On the other hand, the same conductance decreases in exponential way in response to a voltage step decrease. Moreover the time constant and the final level of the conductance are dependent on the value of the voltage. Instead the sodium conductance in response to a step increase of the voltage, first increases, but then decreases again. In order to describe the experimental data through their model they proposed the following expressions for the potassium and the sodium conductances:

- $g_K = \bar{g}_K n^4$, where n, often called the K^+ activation variable, obeys to the differential equation

$$\tau_n(V)\frac{\mathrm{d}n}{\mathrm{d}t} = n_\infty(V) - n, \tag{1.1.3}$$

  for some functions $n_\infty(V)$ and $\tau_n(V)$ that have to be determined from experimental data. Often equation (1.1.3) is written

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha_n(V)(1-n) - \beta_n(V)n,$$

  where

$$n_\infty(V) = \frac{\alpha_n(V)}{\alpha_n(V) + \beta_n(V)} \quad \text{and} \quad \tau_n(V) = \frac{1}{\alpha_n(V) + \beta_n(V)}.$$

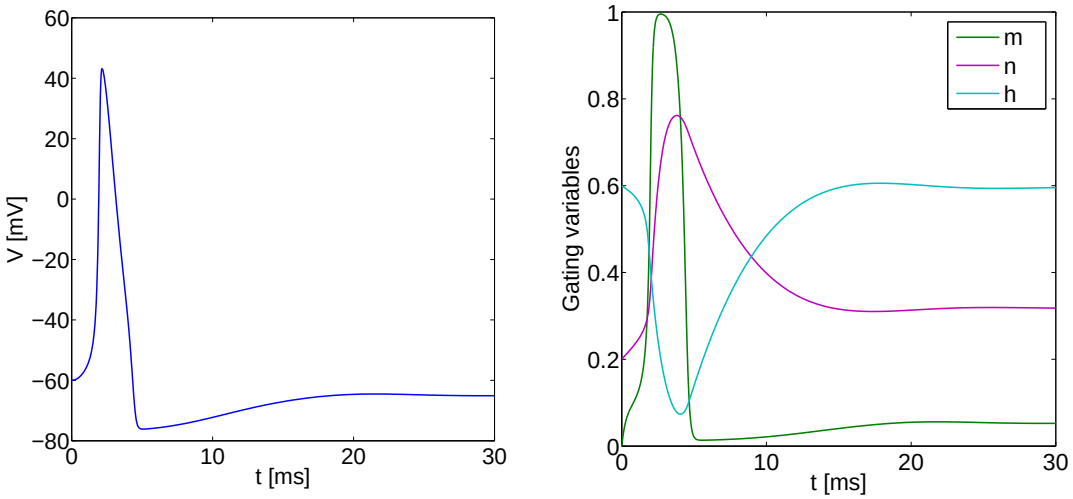
- $g_{Na} = \bar{g}_{Na} m^3 h$, where both m and h obey to the differential equation

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \alpha_w(V)(1-w) - \beta_w(V)w,$$

  where $w = m$ or h. In particular m and h are respectively referred to as the Na^+ activation variable and the Na^+ inactivation variable.

Finally the Hodgkin-Huxley model reads as a system of four first order differential equations

$$\begin{cases} C_m \dfrac{\mathrm{d}V}{\mathrm{d}t} = -\bar{g}_K n^4 (V - V_K) - \bar{g}_{Na} m^3 h (V - V_{Na}) - \bar{g}_L (V - V_L) + I_{app}, \\ \dfrac{\mathrm{d}n}{\mathrm{d}t} = \alpha_n(V)(1-n) - \beta_n(V)n, \\ \dfrac{\mathrm{d}m}{\mathrm{d}t} = \alpha_m(V)(1-m) - \beta_m(V)m, \\ \dfrac{\mathrm{d}h}{\mathrm{d}t} = \alpha_h(V)(1-h) - \beta_h(V)h. \end{cases} \tag{1.1.4}$$

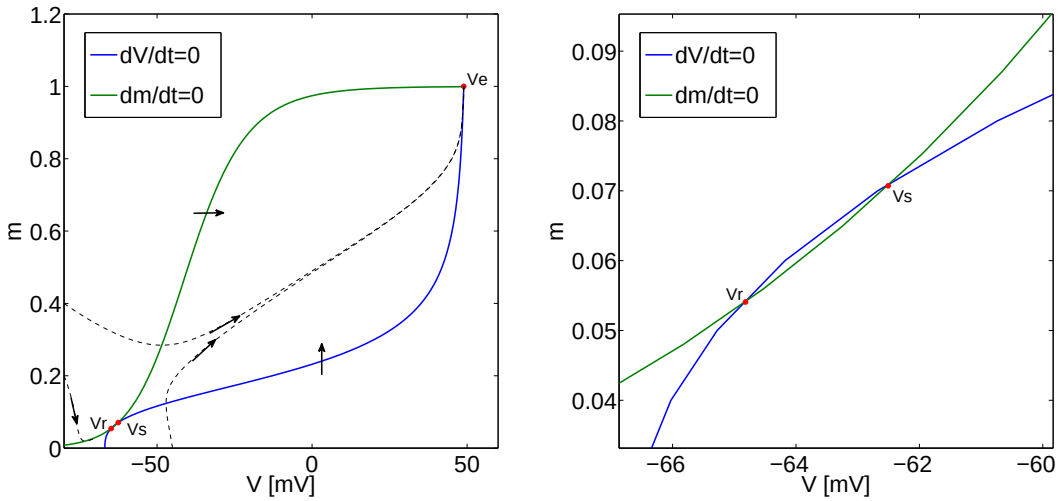


(a) (b)

Figure 1.1: (a): Action potential of the Hodgkin-Huxley model. (b): Change in the gating variables during an action potential.

In order to better understand the behaviour of the model (1.1.4), FitzHugh provided an elegant qualitative analysis of the Hodgkin-Huxley equations [6, 7, 8]. In particular it is important to observe that m is a fast variable while n and h are slow variables, that is m responds more quickly to changes in V than either n or h. Thus in the initial stage, while m and V are varying, n and h can be thought as constant. Then it is possible to provide a mathematical analysis of the fast phase of the system behaviour fixing n and h at their respective resting states n_0 and h_0, and then considering the behaviour of the model as a function only of m and V. The resulting two-dimensional system can be therefore studied in the (V, m) phase plane, a plot of which is reported in Figure 1.2, together with three representative trajectories. In particular it can be observed that the two nullclines $\mathrm{d}V/\mathrm{d}t = 0$ and $\mathrm{d}m/\mathrm{d}t = 0$ intersect three times, giving origin to two stable steady states (respectively the resting state V_r and the excited state V_e) and a saddle point V_s. We can observe that any perturbation from the resting state that is not strong enough to cross the stable manifold of the saddle point, dies and goes back to the resting state. However if the perturbation is larger, and crosses the stable manifold, it goes for a large excursion up to the excited state. If we consider the model as only function of V and m, it is clear that once V reaches the excited state V_e, it will stay there indefinitely. However we have already anticipated that it is expected that V after a certain time goes back to its resting state V_r, the only stable steady state of the full model. In order to capture this behaviour, we must consider the slower variation of h and n, and observe how the change in h and n affects the V-nullcline in the fast phase plane. Indeed as V reaches the excited state, h starts slowly to decrease, inactivating the Na^+ channels, while n begins to increase activating the K^+ channels. Of course different values of h and n give result to different V-nullclines as shown in Figure 1.3. In particular what happens in a longer time scale it is that the two points V_s and V_e get closer as h decreases and n increases, until they both disappear in a saddle-point bifurcation, and finally V_r is the only remaining steady state of the model.

Instead of dropping the slow variables and studying the model only as a function of the fast ones, we can perform another analysis considering one fast variable together with a slow one. For example we can think of the fast variable m as instantaneously equal to its steady value $m_\infty(V)$. Moreover FitzHugh observed that during an action potential $h + n \approx 0.8$. Therefore we can eliminate the h variable by taking $h = 0.8 - n$ which gives us a new model as function of only V and n. The new model can be



(a) (b)

Figure 1.2: (a): Fast phase plane of the Hodgkin-Huxley model. (b): Zoom on V_s and V_r.

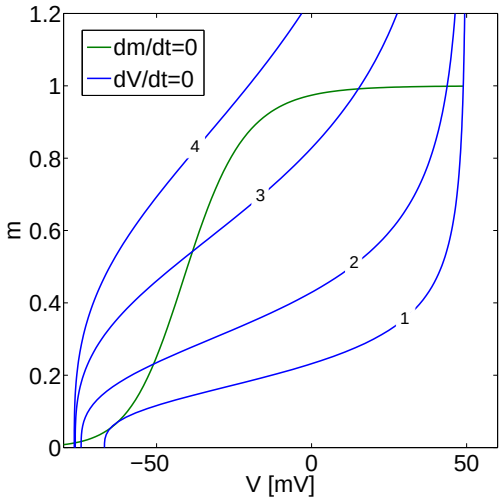


Figure 1.3: V-nullcline as function of the slow variable h and n. For these curves parameter values are: (1) $(h,n,) = (0.596, 318)$, (2) $(h,n) = (0.4, 0.5)$, (3) $(h,n) = (0.2, 0.7)$, (4) $(h,n) = (0.1, 0.8)$.

then studied in the new fast-slow phase plane (V,n) shown in Figure 1.4. The V-nullcline has a cubic shape and it is characterized by three branches. The left and the right branches are referred to as the stable branches, while the middle one as the unstable branch. Far from the stable branches the solution moves horizontally in the plane (V,n) following the direction determined by the V-nullcline. Instead, when close to the sable branches the trajectories follow slowly the V-nullcline in the direction determined by the n-nullcline. In this situation the two variables V and n are called respectively the excitation and recovery variables. Indeed V governs the rise to the excited state, while n the return to the steady resting state. Let us point out that the left, middle and right branches of the V-nullcline correspond respectively to the three points V_r, V_s and V_e observes in the fast phase plane of Figure 1.2.

If a perturbation of the resting state V_r is strong enough to cross the unstable manifold, the trajectory moves to right to reach the excited branch, otherwise it immediately goes back to the steady state.

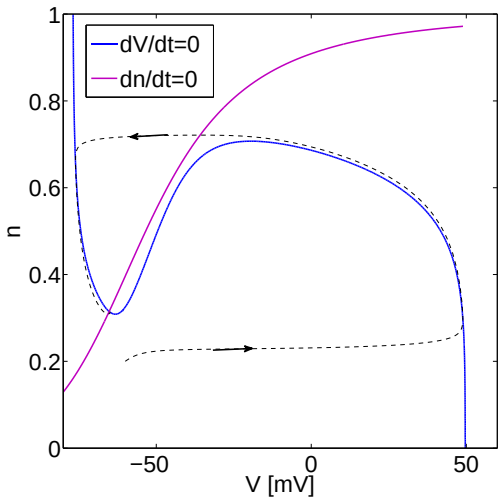


Figure 1.4: Fast-slow phase plane of the Hodgkin-Huxley model.

1.1.2 The FitzHugh-Nagumo Model

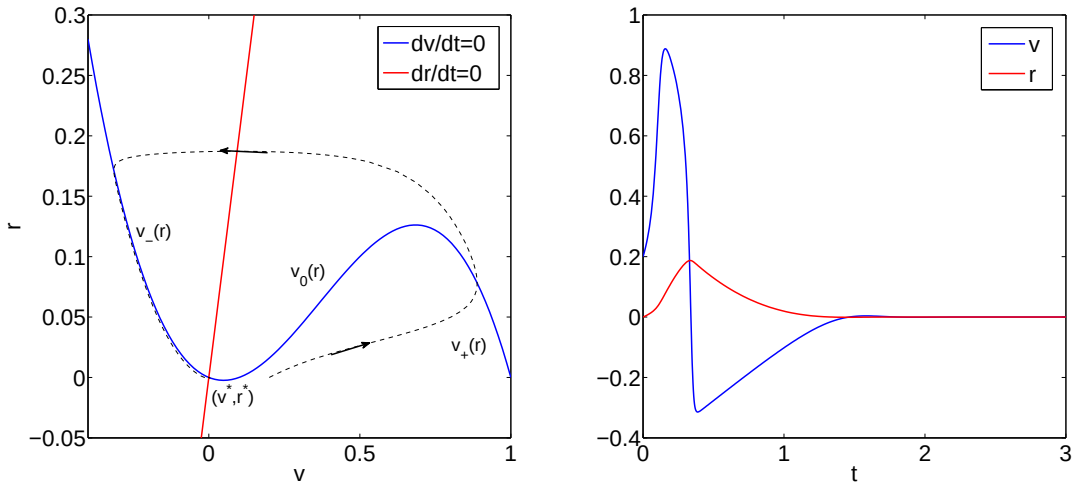
The FitzHugh-Nagumo model extracts the essential behaviour of the Hodgkin-Huxley model fast-slow phase plane making use of simpler equations describing the dynamics. Thus the model has a fast variable v and a slow variable r. Again the nullcline of the fast variable has a cubic shape and it is called the excitation variable, while the slow variable has a nullcline which is monotonically increasing, and it is called the recovery variable. Assuming a cubic nullcline for v and a linear nullcline for r, the traditional FitzHugh-Nagumo equations, involving dimensionless quantities, read as

$$\begin{cases} \epsilon \dfrac{\mathrm{d}v}{\mathrm{d}t} = v(1-v)(v-\alpha) - r, \\ \dfrac{\mathrm{d}r}{\mathrm{d}t} = v - \gamma r. \end{cases} \tag{1.1.5}$$

Usual values for the parameters appearing in the model are $\epsilon = 0.01$, $\alpha = 0.1$ and $\gamma = 0.5$. As for the fast-slow phase plane of the Hodgkin-Huxley model, the equation $f(v,r) = 0$ has three solutions $v = v(r)$ defining the three branches of the v-nullcline. Let us denote these three solutions as v_-, v_0 and v_+, where $v_-(r) \le v_0(r) \le v_+(r)$. Moreover the r-nullcline $g(v,r) = 0$ has exactly one intersection with the v-nullcline, and then only one steady stable state (v^*, r^*) exists. If a perturbation of the steady stable state is small, and does not cross the unstable manifold, any trajectory moves immediately back to (v^*, r^*). On the other hand, if the perturbation is large enough to cross the unstable manifold, the trajectory moves horizontally to right to reach the excited state v_+, and after goes back to the resting state v_- to conclude the path in (v^*, r^*).

1.1.3 The Aliev-Panfilov Model

Incorporating one of the models for excitable membrane into a nonlinear cable equation, it gives rise to travelling waves of electrical excitation, that is a solution of a partial differential equation that travels along the domain, at constant velocity and with fixed shape. We can distinguish between two kind of travelling waves: travelling fronts and travelling pulses. The first type is characterized by the presence



(a) (b)

Figure 1.5: (a): Phase plane of the FizHugh-Nagumo model (1.1.5). (b): Action potential and recovery variable as function of time.

of two stable steady state, a resting state v_r and an excited state v_e. We encounter this situation when the recovery variable is fixed to its steady state, and therefore the system presents two steady stable states (it is bistable). Thus the wave looks like a moving plateu, switching the domain from the resting to the excited state. When instead the recovery variable is allowed to vary, the system presents only one steady stable state, the resting one. Then we get travelling pulses: after excitation, the system goes back to its resting state and therefore the wave looks like a moving bump. Travelling pulses for the FitzHugh-Nagumo model satisfy the equations

$$\begin{cases} \epsilon \dfrac{\partial v}{\partial t} = \epsilon^2 \dfrac{\partial^2 v}{\partial x^2} + f(v,r), & x \in (0,1), t > 0, \\ \dfrac{\partial r}{\partial t} = g(v,r), & x \in (0,1), t > 0, \end{cases} \tag{1.1.6}$$

where ϵ is a positive small number. Finally it is important to emphasize that ϵ does not imply anything on the real magnitude of the physical conductivity coefficient, but can be interpreted simply as a scaling of the space variable. However the model (1.1.6), although successfully describing some qualitative aspects of excitation propagation, fails to simulate several quantitative parameters of cardiac tissue, such as the shape of the action potential.

In 1995 Aliev and Panfilov presented a simple two-variable model to simulate cardiac excitation. It is built up on the two-variables FitzHugh-Nagumo model, retaining its simplicity. It is designed to represent properly the action potential shape and the restitution property of the cardiac tissue. The two equations defining the model read as:

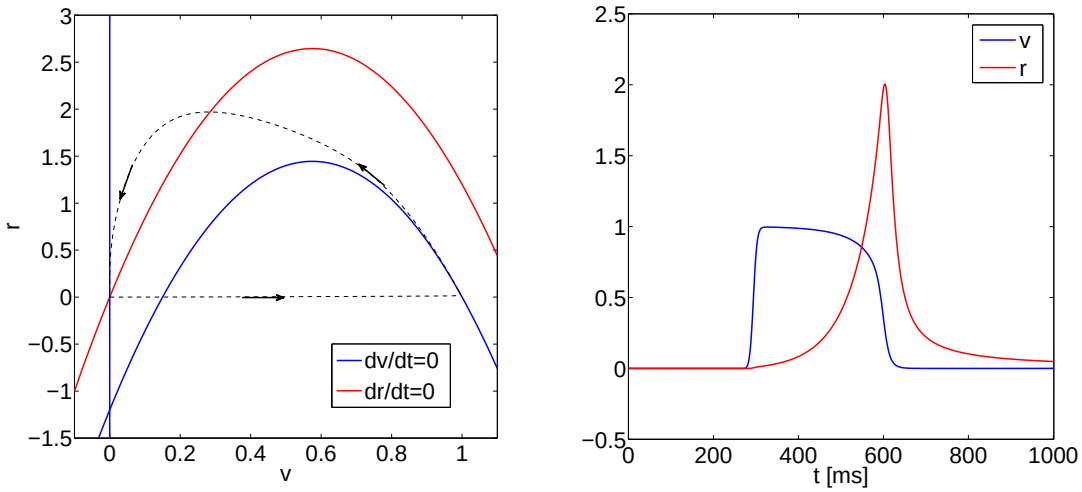
$$\begin{cases} \dfrac{\partial v}{\partial t} = \mu \dfrac{\partial^2 v}{\partial x^2} - kv(v-a)(v-1) - vr + I_{app}, & x \in (0,1), t > 0, \\ \dfrac{\partial r}{\partial t} = \varepsilon(v,r)(-r - kv(v-a-1)), & x \in (0,1), t > 0, \end{cases} \tag{1.1.7}$$

where $\varepsilon(v,r) = \varepsilon_0 + \mu_1 r/(v+\mu_2)$, $k = 8$, $a = 0.15$, $\varepsilon = 0.002$, $\mu \in \mathbb{R}^+$, and I_{app} is the current due to an external stimulus. By adjusting the parameters μ_1 and μ_2 we can then tune the restitution curve to that

experimentally observed. After having computed several restitution curves Aliev and Panfilov observed that the values which were giving the best fit are $\mu_1 = 0.2$ and $\mu_2 = 0.3$. Let us remark that the model involves dimensionless variables v, r and t, while, as already specified for (1.1.6), μ serves only as a scaling for the space variable. The actual transmembrane potential V and time t can be obtained with the formulae

$$V\ [\text{mV}] = 100\,v - 80 \quad \text{and} \quad t\ [\text{ms}] = 12.9\,t. \tag{1.1.8}$$

In this case the resting potential is equal to -80 mV, while the time variable has been rescaled so that the duration of an action potential measured at the level of 90% of repolarization is 330 ms. In Figure 1.6 are shown the nullclines of the model (1.1.7). As for the FitzHugh-Nagumo equations, the v-nullcline has a cubic shape, that is, it presents three solutions $v_-(r)$, $v_0(r)$ and $v_+(r)$ to the equation $\partial v/\partial t = 0$. Differently from (1.1.5), in the first equation of (1.1.7) it appears the term vr instead of simply r. This change allows to better represent the shape of the action potential. In particular, if we compare the two phase plane of Figure 1.5(a) and Figure 1.6, we can observe that in the Aliev-Panfilov representation the potential does not enter the region $v < 0$. This phenomena is known as super-repolarization, it occurs in the FitzHugh-Nagumo model, but does not exists in real myocardium. Another difference



(a) (b)

Figure 1.6: (a): Nullclines of the Aliev-Panfilov equations. (b): Action potential and recovery variable of the Aliev-Panfilov model as function of time.

between the models (1.1.5) and (1.1.7) is that the nullcline of the slow variable in (1.1.7) has quadratic shape instead of linear. Indeed, from comparison with experimentally observed nullclines, it has been observed that such dynamics are more appropriate for heart tissue.

1.2 Numerical Approximation of the Aliev-Panfilov equations

Analytical treatment of the Aliev-Panfilov equations is a difficult task due to the nonlinear coupling between the potential and the recovery variables. We aim at approximate the solutions $v(x,t)$ and $r(x,t)$ to problem (1.1.7) by means of the finite element method. The space variable x belongs to the domain $\Omega = (0,1)$, while $t \in (0,T)$. Moreover we consider homogeneous Neumann boundary conditions for v. Finally we complete the equations by fixing the initial condition, that is $v(x,0) = v_0(x)$

and $r(x,0) = r_0(x)$. Then the problem we want to solve reads as

$$\begin{cases} \dfrac{\partial v}{\partial t} = \mu \dfrac{\partial^2 v}{\partial x^2} - kv(v-a)(v-1) - vr + I_{app}, & x \in (0,1),\ t \in (0,T), \\ \dfrac{\partial r}{\partial t} = \varepsilon(v,r)(-r - kv(v-a-1)), & x \in (0,1),\ t \in (0,T), \\ \dfrac{\partial v}{\partial x}(0,t) = \dfrac{\partial v}{\partial x}(1,t) = 0, & t \in (0,T), \\ v(x,0) = v_0(x), \quad r(x,0) = r_0(x), & x \in (0,1), \end{cases} \tag{1.2.1}$$

where k, a, μ and $\varepsilon(\cdot,\cdot)$ are defined as in the problem (1.1.7). We start rewriting the problem in its weak formulation. Let us call W and S the two functional spaces where respectively v and r lie in. In particular W and S are two functional spaces such that the integrals appearing in the weak formulation are defined in the Lebesgue sense. We multiply for each $t \in (0,T)$ the two differential equations respectively by the test functions $w = w(x)$ and $s = s(x)$, $w \in W$, $s \in S$, and we integrate over Ω. Then for each $t \in (0,T)$ we seek for $v(t) \in W$ and $r(t) \in S$ such that

$$\begin{cases} \displaystyle\int_0^1 \frac{\partial v(t)}{\partial t} w \,\mathrm{d}x = -\int_0^1 \mu \frac{\partial v(t)}{\partial x}\frac{\partial w}{\partial x}\,\mathrm{d}x + \int_0^1 f(v(t), r(t)) w \,\mathrm{d}x & \forall w \in W, \\ \displaystyle\int_0^1 \frac{\partial r(t)}{\partial t} s \,\mathrm{d}x = \int_0^1 g(v(t), r(t)) s \,\mathrm{d}x & \forall s \in S, \end{cases} \tag{1.2.2}$$

with

$$\begin{aligned} f(v(t), r(t)) &= -kv(t)(v(t)-a)(v(t)-1) - v(t)r(t) + I_{app}(t), \\ g(v(t), r(t)) &= \varepsilon(v(t), r(t))(-r(t) - kv(t)(v(t)-a-1)), \end{aligned}$$

and $v(0) = v_0$, $r(0) = r_0$.
We now consider the Galerkin approximation of problem (1.2.2). Thus for each $t \in (0,T)$ we seek for $v_h(t) \in W_h$ and $r_h(t) \in S_h$ such that

$$\begin{cases} \displaystyle\int_0^1 \frac{\partial v_h(t)}{\partial t} w_h \,\mathrm{d}x = -\int_0^1 \mu \frac{\partial v_h(t)}{\partial x}\frac{\partial w_h}{\partial x}\,\mathrm{d}x + \int_0^1 f(v_h(t), r_h(t)) w_h \,\mathrm{d}x & \forall w_h \in W_h, \\ \displaystyle\int_0^1 \frac{\partial r_h(t)}{\partial t} s_h \,\mathrm{d}x = \int_0^1 g(v_h(t), r_h(t)) s_h \,\mathrm{d}x & \forall s_h \in S_h, \end{cases} \tag{1.2.3}$$

with $v_h(0) = v_{0h}$, $r_h(0) = r_{0h}$, where v_{0h} and r_{0h} are convenient finite approximations of v_0 and r_0, while $W_h \subseteq W$ and $S_h \subseteq S$ are two suitable spaces of finite dimension. In particular we identify the two spaces with the same finite dimensional space, and we refer to that simply as W_h. Then we have

$$S_h(\Omega, \mathcal{T}_h) = W_h(\Omega, \mathcal{T}_h) = \left\{ z_h \in \mathscr{C}^0(\overline{\Omega});\ z_h|_K \in \mathscr{P}^r(K)\ \forall K \in \mathcal{T}_h \right\},$$

where h represents the mesh size, while $\mathscr{P}^r$ is the set of polynomials of degree smaller or equal to r. Let us introduce the basis functions $\{\varphi_j\}_{j=1}^{N_h}$ for W_h. The approximate solutions $v_h(t)$ and $r_h(t)$ belong to the subspace W_h, and then can be represented as

$$v_h(t) = \sum_{j=1}^{N_h} v_j(t)\varphi_j, \qquad r_h(t) = \sum_{j=1}^{N_h} r_j(t)\varphi_j,$$

where $\{v_j(t)\}_{j=1}^{N_h}$ and $\{r_j(t)\}_{j=1}^{N_h}$ are the unknowns of the problem. If we denote $\dot{v}_j(t)$ and $\dot{r}_j(t)$ the time derivatives of $v_j(t)$ and $r_j(t)$, (1.2.3) becomes (we omit the dependence on t to simplify the notation)

$$\begin{cases} \displaystyle\int_0^1 \sum_{j=1}^{N_h} \dot{v}_j \varphi_j \varphi_i \,\mathrm{d}x = -\int_0^1 \mu \sum_{j=1}^{N_h} v_j \frac{\partial \varphi_j}{\partial x}\frac{\partial \varphi_i}{\partial x}\,\mathrm{d}x + \int_0^1 f\left(\sum_{j=1}^{N_h} v_j\varphi_j, \sum_{l=1}^{N_h} r_l\varphi_l\right)\varphi_i \,\mathrm{d}x \\ \displaystyle\int_0^1 \sum_{l=1}^{N_h} \dot{r}_l \varphi_l \varphi_k \,\mathrm{d}x = \int_0^1 g\left(\sum_{j=1}^{N_h} v_j\varphi_j, \sum_{l=1}^{N_h} r_l\varphi_l\right)\varphi_k \,\mathrm{d}x, \end{cases}$$

that is

$$\begin{cases} \displaystyle\sum_{j=1}^{N_h} \dot{v}_j \underbrace{\int_0^1 \varphi_j\varphi_i \,\mathrm{d}x}_{m_{ij}} = -\mu \sum_{j=1}^{N_h} v_j \underbrace{\int_0^1 \frac{\partial \varphi_j}{\partial x}\frac{\partial \varphi_i}{\partial x}\,\mathrm{d}x}_{k_{ij}} + \underbrace{\int_0^1 f\left(\sum_{j=1}^{N_h} v_j\varphi_j, \sum_{l=1}^{N_h} r_l\varphi_l\right)\varphi_i \,\mathrm{d}x}_{f_i}, \\ \displaystyle\sum_{l=1}^{N_h} \dot{r}_l \underbrace{\int_0^1 \varphi_l\varphi_k \,\mathrm{d}x}_{m_{kl}} = \underbrace{\int_0^1 g\left(\sum_{j=1}^{N_h} v_j\varphi_j, \sum_{l=1}^{N_h} r_l\varphi_l\right)\varphi_k \,\mathrm{d}x}_{g_k}, \end{cases} \tag{1.2.4}$$

with $i = k = 1, \ldots, N_h$. If we define the vectors of unknowns respectively $\mathbf{v}(t) = (v_1(t), \ldots, v_{N_h}(t))^T$ and $\mathbf{r}(t) = (r_1(t), \ldots, r_{N_h}(t))^T$, the mass matrix $\mathbf{M} = [m_{ij}] = [m_{kl}]$, the stiffness matrix $\mathbf{K} = [k_{ij}]$ and the vectors $\mathbf{f}(\mathbf{v}(t), \mathbf{r}(t)) = [f_i]$ and $\mathbf{g}(\mathbf{v}(t), \mathbf{r}(t)) = [g_k]$, the abstract formulation of the problem reads as

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}(t) \\ \dot{\mathbf{r}}(t) \end{bmatrix} = -\mu \begin{bmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}(t) \\ \mathbf{r}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{f}(\mathbf{v}(t), \mathbf{r}(t)) \\ \mathbf{g}(\mathbf{v}(t), \mathbf{r}(t)) \end{bmatrix}, \tag{1.2.5}$$

which is an ODE system that can be solved using one of the many existing numerical schemes for time integration.

1.2.1 Approximation of Elementwise Integrals

Before going into details on how effectively solve the system (1.2.5), we discuss here how to treat the elementwise integrals appearing in (1.2.4). In particular we refer to the approximation of the integrals involving the nonlinear functions f and g. Let us consider the computation of the entries f_i of the vector $\mathbf{f}$ (the discussion for the vector $\mathbf{g}$ is analogous). We have

$$f_i = \int_0^1 f(v_h(t), r_h(t))\varphi_i \,\mathrm{d}x = \sum_{K\in\mathcal{T}_h} \int_K f(v_h(t), r_h(t))\varphi_i \,\mathrm{d}x, \quad i = 1, \ldots, N_h.$$

There are three standard ways to treat the elementwise integrals by quadrature approximations [22]:

- State Variable Interpolation (SVI):

$$f_i = \sum_{K\in\mathcal{T}_h} \sum_{p=1}^{Q} \omega_p f\left(\sum_{j=1}^{N_h} v_j\varphi_j(q_{p,k}), \sum_{l=1}^{N_h} r_l\varphi_l(q_{p,k})\right)\varphi_i(q_{p,k}).$$

- Ionic Current Interpolation (ICI):

$$\begin{aligned} f_i &= \sum_{K\in\mathcal{T}_h} \sum_{p=1}^{Q} \omega_p \left(\sum_{j=1}^{N_h} f(v_j, r_j)\varphi_j(q_{p,k})\right)\varphi_i(q_{p,k}) \\ &= \sum_{j=1}^{N_h} m_{ij} f(vj, rj). \end{aligned}$$

- Lumped Ionic Current Interpolation (L-ICI):

$$f_i = \sum_{K\in\mathcal{T}_h}\sum_{p=1}^{Q}\omega_p\left(f(v_i,r_i)\sum_{j=1}^{N_h}\varphi_j(q_{p,k})\right)\varphi_i(q_{p,k})$$
$$=\left(\sum_{j=1}^{N_h}m_{ij}\right)f(v_i,r_i),$$

where $q_{p,k}$ is the pth quadrature node in the kth element and ω_p is the corresponding weight. The SVI approach imposes a significant computational cost, since it requires QK evaluations of the function $f(v,r)$, which can become prohibitive for a large number of elements and quadrature nodes. Then in this case the ICI and L-ICI approaches represent a valid alternative to the SVI method. However significant differences arise in the numerical solution, depending on the numerical scheme used, at least when linear finite elements are used. In particular the SVI approach produces an overestimation of the wave velocity, while the ICI and the L-ICI approaches tend to underestimate it, especially when a coarse mesh is used (Figure 1.7). Let us denote with v_h the wave velocity observed for the mesh size h. If we denote with v^* the exact value of the wave velocity, all the three solutions converge to the same value v^* as the mesh size h goes to zero. However it is observed that for values of h too large, the wave does not propagate along the domain when ICI or L-ICI approaches are used. Then there exists a critical value h_c, such that $v_h = 0$ for $h \geq h_c$. Finally in Figure 1.8 we show the activation time

h	v_h^{SVI}/v^*	v_h^{ICI}/v^*	v_h^{LICI}/v^*
0.005	1.009	1.019	0.981
0.01	1.065	1.056	0.926
0.02	1.222	1.074	0.759
0.025	1.324	1.037	0.667
0.033	1.426	0.944	0.472
0.05	2.009	0.648	0
0.067	2.519	0	0
0.1	3.620	0	0

Table 1.1: Numerical velocities computed by using SVI, ICI and LICI approaches for different mesh sizes h. For these results piecewise linear polynomials are used. The domain Ω is defined as $\Omega = (0,1)$ and $\mu = 10^{-4}$. v^* is approximated by taking $h = 0.0001$. The time integration is performed by using the Crank-Nicholson method with $\Delta t = 0.0775$.

($v(t) \geq 0.95$) obtained for different mesh sizes in the cases the SVI, ICI or the L-ICI approach is used.

1.2.2 Time Step Restriction for Explicit Methods

To solve the ODE system (1.2.5) many numerical scheme for time integration can be adopted. Implicit methods are more costly than explicit ones, since being the functions f and g nonlinear, at every time level t^{n+1} we need to solve a nonlinear problem to find (v_h^{n+1}, r_h^{n+1}). On the other hand implicit methods enjoy better stability properties than explicit ones, which are subject to restrictive limitations on the choice of the time step Δt. Limitations on Δt are related to the eigenvalues of the Jacobian associated to the ODE system. Thus in this section we compute limitations on Δt for explicit schemes, by giving an estimation of the eigenvalues of the Jacobian associated to the system (1.2.5).

Remark 1.1. *Stability Analysis for Explicit Runge-Kutta Methods* [10]. Let us consider the generic ODE system $\dot{\mathbf{y}} = \mathbf{f}(t,\mathbf{y})$, and let be $\boldsymbol{\varphi}(t)$ a smooth solution. We linearise $\mathbf{f}$ in its neighbourhood as

$$\dot{\mathbf{y}}(t) = \mathbf{f}(t,\boldsymbol{\varphi}(t)) + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t,\boldsymbol{\varphi}(t))(\mathbf{y}(t)-\boldsymbol{\varphi}(t)) + \dots$$

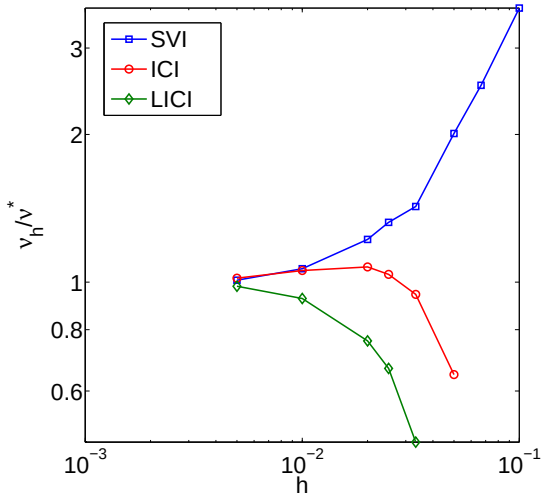


Figure 1.7: Numerical velocities computed by using SVI, ICI and LICI as function of the mesh size h.

and introduce $\bar{\mathbf{y}}(t) = \mathbf{y}(t) - \boldsymbol{\varphi}(t)$ to obtain

$$\dot{\bar{\mathbf{y}}}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \boldsymbol{\varphi}(t))\bar{\mathbf{y}}(t) + \ldots = \mathbf{J}(t, \boldsymbol{\varphi}(t))\bar{\mathbf{y}}(t) \ldots \quad .$$

As first approximation let us assume $\mathbf{J}$ constant. Neglecting the error terms and omitting the bars we arrive at

$$\dot{\mathbf{y}}(t) = \mathbf{J}\mathbf{y}(t). \tag{1.2.6}$$

An explicit Runge-Kutta method applied to (1.2.6) gives

$$\mathbf{y}^{n+1} = R(\Delta t \mathbf{J})\mathbf{y}^n, \tag{1.2.7}$$

where R is called the stability function of the method, and for an s-stage explicit Runge-Kutta method it is a polynomial of degree $\leq s$. In particular if an explicit Runge-Kutta method is of order p, R is a polynomial of the form

$$R(z) = 1 + z + \ldots + \frac{z^p}{p!} + \mathcal{O}(z^{p+1}).$$

We suppose $\mathbf{J}$ diagonalizable with eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ and we write $\mathbf{y}^0 = \boldsymbol{\varphi}(t^0)$ in this basis as

$$\mathbf{y}^0 = \sum_{i=1}^{N} \alpha_i \mathbf{v}_i. \tag{1.2.8}$$

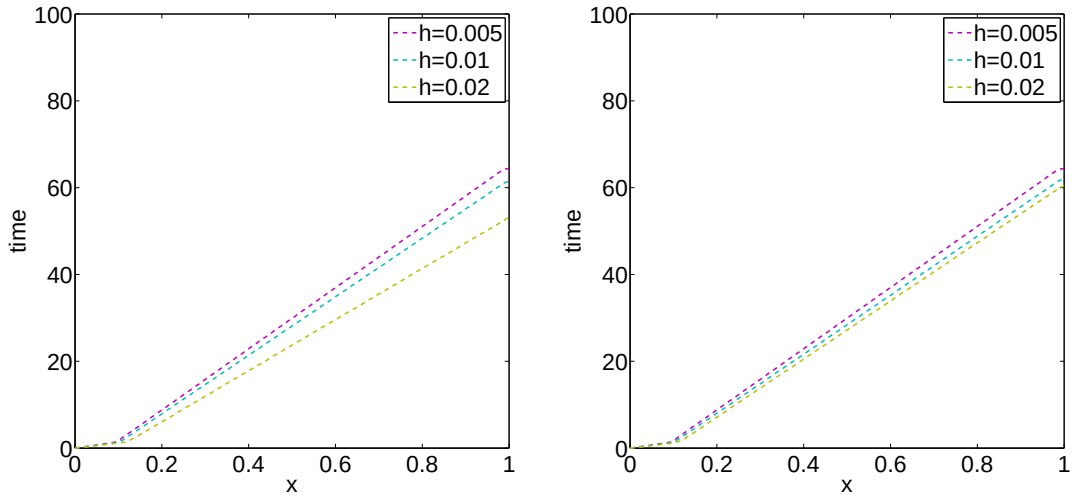
Inserting (1.2.8) into (1.2.7) we finally get

$$\mathbf{y}^{n+1} = \sum_{i=1}^{N} R(\Delta t \lambda_i)^{n+1} \alpha_i \mathbf{v}_i,$$

where λ_i are the eigenvalues associated to $\mathbf{J}$. Clearly $\mathbf{y}^{n+1}$ remains bounded for $m \to \infty$ only if for all the eigenvalues the complex number $z = \Delta t \lambda_i$ lies in the region

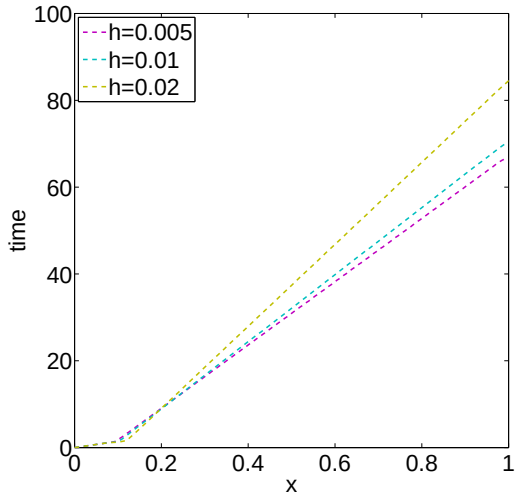
$$S = \{z \in \mathbb{C};\ |R(z)| \leq 1\}, \tag{1.2.9}$$

which is referred to as the stability domain of the method. In the case where $\mathbf{J} = \mathbf{J}(t, \boldsymbol{\varphi}(t))$, we need to ensure that for all the eigenvalues the complex number $z = \Delta t \lambda_i(t, \boldsymbol{\varphi}(t))$ lies in the region $S\ \forall t \in [t^0, T]$.



(a) SVI

(b) ICI



(c) L-ICI

Figure 1.8: Activation time for various mesh sizes($\mu = 10^{-4}$).

In particular if we apply the explicit Euler's scheme, we have $R(z) = 1+z$, and therefore the region (1.2.9) becomes

$$S = \{z \in \mathbb{C};\ |z+1| \leq 1\},$$

which is equivalent to require that the following stability condition holds

$$0 \leq \Delta t \leq 2/\max_{t\in[t^0,T]} \max_i |\lambda_i(t, \boldsymbol{\varphi}(t))|. \tag{1.2.10}$$

In order to give an estimation of the maximum eigenvalues associated to the ODE system (1.2.5), we start considering problem (1.2.1) and we write its finite differences approximation on a grid of N_h

points, $1 \le i \le N_h$, $h = 1/(N_h+1)$, $x_i = i/(N_h+1)$, to obtain

$$\begin{cases} \dot{v}_i = \dfrac{\mu}{h^2}(v_{i-1} - 2v_i + v_{i+1}) - kv_i(v_i - a)(v_i - 1) - v_i r_i, & 1 \le i \le N_h,\ t \in (0,T), \\ \dot{r}_i = \varepsilon(v_i, r_i)(-r_i - kv_i(v_i - a - 1)), & 1 \le i \le N_h,\ t \in (0,T), \\ v_0(t) = r_1(t), \quad v_{N_h+1}(t) = v_{N_h}(t), & t \in (0,T), \\ r_0(t) = r_1(t), \quad r_{N_h+1}(t) = r_{N_h}(t), & t \in (0,T), \\ v_i(0) = v_0(x_i), \quad r_i(0) = r_0(x_i), & 0 \le i \le N_h + 1. \end{cases} \tag{1.2.11}$$

The Jacobian associated to system (1.2.11) is composed of two parts, one associated to the diffusion term, and the other associated to the reaction terms

$$\mathbf{J} = -\mu \begin{bmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \operatorname{diag}\left(\dfrac{\partial f(v_i, r_i)}{\partial v_i}\right) & \operatorname{diag}\left(\dfrac{\partial f(v_i, r_i)}{\partial r_i}\right) \\ \operatorname{diag}\left(\dfrac{\partial g(v_i, r_i)}{\partial v_i}\right) & \operatorname{diag}\left(\dfrac{\partial g(v_i, r_i)}{\partial r_i}\right) \end{bmatrix}. \tag{1.2.12}$$

The eigenvalues of $\mathbf{K}$ are known, and they are given by the formula

$$\lambda_i = -\frac{4}{h^2}\left(\sin\frac{\pi(i-1)}{2(N_h-1)}\right)^2, \quad i = 1 \dots N_h.$$

Therefore the eigenvalues of the matrix associated to the diffusion part are located between $-4\mu/h^2$ and 0. For which regards the second matrix, we look at an approximation of it by neglecting the off diagonal entries. Moreover let us remark that $(v(t), r(t)) \in [0,1] \times [0,2]\ \forall t \in [t^0, T]$. We have that

$$\frac{\partial f(v,r)}{\partial v} = -k(3v^2 - 2(a+1)v + a) - r,$$

and its absolute value reaches its maximum in $(v,r) = (1,2)$. Then we obtain

$$\max_{t\in[t^0,T]} \max_i \left|\frac{\partial f(v_i, r_i)}{\partial v_i}\right| \le k(1-a) + 2.$$

On the other hand

$$\frac{\partial g(v,r)}{\partial r} = -\varepsilon_0 + \frac{\mu_1}{\mu_2 + v}(-2r - kv(v - a - 1)),$$

whose absolute value reaches its maximum for $(v,r) = (0,2)$ and then

$$\max_{t\in[t^0,T]} \max_i \left|\frac{\partial g(v_i, r_i)}{\partial r_i}\right| \le \varepsilon_0 + 4\frac{\mu_1}{\mu_2}.$$

If we call λ_i^v the eigenvalues relative to the first equation in (1.2.1), and λ_i^r the eigenvalues relative to the second one, we finally have

$$\begin{aligned} \max_{t\in[t^0,T]} \max_i |\lambda_i(t)| &\le \max_{t\in[0,T]} \max_i \max\left\{|\lambda_i^v(t)|, |\lambda_i^r(t)|\right\} \\ &\le \max\left\{4\mu/h^2 + k(1-a) + 2, \varepsilon_0 + 4\mu_1/\mu_2\right\} \\ &= 4\mu/h^2 + k(1-a) + 2. \end{aligned} \tag{1.2.13}$$

However some important remarks have to be made on the bound on $\lambda_i(t)$ we just found. Firstly, we have obtained it starting from a first order finite differences approximation of (1.2.1), while the system we want to solve, (1.2.5), it is obtained through finite element approximation. In particular, if we use piecewise linear polynomials, $|\lambda_i(\mathbf{M}^{-1}\mathbf{K})|$ are not bound by $4/h^2$, but by $12/h^2$ due to the presence of the mass matrix. Secondly, in our calculations we ignored the off-diagonal entries of the reaction matrix. Then, to check if (1.2.13) is verified $\forall t \in [t^0, T]$ we proceed as follows. We solve the system (1.2.5) using

h	Δt^{th}	Δt^*
0.01	0.0155	0.0159
0.005	0.00410	0.00411
0.002	6.647e−04	6.655e−04

Table 1.2: Comparison between Δt^{th} and Δt^* for explicit Euler's method ($\mu = 10^{-3}$).

the explicit Euler's method, and we compare the theoretical stable time step Δt^{th} given by (1.2.13) with Δt^*, which is the largest time step for which experimentally the numerical solution shows stability. The results are reported in Table 1.2 and are consistent with (1.2.13). Thus, if we want to solve the system (1.2.5) using explicit Euler's method, the largest restriction on time step is due to the diffusion part, which makes the stable time step decrease as h^2. Then for a significant number of elements it can become prohibitive to solve the system using an explicit method, and implicit ones are usually preferred. However the functions responsible of the reaction part are nonlinear, and then for every time level t^{n+1} we must solve a nonlinear problem to find (v_h^{n+1}, r_h^{n+1}), and it can be done by using the Newton's algorithm.

1.2.3 Operator Splitting

For many PDEs, especially in higher space dimension, such as the advection-diffusion-reaction problem, it is in general inefficient to apply the same integration rule to all the parts of the system. Moreover, even if the system (1.2.5) involves only two variables, solving it for a large number of elements using a single implicit integration formula can be already very costly, since for each time step we need to solve a nonlinear system. A valid alternative is to split the problem's equations into a PDE carrying on the diffusion of the variable v and a set of two ODEs taking care of the nonliner reaction terms [22, 16]. Then these equations have to be solved in an alternating way. The overall procedure reads as follows:

- **Step 1.** With (v_h^n, r_h^n) as initial conditions, we solve for half time step on each grid point of the mesh the two ODEs

$$\begin{cases} \dfrac{\partial v_i(t)}{\partial t} = f(v_i(t), r_i(t)), & i = 1, \dots, N_h, \quad t \in [t^n, t^n + \Delta t/2], \\ \dfrac{\partial r_i(t)}{\partial t} = g(v_i(t), r_i(t)), & i = 1, \dots, N_h, \quad t \in [t^n, t^n + \Delta t/2], \end{cases}$$

  to compute $(v_h^{1/2}, r_h^{1/2})$.

- **Step 2.** Set $v_h^n = v_h^{1/2}$ and solve the system

$$\mathbf{M}\dot{\mathbf{v}}(t) = -\mu \mathbf{K}\mathbf{v}(t), \quad t \in [t^n, t^{n+1}],$$

  to compute the intermediate solution v_{h*}^{n+1}.

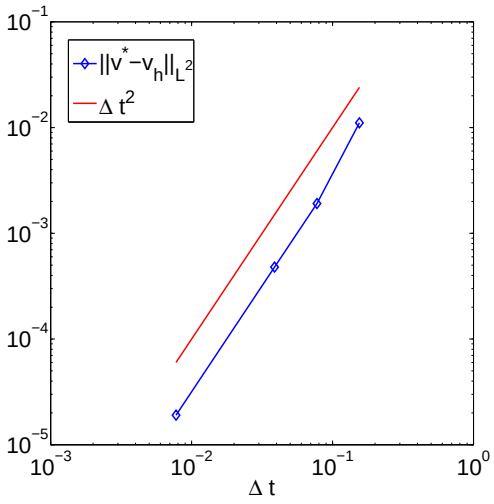
- **Step 3.** With $(v_{h*}^{1/2}, r_h^{1/2})$ as initial conditions, we solve for half time step on each grid point of the mesh the two ODEs

$$\begin{cases} \dfrac{\partial v_i(t)}{\partial t} = f(v_i(t), r_i(t)), & i = 1, \dots, N_h, \quad t \in [t^n + \Delta t/2, t^{n+1}], \\ \dfrac{\partial r_i(t)}{\partial t} = g(v_i(t), r_i(t)), & i = 1, \dots, N_h, \quad t \in [t^n + \Delta t/2, t^{n+1}], \end{cases}$$

  to compute (v_h^{n+1}, r_h^{n+1}).

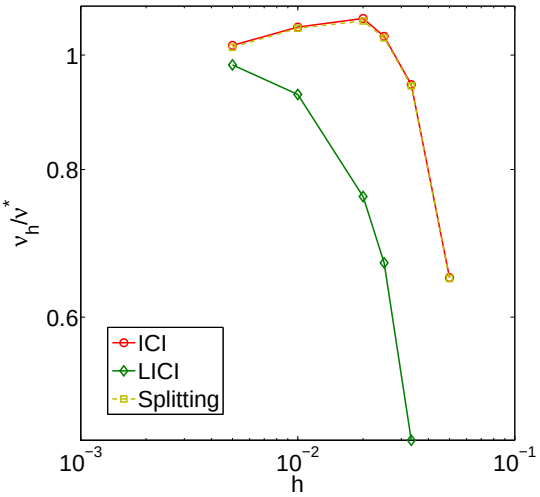
A nice property of the algorithm above is that if we use a scheme of second order accuracy in time for all the three steps, the overall procedure preserves second order accuracy. In particular we solve the

ODEs in Step 1 and 3 explicitly by using Heun's method, since there are not significant restrictions on the choice of the time step, while in Step 2 we use Crank-Nicholson method. Let us observe that solving implicitly the system in Step 2 reduces simply to solving a linear system, since all the nonlinear terms are carried by the ODEs in Step 1 and 3. In Figure 1.10 we show the numerical error as $\Delta t \to 0$. In Step 1 and 3 we solve the ODEs only on the grid points of the mesh, while in Step 2 the elementwise integral are approximated by numerical quadrature on the Gauss points. It is then expected that the front velocity converges as $h \to 0$ in a similar way as when the ICI or L-ICI approaches are used to solve (1.2.5). In particular the numerical results show almost identical convergence behaviour between operator splitting method and ICI approach (Figure 1.11).



Δt	$\|\|v^* - v_h\|\|_{L^2}$
0.1550	0.0111
0.0775	0.0019
0.0388	4.7932e−04
7.75e−03	1.9055e−05

Figure 1.9: $||v^* - v_h||_{L^2}$ as function of Δt using the operator splitting method. For these results $h = 0.001$, while v^* is computed by taking Δt = 7.75e−04



h	v_h^{OS}/v^*
0.005	1.015
0.01	1.054
0.02	1.069
0.025	1.034
0.033	0.942
0.05	0.647
0.067	0
0.1	0

Figure 1.10: Numerical velocities computed using the operator splitting method, as function of the mesh size h.

1.2.4 Stabilized Explicit Runge-Kutta Methods

In this section we discuss explicit methods, which possess extended stability domains along the negative real axis, which therefore are especially suited for time integration of parabolic partial differential problems [33, 10, 16]. These can be very efficient for many problems, usually not very stiff, of large dimension, and with eigenvalues known to lie in a certain region, since they allow to avoid algebraic system solutions. Moreover they are particularly attractive since the length of their stability region is proportional to s^2, where s is the number of stages of the method. The main ingredient for the construction of such methods are the Chebyshev polynomials

$$T_s(x) = \cos(s \arccos(x))$$

or

$$T_s(x) = 2xT_{s-1}(x) - T_{s-2}(x), \quad T_0(x) = 1, \quad T_1(x) = x, \tag{1.2.14}$$

and thus these methods are referred to as Runge-Kutta-Chebyshev methods (RKC). In particular we focus on the realization of second order RKC proposed by van der Houwen & Sommeijer [14]. Let us call β_R the real stability boundary of any explicit Runge-Kutta method. By definition $[-\beta_R, 0]$ is the largest segment of the negative real axis contained in the stability region $S = \{z \in \mathbb{C};\ |R(z)| \le 1\}$, as determined by the stability polynomial $R(z)$. We want to find, for a given s, a polynomial of the form $R_s(z) = a_0 + a_1 z + a_2 z^2 + \ldots + a_s z^s$ such that the stability domain is as large as possible in the direction of the negative real axis. In particular it can be proven that for any explicit, consistent Runge-Kutta method we have $\beta_R \le 2s^2$ ([16], Ch. 5, Theorem 1.1). In the case of first order approximation, for consistency one must have $a_0 = a_1 = 1$. Observing that $T_s(x)$ remains bounded for $-1 \le x \le 1$ between -1 and $+1$, and that among these polynomials has the largest possible derivative $T_s'(1) = s^2$, we set

$$R_s(z) = T_s(1 + z/s^2)$$

so that $R(0) = 1$, $R'(0) = 1$, and $|R(z)| \le 1$ for $-2s^2 \le z \le 0$.
However in the points where $T_s(1 + z/s^2) = \pm 1$, the stability domain has zero width. We therefore choose a small damping parameter $\varepsilon > 0$ and set

$$R_s(z) = \frac{1}{T_s(\omega_0)} T_s(\omega_0 + \omega_1 z), \quad \omega_0 = 1 + \frac{\varepsilon}{s^2}, \quad \omega_1 = \frac{T_s(\omega_0)}{T_s'(\omega_0)}.$$

These polynomials now oscillate approximately between $-1 + \varepsilon$ and $+1 - \varepsilon$, and again satisfy $R(z) = 1 + z + \mathcal{O}(z^2)$. The stability domain is a bit shorter (by $(4\varepsilon/3)s^2$), but at least the boundary is in a safe distance from the real axis.
The next step is to actually build Runge-Kutta methods which realize these stability polynomials. Houwen and Sommeijer [14] set

$$R_s(z) = a_s + b_s T_s(\omega_0 + \omega_1 z), \quad \omega_0 = 1 + \frac{\varepsilon}{s^2}, \quad \varepsilon \approx 0.15.$$

However in this formulation ε does not represent the damping parameter, though it is still related to it. Conditions for second order

$$R_s(0) = 1, \quad R_s'(0) = 1, \quad R_s''(0) = 1,$$

lead to

$$\omega_1 = \frac{T_s'(\omega_0)}{T_s''(\omega_0)}, \quad b_s = \frac{T_s''(\omega_0)}{(T_s'(\omega_0))^2}, \quad a_s = 1 - b_s T_s, (\omega_0)$$

with a damping equal to $a_s + b_s \approx 1 - \varepsilon/3$. For which regards the internal stages we define

$$R_j(z) = a_j + b_j T_j(\omega_0 + \omega_1 z), \quad j = 0, 1, \ldots, s-1.$$

Sommeijer discovered that these $R_j(z)$ can, for $j \ge 2$, be approximations of second order at certain points $t_0 + c_j \Delta t$ if

$$R_j(0) = 0, \quad R_j'(0) = c_j, \quad R_j''(0) = c_j^2$$

	h	Δt^{th}	# fct. evals.
Heun	0.01	0.0155	$\approx$ 1e+04
RKC ($s=9$)	0.01	0.4115	$\approx$ 1.7e+03
Heun	0.005	0.00410	$\approx$ 3.8e+04
RKC ($s=9$)	0.005	0.1084	$\approx$ 6.4e+03
Heun	0.002	6.647e−04	$\approx$ 2.3e+05
RKC ($s=9$)	0.002	0.0176	$\approx$ 4e+04
Heun	0.001	1.665e−04	$\approx$ 9.3e+05
RKC ($s=9$)	0.001	0.0044	$\approx$ 1.6e+05

Table 1.3: Comparison between the second order Heun's method and RKC method ($s=9$) in terms of theoretical largest stable time step, and thus of number of function evaluations needed for an hypothetical simulation with $t \in (0, 77.5)$. ($\mu = 10^{-3}$).

which gives

$$R_j(z) - 1 = b_j(T_j(\omega_0 + \omega_1 z) - T_j(\omega_0)), \quad b_j = \frac{T_j''(\omega_0)}{(T_j'(\omega_0))^2}.$$

The recurrence relation (1.2.14) leads to

$$R_j(z) - 1 = \mu_j(R_{j-1}(z) - 1) + \nu_j(R_{j-2}(z) - 1) + \kappa_j z(R_{j-1}(z) - a_{j-1}),$$

where

$$\mu_j = \frac{2b_j\omega_0}{b_j - 1}, \quad \nu_j = \frac{-b_j}{b_{j-2}}, \quad \kappa_j = \frac{2b_j\omega_1}{b_{j-1}}, \quad j = 2, 3, \ldots, s.$$

This formulation finally allows to define the scheme

$$\begin{aligned}
&g_0 - y_0 = 0,\\
&g_1 - y_0 = \kappa_1 \Delta t f(g_0),\\
&g_j - y_0 = \mu_j(g_{j-1} - y_0) + \nu_j(g_{j-2} - y_0) + \kappa_j \Delta t f(g_{j-1}) - a_{j-1}\kappa_j \Delta t f(g_0)
\end{aligned}$$

which is of second order. For the starting coefficients Sommeijer & Verwer suggest to put

$$b_0 = b_2, \quad b_1 = b_2 \quad \text{which gives} \quad \kappa_1 = c_1 = \frac{c_2}{T_2'(\omega_0)} \approx \frac{c_2}{4}.$$

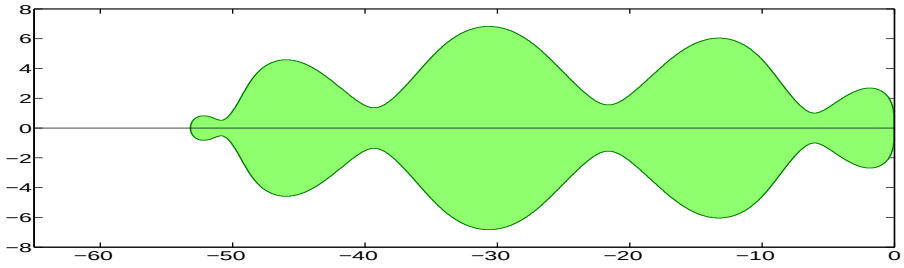


Figure 1.11: Stability domain for the RKC method, $s=9$, following the approach of van der Houwen and Sommeijer.

Chapter 2

Model of Muscle Contraction

2.1 Organization of Striated Muscles

This chapter aims at showing some historically important models that were designed to describe the contractile mechanism having place in striated muscles. We refer to [19] for more details on the description of striated muscles.

Muscle cells have the ability to convert the electrical signal into a mechanical contraction, which enables the muscle cells to perform work. Skeletal and cardiac cells have a banded appearance (they are called striated muscles), and possess similar contractile mechanisms. Each cell is made up of numerous cylindrical structures, called myofibrils, surrounded by the membranous channels of the sarcoplasmic reticulum (Figure 2.1). Myofibrils represent the functional units of striated muscle, and they contain protein filaments that make up the contractile units, the sarcomeres. Each sarcomere, which is about 2.5 μm long, is made up primarily of two types of parallel filaments, thin and thick filaments. A sarcomere is schematically illustrated in Figure 2.2. Each sarcomere is demarcated by two lines called Z-lines. Thin filaments, which are composed primarily of the protein actin, extend from the Z-lines toward the center of the sarcomere, where they overlap with the thick filaments. A cross-sectional view of the sarcomere shows that six thin filaments are placed around each thick filament in a hexagonal arrangement. Viewing a sarcomere along its length allows us to distinguish three different bands, defined by the overlapping or nonoverlapping of thin and thick filaments. The regions where there is no overlap are called I-bands. The area between two I-bands is the A-band, which contains thick filaments composed primarily of the protein myosin. At the ends of this area thin filaments overlap a portion of the thick filaments. Finally, the central region of the sarcomere is called the H-band, and contains only thick filaments. During contraction, both the H-band and the I-bands shorten as the overlap between thin and thick filaments increases. The contractile mechanism of striated muscles is initiated by an action potential transmitted across a synapse from a neuron. This electrical signal spreads rapidly across the muscle membrane, and reaches quickly the cell interior. In cardiac muscles, the action potential causes the voltage-gated Ca^{2+} channels to open, and the Ca^{2+} enters the cell, initiating the release of additional Ca^{2+} from the sarcoplasmic reticulum. The resulting high intracellular concentration of Ca^{2+} causes a change in the myofilament structure that allows the thick filaments to bind and pull on the thin filaments, resulting in muscle contraction. Thick filaments contain myosin, which is a large protein with a globular head. These heads constitute the crossbridges that interact with the thin filaments to form bonds that act in ratchet-like fashion to pull on the thin filaments. Contraction takes place when the crossbridges bind and generate a force causing the thin filaments to slide long the thick ones.

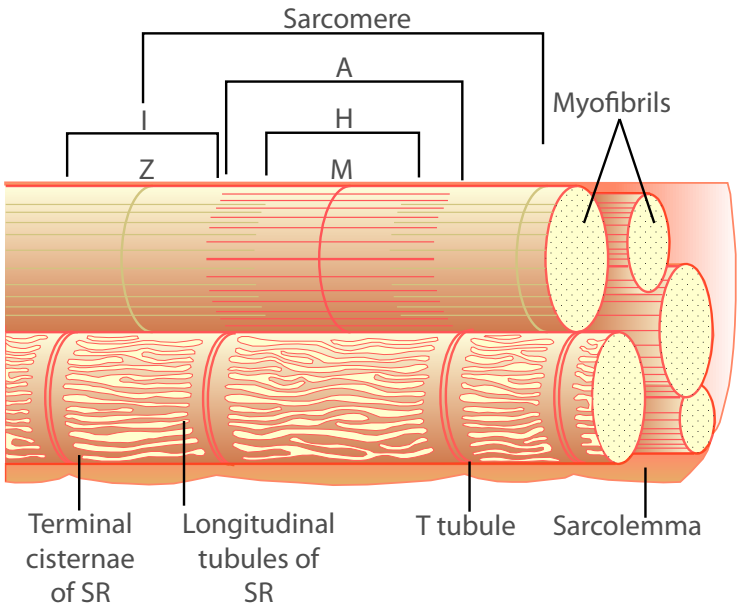


Figure 2.1: Schematic diagram of a skeletal muscle cell. (*Berne & Levy Physiology*, 2010, p. 235, Fig. 12-3 (B)).

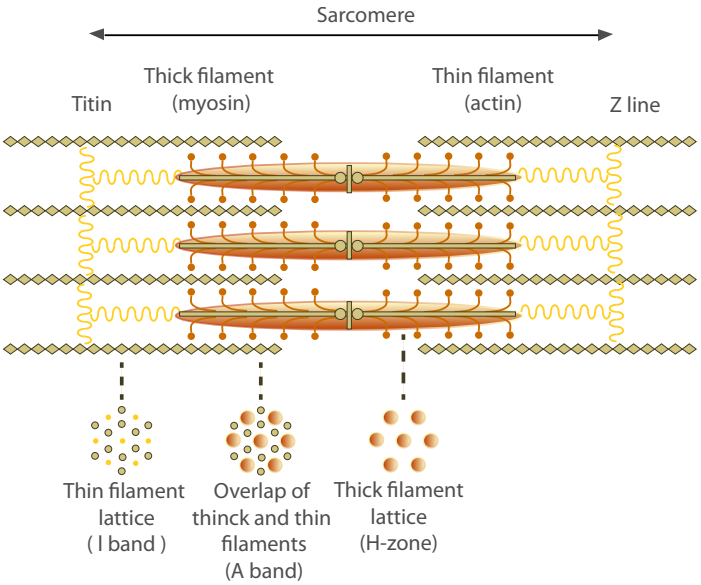


Figure 2.2: Organization of the protein filaments within a single sarcomere. The cross-sectional arrangement of the proteins is also illustrated. (*Berne & Levy Physiology*, 2010, p. 235, Fig. 12-3 (C)).

2.2 Review of Muscle Models

2.2.1 The Force-Velocity Relation: The Hill Model

One of the earliest model muscles activity was proposed by Hill in 1938 [12, 19], before the sarcomere anatomy was known. It is based on the experimental observation that when a muscle contracts against

a constant load (isotonic contraction), the relationship between the rate of shortening v and the load σ is well described by the force-velocity equation

$$(\sigma + a)v = b(\sigma_i - \sigma) \tag{2.2.1}$$

where a and b are two parameters which have to be determined y fitting experimental data. When $v = 0$, we have $\sigma = \sigma_i$, and then σ_i represents the force generated by the muscle when its length is held fixed, and it is called isometric force. On the other hand, when $\sigma = 0$, $v = b\sigma_0/a$, which is the maximum velocity at which the muscle is able to shorten. Then a muscle fibre is modelled as an elastic element, whose length is denoted as x, in series with a contractile element, whose length is denoted as l, so that $L = x + l$ is the total length of the fibre. We denote by v the velocity of contraction of the contractile element ($v = -dl/dt$), which by assumption is related to the load on the muscle by (2.2.1). To derive a differential equation for the time dependence of σ, we start by observing that since x and l are in series, they experience the same force. Moreover we assume that the force σ generated by the elastic element is a function of its length, so that $\sigma = \sigma(x)$, and then by using the chain rule and (2.2.1) we obtain

$$\begin{aligned}\frac{\partial \sigma}{\partial t} &= \frac{\partial \sigma}{dx}\frac{\partial x}{dt} \\ &= \frac{\partial \sigma}{\partial x}\left(\frac{\partial L}{\partial t} - \frac{\partial l}{\partial t}\right) \\ &= \frac{\partial \sigma}{\partial x}\left(\frac{\partial L}{\partial t} + v\right) \\ &= \frac{\partial \sigma}{\partial x}\left(\frac{\partial L}{\partial t} + \frac{b(\sigma_i - \sigma)}{\sigma + a}\right).\end{aligned} \tag{2.2.2}$$

It only remains to choose a law for $\sigma(x)$. Hill made the simplest choice, by assuming that the elastic element is linear, and thus

$$\sigma(x) = \alpha(x - x_0),$$

where x_0 represents its resting length. Then the differential equation for σ becomes

$$\frac{\partial \sigma}{\partial t} = \alpha\left(\frac{\partial L}{\partial t} + \frac{b(\sigma - \sigma_i)}{\sigma + a}\right). \tag{2.2.3}$$

Hill's model, although historically important, was shown to have many important defects. In particular the fact that the force-velocity relation is satisfied immediately after a change in tension is a probable major source of error. Moreover the discovery of the sarcomere anatomy motivated the construction of new completely different models, based on the kinetics of crossbridges rather than on heuristic elastic and contractile elements. The first model of this type was proposed by Huxley in 1957 [17, 18] and it is the subject of the next section.

2.2.2 The Huxley Crossbridge Model

To construct quantitative models of crossbridge binding it is necessary to know how many binding sites are available to a single crossbridge. One possibility is that at any time only one single binding site is available to each crossbridge. This is the assumption behind the Huxley model. It is supposed that a crossbridge can bind to an actin site at position x, where x denotes the distance along the thin filament to a binding site from the crossbridge's hinge. At $x = 0$ the bound crossbridge exerts no force during the power stroke on the thin filament. Crossbridges can be bound at $x > 0$, in which case they exert a contractile force, or at $x < 0$, in which case they exert a force that oppose contraction. Saying that each crossbridge can be bound with one and only one binding site is physically equivalent to making the assumption that the acting binding sites are sufficiently far apart. Therefore each crossbridge, whether bound or not, can be associated with a unique value of x. Let M denote the number of crossbridges (either bound or unbound) with displacement x. It is assumed that the binding is restricted to occur within some limited interval, $-l \le x \le l$, and that M is a constant independent on that interval. Thus,

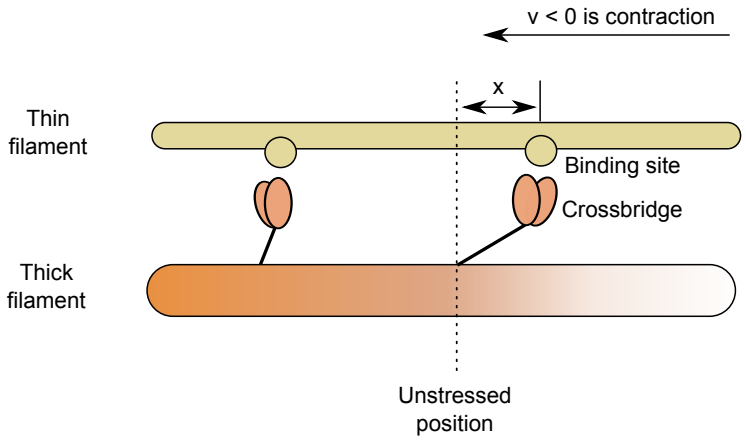


Figure 2.3: Schematic diagram of the Huxley crossbridge model. (Keener, *Mathematical Physiology*, 2008, p.730, Fig. 15.12).

for each displacement x, the total number of crossbridges with that displacement is conserved, and the case that all the crossbridges could end up with the same displacement is excluded. We denote as $n(x,t)$ the fraction of crossbridges with displacement x that are bound at time t. We assume that each crossbridge can be in one of two states, namely either unbound (U), or bound (B) and thereby generating a force, and that the binding and unbinding of crossbridges is described by the following reaction scheme

$$\mathrm{U} \xrightarrow{f(x,t)} \mathrm{B},$$
$$\mathrm{U} \xleftarrow{g(x,t)} \mathrm{B},$$

where $f(x,t)$ and $g(x,t)$ are the positive rates at which crossbridges bound and unbound respectively. To derive a conservation law for the fraction of bound crossbridges, we start by considering the total number of crossbridges that are bound with displacement x within an arbitrary interval $[a,b]$. This total number is given by

$$M\int_a^b n(x,t)\,dx.$$

The rate of change of this total number is given by the reactions of the crossbridges with the actin filaments as well as by their flux across the boundaries of the interval $[a,b]$. Let us denote with ε the strain, and let us consider $\dot{\varepsilon}$ as the velocity of the thin actin filament, relative to the thick filament ($\dot{\varepsilon}<0$ means contraction). During contraction, at $x=a$ the flux of crossbridges out of the domain is $-M\dot{\varepsilon}(t)n(a,t)$, while at $x=b$ the flux of crossbridges that enter the domain is $-M\dot{\varepsilon}(t)n(b,t)$. The conservation of crossbridges gives

$$M\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b n(x,t)\,\mathrm{d}x = -M\dot{\varepsilon}(t)n(b,t) + M\dot{\varepsilon}(t)n(a,t) + M\int_a^b [f(x,t)(1-n(x,t)) - g(x,t)n(x,t)]\,\mathrm{d}x,$$

and so

$$\int_a^b \frac{\partial}{\partial t}n(x,t)\,\mathrm{d}x = -\int_a^b \dot{\varepsilon}(t)\frac{\partial}{\partial x}n(x,t)\,\mathrm{d}x + \int_a^b [f(x,t)(1-n(x,t)) - g(x,t)n(x,t)]\,\mathrm{d}x.$$

Since a and b are arbitrary, the integral can be dropped to finally obtain

$$\frac{\partial}{\partial t}n(x,t) = -\dot{\varepsilon}(t)\frac{\partial}{\partial x}n(x,t) + f(x,t)(1-n(x,t)) - g(x,t)n(x,t). \tag{2.2.4}$$

It is also supposed that a bound crossbridge acts like a spring, generating a restoring force $r(x) = -\partial W/\partial x$ related to its displacement, where W is the elastic free energy of the actin-myosin interaction responsible of muscle contraction. Then the total force generated by the muscle is defined as

$$\sigma(t) = -\int_{-\infty}^{+\infty} \frac{\partial W(x)}{\partial x} n(x,t)\,\mathrm{d}x. \tag{2.2.5}$$

To find the force-velocity relationship, we assume that the fiber moves with constant velocity so that $\dot{\varepsilon}(t) = v$, and that $n(x,t)$ is equilibrated so that $\partial n/\partial t = 0$. We then seek for steady solutions of (2.2.4), after having made a reasonable guess on the form of the functions f and g. Huxley chose the following expression for $f(x)$ and $g(x)$:

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ f_1 x/h & \text{if } 0 \le x \le h, \\ 0 & \text{if } x > h, \end{cases} \qquad g(x) = \begin{cases} g_2 & \text{if } x \le 0, \\ g_1 x/h & \text{if } x > 0, \end{cases}$$

where f_1, g_1, g_2 and h are constant parameters. In particular we assume that for a certain value h the rate of crossbridge attachment falls to zero, since crossbridges can not attach to binding sites that are too far away. In [3] it is stated that using $f_1 = 65\ \mathrm{s}^{-1}$, $g_1 = 15\ \mathrm{s}^{-1}$, $g_2 = 313.5\ \mathrm{s}^{-1}$ and $h = 10$ nm gives numerical results in good agreement with experimental data. Moreover one can show that for these parameters $|v|_{max} \approx 2(f_1 + g_1)h$. Assuming that the crossbridges behaves like a linear spring, so that $r(x) = \sigma_0 x$ for some constant σ_0, the total force exterted by the muscle can be calculated as function of v. Let us note that for this particular choice, as already stated in [34], the total stiffness k and total stress σ are respectively proportional to the zero and first-order momentum of n:

$$k(t) = k_0 \int_{-\infty}^{+\infty} n(x,t)\,\mathrm{d}x \quad \text{and} \quad \sigma(t) = \sigma_0 \int_{-\infty}^{+\infty} x n(x,t)\,\mathrm{d}x. \tag{2.2.6}$$

Some steady distributions of n are shown in Figure 2.4 while in Figure 2.5 is shown the force-velocity curve of the Huxley model.

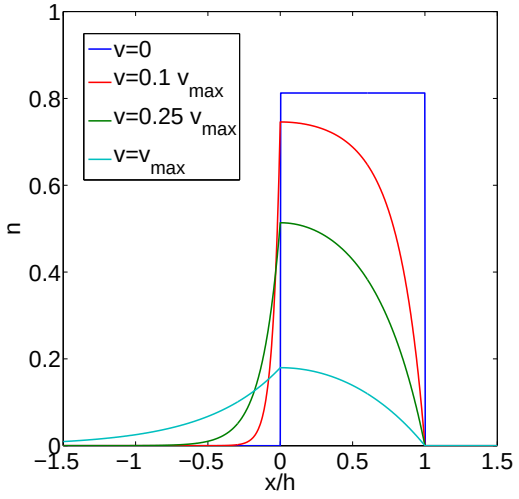


Figure 2.4: Steady state solution of n in the Huxley model, for different values of v, aas function of dimensionless space x/h.

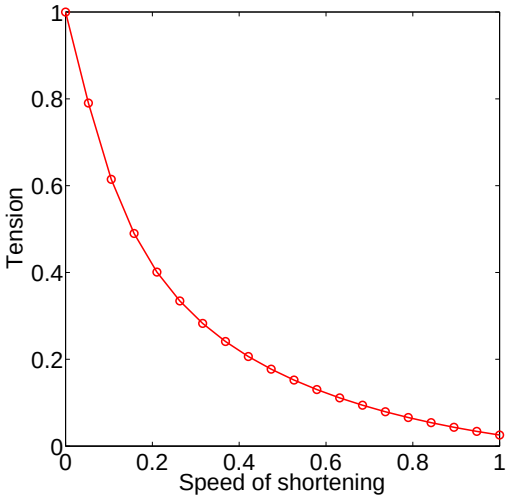


Figure 2.5: The force-velocity curve of the Huxley model. Here σ_0 is determined by enforcing $\sigma(0) = 1$. Further, the parameter have been scaled so that $v_{max} = 1$.

Let us remark that the Huxley model was designed to describe the functioning of muscle contraction on the microscopic scale, i.e., the sarcomere scale. Moreover there is no physiological justification on the form used so far for the functions f and g, and the model, as it was just presented, is not directly coupleable with the other electromechanical phenomena happening. We discuss then in the next section a model of muscle contraction, proposed by Bestel Clément and Sorine [2], based on multi-scale analysis of the Huxley model, which provides new expressions for f and g, and which is devoted to describing the crossbridges kinetics on the myofibre scale.

2.3 The Bestel-Clément-Sorine Excitation-Contraction Model

Bestel, Clément and Sorine started by making the following choice for the two positive rates at which respectively crossbridges bound and unbound

$$f(x,t) = c(t)^+ \text{ if } x \in [0,1] \ (= 0 \text{ elsewhere}) \quad \text{and} \quad g(x,t) = |c(t)| + |\dot{\varepsilon}(t)| - f(x,t), \tag{2.3.1}$$

where x denotes again the crossbridges displacement, c represents a chemical input mainly depending on calcium concentration (and therefore on the electrical signal), and

$$c^+ = \begin{cases} c & \text{if } c \geq 0, \\ 0 & \text{if } c < 0. \end{cases}$$

As presented in the previous section the contraction of the sarcomere can be described by the Huxley crossbridges model (2.2.4). The sarcomere represents the contractile unit of the whole muscle, which can be thought as a multi-scale structure mainly composed of myocites (the muscular cells). Each myocite in turn is made of numerous cylindrical structures, the myofibrils, which are divided into sarcomeres by the Z-lines (Figure 2.1). Statistical analysis allows to describe the contraction at the macroscopic scale, by averaging over all volume's crossbridges the Huxley equation, resulting in a set of ODEs modeling the control of the total stiffness k and total stress σ by the chemical input c and the strain ε. With the new formulations for f and g, the model (2.2.4) can be rewritten as

$$\frac{\partial}{\partial t} n(x,t) = -\dot{\varepsilon}(t) \frac{\partial}{\partial x} n(x,t) - (|c(t)| + |\dot{\varepsilon}(t)|) n(x,t) + c(t)^+. \tag{2.3.2}$$

We start by deriving the ODE for the total stiffness k. Integration over $\mathbb{R}$ produces

$$\begin{aligned}\int_{-\infty}^{+\infty} \frac{\partial}{\partial t} n(x,t)\,\mathrm{d}x &= -\int_{-\infty}^{+\infty} \dot{\varepsilon}(t)\frac{\partial}{\partial x} n(x,t)\,\mathrm{d}x - \int_{-\infty}^{+\infty} (|c(t)| + |\dot{\varepsilon}(t)|)\, n(x,t)\,\mathrm{d}x + \int_{-\infty}^{+\infty} c(t)^+\,\mathrm{d}x \\ &= 0 - \int_{-\infty}^{+\infty} (|c(t)| + |\dot{\varepsilon}(t)|)\, n(x,t)\,\mathrm{d}x + c(t)^+.\end{aligned}$$

We assume that it is allowed to carry out the time derivation from the integral operator. Then it suffices to multiply the equation by k_0 and to use (2.2.6) to obtain

$$\frac{\mathrm{d}k(t)}{\mathrm{d}t} = -(|c(t)| + |\dot{\varepsilon}(t)|)\,k(t) + k_0\, c(t)^+. \tag{2.3.3}$$

For the total stiffness we proceed in a similar way. We multiply by x equation (2.3.2), and again we integrate over $\mathbb{R}$, to obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} x\frac{\partial}{\partial t} n(x,t)\,\mathrm{d}x &= -\int_{-\infty}^{+\infty} x\dot{\varepsilon}(t)\frac{\partial}{\partial x} n(x,t)\,\mathrm{d}x - \int_{-\infty}^{+\infty} x(|c(t)| + |\dot{\varepsilon}(t)|)\, n(x,t)\,\mathrm{d}x + \int_{-\infty}^{+\infty} x c(t)^+\,\mathrm{d}x \\ &= 0 + \int_{-\infty}^{+\infty} \dot{\varepsilon}(t)\, n(x,t)\,\mathrm{d}x - \int_{-\infty}^{+\infty} x(|c(t)| + |\dot{\varepsilon}(t)|)\, n(x,t)\,\mathrm{d}x + \frac{c(t)^+}{2}.\end{aligned}$$

Then, as before, we carry out the time derivation from the integral operator, and we multiply the equation by σ_0. By using (2.2.6) we finally have

$$\frac{\mathrm{d}\sigma(t)}{\mathrm{d}t} = \frac{\sigma_0}{k_0}\dot{\varepsilon}(t)k(t) - (|c(t)| + |\dot{\varepsilon}(t)|)\sigma(t) + \sigma_0\frac{c(t)^+}{2}. \tag{2.3.4}$$

Then as final result, we obtained, as constitutive relation between the total stress σ and the strain ε, the following set of ODEs:

$$\begin{cases} \dfrac{\mathrm{d}k(t)}{\mathrm{d}t} = -(|c(t)| + |\dot{\varepsilon}(t)|)\,k(t) + k_0\, c(t)^+, & t \geq 0, \\ \dfrac{\mathrm{d}\sigma(t)}{\mathrm{d}t} = \dfrac{\sigma_0}{k_0}\dot{\varepsilon}(t)k(t) - (|c(t)| + |\dot{\varepsilon}(t)|)\sigma(t) + \sigma_0\dfrac{c(t)^+}{2}, & t \geq 0, \\ k(0) = \sigma(0) = 0. \end{cases} \tag{2.3.5}$$

In particular this system is capable of accounting for: shortening from resting condition ($\dot{\varepsilon}(t) = 0$) in response to $c(t)$, and static relation ($\mathrm{d}\sigma/\mathrm{d}t = 0$) between $\dot{\varepsilon}$ and σ as in Hill's experimental model (Figure 2.6).

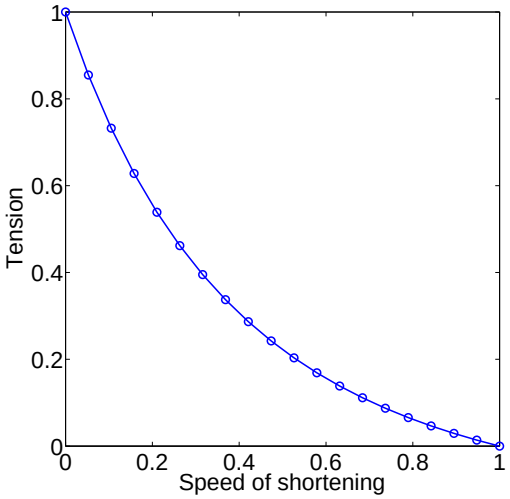


Figure 2.6: The force-velocity curve of the Bestel-Clément-Sorine model. For these figure we used $c = 1$, $\sigma_0 = 2$, $k_0 = 2$.

Chapter 3

Review of Continuum Mechanics

3.1 Kinematics

3.1.1 Description of Motion of Material Points in a Body

Let us consider a physical body, namely $\mathscr{B}$, and we idealize it as composed of as a continuous set of material points. We embed $\mathscr{B}$ in a three-dimensional Euclidean space, and we use $\mathscr{R}(t)$ to denote the body's configuration at time t in such space. However we choose a particular configuration, let us call it $\mathscr{R}_0$, that serves as a reference configuration, enabling us to uniquely identify an arbitrary point P in $\mathscr{B}$ by its position $\mathbf{X}$ in this configuration. As example we could choose $\mathscr{R}_0$ as the body's configuration at time zero. During motion an arbitrary particle P, located at $\mathbf{X}$ in the reference configuration $\mathscr{R}_0$, moves to position $\mathbf{x}$ in the configuration $\mathscr{R}(t)$. Let us call $\mathbf{X}$ as referential position and $\mathbf{x}$ as current position. It is assumed that the body's motion can be described by a relationship of the form

$$\mathbf{x} = \boldsymbol{\phi}(\mathbf{X}, t). \tag{3.1.1}$$

With respect to the function (3.1.1), we make the following mathematical assumptions:

1. $\boldsymbol{\phi}(\mathbf{X}, t)$ is continuously differentiable in all variables, at least up through the second order.

2. At each time $t \geq 0$, the following property holds: for any $\mathbf{X}$ and corresponding $\mathbf{x}$, there are open balls $B_{\mathbf{X}}$ and $B_{\mathbf{x}}$ (respectively centred at $\mathbf{X}$ and $\mathbf{x}$), both contained in $\mathscr{B}$, such that points of $B_{\mathbf{X}}$ are in one-to-one correspondence with points of $B_{\mathbf{x}}$.

In particular the first assumption allows us to define quantities such as velocity and acceleration, while the second one implies that the body is not penetrable, and no voids can be created in the body. The second assumption implies also that the Jacobian of the transformation (3.1.1),

$$J = \det\left(\frac{\partial \boldsymbol{\phi}(\mathbf{X}, t)}{\partial \mathbf{X}}\right),$$

is non-zero for all times $t \geq 0$. If at any time t^*, $\mathscr{R}(t^*) = \mathscr{R}_0$, then at this time $\mathbf{x} = \mathbf{X}$, and therefore $J(t^*) = 1$. However, as already stated, J is continuous and can never vanish, and then if J is positive once, it will stay positive for all times:

$$J(t) > 0 \quad \forall\, t \geq 0.$$

Furthermore we can introduce the inverse of transformation (3.1.1)

$$\mathbf{X} = \boldsymbol{\phi}^{-1}(\mathbf{x}, t).$$

3.1.2 Referential and Spatial Description

As already stated, it is assumed that the position vectors $\mathbf{X}$ and $\mathbf{x}$ are in one-to-one correspondence. Therefore field variables, such as the density, can be either written as function of $\mathbf{X}$ and t (referential or Lagrangian description) or as function of $\mathbf{x}$ and t (spatial or Eulerian description), that is

$$\rho = \bar{\rho}(\mathbf{X}, t) = \hat{\rho}(\mathbf{x}, t).$$

The two description can be related between them. Indeed we have

$$\hat{\rho}(\mathbf{x}, t) = \hat{\rho}(\boldsymbol{\phi}(\mathbf{X}, t), t) = \bar{\rho}(\mathbf{X}, t), \quad \text{and} \quad \bar{\rho}(\mathbf{X}, t) = \bar{\rho}(\boldsymbol{\phi}^{-1}(\mathbf{x}, t), t) = \hat{\rho}(\mathbf{x}, t).$$

When discussing quantities that involve spatial derivatives, such as gradient, divergence or curl, it is convenient to define some notation that allows to distinguish the cases when we are using referential or current spatial variables. We decide to use the subscript 0 to mean the use of the referential spatial variables, and then make distinction from current spatial ones. As example, for a generic field variable $f = \bar{f}(\mathbf{X}, t) = \hat{f}(\mathbf{x}, t)$ we have,

$$\nabla_0 f = \frac{\partial \bar{f}(\mathbf{X}, t)}{\partial \mathbf{X}} \quad \text{and} \quad \nabla f = \frac{\partial \hat{f}(\mathbf{x}, t)}{\partial \mathbf{x}}.$$

3.1.3 Displacement, Velocity and Acceleration

The displacement $\mathbf{u}$ of a particle P at time t it is defined as

$$\mathbf{u} = \mathbf{x} - \mathbf{X}.$$

In particular we can either write the displacement using the Eulerian description, which reads

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{x} - \boldsymbol{\phi}^{-1}(\mathbf{x}, t),$$

or, alternatively, using the Lagrangian description

$$\mathbf{u}(\mathbf{X}, t) = \boldsymbol{\phi}(\mathbf{X}, t) - \mathbf{X}.$$

The velocity $\mathbf{v}$ of a material point P, it is obtained as the time derivative of the displacement. Using the Lagrangian description we have

$$\mathbf{v}(\mathbf{X}, t) = \frac{\partial \mathbf{u}(\mathbf{X}, t)}{\partial t} = \frac{\partial \boldsymbol{\phi}(\mathbf{X}, t)}{\partial t}.$$

The acceleration $\mathbf{a}$ is defined as the time derivative of the velocity, and then we have

$$\mathbf{a}(\mathbf{X}, t) = \frac{\partial \mathbf{v}(\mathbf{X}, t)}{\partial t} = \frac{\partial^2 \boldsymbol{\phi}(\mathbf{X}, t)}{\partial t^2}.$$

Of course, both velocity and acceleration can be written using the Eulerian description. For example for the velocity we have

$$\mathbf{v} = \bar{\mathbf{v}}(\mathbf{X}, t) = \bar{\mathbf{v}}(\boldsymbol{\phi}^{-1}(\mathbf{x}, t), t) = \hat{\mathbf{v}}(\mathbf{x}, t).$$

The material derivative of a field such as the density is defined as the partial derivative of such variable with respect to time, holding the material point P fixed

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\partial \bar{\rho}(\mathbf{X}, t)}{\partial t}. \tag{3.1.2}$$

Frequently we work with the Eulerian description of functions without knowledge of the deformation (3.1.1), and therefore it is not possible to evaluate directly (3.1.2). However, by using the chain

rule, the material derivative can be written for the spatial formulation of a field variable. Indeed for the density we obtain

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\partial\hat{\rho}(\mathbf{x},t)}{\partial t} + \mathbf{v}(\mathbf{x},t)\cdot\nabla\rho.$$

Therefore, depending on the type of description we are using, we have two choices for the material derivative

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \begin{cases} \dfrac{\partial\bar{\rho}(\mathbf{X},t)}{\partial t} & \text{Lagrangian Representation,} \\ \dfrac{\partial\hat{\rho}(\mathbf{x},t)}{\partial t} + \mathbf{v}(\mathbf{x},t)\cdot\nabla\rho & \text{Eulerian Representation.} \end{cases}$$

Then the acceleration, which is defined as the material derivative of the velocity field, can be written in the two following forms:

$$\mathbf{a} = \begin{cases} \dfrac{\partial\mathbf{v}(\mathbf{X},t)}{\partial t}, \\ \dfrac{\partial\mathbf{v}(\mathbf{x},t)}{\partial t} + (\nabla\mathbf{v}(\mathbf{x},t))\mathbf{v}(\mathbf{x},t). \end{cases}$$

3.1.4 Deformation Gradient of the Motion

Given a deformation $\boldsymbol{\phi}$, it is possible to define the deformation gradient tensor, that we denote as $\mathbf{F}$, as

$$\mathbf{F} = \frac{\partial\boldsymbol{\phi}(\mathbf{X},t)}{\partial\mathbf{X}} = \frac{\partial\mathbf{x}}{\partial\mathbf{X}} \quad \text{or, using index notation,} \quad F_{ij} = \frac{\partial\phi_i(\mathbf{X},t)}{\partial X_j} = \frac{\partial x_i}{\partial X_j}. \tag{3.1.3}$$

The inverse gradient tensor $\mathbf{F}^{-1}$ is given by

$$\mathbf{F}^{-1} = \frac{\partial\boldsymbol{\phi}^{-1}(\mathbf{x},t)}{\partial\mathbf{x}} = \frac{\partial\mathbf{X}}{\partial\mathbf{x}} \quad \text{or} \quad F^{-1}_{ij} = \frac{\partial\phi_i^{-1}(\mathbf{x},t)}{\partial\mathbf{x}} = \frac{\partial X_i}{\partial x_j}.$$

Using this notation, it can be shown that during motion each infinitesimal vector $\mathbf{dX}$ in the reference configuration is transformed in the infinitesimal vector $\mathbf{dx}$ through the following relation

$$\mathbf{dx} = \mathbf{F}\mathbf{dX} \quad \text{or} \quad \mathrm{d}x_i = F_{ij}\mathrm{d}X_j. \tag{3.1.4}$$

Then the deformation gradient tensor $\mathbf{F}$ determines how the infinitesimal vector are deformed during motion. $\mathbf{F}$ and $\mathbf{F}^{-1}$ are related to the gradient of the displacement through the following relations:

$$\mathbf{F} = \nabla_0\mathbf{u} + \mathbf{I} \quad \text{and} \quad \mathbf{F}^{-1} = \mathbf{I} - \nabla\mathbf{u}. \tag{3.1.5}$$

3.1.5 Transformations of Infinitesimal Areas and Volumes During Deformation

The relationship (3.1.4) can be used to calculate the changes in material area and volume during the deformation. If we consider two distinct material line elements $\mathbf{dX}$ and $\mathbf{dY}$ at $\mathbf{X}$ in the reference configuration $\mathscr{R}_0$, the infinitesimal material areas $\mathbf{dA}$ in $\mathscr{R}_0$ and $\mathbf{da}$ in $\mathscr{R}(t)$ are defined as

$$\mathbf{dA} = \mathbf{dY}\times\mathbf{dX} \quad \text{and} \quad \mathbf{da} = \mathbf{dy}\times\mathbf{dx}.$$

From (3.1.4) follows the so-called Nanson's Formula:

$$\mathbf{da} = J\mathbf{F}^{-T}\mathbf{dA}. \tag{3.1.6}$$

In order to find a formulation describing changes in material volume, we consider three material line elements $\mathbf{dX}$, $\mathbf{dY}$ and $\mathbf{dZ}$, and the infinitesimal volumes $\mathbf{dV}$ in $\mathscr{R}_0$ and $\mathbf{dv}$ in $\mathscr{R}(t)$ are defined by

$$\mathrm{d}V = \mathbf{dX}\cdot(\mathbf{dY}\times\mathbf{dZ}) \quad \text{and} \quad \mathrm{d}v = \mathbf{dx}\cdot(\mathbf{dy}\times\mathbf{dz}).$$

Again using (3.1.4) we obtain that the transformation of infinitesimal volumes is given by

$$\mathrm{d}v = J\mathrm{d}V.$$

3.1.6 Measure of Stretch and Strain

Let $\mathbf{dX}$ be an infinitesimal material element in the configuration $\mathscr{R}_0$, so that $\mathbf{dX} = \mathrm{d}S\mathbf{a}_0$, where $\mathbf{a}_0$ is a unit vector. During deformation $\mathbf{dX}$ is mapped to $\mathbf{dx} = \mathrm{d}s\mathbf{a}$, where also $\mathbf{a}$ denotes a unit vector. The stretch undergone by the material is defined as $\lambda = |\mathbf{dx}|/|\mathbf{dX}| = \mathrm{d}s/\mathrm{d}S$. Using (3.1.4) we can rewrite the magnitude $\mathrm{d}s$ of the element $\mathbf{dx}$ as

$$\mathrm{d}s^2 = |\mathbf{dx}|^2 = \mathbf{dx}\cdot\mathbf{dx} = \mathbf{a}_0\cdot(\mathbf{F}^T\mathbf{F}\mathbf{a}_0)\mathrm{d}S^2.$$

Recalling the definition of the right Cauchy-Green tensor,

$$\mathbf{C} = \mathbf{F}^T\mathbf{F} \quad \text{or} \quad C_{ij} = F_{ki}F_{kj}, \tag{3.1.7}$$

we therefore have that

$$\lambda^2 = \mathbf{a}_0\cdot(\mathbf{C}\mathbf{a}_0).$$

Then the stretch of a material element located at $\mathbf{X}$ in $\mathscr{R}_0$, can be calculated at any time t only with the knowledge of its orientation in $\mathscr{R}_0$ and the right Cauchy-Green tensor $\mathbf{C}$. It can be shown that $\mathbf{C}$ is symmetric and positive definite, that is

$$\mathbf{u}\cdot(\mathbf{C}\mathbf{u}) \geq 0, \quad \forall\, \mathbf{u} \neq \mathbf{0}.$$

If we consider the special case $\mathbf{dX} = \mathrm{d}S\mathbf{e}_i$, where $\mathbf{e}_i$ is a unit vector aligned with one of the axis of the Euclidean space in which the material body is embed in, it follows that $\lambda^2 = C_{ii}$. Then each diagonal element of $\mathbf{C}$ represents the stretch squared of an infinitesimal element $\mathbf{dX}$, if it was aligned with the axis $\mathbf{e}_i$ in the reference configuration $\mathscr{R}_0$.

To understand the physical significance of the off-diagonal entries of $\mathbf{C}$, we consider two infinitesimal elements $\mathbf{dX} = \mathrm{d}S^x\mathbf{a}_0^x$ and $\mathbf{dY} = \mathrm{d}S^y\mathbf{a}_0^y$. In the current configuration $\mathscr{R}(t)$, these two elements are mapped respectively to $\mathbf{dx} = \mathrm{d}s^x\mathbf{a}^x$ and $\mathbf{dy} = \mathrm{d}s^y\mathbf{a}^y$. If we denote as α the angle between the two elements in the current configuration, we have that

$$\cos(\alpha) = \frac{\mathbf{dx}\cdot\mathbf{dy}}{\mathrm{d}s^x\mathrm{d}s^y} = \frac{C_{ij}a_{0,i}^x a_{0,j}^y}{\lambda^x\lambda^y}.$$

If we consider now the special case in which the two elements are perpendicular between them and aligned with two axis of the Euclidean space in $\mathscr{R}_0$, for example $\mathbf{a}_0^x = \mathbf{e}_1$ and $\mathbf{a}_0^y = \mathbf{e}_2$, we get

$$\cos(\alpha) = \frac{C_{12}}{\sqrt{C_{11}C_{22}}}.$$

In addition to the stretch undergone by a material element, we can consider its strain. There is no unique definition for the strain, but a measure of interest is the change in the square of the magnitude of the infinitesimal material element, relative to its magnitude squared in $\mathscr{R}_0$, that is

$$\frac{\mathrm{d}s^2 - \mathrm{d}S^2}{\mathrm{d}S^2}.$$

This quantity can be related to the right Cauchy-Green tensor as

$$\frac{\mathrm{d}s^2 - \mathrm{d}S^2}{\mathrm{d}S^2} = \lambda^2 - 1 = \mathbf{a}_0\cdot(\mathbf{C} - \mathbf{I})\mathbf{a}_0.$$

Alternatively, we can consider stretch and strain of material elements, calculated solely from knowledge of their direction in $\mathscr{R}(t)$. As example, a measure of strain relative to $\mathscr{R}(t)$ is given by

$$\frac{\mathrm{d}s^2 - \mathrm{d}S^2}{\mathrm{d}s^2}.$$

Let us note that dS can be written in terms of ds and $\mathbf{a}$, defined in $\mathscr{R}(t)$. Indeed we have

$$dS^2 = |\mathbf{dX}| = \mathbf{dX}\cdot\mathbf{dX} = \mathbf{a}\cdot(\mathbf{F}^{-T}\mathbf{F}^{-1}\mathbf{a})ds^2 = \mathbf{a}\cdot(\mathbf{B}^{-1}\mathbf{a})ds^2, \tag{3.1.8}$$

where $\mathbf{B}$ is the left Cauchy-Green tensor

$$\mathbf{B} = \mathbf{F}\mathbf{F}^T \quad \text{or} \quad b_{ij} = F_{ik}F_{jk}.$$

Also $\mathbf{B}$ is symmetric and positive definite. From (3.1.8) we get

$$\frac{1}{\lambda^2} = \mathbf{a}\cdot\mathbf{B}^{-1}\mathbf{a},$$

which enables us to calculate the stretch undergone by an infinitesimal material element only with the knowledge of its orientation $\mathbf{a}$ in $\mathscr{R}(t)$ and $\mathbf{B}$.

3.1.7 Polar Decomposition of the Deformation Gradient Tensor

The Polar Decomposition Theorem states that an arbitrary non singular tensor can be decomposed as $\mathbf{F} = \mathbf{RU} = \mathbf{VR}$, where $\mathbf{U}$ and $\mathbf{V}$ are symmetric and positive definite, while $\mathbf{R}$ is an orthogonal tensor. When this theorem is applied to the Deformation Gradient Tensor, $\mathbf{R}$ is called the rotation tensor. On the other hand $\mathbf{U}$ and $\mathbf{V}$ are respectively referred to as the right and left stretch tensor. Then it follows that for the right and left Cauchy-Green tensors we have

$$\mathbf{C} = \mathbf{F}^T\mathbf{F} = \mathbf{U}^2 \quad \text{and} \quad \mathbf{B} = \mathbf{F}\mathbf{F}^T = \mathbf{V}^2.$$

Noting that

$$\mathbf{dx} = \mathbf{R}(\mathbf{UdX}),$$

if we denote as $\mathbf{dX}'$ the product $\mathbf{UdX}$, we have that

$$\mathbf{dx} = \mathbf{RdX}',$$

and therefore

$$|\mathbf{dx}| = |\mathbf{RdX}'| = \sqrt{\mathbf{dX}'^T\mathbf{R}^T\mathbf{RdX}'} = |\mathbf{dX}'|.$$

Then it is clear that $\mathbf{R}$ does not contribute to the stretch of the infinitesimal material element, but only to the rotation. The two tensors $\mathbf{U}$ and $\mathbf{V}$ are related through

$$\mathbf{U} = \mathbf{R}^T\mathbf{VR}.$$

3.1.8 Velocity Gradient

We denote as $\mathbf{L}$ the gradient of spatial form of the velocity vector, so that the entries of $\mathbf{L}$ are given by

$$L_{ij} = \frac{\partial v_i(\mathbf{x},t)}{\partial x_j}.$$

Recalling that any second order tensor can be decomposed into the sum of a symmetric and skew symmetric second order tensor, we can represent $\mathbf{L}$ as

$$L_{ij} = \underbrace{\frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right)}_{D_{ij}} + \underbrace{\frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} - \frac{\partial v_j}{\partial x_i}\right)}_{W_{ij}}.$$

We denote the symmetric part of $\mathbf{L}$ as $\mathbf{D}$,

$$\mathbf{D} = \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T). \tag{3.1.9}$$

D is often referred to as the rate of deformation tensor or the rate of strain tensor. On the other hand we denote as **W** the skew symmetric part of **L**,

$$\mathbf{W} = \frac{1}{2}(\nabla\mathbf{v} - \nabla\mathbf{v}^T).$$

W is called spin tensor or vorticity tensor.

To study the physical significance of **D**, we consider the rate of change in magnitude of an infinitesimal material element **dx** of length $\mathrm{d}s$. For the rate of change of $\mathrm{d}s^2$ we have that

$$\frac{\mathrm{d}(\mathrm{d}s^2)}{\mathrm{d}t} = 2\mathrm{d}x_i \frac{\mathrm{d}(\mathrm{d}x_i)}{\mathrm{d}t}.$$

Furthermore we have that the rate of change in the infinitesimal material element $\mathrm{d}x$ is

$$\frac{\mathrm{d}(\mathrm{d}x_i)}{\mathrm{d}t} = L_{ij}\mathrm{d}x_j,$$

and so we find

$$\frac{\mathrm{d}(\mathrm{d}s^2)}{\mathrm{d}t} = 2\mathbf{dx}\cdot(\mathbf{Ddx}).$$

Thus the last equation implies

$$\frac{\mathrm{d}(\mathrm{d}s)}{\mathrm{d}t} = \frac{\mathbf{dx}\cdot(\mathbf{Ddx})}{\mathrm{d}s}.$$

Now by employing special choices for **dx** we can interpret the meaning of each entry of **D**. By choosing $\mathbf{dx} = \mathrm{d}s\mathbf{e}_1$ we get

$$\frac{\mathrm{d}(\mathrm{d}s)}{\mathrm{d}t} = D_{11}\mathrm{d}s.$$

Then D_{11} is the rate of change of the magnitude $\mathrm{d}s$, divided by $\mathrm{d}s$, of a material element which is aligned with the axis $\mathbf{e}_1$ at time t. On the other hand let us consider two infinitesimal material elements **dx** and **dy**, respectively of length $\mathrm{d}s^x$ and $\mathrm{d}s^y$. Let us call α the angle they intersect. We have that

$$\cos(\alpha) = \frac{\mathbf{dx}\cdot\mathbf{dy}}{\mathrm{d}s^x\mathrm{d}s^y},$$

and therefore

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\cos(\alpha) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{dx}\cdot\mathbf{dy}}{\mathrm{d}s^x\mathrm{d}s^y}\right)\\ &= 2\frac{\mathbf{dy}\cdot(\mathbf{Ddx})}{\mathrm{d}s^x\mathrm{d}s^y} - \frac{\mathbf{dx}\cdot\mathbf{dy}}{(\mathrm{d}s^x\mathrm{d}s^y)^2}\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{d}s^x\mathrm{d}s^y).\end{aligned}$$

By choosing **dx** and **dy** respectively aligned as $\mathbf{e}_1$ and $\mathbf{e}_2$ at time t, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\cos(\alpha) = -2D_{12},$$

and then D_{12} is proportional to the rate of change of the angle between two elements aligned respectively as $\mathbf{e}_1$ and $\mathbf{e}_2$ at time t.

3.2 Governing Equations

3.2.1 The Transport Theorem

Before discussing the governing equations of the mechanics, it is useful to review the transport theorem. The transport theorem for an arbitrary scalar function φ of position **x** and time t is

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)}\varphi(\mathbf{x},t)\,\mathrm{d}v = \int_{\mathcal{V}(t)}\left(\frac{\partial\varphi(\mathbf{x},t)}{\partial t} + \varphi(\mathbf{x},t)\nabla\cdot\mathbf{v}(\mathbf{x},t)\right)\mathrm{d}v,$$

where **v** is the velocity vector and $\mathcal{V}(t)$ is an arbitrary material volume of the body.

3.2.2 Conservation of Mass

The mass $\mathscr{M}$ of a volume section $\mathscr{V}$ at time t is given by

$$\mathscr{M}(t) = \int_{\mathscr{V}(t)} \rho(\mathbf{x}, t)\, \mathrm{d}v.$$

The mass $\mathscr{M}_0$ of the same material points in configuration $\mathscr{R}_0$ is given by

$$\mathscr{M}_0 = \int_{\mathscr{V}_0} \rho_0(\mathbf{X})\, \mathrm{d}V,$$

where $\rho_0(\mathbf{X})$ denotes the material density in the reference configuration. The principle of conservation of mass states that we must have $\mathscr{M}(t) = \mathscr{M}_0$ for all times t. It follows that we must have

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathscr{V}(t)} \rho(\mathbf{x}, t)\, \mathrm{d}v = 0.$$

Making use of the Transport Theorem we obtain

$$\int_{\mathscr{V}(t)} \left(\frac{\partial \rho(\mathbf{x}, t)}{\partial t} + \rho(\mathbf{x}, t) \nabla \cdot \mathbf{v}(\mathbf{x}, t) \right) \mathrm{d}v = 0,$$

and since the last relation must hold for any volume $\mathscr{V}(t)$, it is required that

$$\frac{\partial \rho(\mathbf{x}, t)}{\partial t} + \rho(\mathbf{x}, t) \nabla \cdot \mathbf{v}(\mathbf{x}, t) = 0 \quad \forall\, \mathbf{x} \in \mathscr{R}(t),$$

which is ensured under suitable continuity assumptions on the field variables.

Conservation of mass leads to some important implications. First of all it implies that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathscr{M}(t) = \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathscr{V}_0} \rho_0(\mathbf{X}, t)\, \mathrm{d}V = \int_{\mathscr{V}_0} \frac{\mathrm{d}}{\mathrm{d}t} \rho_0(\mathbf{X}, t)\, \mathrm{d}V = 0,$$

and thus

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho_0(\mathbf{X}, t) = 0 \quad \forall\, \mathbf{X} \in \mathscr{R}_0.$$

On the other hand

$$\mathscr{M}(t) = \mathscr{M}_0 \Rightarrow \int_{\mathscr{V}_0} J\rho(\boldsymbol{\phi}(\mathbf{X}, t), t)\, \mathrm{d}V = \int_{\mathscr{V}_0} \rho_0(\mathbf{X}, t)\, \mathrm{d}V \Rightarrow J\rho = \rho_0.$$

Then it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho_0 = \frac{\mathrm{d}J}{\mathrm{d}t} \rho + \frac{\mathrm{d}\rho}{\mathrm{d}t} J = 0. \tag{3.2.1}$$

Then since

$$\frac{1}{\rho} \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\frac{1}{J} \frac{\mathrm{d}J}{\mathrm{d}t} = -\nabla \cdot \mathbf{v},$$

we have three equivalent necessary and sufficient conditions for incompressible motions:

$$J = 1, \quad \frac{\mathrm{d}J}{\mathrm{d}t} = 0, \quad \nabla \cdot \mathbf{v} = 0.$$

3.2.3 Balance of Linear Momentum

The postulate of balance of linear momentum is that the rate of change of linear momentum of a fixed mass of the body is equal to the sum of all the forces acting on that body. Then we must have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)}\rho\mathbf{v}\,\mathrm{d}v=\int_{\mathcal{V}(t)}\rho\mathbf{b}\,\mathrm{d}v+\int_{\partial\mathcal{V}(t)}\mathbf{t}\,\mathrm{d}a,$$

where $\mathbf{b}$ represent the body force per unit of mass, $\mathbf{t}=\mathbf{t}(\mathbf{x},t;\mathbf{n})$ is the surface tension acting on the body in the current configuration per unit of area, and $\mathbf{n}$ is the unit normal vector to the surface $\partial\mathcal{V}(t)$ in $\mathbf{x}$. It can be shown that, under certain conditions, the dependence of the stress vector on the surface under consideration is linear. Indeed Cauchy's lemma states that the stress vectors acting on opposite sides of the same surface at a given point and time are equal in magnitude and opposite in sign, that is

$$\mathbf{t}(\mathbf{x},t;\mathbf{n})=-\mathbf{t}(\mathbf{x},t;-\mathbf{n}).$$

Furthermore it can be proven that there exists a second order tensor $\mathbf{T}$, such that

$$\mathbf{t}(\mathbf{x},t;\mathbf{n})=\mathbf{T}(\mathbf{x},t)\cdot\mathbf{n}\quad\text{or}\quad t_k(\mathbf{x},t;\mathbf{n})=T_{ki}(\mathbf{x},t)n_i.$$

Then the stress vector depends linearly on the normal to the surface under consideration. $\mathbf{T}$ is called the Cauchy stress tensor, and does not depend on $\mathbf{n}$. Making use of $\mathbf{T}$, the balance of linear momentum equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}(t)}\rho\mathbf{v}\,\mathrm{d}v=\int_{\mathcal{V}(t)}(\rho\mathbf{b}+\nabla\cdot\mathbf{T})\,\mathrm{d}v,$$

and by using the Transport Theorem we get its local formulation, which reads

$$\rho\mathbf{a}=\rho\mathbf{b}+\nabla\cdot\mathbf{T}\quad\forall\mathbf{x}\in\mathscr{R}(t),\tag{3.2.2}$$

where $\mathbf{a}$ is the acceleration vector. It is often convenient to represent $\mathbf{T}$ as the sum of two distinct components, a deviatoric and a spherical part:

$$\mathbf{T}=\boldsymbol{\tau}+\kappa\mathbf{I},\tag{3.2.3}$$

where

$$\tau_{ii}=0\quad\text{and}\quad\kappa=\frac{1}{3}T_{ii}.$$

When the Cauchy stress tensor is decomposed in this way, $-\kappa$ is often called pressure and it is denoted by p. Plugging (3.2.3) into (3.2.2), we get

$$\rho\mathbf{a}=-\nabla p+\nabla\cdot\boldsymbol{\tau}+\rho\mathbf{b}\quad\forall\mathbf{x}\in\mathscr{R}(t).$$

Now let us introduce $\mathbf{p}$, which denotes the surface force acting on the body in the current configuration per unit area of $\partial\mathcal{V}_0$. $\mathbf{p}$ and the stress vector $\mathbf{t}$ are related by the following equation

$$\int_{\mathcal{V}_0}\mathbf{p}\,\mathrm{d}A=\int_{\mathcal{V}(t)}\mathbf{t}\,\mathrm{d}a.$$

Then the balance of linear momentum can be written also with respect to $\partial\mathcal{V}_0$ and $\mathcal{V}_0$, by using $\mathrm{d}v=J\mathrm{d}V$ and $\rho J=\rho_0$ as

$$\int_{\mathcal{V}_0}\rho_0\mathbf{a}\,\mathrm{d}V=\int_{\mathcal{V}_0}\rho_0\mathbf{b}\,\mathrm{d}V+\int_{\partial\mathcal{V}_0}\mathbf{p}\,\mathrm{d}A.\tag{3.2.4}$$

Similarly as already done for $\mathbf{t}$, it can be shown that

$$\mathbf{p}(\mathbf{X},t;\mathbf{N})=-\mathbf{p}(\mathbf{X},t;-\mathbf{N}),$$

and that there exists a tensor **P** such that

$$\mathbf{p}(\mathbf{X}, t; \mathbf{N}) = \mathbf{P}(\mathbf{X}, t) \cdot \mathbf{N}, \tag{3.2.5}$$

where **N** is the normal unit vector to the surface $\partial \mathcal{V}_0$ at **X**, and **P** is called the first Piola-Kirchoff tensor. Then from (3.2.5) and the relation (3.1.6) it follows that

$$\mathbf{P} = J\mathbf{T}\mathbf{F}^{-T}.$$

In general the first Piola-Kirchoff tensor is not symmetric, and often it is convenient to define a symmetric tensor **S**, called the second Piola-Kirchoff tensor, which is related to **P** and **T** through the equation

$$\mathbf{S} = \mathbf{F}^{-1}\mathbf{P} \quad \text{and} \quad \mathbf{S} = J\mathbf{F}^{-1}\mathbf{T}\mathbf{F}^{-T}. \tag{3.2.6}$$

Combining (3.2.4) and (3.2.5) we get

$$\int_{\mathcal{V}_0} \rho_0 \mathbf{a}\, \mathrm{d}V = \int_{\mathcal{V}_0} (\rho_0 \mathbf{b} + \nabla_0 \cdot \mathbf{P})\, \mathrm{d}V,$$

whose local form reads

$$\rho_0 \mathbf{a} = \rho_0 \mathbf{b} + \nabla_0 \cdot \mathbf{P} \quad \forall \mathbf{X} \in \mathscr{R}_0. \tag{3.2.7}$$

3.2.4 Mechanical Energy Equation

The Mechanical Energy Equation is obtained from the equation of balance of linear momentum and the principle of mass conservation. If we take the inner product between the velocity vector and the equation (3.2.2), we obtain (using index notation)

$$\frac{1}{2}\rho \frac{\mathrm{d}}{\mathrm{d}t}(v_i v_i) = \frac{\partial T_{ij}}{\partial x_j} v_i + \rho b_i v_i \quad \forall \mathbf{x} \in \mathscr{R}(t).$$

An integral form of this equation can be obtained by integrating over a fixed region of the body occupying the volume $\mathcal{V}(t)$ and delimited by the surface $\partial\mathcal{V}(t)$ to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_{\mathcal{V}(t)} \frac{1}{2}\rho \mathbf{v}\cdot\mathbf{v}\, \mathrm{d}v}_{\mathscr{K}} = \underbrace{\int_{\partial\mathcal{V}(t)} \mathbf{t}\cdot\mathbf{v}\, \mathrm{d}a}_{R_c} - \underbrace{\int_{\mathcal{V}(t)} \mathbf{T}:\mathbf{D}\, \mathrm{d}v}_{\mathscr{P}_{int}} + \underbrace{\int_{\mathcal{V}(t)} \rho \mathbf{b}\cdot\mathbf{v}\, \mathrm{d}v}_{R_b}, \tag{3.2.8}$$

where **D** is the symmetric part of the velocity gradient tensor defined in (3.1.9), and we have made use of the divergence theorem and the conservation of mass. In (3.2.8) we have introduced some notation that we will use in what follows. In particular we have denoted as

- $\mathscr{K}$, the kinetic energy in $\mathcal{V}(t)$,
- R_c, the rate of work done by the surface forces on $\partial\mathcal{V}(t)$,
- R_b, the rate of work done on the material volume $\mathcal{V}(t)$ by the body forces,
- $\mathscr{P}_{int}$, the rate of internal work in the material volume $\mathcal{V}(t)$.

Then the Mechanical Energy Equation can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{K} + \mathscr{P}_{int} = R_c + R_b. \tag{3.2.9}$$

The Mechanical Energy Equation can be written also with respect to the reference configuration $\mathscr{R}_0$. It can be shown that we obtain the same formulation as in (3.2.9), with

$$\mathscr{K} = \int_{\mathcal{V}_0} \frac{1}{2}\rho_0 \mathbf{v}\cdot\mathbf{v}\, \mathrm{d}V, \quad R_c = \int_{\partial\mathcal{V}_0} \mathbf{p}\cdot\mathbf{a}\, \mathrm{d}A, \quad R_b = \int_{\mathcal{V}_0} \rho_0 \mathbf{b}\cdot\mathbf{v}\, \mathrm{d}V, \quad \mathscr{P}_{int} = \int_{\mathcal{V}_0} \mathbf{P}:\dot{\mathbf{F}}\, \mathrm{d}V.$$

3.3 Constitutive Equations for Hyperelastic Materials

We assume that for hyperelastic materials the stress power can be represented as

$$\mathbf{T} : \mathbf{D} = \rho \frac{\mathrm{d}\Sigma}{\mathrm{d}t}, \tag{3.3.1}$$

where Σ is called the strain energy density function or the stored energy per unit mass. In other words the rate of change in strain energy per unit mass arises from the work done on the body by internal stresses. The total strain energy is denoted by $\mathscr{U}$ and is given by

$$\mathscr{U} = \int_{\mathscr{V}(t)} \rho \Sigma \, \mathrm{d}v.$$

Integrating (3.3.1) over an arbitrary material volume $\mathscr{V}(t)$, and using (3.2.8), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathscr{K} + \mathscr{U}) = \int_{\partial\mathscr{V}(t)} \mathbf{t} \cdot \mathbf{v} \, \mathrm{d}a + \int_{\mathscr{V}(t)} \rho \mathbf{b} \cdot \mathbf{v} \, \mathrm{d}v.$$

This equation states that the work done by external forces on the body is directly converted in kinetic energy or stored strain energy. In classical hyperelasticity it is assumed that the strain energy density function at each material point and for all time depends on the deformation gradient at that point and time:

$$\Sigma = \Sigma(\mathbf{F}, \mathbf{X}).$$

However for homogeneous materials the function Σ must have the same form for all the material points $\mathbf{X}$, and thus $\Sigma = \Sigma(\mathbf{F})$. Usually a normalization condition is applied to the strain energy density function so that Σ vanishes when the body is in the reference configuration. Since in $\mathscr{R}_0$ we have $\mathbf{F} = \mathbf{I}$, we require $\Sigma(\mathbf{I}) = 0$. Often also a strain energy per unit volume in the reference configuration is introduced. It is called W, and we have

$$\mathscr{U} = \int_{\mathscr{V}_0} W \, \mathrm{d}V.$$

Then $W = \rho J \Sigma = \rho_0 \Sigma$.

We aim at evaluating the work done on a hyperelastic material by the internal stresses, during a closed dynamic process. The dynamic process is defined by $\mathbf{T}$ and $\mathbf{x} = \boldsymbol{\phi}(\mathbf{X}, t)$, and we assume that it takes place during the time interval $t \in [t_i, t_f]$. A dynamic process is said to be closed if

$$\mathbf{F}(\mathbf{X}, t_i) = \mathbf{F}(\mathbf{X}, t_f) \quad \forall \mathbf{X} \in \mathscr{R}_0.$$

The work done on a hyperelastic material as defined by (3.3.1) during $t \in [t_i, t_f]$ at the material point $\mathbf{X}$

$$\begin{aligned}
\mathscr{W} &= \int_{t_i}^{t_f} \mathbf{T} : \mathbf{D} \, \mathrm{d}t \\
&= \int_{t_i}^{t_f} \rho \frac{\partial \Sigma(\mathbf{F}(\mathbf{X}, t))}{\partial t} \, \mathrm{d}t \\
&= \rho(\Sigma(\mathbf{F}(\mathbf{X}, t_f)) - \Sigma(\mathbf{F}(\mathbf{X}, t_i))) \\
&= 0.
\end{aligned}$$

Therefore the work done by the stress on a hyperelastic material during a closed dynamic process is zero, and it is not dependent on the deformation undergone by the body during the time interval $t \in [t_i, t_f]$.

The constitutive law of the strain energy function should not depend on the frame of reference, and therefore this invariance requirement restricts the form of the functional dependence of W (or Σ) on $\mathbf{F}$. We must have

$$W(\mathbf{F}) = W(\mathbf{QF})$$

for all proper orthogonal tensors $\mathbf{Q}$. In particular for $\mathbf{Q} = \mathbf{R}^T$, where $\mathbf{F} = \mathbf{RU}$ is the polar decomposition, we get

$$W(\mathbf{F}) = W(\mathbf{U}).$$

Without loss of generality, the most general form of strain energy function that satisfies invariance can be written as $W = W(\mathbf{C})$, where $\mathbf{C}$ is the right Cauchy-Green tensor. Any function of $\mathbf{U}$ can be written as function of $\mathbf{C}$, and in particular it follows that:

$$\begin{aligned}
\mathbf{T} &= \frac{1}{J}\mathbf{F}\left(\frac{\partial W}{\partial \mathbf{C}} + \frac{\partial W}{\partial \mathbf{C}^T}\right)\mathbf{F}^T, \\
\mathbf{P} &= \mathbf{F}\left(\frac{\partial W}{\partial \mathbf{C}} + \frac{\partial W}{\partial \mathbf{C}^T}\right), \\
\mathbf{S} &= \left(\frac{\partial W}{\partial \mathbf{C}} + \frac{\partial W}{\partial \mathbf{C}^T}\right).
\end{aligned} \tag{3.3.2}$$

In the case of isotropic hyperelastic materials we have that

$$W(\mathbf{C}) = W(\mathbf{QCQ}^T)$$

for all orthogonal tensors $\mathbf{Q}$. This implies that W has to be a scalar valued isotropic function of $\mathbf{C}$. Then it follows from a representation theorem for isotropic scalar functions that the strain energy function for a hyperelastic isotropic material may be expressed by as a function of the scalar invariants of $\mathbf{C}$, denoted as $\mathscr{I}_{1C}$, $\mathscr{I}_{2C}$ and $\mathscr{I}_{3C}$, where

$$\begin{aligned}
\mathscr{I}_{1C} &= \mathrm{tr}(\mathbf{C}), \\
\mathscr{I}_{2C} &= \frac{1}{2}(\mathrm{tr}(\mathbf{C})^2 - \mathrm{tr}(\mathbf{C}^2)), \\
\mathscr{I}_{3C} &= \det(\mathbf{C}).
\end{aligned}$$

Chapter 4

Cardiac Model

4.1 Constitutive Mechanical Law for Cardiac Tissue

The aim of this chapter is to describe a fully coupled model describing the activity of the cardiac muscle. Thus as a first step we present the mechanical constitutive law we choose to describe the response of the cardiac tissue to stresses. It is important to underline that we consider only one-dimensional models in space, that is, we suppose that the forces acting on the fibre are directed as the longitudinal axis of the fibre. This consideration implies that the tensor **S** defined in (3.2.6) is such that

$$S_{22} = S_{33} = 0.$$

Moreover we assume equal to zero also the off-diagonal entries of **S**, and therefore the second Piola-Kirchoff tensor is given by the only component S_{11}. Then, in absence of body forces, the equation (3.2.7) in our case reduces to

$$\rho_0 \frac{\partial^2 u_1}{\partial t^2} = \frac{\partial}{\partial X_1}(S_{11}F_{11}), \tag{4.1.1}$$

where u_1 is the fiber's displacement in the longitudinal direction, and we have assumed that the off-diagonal entries of **F** are small such that they can be neglected and considered as equal to 0. Furthermore we assume incompressibility, that implies $\det(\mathbf{F}) = 1$. Then since $\mathbf{F} = \nabla_0 \mathbf{u} + \mathbf{I}$, we assume that **F** has the form

$$\mathbf{F} = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \frac{1}{\sqrt{\lambda}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\lambda}} \end{bmatrix},$$

where $\lambda = \partial u_1/\partial X_1 + 1$. The tensor **S** is given by (3.3.2), and then we must choose a formulation for the strain energy function. We choose to adopt the strain energy function for an neo-Hookean incompressible solid, proposed by R. Rivlin in 1948, that is

$$W(\mathbf{C}) = c_1(\mathscr{I}_{1C} - 3),$$

where $\mathscr{I}_{1C}$ is the first invariant of the right Cauchy-Green tensor, and in our case

$$\mathscr{I}_{1C} = \lambda^2 + \frac{2}{\lambda}.$$

Then now we can define S_{11}. Using the chain rule we obtain

$$\begin{aligned} S_{11} &= 2\frac{\partial W}{\partial C_{11}} \\ &= 2\frac{\partial W}{\partial \mathscr{I}_{1C}}\frac{\partial \mathscr{I}_{1C}}{\partial C_{11}} \\ &= 2c_1\left(1-\frac{1}{\lambda^3}\right). \end{aligned}$$

Finally, the mechanical model governing the fiber's displacement u_1 is given by

$$\rho_0\frac{\partial^2 u_1}{\partial t^2} = \frac{\partial}{\partial X_1}\left(\left[2c_1\left(1-\frac{1}{\left(1+\frac{\partial u_1}{\partial X_1}\right)^3}\right)\right]\left(1+\frac{\partial u_1}{\partial X_1}\right)\right). \tag{4.1.2}$$

In what follows, in order to simplify the notation, we will denote the displacement u_1 simply as u, and the reference position X_1 simply as x.

4.1.1 Active Stress Coupling

The Bestel-Clément-Sorine model (2.3.5) (BCS) of electro-mechanical activity presented in Section 2.3, allows to compute the total stress and the total stiffness in the myofibre direction by solving two ODEs, which depend on the action potential and the strain. We consider the solution of the fully coupled problem for the one-dimensional domain $\Omega = (0,1)$. Then both the stress and stiffness have to be evaluated for each point $x \in \Omega$. The stress σ has then to enter in the model governing the mechanics of the fibre as active stress. However instead of plugging directly σ into (4.1.2), we adopt an active contraction model similar to the one presented in [30]. In particular we define a new internal variable H, such that $0 \le H \le 1$, and as active stress we consider the quantity $\sigma_{max}H^2$, where σ_{max} represents the maximum active stress observed in the fibre. The evolution law of the internal variable H depends on the stress σ and the displacement u, and reads as

$$\begin{cases} \dfrac{\partial H}{\partial t} = \dfrac{\nu}{\alpha}\sigma - \dfrac{1}{2}\dfrac{\sigma_{max}}{\alpha}H\left(1+\dfrac{\partial u}{\partial x}\right)^2, & x \in (0,1),\ t \ge 0, \\ H(x,0) = 0, & \forall\, x \in (0,1), \end{cases} \tag{4.1.3}$$

where ν and α are two model parameters which have to be properly chosen. On the other hand the mechanical model we need to solve becomes

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dot{u}, & x \in (0,1),\ t \ge 0, \\ \dfrac{\partial \dot{u}}{\partial t} = \dfrac{1}{\rho_0}\dfrac{\partial}{\partial x}\left(\left[2c_1\left(1-\dfrac{1}{\left(1+\frac{\partial u}{\partial x}\right)^3}\right)+\sigma_{max}H^2\right]\left(1+\dfrac{\partial u}{\partial x}\right)\right), & x \in (0,1),\ t \ge 0, \\ u(x,0) = 0, \quad \dot{u}(x,t) = 0, & \forall\, x \in (0,1), \\ + \text{ b. c.}, \end{cases} \tag{4.1.4}$$

where we have introduced the additional variable $\dot{u}$ representing the displacement's velocity, to transform the second order differential equation (4.1.2) into a system of two first order differential equations.

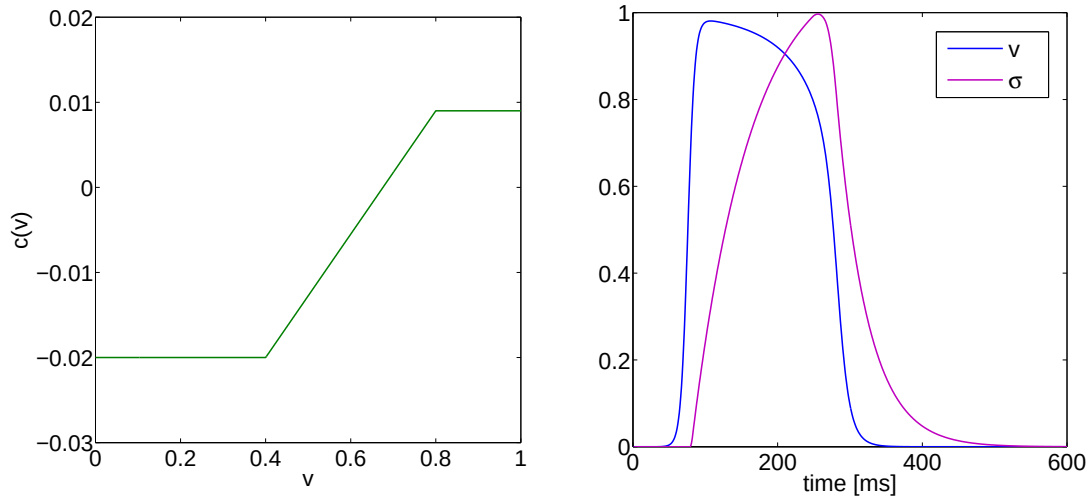
4.2 Electromechanical Coupling

The equations defining the BCS model (2.3.5) depend on the time derivative of the strain and on the function c representing the chemical input. Since we do not know $\dot{\varepsilon} = \partial/\partial t(\partial u/\partial x)$ from (4.1.4), we

assume $u(x,t) \in \mathscr{C}^2(\Omega \times (0,\infty))$, so that we can replace $\dot{\varepsilon}$ in (2.3.5) by $\partial \dot{u}/\partial x$, where $\dot{u}$ is the solution of model (4.1.4). Lastly we need to discuss how the function c governing the chemical input is defined. We assume that c depends directly on the electrical activation model, and therefore we neglect calcium dynamics. We choose the following formulation for c:

$$c = c(v(x,t)) = \begin{cases} -k_{rs} & \text{if } v < v_{a1}, \\ -k_{rs} + \dfrac{k_{atp} + k_{rs}}{v_{a2} - v_{a1}}(v - v_{a1}) & \text{if } v_{a1} \le v < v_{a2}, \\ k_{atp} & \text{if } v \ge v_{a2}, \end{cases} \tag{4.2.1}$$

where k_{atp} is the rate of the myosin ATPase activity controlling the contraction rate and k_{rs} is the rate of sarcoplasmic reticulum calcium re-uptake controlling the relaxation rate [23], while the values v_{a1} and v_{a2} define the activation threshold.



(a) (b)

Figure 4.1: (a): Chemical input function c as function of the action potential v. For this figure $k_{rs} = 0.02$ ms^{-1}, $k_{atp} = 0.009$ ms^{-1}, $v_{a1} = 0.4$, $v_{a2} = 0.8$. (b): Normalized stress σ as function of time for the chemical input function showed in (a).

4.2.1 Mechanoelectrical Feedback

In [24, 25] it is pointed out that the mechanical deformation of the tissue membrane can affect the process of wave propagation through the medium, resulting in a complex global feedback phenomenon referred to as mechanoeletrical feedback. Moreover an electromechanical model is presented, which is capable of inducing the formation of self-organized pacemakers. In particular the stretching of the fibre can give origin to currents that in turn can cause contraction. Here we intend only to present how the model is defined, while in the next chapter we will give some numerical results to actually show how the stretch-activated currents can be important in the termination of re-entrant waves. Then we

replace (1.2.1) with the following model:

$$\begin{cases} \dfrac{\partial v}{\partial t} = \dfrac{\partial}{\partial x}\left(D(u)\dfrac{\partial v}{\partial x}\right) - kv(v-a)(v-1) - vr + I_{app} - I_s(v,u), & x \in (0,1),\, t>0, \\ \dfrac{\partial r}{\partial t} = \varepsilon(v,r)(-r - kv(v-a-1)), & x \in (0,1),\, t>0, \\ v(x,0) = v_0(x), \quad r(x,0) = r_0(x), & \forall\, x \in (0,1), \\ + \text{b. c.}, \end{cases} \tag{4.2.2}$$

where D is the new conductivity function and I_s represents the stretch-activated current. It is supposed that contraction should induce an increase of the conductivity coefficient in the reference configuration. Indeed in the current configuration the electrical signal travels at the same velocity, but when mapped to the reference configuration we have as result an increase of the conductivity coefficient during contraction. Then D is defined as

$$D = D(u(x,t)) = \frac{\mu}{(1+\partial u/\partial x)^2}. \tag{4.2.3}$$

On the other hand the stretch-activated current I_s is described as

$$I_s = I_s(v(x,t), u(x,t)) = \begin{cases} G_s(\partial u/\partial x)(u - E_s) & \text{if } \partial u/\partial x > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{4.2.4}$$

which makes I_s directly dependent on dilatation. G_s and E_s are two model parameters and are usually set respectively to 0.5 and 1.

4.3 Numerical Approximation of the Fully Coupled Cardiac Model

We aim at solving the fully coupled cardiac model using the finite element method. To summarize, the governing equations characterizing the problem are:

$$\left.\begin{cases} \dfrac{\partial v}{\partial t} = \dfrac{\partial}{\partial x}\left(D(u)\dfrac{\partial v}{\partial x}\right) - kv(v-a)(v-1) - vr + I_{app} - I_s(v,u), & x \in (0,1),\, t>0, \\ \dfrac{\partial r}{\partial t} = \varepsilon(v,r)(-r - kv(v-a-1)), & x \in (0,1),\, t>0, \\ v(x,0) = v_0(x), \quad r(x,0) = r_0(x), & \forall\, x \in (0,1), \\ + \text{b. c.}, \end{cases}\right\} \textbf{AP MODEL}$$

$$\left.\begin{cases} \dfrac{\partial k}{\partial t} = -\left(|c(v)| + \left|\dfrac{\partial \dot{u}}{\partial x}\right|\right)k + k_0 c(v)^+, & x \in (0,1),\ t>0, \\ \dfrac{\partial \sigma}{\partial t} = \dfrac{\sigma_0}{k_0}\dfrac{\partial \dot{u}}{\partial x}k - \left(|c(v)| + \left|\dfrac{\partial \dot{u}}{\partial x}\right|\right)\sigma + \sigma_0\dfrac{c(v)^+}{2}, & x \in (0,1),\ t>0, \\ k(x,0) = 0, \quad \sigma(x,0) = 0, & \forall\, x \in (0,1), \end{cases}\right\} \textbf{BCS MODEL}$$

$$\left.\begin{cases} \dfrac{\partial H}{\partial t} = \dfrac{\nu}{\alpha}\sigma - \dfrac{1}{2}\dfrac{\sigma_{max}}{\alpha}H\left(1+\dfrac{\partial u}{\partial x}\right)^2, & x \in (0,1),\ t>0, \\ H(x,0) = 0, & \forall\, x \in (0,1), \end{cases}\right\} \textbf{INTERNAL VARIABLE MODEL}$$

$$\left.\begin{cases} \dfrac{\partial u}{\partial t} = \dot{u}, & x \in (0,1),\ t>0, \\ \dfrac{\partial \dot{u}}{\partial t} = \dfrac{1}{\rho_0}\dfrac{\partial}{\partial x}\left(S(H,u)\left(1+\dfrac{\partial u}{\partial x}\right)\right), & x \in (0,1),\ t>0, \\ u(x,0) = 0, \quad \dot{u}(x,t) = 0, & \forall\, x \in (0,1), \\ +\ \text{b. c.}, \end{cases}\right\} \textbf{MECHANICAL MODEL} \tag{4.3.1}$$

where $c(v)$, $D(u)$ and $I_s(v,u)$ are the functions described in the previous section, and S, which represents the sum of both passive and active stresses, is

$$S(H(x,t),u(x,t)) = \left[2c_1 \left(1 - \frac{1}{\left(1+\frac{\partial u}{\partial x}\right)^3} \right) + \sigma_{max} H^2 \right].$$

For which regards the choice of the boundary conditions, we make the following choice:

- AP MODEL:
$$\frac{\partial v(0,t)}{\partial x} = 0 \quad \forall\, t>0,$$
$$\frac{\partial v(1,t)}{\partial x} = 0 \quad \forall\, t>0.$$
- MECHANICAL MODEL:
$$\dot{u}(0,t) = 0 \qquad \forall\, t>0,$$
$$S(H(1,t),u(1,t)) = 0 \quad \forall\, t>0.$$

As already done in Section 1.2 we start by writing the weak formulation of the whole problem. We seek for solutions lying in suitable functional spaces so that the integrals appearing in the weak formulation are defined in the Lebesgue sense. Then we introduce the approximation spaces of finite dimension N_h, where we seek for the approximated solutions. In particular, for each $t \in (0,\infty)$, we seek for $v_h(t)$,$r_h(t)$,$k_h(t)$,$\sigma_h(t)$,$H_h(t)$ and $u_h(t)$ belonging to the finite dimensional space

$$W_h(\Omega,\mathcal{T}_h) = \left\{ z_h \in \mathscr{C}^0(\overline{\Omega});\, z_h|_K \in \mathscr{P}^1(K) \forall\, K \in \mathcal{T}_h \right\},$$

while $\dot{u}_h(t)$ has to be sought in

$$V_h = \{ z_h \in W_h(\Omega,\mathcal{T}_h);\, z_h(0) = 0 \}.$$

Let be $\{\varphi_j\}_{j=1}^{N_h}$ the basis functions for W_h. Each function f_h lying in W_h, and then each of the approximated solutions we are looking for, can be represented as

$$f_h(t) = \sum_{j=1}^{N_h} f_j(t)\varphi_j.$$

The unknowns of the coupled problem can then be grouped in the vector

$$\mathbf{y}(t) = \begin{bmatrix} \mathbf{v}(t) \\ \mathbf{r}(t) \\ \mathbf{k}(t) \\ \boldsymbol{\sigma}(t) \\ \mathbf{H}(t) \\ \mathbf{u}(t) \\ \dot{\mathbf{u}}(t) \end{bmatrix} = \begin{bmatrix} (v_1(t),\dots,v_{N_h}(t))^T \\ (r_1(t),\dots,r_{N_h}(t))^T \\ (k_1(t),\dots,k_{N_h}(t))^T \\ (\sigma_1(t),\dots,\sigma_{N_h}(t))^T \\ (H_1(t),\dots,H_{N_h}(t))^T \\ (u_1(t),\dots,u_{N_h}(t))^T \\ (\dot{u}_2(t),\dots,\dot{u}_{N_h}(t))^T \end{bmatrix}, \quad \mathbf{y}(t) \in \mathbb{R}^M,$$

where $M = 7N_h - 1$. Then after discretization in space we end up with the following ODE system

$$\mathbf{M}^B \dot{\mathbf{y}}(t) = \mathscr{L}(\mathbf{y}(t)), \tag{4.3.2}$$

where $\mathbf{M}^B \in \mathbb{R}^{M\times M}$ is the block diagonal mass matrix

$$\mathbf{M}^B = \begin{bmatrix} \mathbf{M}_{N_h} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{M}_{N_h} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{M}_{N_h-1} \end{bmatrix}, \quad \mathbf{M}_{N_h} \in \mathbb{R}^{N_h\times N_h},$$

while $\mathscr{L}(\mathbf{y}(t)) \in \mathbb{R}^M$ is defined as

$$\mathscr{L}(\mathbf{y}(t)) = \begin{bmatrix} \mathscr{L}_1(\mathbf{y}(t)) \\ \mathscr{L}_2(\mathbf{y}(t)) \\ \mathscr{L}_3(\mathbf{y}(t)) \\ \mathscr{L}_4(\mathbf{y}(t)) \\ \mathscr{L}_5(\mathbf{y}(t)) \\ \mathscr{L}_6(\mathbf{y}(t)) \\ \mathscr{L}_7(\mathbf{y}(t)) \end{bmatrix} = \begin{bmatrix} (\mathscr{L}_{1,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{1,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{2,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{2,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{3,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{3,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{4,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{4,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{5,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{5,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{6,1}(\mathbf{y}(t)), \dots, \mathscr{L}_{6,N_h}(\mathbf{y}(t)))^T \\ (\mathscr{L}_{7,2}(\mathbf{y}(t)), \dots, \mathscr{L}_{7,N_h}(\mathbf{y}(t)))^T \end{bmatrix},$$

and

$$\mathscr{L}_{1,i}(\mathbf{y}(t)) = \mathscr{L}_{1,i}(v_h, r_h, u_h) = -\int_0^1 D(u_h)\frac{\partial v_h}{\partial x}\frac{\partial \varphi_i}{\partial x}\,\mathrm{d}x + \int_0^1 (f(v_h, r_h) - I_s(v_h, u_h))\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{2,i}(\mathbf{y}(t)) = \mathscr{L}_{2,i}(v_h, r_h) = \int_0^1 g(v_h, r_h)\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{3,i}(\mathbf{y}(t)) = \mathscr{L}_{3,i}(v_h, k_h, \dot{u}_h) = \int_0^1 \left[-\left(|c(v_h)| + \left|\frac{\partial \dot{u}_h}{\partial x}\right|\right)k_h + k_0 c(v_h)^+\right]\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{4,i}(\mathbf{y}(t)) = \mathscr{L}_{4,i}(v_h, k_h, \sigma_h, \dot{u}_h) = \int_0^1 \left[\frac{\sigma_0}{k_0}\frac{\partial \dot{u}_h}{\partial x}k_h - \left(|c(v_h)| + \left|\frac{\partial \dot{u}_h}{\partial x}\right|\right)\sigma_h + \sigma_0\frac{c(v_h)^+}{2}\right]\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{5,i}(\mathbf{y}(t)) = \mathscr{L}_{5,i}(\sigma_h, H_h, u_h) = \int_0^1 \left[\frac{\nu}{\alpha}\sigma_h - \frac{1}{2}\frac{\sigma_{max}}{\alpha}H_h\left(1 + \frac{\partial u_h}{\partial x}\right)^2\right]\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{6,i}(\mathbf{y}(t)) = \mathscr{L}_{6,i}(\dot{u}_h) = \int_0^1 \dot{u}_h\varphi_i\,\mathrm{d}x,$$

$$\mathscr{L}_{7,i}(\mathbf{y}(t)) = \mathscr{L}_{7,i}(H_h, u_h) = -\int_0^1 \frac{1}{\rho_0}S(H_h, u_h)\left(1 + \frac{\partial u_h}{\partial x}\right)\frac{\partial \varphi_i}{\partial x}\,\mathrm{d}x.$$

Note that in writing $\mathscr{L}_{1,i}$ and $\mathscr{L}_{7,i}$ we have used integration by parts, and that all the boundary terms cancel out.

We decide to solve the ODE system (4.3.2) implicitly by using the θ-method. Then for each time point $t^{n+1} = (n+1)\Delta t$, we need to solve the following non-linear system ($\theta \neq 0$):

$$\mathscr{F}(\mathbf{y}^{n+1}) = \mathbf{M}^B\frac{\mathbf{y}^{n+1} - \mathbf{y}^n}{\Delta t} - \mathscr{L}(\theta\mathbf{y}^{n+1} + (1-\theta)\mathbf{y}^n) = \mathbf{0}, \tag{4.3.3}$$

where $\mathbf{y}^n \approx \mathbf{y}(t^n)$. To solve the non-linear system (4.3.3) we use the Newton's algorithm. For each time point t^{n+1}, we set $\mathbf{y}^{n+1,0} = \mathbf{y}^n$. Then for $k = 0, 1, \dots$ until convergence

1. solve $\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})\boldsymbol{\delta}\mathbf{y}^{n+1,k} = -\mathscr{F}(\mathbf{y}^{n+1,k})$,
2. set $\mathbf{y}^{n+1,k+1} = \mathbf{y}^{n+1,k} + \boldsymbol{\delta}\mathbf{y}^{n+1,k}$,

where $\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})$ is the Jacobian of $\mathscr{F}$, evaluated at $\mathbf{y}^{n+1,k}$. We base the stop criterion on the relative residual, which means we say we reach convergence if

$$\frac{||\mathscr{F}(\mathbf{y}^{n+1,k+1})||_2}{||\mathscr{F}(\mathbf{y}^{n+1,0})||_2} < \epsilon_N,$$

where ϵ_N is a tolerance factor.

4.3.1 Linearisation of the Fully Coupled Cardiac Model

The Newton's method requires that for each time point t^{n+1}, at each step k, we solve a linear system with the matrix $\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})$. The linear system has the following block structure:

$$\begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & & & & \mathbf{J}_{16} & \\ \mathbf{J}_{21} & \mathbf{J}_{22} & & & & & \\ \mathbf{J}_{31} & & \mathbf{J}_{33} & & & & \mathbf{J}_{37} \\ \mathbf{J}_{41} & & \mathbf{J}_{43} & \mathbf{J}_{44} & & & \mathbf{J}_{47} \\ & & & \mathbf{J}_{54} & \mathbf{J}_{55} & \mathbf{J}_{56} & \\ & & & & & \mathbf{J}_{66} & \mathbf{J}_{67} \\ & & & & \mathbf{J}_{75} & \mathbf{J}_{76} & \mathbf{J}_{77} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{v}^k \\ \boldsymbol{\delta}\mathbf{r}^k \\ \boldsymbol{\delta}\mathbf{k}^k \\ \boldsymbol{\delta}\boldsymbol{\sigma}^k \\ \boldsymbol{\delta}\mathbf{H}^k \\ \boldsymbol{\delta}\mathbf{u}^k \\ \boldsymbol{\delta}\dot{\mathbf{u}}^k \end{bmatrix} = - \begin{bmatrix} \mathscr{F}_1(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_2(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_3(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_4(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_5(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_6(\mathbf{y}^{n+1,k}) \\ \mathscr{F}_7(\mathbf{y}^{n+1,k}) \end{bmatrix}, \tag{4.3.4}$$

where

$$\mathscr{F}_1(\mathbf{y}^{n+1,k}) = \mathbf{M}_{N_h} \frac{\mathbf{v}^{n+1,k} - \mathbf{v}^n}{\Delta t} - \mathscr{L}_1(\theta \mathbf{y}^{n+1} + (1-\theta)\mathbf{y}^n), \quad \mathscr{F}_1(\mathbf{y}^{n+1,k}) \in \mathbb{R}^{N_h},$$

and so on for the resting $\mathscr{F}_i$, $i = 2, \dots, 7$. The following subsection is devoted to the linearisation of the problem, and therefore to give the exact formulation of each entry $J_{kl,ij}$ of the 20 blocks $\mathbf{J}_{kl}$ that define the whole Jacobian matrix $\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})$. In what follows we will omit the step index k relative to the Newton's method iteration in order to simplify the notation. Moreover we will use the notation ∂_f to denote the partial differentiation with respect to the arbitrary quantity f.
At first let us observe that, due to the structure of the residual, each diagonal block $\mathbf{J}_{ii}$ can be written as the sum of two contributions, one due to time discretization and the other to the linearisation of the operator $\mathscr{L}_i$, so that

$$\mathbf{J}_{ii} = \frac{\mathbf{M}_{N_h}}{\Delta t} - \mathbf{J}_{ii,\mathscr{L}_i}.$$

- $J_{11,ij}$. We have

  $\partial_v f = -k(3v^2 - 2(a+1)v + a) - r$ and
  $\partial_v I_s = G_s \partial_x u$ if $\partial_x u > 0$.

  Moreover let be

  $F(v,u) = D(u)\partial_x v$ and thus
  $\partial_{\partial_x v} F = D(u)$.

$$\begin{aligned} \lim_{\epsilon\to 0} \frac{\mathscr{F}_{1,i}(v_h^{n+1} + \epsilon\delta v_h) - \mathscr{F}_{1,i}(v_h^{n+1})}{\epsilon} &= \sum_{j=1}^{N_h} \delta v_j \frac{1}{\Delta t} \int_0^1 \varphi_j \varphi_i \,\mathrm{d}x \\ &\quad - \lim_{\epsilon\to 0} \frac{\mathscr{L}_{1,i}(\theta(v_h^{n+1} + \epsilon\delta v_h^{n+1}) + (1-\theta)v_h^n) - \mathscr{L}_{1,i}(\theta v_h^{n+1} + (1-\theta)v_h^n)}{\epsilon} \\ &= \sum_{j=1}^{N_h} \delta v_j \frac{1}{\Delta t} \int_0^1 \varphi_j \varphi_i \,\mathrm{d}x + \theta \sum_{j=1}^{N_h} \delta v_j \int_0^1 D(\theta u_h^{n+1} + (1-\theta)u_h^n) \partial_x \varphi_j \partial_x \varphi_i \,\mathrm{d}x \\ &\quad - \theta \sum_{j=1}^{N_h} \delta v_j \int_0^1 \partial_v f(\theta v_h^{n+1} + (1-\theta)v_h^n, \theta r_h^{n+1} + (1-\theta)r_h^n) \varphi_j \varphi_i \,\mathrm{d}x \\ &\quad + \theta \sum_{j=1}^{N_h} \delta v_j \int_0^1 \partial_v I_s(\theta u_h^{n+1} + (1-\theta)u_h^n) \varphi_j \varphi_i \,\mathrm{d}x \\ &= \sum_{j=1}^{N_h} \delta v_j J_{11,ij}. \end{aligned}$$

- $J_{12,ij}$.

  $\partial_r f = -v$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{1,i}(r_h^{n+1}+\epsilon\delta r_h)-\mathscr{F}_{1,i}(r_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta r_j\int_0^1(\theta v_h^{n+1}+(1-\theta)v_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta r_j J_{12,ij}.$$

- $J_{16,ij}$.

  $\partial_{\partial_x u}F = -\dfrac{2\mu}{(1+\partial_x u)^3}\partial_x v$ and
  $\partial_{\partial_x u}I_s = G_s(v-E_s)$ if $\partial_x u > 0$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{1,i}(u_h^{n+1}+\epsilon\delta u_h)-\mathscr{F}_{1,i}(u_h^{n+1})}{\epsilon} = \theta\sum_{j=1}^{N_h}\delta u_j\int_0^1\partial_{\partial_x u}F(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta u_h^{n+1}+(1-\theta)u_h^n)\partial_x\varphi_j\partial_x\varphi_i\,\mathrm{d}x$$
$$+\theta\sum_{j=1}^{N_h}\delta u_j\int_0^1\partial_{\partial_x u}I_s(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta u_h^{n+1}+(1-\theta)u_h^n)\partial_x\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta u_j J_{16,ij}.$$

- $J_{21,ij}$.

$$\partial_v g = \frac{\mu_1 r}{(v+\mu_2)^2}(r+kv(v-a-1)) - \left(\varepsilon_0+\frac{\mu_1 r}{v+\mu_2}\right)(2kv-k(a+1)).$$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{2,i}(v_h^{n+1}+\epsilon\delta v_h)-\mathscr{F}_{2,i}(v_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta v_j\int_0^1\partial_v g(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta r_h^{n+1}+(1-\theta)r_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta v_j J_{21,ij}.$$

- $J_{22,ij}$.

$$\partial_r g = -\varepsilon_0 + \frac{\mu_1}{\mu_1+\mu_2}(-2r-kv(v-a-1)).$$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{2,i}(r_h^{n+1}+\epsilon\delta r_h)-\mathscr{F}_{2,i}(r_h^{n+1})}{\epsilon} = \sum_{j=1}^{N_h}\delta r_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x$$
$$-\theta\sum_{j=1}^{N_h}\delta r_j\int_0^1\partial_r g(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta r_h^{n+1}+(1-\theta)r_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta r_j J_{22,ij}.$$

- $J_{31,ij}$. Let be

  $F_k(v,k,\dot{u}) = -(|c(v)| + |\partial_x \dot{u}|)k + k_0 c(v)^+$ and thus
  $\partial_v F_k = -\text{sign}(c(v))\partial_v c(v)k + k_0 \partial_v c(v)^+$.

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{3,i}(v_h^{n+1}+\epsilon\delta v_h)-\mathscr{F}_{3,i}(v_h^{n+1})}{\epsilon} &= -\theta\sum_{j=1}^{N_h}\delta v_j\int_0^1 \partial_v F_k(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta k_h^{n+1}+(1-\theta)k_h^n)\varphi_j\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta v_j J_{31,ij}.\end{aligned}$$

- $J_{33,ij}$.

  $\partial_k F_k = -(|c(v)| + |\partial_x \dot{u}|)$.

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{3,i}(k_h^{n+1}+\epsilon\delta k_h)-\mathscr{F}_{3,i}(k_h^{n+1})}{\epsilon} &= \sum_{j=1}^{N_h}\delta k_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x\\ &\quad -\theta\sum_{j=1}^{N_h}\delta k_j\int_0^1\partial_k F_k(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta\dot{u}_h^{n+1}+(1-\theta)\dot{u}_h^n)\varphi_j\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta k_j J_{33,ij}.\end{aligned}$$

- $J_{37,ij}$.

  $\partial_{\partial_x\dot{u}} F_k = -\text{sign}(\partial_x\dot{u})k$.

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{3,i}(\dot{u}_h^{n+1}+\epsilon\delta\dot{u}_h)-\mathscr{F}_{3,i}(\dot{u}_h^{n+1})}{\epsilon} &= -\theta\sum_{j=1}^{N_h}\delta\dot{u}_j\int_0^1\partial_{\partial_x\dot{u}}F_k(\theta k_h^{n+1}+(1-\theta)k_h^n,\theta\dot{u}_h^{n+1}+(1-\theta)\dot{u}_h^n)\partial_x\varphi_j\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta\dot{u}_j J_{37,ij}.\end{aligned}$$

- $J_{41,ij}$. Let be

  $F_\sigma(v,k,\sigma,\dot{u}) = \frac{\sigma_0}{k_0}\partial_x\dot{u}k - (|c(v)| + |\partial_x\dot{u}|)\sigma + \sigma_0\frac{c(v)^+}{2}$ and thus
  $\partial_v F_\sigma = -\text{sign}(c(v))\partial_v c(v)\sigma + \frac{\sigma_0}{2}\partial_v c(v)^+$.

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{4,i}(v_h^{n+1}+\epsilon\delta v_h)-\mathscr{F}_{4,i}(v_h^{n+1})}{\epsilon} &= -\theta\sum_{j=1}^{N_h}\delta v_j\int_0^1\partial_v F_\sigma(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta k_h^{n+1}+(1-\theta)k_h^n)\varphi_j\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta v_j J_{41,ij}.\end{aligned}$$

- $J_{43,ij}$.

  $\partial_k F_\sigma = \frac{\sigma_0}{k_0}\partial_x\dot{u}$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{4,i}(k_h^{n+1}+\epsilon\delta k_h)-\mathscr{F}_{4,i}(k_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta v_j\int_0^1 \partial_k F_\sigma(\theta\dot{u}_h^{n+1}+(1-\theta)\dot{u}_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta k_j J_{43,ij}.$$

- $J_{44,ij}$.

  $\partial_\sigma F_\sigma = -(|c(v)|+|\partial_x\dot{u}|)$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{4,i}(\sigma_h^{n+1}+\epsilon\delta\sigma_h)-\mathscr{F}_{4,i}(\sigma_h^{n+1})}{\epsilon} = \sum_{j=1}^{N_h}\delta\sigma_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x$$
$$-\theta\sum_{j=1}^{N_h}\delta\sigma_j\int_0^1\partial_\sigma F_\sigma(\theta v_h^{n+1}+(1-\theta)v_h^n,\theta\dot{u}_h^{n+1}+(1-\theta)\dot{u}_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta\sigma_j J_{44,ij}.$$

- $J_{47,ij}$.

  $\partial_{\partial_x\dot{u}}F_\sigma = \frac{\sigma_0}{k_0}k - \mathrm{sign}(\partial_x\dot{u})\sigma$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{4,i}(\dot{u}_h^{n+1}+\epsilon\delta\dot{u}_h)-\mathscr{F}_{4,i}(\dot{u}_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta\dot{u}_j\int_0^1\partial_{\partial_x\dot{u}}F_\sigma(\theta k_h^{n+1}+(1-\theta)k_h^n,\theta\sigma_h^{n+1}+(1-\theta)\sigma_h^n,\dots$$
$$\dots\theta\dot{u}_h^{n+1}+(1-\theta)\dot{u}_h^n)\partial_x\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta\dot{u}_j J_{47,ij}.$$

- $J_{54,ij}$. Let be

  $F_H(\sigma,H,u) = \frac{\nu}{\alpha}\sigma - \frac{\sigma_{max}}{2\alpha}H(1+\partial_x u)^2$ and thus
  $\partial_\sigma F_H = \frac{\nu}{\alpha}$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{5,i}(\sigma_h^{n+1}+\epsilon\delta\sigma_h)-\mathscr{F}_{5,i}(\sigma_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta\sigma_j\int_0^1\partial_\sigma F_H\varphi_j\varphi_i\,\mathrm{d}x$$
$$= \sum_{j=1}^{N_h}\delta\sigma_j J_{54,ij}.$$

- $J_{55,ij}$.

  $\partial_H F_H = -\frac{\sigma_{max}}{2\alpha}(1+\partial_x u)^2$.

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{5,i}(H_h^{n+1}+\epsilon\delta H_h)-\mathscr{F}_{5,i}(H_h^{n+1})}{\epsilon} = \sum_{j=1}^{N_h}\delta H_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x$$
$$-\theta\sum_{j=1}^{N_h}\delta H_j\int_0^1\partial_H F_H(\theta u_h^{n+1}+(1-\theta)u_h^n)\varphi_j\varphi_i\,\mathrm{d}x$$
$$=\sum_{j=1}^{N_h}\delta H_j J_{55,ij}.$$

- $J_{56,ij}$.

  $\partial_{\partial_x u}F_H = -\dfrac{\sigma_{max}}{\alpha}H(1+\partial_x u).$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{5,i}(u_h^{n+1}+\epsilon\delta u_h)-\mathscr{F}_{5,i}(u_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta u_j\int_0^1\partial_{\partial_x u}F_H(\theta H_h^{n+1}+(1-\theta)H_h^n,\theta u_h^{n+1}+(1-\theta)u_h^n)\partial_x\varphi_j\varphi_i\,\mathrm{d}x$$
$$=\sum_{j=1}^{N_h}\delta u_j J_{56,ij}.$$

- $J_{66,ij}$. Let be

  $F_u(\dot{u}) = \dot{u}$ and thus
  $\partial_u F_u = 0.$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{6,i}(u_h^{n+1}+\epsilon\delta u_h)-\mathscr{F}_{6,i}(u_h^{n+1})}{\epsilon} = \sum_{j=1}^{N_h}\delta u_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x$$
$$=\sum_{j=1}^{N_h}\delta u_j J_{66,ij}.$$

- $J_{67,ij}$.

  $\partial_{\dot{u}}F_u = 1.$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{6,i}(\dot{u}_h^{n+1}+\epsilon\delta \dot{u}_h)-\mathscr{F}_{6,i}(\dot{u}_h^{n+1})}{\epsilon} = -\theta\sum_{j=1}^{N_h}\delta \dot{u}_j\int_0^1\varphi_j\varphi_i\,\mathrm{d}x$$
$$=\sum_{j=1}^{N_h}\delta \dot{u}_j J_{67,ij}.$$

- $J_{75,ij}$. Let be

  $F_{\dot{u}}(H,u) = \dfrac{1}{\rho_0}\left[2c_1\left(1-\dfrac{1}{(1+\partial_x u)^3}\right)+\sigma_{max}H^2\right](1+\partial_x u)$ and thus
  $\partial_H F_{\dot{u}} = \dfrac{2}{\rho_0}\sigma_{max}H(1+\partial_x u).$

$$\lim_{\epsilon\to 0}\frac{\mathscr{F}_{7,i}(H_h^{n+1}+\epsilon\delta H_h)-\mathscr{F}_{7,i}(H_h^{n+1})}{\epsilon} = \theta\sum_{j=1}^{N_h}\delta H_j\int_0^1\partial_H F_{\dot{u}}(\theta H_h^{n+1}+(1-\theta)H_h^n,\theta u_h^{n+1}+(1-\theta)u_h^n)\varphi_j\partial_x\varphi_i\,\mathrm{d}x$$
$$=\sum_{j=1}^{N_h}\delta H_j J_{75,ij}.$$

- $J_{76,ij}$.

  $$\partial_{\partial_x u} F_{\dot{u}} = \frac{1}{\rho_0}\left[2c1\left(1+\frac{2}{(1+\partial_x u)^3}\right)+\sigma_{max}H^2\right].$$

  $$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{7,i}(u_h^{n+1}+\epsilon\delta u_h)-\mathscr{F}_{7,i}(u_h^{n+1})}{\epsilon} &= \theta\sum_{j=1}^{N_h}\delta u_j\int_0^1 \partial_{\partial_x u}F_{\dot{u}}(\theta H_h^{n+1}+(1-\theta)H_h^n,\theta u_h^{n+1}+(1-\theta)u_h^n)\dots\\ &\quad\dots\partial_x\varphi_j\partial_x\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta u_j J_{76,ij}.\end{aligned}$$

- $J_{77,ij}$.

  $$\partial_{\dot{u}} F_{\dot{u}} = 0.$$

  $$\begin{aligned}\lim_{\epsilon\to 0}\frac{\mathscr{F}_{7,i}(\dot{u}_h^{n+1}+\epsilon\delta\dot{u}_h)-\mathscr{F}_{7,i}(\dot{u}_h^{n+1})}{\epsilon} &= \sum_{j=1}^{N_h}\delta\dot{u}_j\frac{1}{\Delta t}\int_0^1\varphi_j\varphi_i\,\mathrm{d}x\\ &= \sum_{j=1}^{N_h}\delta\dot{u}_j J_{77,ij}.\end{aligned}$$

Chapter 5

Numerical Results

5.1 Problem Settings

Before starting with the numerical experiments, it is necessary to make some comments on the choice of the several parameters characterizing the whole model. We solve the problem for the one-dimensional domain $\Omega = (0,1)$ [cm], and for $t \in (0,T)$ [ms], where the final time is chosen depending on what we aim at computing. If not specified otherwise, in the numerical results we show we have always $T = 750$ [ms]. The mesh size h is set to 0.01 [cm]. We will show results for different choices of Δt [ms]. However the AP model, as it was presented so far, involves only dimensionless quantities. If we do not intend to change its formulation, then simply the time step associated to the AP model, which we call Δt_{AP}, has to be rescaled accordingly to (1.1.8)

$$\Delta t_{AP} = \Delta t / 12.9 \,[\text{t. u.}],$$

where [t. u.] denotes adimensional time units. In Table 5.1 we report all the parameters values adopted for each model.

AP model	Value	BCS/Mech. Model	Value
μ	0.001	k_0	$0.6\,[\text{g cm}^{-1}\text{ms}^{-2}]$
k	8	σ_0	$0.7\,[\text{g cm}^{-1}\text{ms}^{-2}]$
a	0.15	k_{rs}	$0.02\,[\text{ms}^{-1}]$
G_s	0.5	k_{atp}	$0.009\,[\text{ms}^{-1}]$
E_s	1	v_{a1}	0.8
ε_0	0.002	v_{a2}	0.95
μ_1	0.2	σ_{max}	$0.5\,[\text{g cm}^{-1}\text{ms}^{-2}]$
μ_2	0.3	c_1	$0.02\,[\text{g cm}^{-1}\text{ms}^{-2}]$
		ρ_0	$1\,[\text{g cm}^{-3}]$

Table 5.1: Parameter values chosen for the numerical experiments.

The input current I_{app} in the AP model is defined as

$$I_{app}(t_{AP},x) = \begin{cases} I_m(t_{AP}-t_0)/t_\varepsilon & \text{if } t_0 \le t_{AP} < t_0 + t_\varepsilon,\ x \le 0.1, \\ I_m & \text{if } t_0 + t_\varepsilon \le t_{AP} < t_1 - t_\varepsilon,\ x \le 0.1, \\ I_m(t_1 - t_{AP})/t_\varepsilon & \text{if } t_1 - t_\varepsilon \le t_{AP} < t_1,\ x \le 0.1, \end{cases}$$

where we take $I_m = 0.05$, $t_0 = 0$, $t_1 = 0.7$ and $t_\varepsilon = 0.1$. It remains to discuss how to choose the parameters ν and α characterizing the evolution law of the internal variable H. First of all, due to the definition of H, they have to be calibrated so that $0 \le H \le 1$. Particular attention has to be given to the choice of the parameter α, which represents the time constant of the evolution law chosen for H. If α is very small,

H changes rapidly, and numerically instabilities can occur if Δt is not small enough. We choose to use $\alpha = 5$ and $\nu = 0.8$.

We know from theory that the stability of the θ-method depends on the choice of the parameter $\theta \in [0,1]$. Particular cases are

- $\theta = 0$ for which we obtain the explicit Euler's method,
- $\theta = 1$ for which we obtain the implicit Euler's method,
- $\theta = 0.5$ for which we obtain the Crank-Nicolson method.

If we consider as model problem the equation

$$\dot{y}(t) = -\lambda^m y(t),$$

where λ^m is the maximum eigenvalue associated to our system, at each time point t^{n+1} the θ-method produces approximations of the form

$$y^{n+1} = Ay^n \quad \text{where} \quad A = \frac{1-(1-\theta)\Delta t\lambda^m}{1+\theta\Delta t\lambda^m},$$

and A is sometimes called the amplification factor. Solving the problem by using the explicit Euler's method is not efficient, since it requires a very small time step. For our settings we observe that, even using $\Delta t = 0.01$ [ms], the solution explodes after few time iterations. It is known that the θ-method is unconditionally stable for $\theta \geq 0.5$, and therefore for stability purposes we restrict our choice between $\theta = 1$ and $\theta = 0.5$. Of course, since the method for $\theta = 0.5$ is of order 2, $\theta = 0.5$ is usually preferred. However, if Δt is not small enough, with $\theta = 0.5$ the method can produce oscillatory solutions, and lose quadratic convergence. The value of $\Delta t\lambda^m$ for which $A = 0$, is called the oscillatory limit because for greater values, the sign of y^n changes from step to step, and then the scheme produces oscillatory approximations. Moreover, since for each time point we solve a non-linear system using the Newton's method, and the initial guess $\mathbf{y}^{n+1,0}$ is chosen as $\mathbf{y}^n$, Newton's method can have also difficulties to converge.

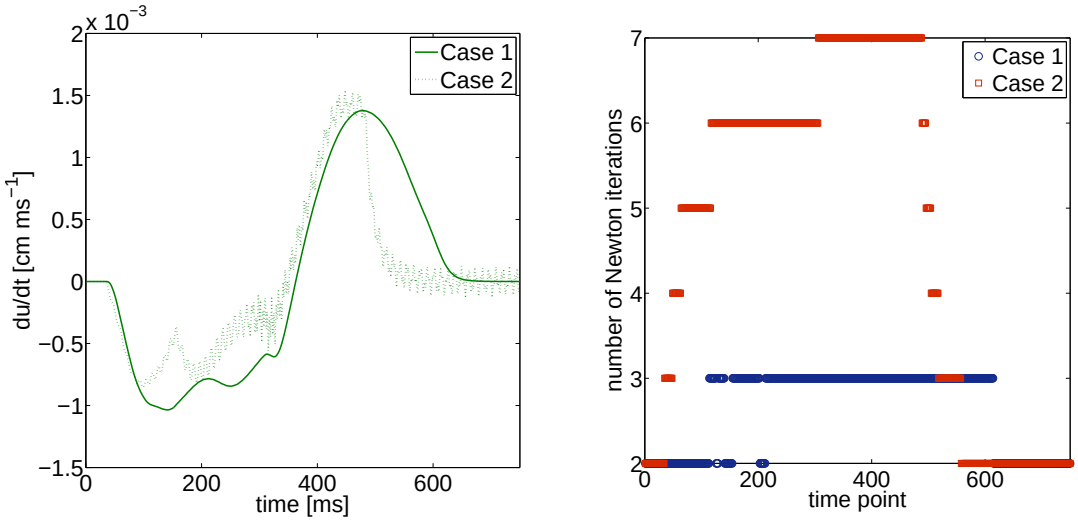


Figure 5.1: Comparison between numerical solutions obtained for Case 1 and Case 2, using $\Delta = 1$ [ms] and $\theta = 0.5$.

To illustrate the phenomena in Figure 5.1 we show the numerical approximations of $\dot{u}$ we get by solving the problem using $\Delta t = 1$ [ms] and $\theta = 0.5$, for two different set of values of α and ν:

- Case 1: $\alpha = 5$, $\nu = 0.8$.
- Case 2: $\alpha = 0.2$, $\nu = 0.5$.

The numerical solutions are evaluated in the last node of the grid. Also shown is the number of Newton's iterations needed for each time step. Newton's stopping tolerance is $\epsilon_N = 10^{-6}$ and the criterion is based on relative residual. The maximum number of Newton's iterations is set to 20. For both set of values H stays bounded between 0 and 1. From the figure we can observe that for Case 2, due to the smaller value chosen for α, the time-step chosen appears to be too large if we want to solve the problem properly by using $\theta = 0.5$. Indeed the internal variable changes rapidly, numerical oscillations can be observed in the approximated solution of $\dot{u}$, and the Newton's method has difficulties to converge.

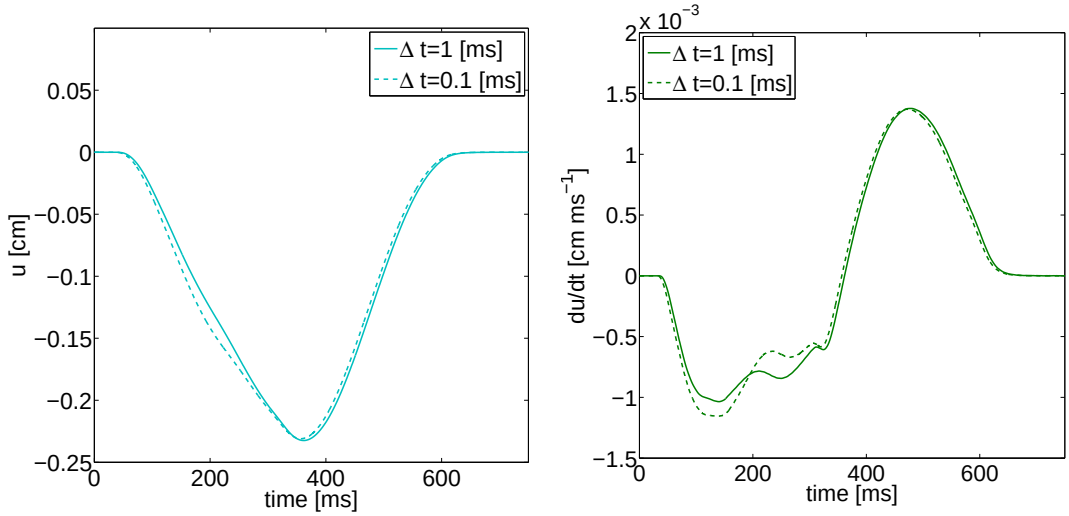


Figure 5.2: Comparison between numerical approximations of the solution to the mechanical model obtained for different time steps and using $\theta = 0.5$.

However we point out that, even if the solution obtained in Case 1 does not present numerical instabilities, if we compare it to the ones obtained by using $\Delta t = 0.1$ [ms], it seems far from being sufficiently accurate. Main differences are observed in the solutions to the mechanical model, which is the one more prone to numerical instabilities. In Figure 5.2 we compare the approximations of u and $\dot{u}$ we get by using $\theta = 0.5$ and $\Delta t \in \{1, 0.1\}$ [ms].

To conclude, for sake of completeness, in Figure 5.3 is shown the numerical solution obtained using $\theta = 0.5$ and $\Delta t = 0.1$ [ms], evaluated in four different nodes of the grid, $x \in \{0.25, 0.5, 0.75, 1\}$ [cm].

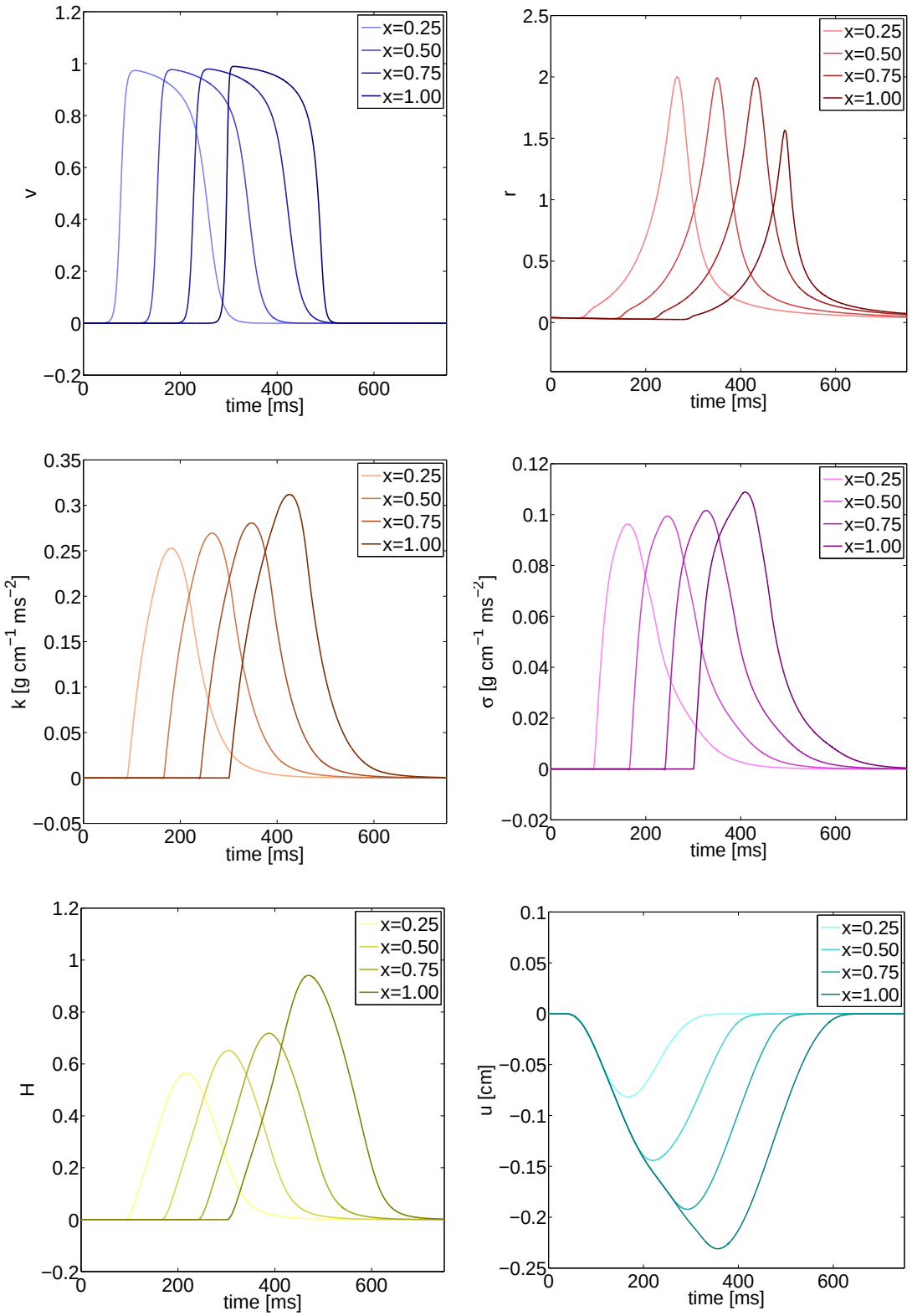


Figure 5.3: (a) Numerical solutions obtained with $\Delta t = 0.1$ [ms] evaluated in $x = 0.25$, $x = 0.5$, $x = 0.75$ and $x = 1$.

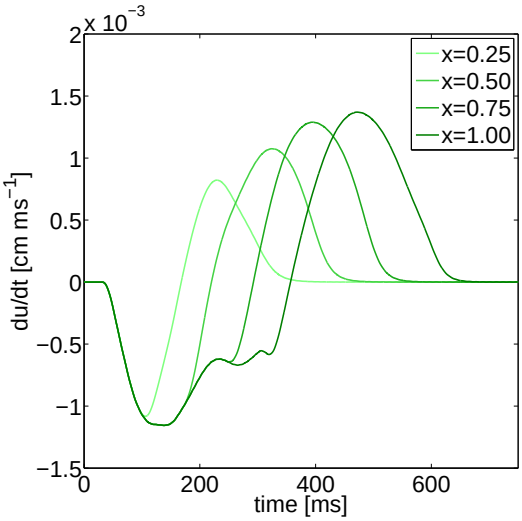


Figure 5.3: (b) Numerical solutions obtained with $\Delta t = 0.1$ [ms] evaluated in $x = 0.25$, $x = 0.5$, $x = 0.75$ and $x = 1$.

5.2 Treating Some of the Coupling Terms Explicitly

So far we have used the Crank-Nicolson method to solve the ODE system (4.3.2). However it would be more efficient to use implicit solvers only when needed, and to treat explicitly the terms that do not suffer from stability issues. As example, for now let us focus only on the AP model, and let us consider only the coupling between v and r. Then we will extend to the fully coupled problem. The ODE system associate to the AP model is

$$\begin{bmatrix} \mathbf{M}_{N_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{N_h} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}(t) \\ \dot{\mathbf{r}}(t) \end{bmatrix} = \begin{bmatrix} \mathscr{L}_1'(\mathbf{v}(t), \mathbf{r}(t)) \\ \mathscr{L}_2(\mathbf{v}(t), \mathbf{r}(t)) \end{bmatrix}, \tag{5.2.1}$$

where

$$\mathscr{L}_{1,i}'(\mathbf{v}(t), \mathbf{r}(t)) = -\int_0^1 \mu \frac{\partial v_h}{\partial x} \frac{\partial \varphi_i}{\partial x} \, \mathrm{d}x + \int_0^1 f(v_h, r_h)\varphi_i \, \mathrm{d}x.$$

The θ-method applied to (5.2.1) produces

$$\frac{1}{\Delta t} \begin{bmatrix} \mathbf{M}_{N_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{N_h} \end{bmatrix} \begin{bmatrix} \mathbf{v}^{n+1} - \mathbf{v}^n \\ \mathbf{r}^{n+1} - \mathbf{r}^n \end{bmatrix} = \begin{bmatrix} \mathscr{L}_1'(\theta \mathbf{v}^{n+1} + (1-\theta)\mathbf{v}^n, \theta \mathbf{r}^{n+1} + (1-\theta)\mathbf{r}^n) \\ \mathscr{L}_2(\theta \mathbf{v}^{n+1} + (1-\theta)\mathbf{v}^n, \theta \mathbf{r}^{n+1} + (1-\theta)\mathbf{r}^n) \end{bmatrix}, \tag{5.2.2}$$

which is a non-linear system which can be solved, as already seen, with the Newton's method. In Chapter 1 we discussed about stability when solving explicitly (5.2.1), and we have seen that time step restriction is mainly due to the diffusion term characterizing the evolution law of v. Then we can think of treating only that part of the problem implicitly, by using the Crank-Nicolson or the implicit Euler's method, while the other terms are solved explicitly by using the explicit Euler's method. A simply way of doing that is to choose $\theta = 0.5$ or $\theta = 1$ for the time discretization of $\mathbf{v}(t)$, input argument of $\mathscr{L}_1'$ in (5.2.2), and $\theta = 0$ for all the others arguments. The choice $\theta = 0.5$ produces

$$\frac{1}{\Delta t} \begin{bmatrix} \mathbf{M}_{N_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{N_h} \end{bmatrix} \begin{bmatrix} \mathbf{v}^{n+1} - \mathbf{v}^n \\ \mathbf{r}^{n+1} - \mathbf{r}^n \end{bmatrix} = \begin{bmatrix} \mathscr{L}_1'(0.5(\mathbf{v}^{n+1} + \mathbf{v}^n), \mathbf{r}^n) \\ \mathscr{L}_2(\mathbf{v}^n, \mathbf{r}^n) \end{bmatrix},$$

and, after linearisation, the linear system we need to solve is

$$\begin{bmatrix} \mathbf{J}_{11} & \\ & \mathbf{M}_{N_h}/\Delta t \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{v}^k \\ \boldsymbol{\delta}\mathbf{r}^k \end{bmatrix} = -\begin{bmatrix} \mathscr{F}_1'(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) \\ \mathscr{F}_2(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) \end{bmatrix}. \tag{5.2.3}$$

where

$$\begin{aligned} \mathscr{F}_1'(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) &= \mathbf{M}_{N_h} \frac{\mathbf{v}^{n+1,k} - \mathbf{v}^n}{\Delta t} - \mathscr{L}_1'(0.5(\mathbf{v}^{n+1,k} + \mathbf{v}^n), \mathbf{r}^n), \\ \mathscr{F}_2(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) &= \mathbf{M}_{N_h} \frac{\mathbf{r}^{n+1,k} - \mathbf{r}^n}{\Delta t} - \mathscr{L}_2(\mathbf{v}^n, \mathbf{r}^n). \end{aligned}$$

Let us remark that the off-diagonal blocks in the Jacobian matrix have disappeared. Furthermore, the only contribution to $\mathbf{J}_{22}$ is the one relative to the time discretization, while

$$\mathbf{J}_{11} = \frac{\mathbf{M}_{N_h}}{\Delta t} - \mathbf{J}_{11,\mathscr{L}_1'}$$

is computed similarly as shown in Section 4.3.1.

An elegant way to implement a method that allows to choose which one of the several contributions we want to solve explicitly, it is to write (5.2.2) as

$$\frac{1}{\Delta t} \begin{bmatrix} \mathbf{M}_{N_h} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{N_h} \end{bmatrix} \begin{bmatrix} \mathbf{v}^{n+1} - \mathbf{v}^n \\ \mathbf{r}^{n+1} - \mathbf{r}^n \end{bmatrix} = \begin{bmatrix} \mathscr{L}_1'(a_{11}\theta \mathbf{v}^{n+1} + (1-a_{11}\theta)\mathbf{v}^n, a_{12}\theta \mathbf{r}^{n+1} + (1-a_{12}\theta)\mathbf{r}^n) \\ \mathscr{L}_2(a_{21}\theta \mathbf{v}^{n+1} + (1-a_{21}\theta)\mathbf{v}^n, a_{22}\theta \mathbf{r}^{n+1} + (1-a_{22}\theta)\mathbf{r}^n) \end{bmatrix}.$$

Then, after linearisation, the linear system we need to solve is

$$\begin{bmatrix} \mathbf{M}_{N_h}/\Delta t - a_{11}\mathbf{J}_{11,\mathscr{L}_1'} & a_{12}\mathbf{J}_{12} \\ a_{21}\mathbf{J}_{21} & \mathbf{M}_{N_h}/\Delta t - a_{22}\mathbf{J}_{22,\mathscr{L}_2'} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{v}^k \\ \boldsymbol{\delta}\mathbf{r}^k \end{bmatrix} = - \begin{bmatrix} \mathscr{F}_1'(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) \\ \mathscr{F}_2(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) \end{bmatrix},$$

where

$$\mathscr{F}_1'(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) = \mathbf{M}_{N_h} \frac{\mathbf{v}^{n+1,k} - \mathbf{v}^n}{\Delta t} - \mathscr{L}_1'(a_{11}\theta\mathbf{v}^{n+1,k} + (1 - a_{11}\theta)\mathbf{v}^n), a_{12}\theta\mathbf{r}^{n+1} + (1 - a_{12}\theta)\mathbf{r}^n),$$
$$\mathscr{F}_2(\mathbf{v}^{n+1,k}, \mathbf{r}^{n+1,k}) = \mathbf{M}_{N_h} \frac{\mathbf{r}^{n+1,k} - \mathbf{r}^n}{\Delta t} - \mathscr{L}_2(a_{21}\theta\mathbf{v}^{n+1,k} + (1 - a_{21}\theta)\mathbf{v}^n), a_{22}\theta\mathbf{r}^{n+1} + (1 - a_{22}\theta)\mathbf{r}^n).$$

Therefore, for each contribution, we can choose if treating it implicitly or explicitly by only setting a_{ij} respectively to 1 or 0. As example, the system (5.2.3) is obtained with the choice

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Also in the fully coupled problem we can choose to solve some of the several contributions explicitly. We define the matrix $\mathbf{A}$ of coefficients a_{ij}, one coefficient for each of the block of the Jacobian matrix in (4.3.4),

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & & & & a_{16} & \\ a_{21} & a_{22} & & & & & \\ a_{31} & & a_{33} & & & & a_{37} \\ a_{41} & & a_{43} & a_{44} & & & a_{47} \\ & & & a_{54} & a_{55} & a_{56} & \\ & & & & & a_{66} & a_{67} \\ & & & & a_{75} & a_{76} & a_{77} \end{bmatrix},$$

and then we do some numerical experiments to find which choices of $a_{ij} = 0$ can cause numerical instabilities.

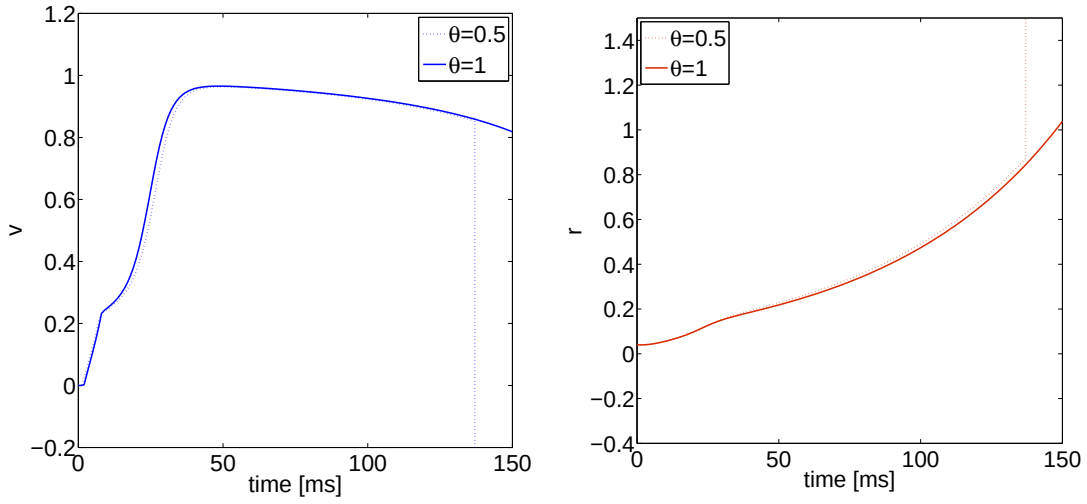


Figure 5.4: (a) Numerical solutions evaluated at $x = 1$, obtained by setting $a_{47} = a_{54} = a_{75} = 0$, $\Delta t = 1$ [ms] and by using respectively $\theta = 0.5$ and $\theta = 1$.

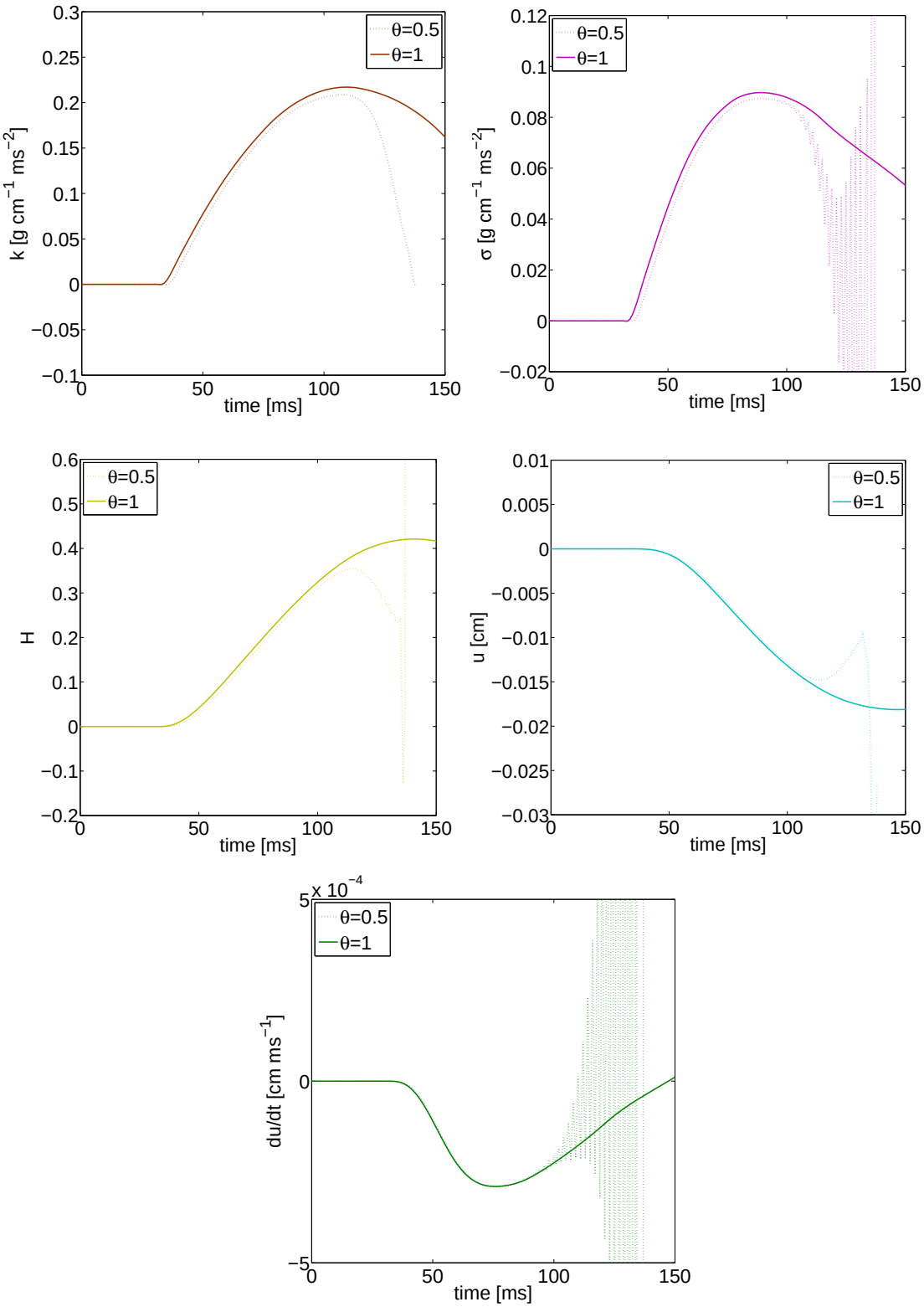


Figure 5.4: (b) Numerical solutions evaluated at $x = 1$, obtained by setting $a_{47} = a_{54} = a_{75} = 0$, $\Delta t = 1$ [ms] and by using respectively $\theta = 0.5$ and $\theta = 1$.

We start by choosing to solve implicitly the mechanical model and the diffusion term in the AP model. Therefore $a_{11} = a_{66} = a_{67} = a_{76} = a_{77} = 1$. Then we consider considering of treating explicitly some of the different coupling terms. In particular we find that when $\theta = 0.5$, having at the same time $a_{47} = a_{54} = a_{75} = 0$ can produce numerical instabilities if Δt is not sufficiently small. This is not the case if one of the coefficients among a_{47}, a_{54} and a_{75} is in turn set to 1. In particular the physical significances of the three coupling terms are the following:

- $a_{47}: \sigma \leftarrow \dot{u}$, is the coupling term responsible of the effect of the strain velocity on the control law of the microscopic stress σ.
- $a_{54}: H \leftarrow \sigma$, is the coupling term which correlates the microscopic stress σ and the macroscopic internal variable H.
- $a_{75}: \dot{u} \leftarrow H$, is the coupling term which correlates the active stress and the mechanical equations.

We obtain that, when $a_{47} = a_{54} = a_{75} = 0$, $\Delta t = 1$ [ms] and $\theta = 0.5$, the numerical scheme does not converge and the solution explodes after few time iterations. However if we decrease Δt to 0.1 [ms] the scheme converges properly and the solution obtained is almost identical to the one compute by solving the problem fully implicitly. On the other hand we find also that, if the implicit Euler's method is used for the parts of the problem that are resolved implicitly, no numerical instabilities occur, even if the only coefficients equal to 1 are a_{11}, a_{66}, a_{67}, a_{76} and a_{77}. Indeed the implicit Euler's method manages to dissipate the occurring numerical oscillations, and the whole system stays stable. In Figure 5.4 we compare the numerical solutions obtained by setting $a_{ij} = 1$ for all $a_{ij} \in \mathbf{A}$ except to a_{47}, a_{54} and a_{75}, which are set to 0, when $\theta = 0.5$ and when $\theta = 1$. The time step is $\Delta t = 1$ [ms] for both cases. The solutions showed are evaluated at $x = 0.05$. The problem settings are the same as defined in Section 5.1. From the figure is possible to observe that the numerical scheme with $\theta = 0.5$ does not converge and the numerical solution explodes after 137 time iterations, while it stays stable when $\theta = 1$. In particular the oscillations start when the system enters the depolarization phase.

5.3 Solving the Linearised System Iteratively

In this section we discuss how to combine the Newton's method with an iterative solution method. Indeed Newton's method requires at each Newton's iteration the construction of a linearised system, whose solution is often obtained by using direct solvers. However it can be thought to adopt iterative solvers in order to reduce the solution cost of the linear system. Then, if we use iterative methods we need a convergence criterion to stop the process, and choosing a suitable resolution precision as well as a good preconditioner are key points of this procedure. Let us consider the solution to system (4.3.4) which we rewrite here as

$$\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})\boldsymbol{\delta}\mathbf{y}^{n+1,k} = -\mathscr{F}(\mathbf{y}^{n+1,k}), \tag{5.3.1}$$

where n and k are the indexes relative respectively to time iteration and Newton's iteration. The iterative method we use is defined by the following iteration:

$$\begin{aligned} \boldsymbol{\delta}\mathbf{y}^{n+1,k,p+1} &= \mathbf{B}\boldsymbol{\delta}\mathbf{y}^{n+1,k,p} + \mathbf{g} \\ &= \mathbf{P}^{-1}\left(\mathbf{P} - \mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})\right)\boldsymbol{\delta}\mathbf{y}^{n+1,k,p} - \mathbf{P}^{-1}\mathscr{F}(\mathbf{y}^{n+1,k}), \end{aligned} \tag{5.3.2}$$

where with p denotes the iteration's index relative to the iterative solver and $\mathbf{P}$ is the preconditioner. The matrix $\mathbf{B}$ is called the iteration matrix and the iterative method converges for any initial guess if and only if its spectral radius is smaller than 1. Then the expression of this matrix fully defines the iterative method. Special instances of iterative methods are obtained by choosing $\mathbf{P} = \mathbf{D}_B$ and $\mathbf{P} = \mathbf{D}_B + \mathbf{L}_B$, where $\mathbf{D}_B$ and $\mathbf{L}_B$ are respectively the block diagonal part and the strict lower block triangular part of $\mathbf{J}_{\mathscr{F}}$. In the first case we get a block version of the Jacobi method (BJ), while the second choice gives a

block version of the Gauss-Seidel method (BGS). The stop criterion of the iterative scheme (5.3.2) is usually based on the relative decrease of the linear residual to a fixed threshold ϵ_L, typically 10^{-6}, that is

$$\frac{||\mathbf{r}^p||_2}{||\mathbf{r}^0||_2} = \frac{||\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})\boldsymbol{\delta}\mathbf{x}^{n+1,k,p} + \mathscr{F}(\mathbf{y}^{n+1,k})||_2}{||\mathbf{J}_{\mathscr{F}}(\mathbf{y}^{n+1,k})\boldsymbol{\delta}\mathbf{x}^{n+1,k,0} + \mathscr{F}(\mathbf{y}^{n+1,k})||_2} < \epsilon_L.$$

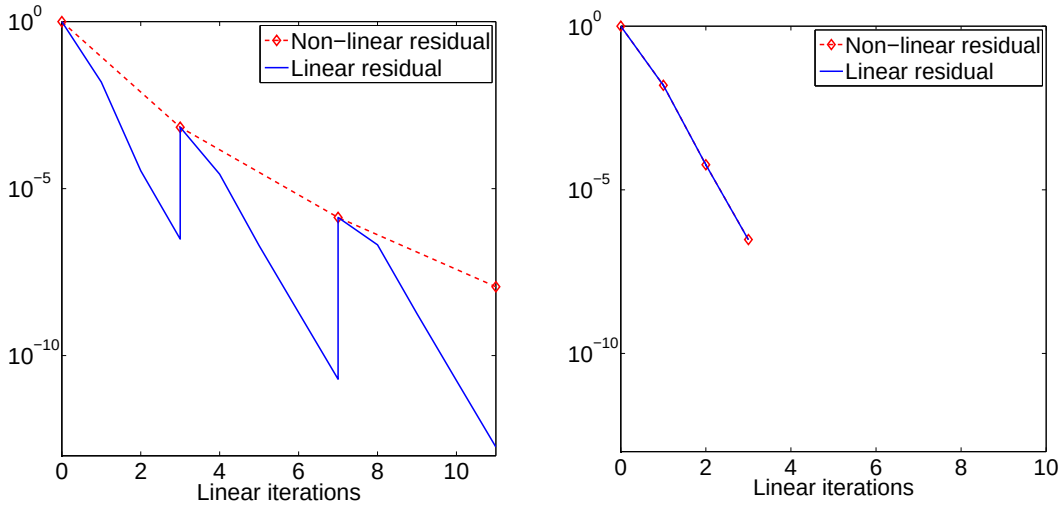


Figure 5.5: (a): Decrease of non-linear and linear residuals with a fixed stopping tolerance $\epsilon_L = 10^{-6}$. (b): Decrease of non-linear and linear residuals when ϵ_L is computed according to (5.3.3), (5.3.4) and (5.3.5). In both cases BGS is used.

However solving the linear system (5.3.1) with an excessively high precision can lead to an unnecessary number of linear iterations, and a good choice of ϵ_L is essential for a good performance of the global non-linear process. The basic idea is the following: when far from the non-linear solution, it is not necessary to compute a precise linear solution, since the same decrease of the non-linear residual can be often obtained with an inexact linear solve as well as a precise one; when getting closer to the non-linear solution, in order to recover the quadratic convergence of the Newton's method, the linear solution must be precise. The tolerance value ϵ_L is then adjusted at each Newton's iteration in order to have a decrease of the non-linear residual proportional to the decrease of the linear residual. In [20, 5] the following adaptive algorithm is proposed

$$\epsilon_L^{k+1} = \epsilon_{res}^{k+1} = \gamma \left(\frac{||\mathscr{F}(\mathbf{y}^{n+1,k})||_2}{||\mathscr{F}(\mathbf{y}^{n+1,k-1})||_2} \right)^{\alpha}, \tag{5.3.3}$$

where $\gamma \in (0,1)$ and $\alpha \in (1,2]$ are two user-defined parameters. In particular, to ensure quadratic convergence, α must be set equal to 2. However, as pointed out in [20, 32], to avoid a too fast decrease of the sequence ϵ_L^{k+1}, causing an unnecessary resolution, or to get unacceptable values of $\epsilon_L^k > 1$, equation (5.3.3) is reinforced with the following condition

$$\epsilon_L^{k+1} = \min\left(\epsilon_{max}, \max\left(\epsilon_{safe}, \frac{\epsilon_N}{2\gamma}\epsilon_{res}^{k+1}\right)\right), \tag{5.3.4}$$

where ϵ_N is the tolerance of the Newton's method, ϵ_{max} is an user-defined parameter, and ϵ_{safe} is given by [32]:

$$\epsilon_{safe} = \begin{cases} \epsilon_{max} & \text{if } k = 0, \\ \max\left(\min\left(\frac{\epsilon_L^k}{2}, \epsilon_{res}^{k+1}\right), \epsilon_{min}\right) & \text{if } k > 0,\ \gamma\left(\epsilon_L^k\right)^2 \leq 0.1, \\ \min\left(\frac{\epsilon_L^k}{2}, \max\left(\epsilon_{res}^{k+1}, \gamma\left(\epsilon^k\right)^2\right)\right) & \text{if } k > 0,\ \gamma\left(\epsilon_L^k\right)^2 > 0.1, \end{cases} \tag{5.3.5}$$

where ϵ_{min} is another user-defined parameter. In particular we choose to use $\gamma = 0.9$, $\epsilon_{max} = 0.9$, $\epsilon_{min} = 10^{-4}$, which provide safe convergence and little cost of linear solving. In Figure 5.5 we compare the decrease of the non-linear relative residual (red dotted line) when ϵ_L is held fixed to 10^{-6} and when it is adapted according to (5.3.3), (5.3.4) and (5.3.5). The solid lines are the values of the linear residual divided by $||\mathscr{F}(\mathbf{y}^{n+1,0})||_2$. The plot is relative to one time step of the numerical simulation and the BGS method is used as linear solver. It is clear from the plot that $\epsilon_L = 10^{-6}$ is unnecessary, since the non-linear residual does not decrease more despite the linear solution is more accurate.

What remains to be discussed is how to choose an efficient preconditioner for solving the linear system (5.3.1). From (5.3.2) it is easy to realize that requirements for $\mathbf{P}$ are being non-singular and easily invertible to reduce the cost of solving (5.3.2). Moreover let us underline that if $\mathbf{P} = \mathbf{J}_{\mathscr{F}}$, the iterative method will converge in one iteration, but at the same cost of an exact solver. The matrix $\mathbf{J}_{\mathscr{F}}$ is block structured and we think of it as a 6×6 block-matrix

$$\mathbf{J}_{\mathscr{F}} = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & & & & \mathbf{J}_{16}^* \\ \mathbf{J}_{21} & \mathbf{J}_{22} & & & & \\ \mathbf{J}_{31} & & \mathbf{J}_{33} & & & \mathbf{J}_{36}^* \\ \mathbf{J}_{41} & & \mathbf{J}_{43} & \mathbf{J}_{44} & & \mathbf{J}_{46}^* \\ & & & \mathbf{J}_{54} & \mathbf{J}_{55} & \mathbf{J}_{56}^* \\ & & & & \mathbf{J}_{65}^* & \mathbf{J}_{66}^* \end{bmatrix}$$

where $\mathbf{J}_{ij}$ are the matrices defined in Section 4.3.1, and we have grouped together the blocks relative to the mechanical model, so that

$$\mathbf{J}_{66}^* = \begin{bmatrix} \mathbf{J}_{66} & \mathbf{J}_{67} \\ \mathbf{J}_{76} & \mathbf{J}_{77} \end{bmatrix}$$

and so on. We consider choices for the preconditioner $\mathbf{P}$ such that $\mathbf{P}$ is a block lower triangular matrix or a block upper triangular matrix. Then the linear system (5.3.2) can be easily solved by a block version of the forward substitutions algorithm (Algorithm 5.1.) or a block version of the backward substitutions algorithm (Algorithm 5.2.).

Algorithm 5.1. *Block Forward Substitutions (BFS).* Let be $\mathbf{L}$ a block lower triangular matrix, and let us consider the block structured linear system

$$\mathbf{L}\mathbf{y} = \mathbf{b},$$

so that

$$\mathbf{L} = [\mathbf{L}_{ij}], \quad \mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T, \quad \mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T, \quad i = 1, \ldots, n,\ j = 1, \ldots, i.$$

The procedure to solve the system is the following:

1. Solve

$$\mathbf{L}_{11}\mathbf{y}_1 = \mathbf{b}_1. \tag{5.3.6}$$

2. Solve

$$\mathbf{L}_{ii}\mathbf{y}_i = \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{L}_{ij}\mathbf{y}_j, \quad i = 2, \ldots, n. \tag{5.3.7}$$

Algorithm 5.2. *Block Backward Substitutions (BBS).* Let be $\mathbf{U}$ a block upper triangular matrix, and let us consider the block structured linear system

$$\mathbf{U}\mathbf{y} = \mathbf{b},$$

so that

$$\mathbf{U} = [\mathbf{U}_{ij}], \quad \mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T, \quad \mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)^T, \quad i = 1, \dots, n,\ j = i, \dots, n.$$

The procedure to solve the system is the following:

1. Solve

$$\mathbf{U}_{nn}\mathbf{y}_n = \mathbf{b}_n. \tag{5.3.8}$$

2. Solve

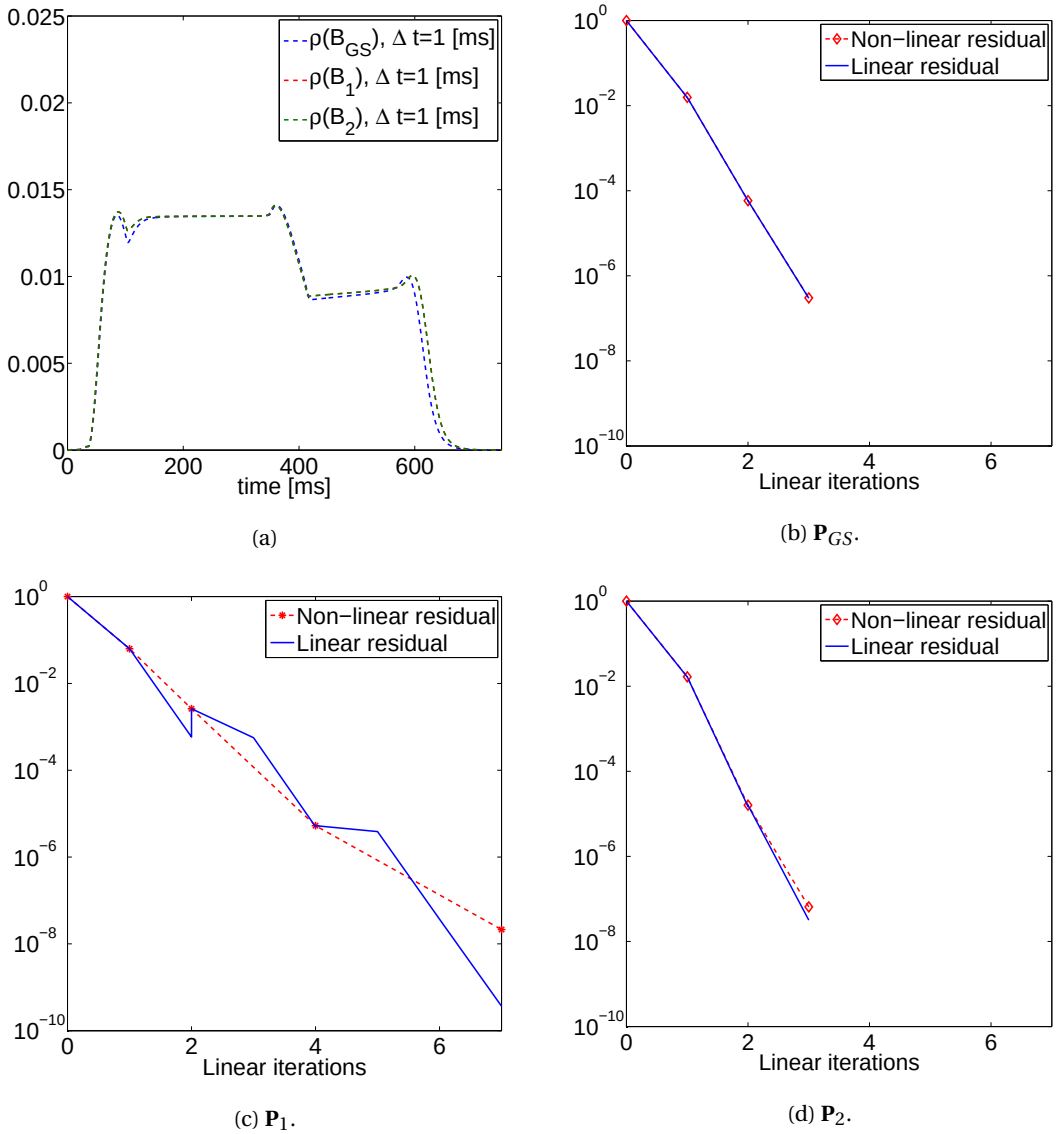
$$\mathbf{U}_{ii}\mathbf{y}_i = \mathbf{b}_i - \sum_{j=i+1}^{n} \mathbf{U}_{ij}\mathbf{y}_j, \quad i = n-1, \dots, 1. \tag{5.3.9}$$

We compare different alternatives for $\mathbf{P}$ in terms of total number of Newton' iterations and total number of linear iterations performed during a possible simulation. The first two choices for $\mathbf{P}$ are given by the BJ method and the BGS method. We call the two preconditioners respectively $\mathbf{P}_J$ and $\mathbf{P}_{GS}$. Both choices allow for a resolution of (5.3.2) by BFS. From theory it can be guessed that BGS will show faster convergence, reducing the number of linear iterations. On the other hand the updating of the solutions in BGS has to be made sequentially, while BJ allows for a simultaneous updating which could be performed in parallel. The preconditioner used in BGS is the block lower triangular part of $\mathbf{J}_{\mathscr{F}}$ and it is constituted of 12 blocks. Then we find other choices for $\mathbf{P}$ by looking for preconditioners which possess more than 12 blocks and that can be placed into a block lower/upper triangular matrix by simultaneous row/column permutations, and we check if with such choices we can reach faster convergence. In particular, after permutation of the system (5.3.1), we consider the following two alternatives for $\mathbf{P}$,

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{J}^*_{66} & & & & & \\ \mathbf{J}^*_{16} & \mathbf{J}_{11} & & & & \\ & \mathbf{J}_{21} & \mathbf{J}_{22} & & & \\ \mathbf{J}^*_{36} & \mathbf{J}_{31} & & \mathbf{J}_{33} & & \\ \mathbf{J}^*_{46} & \mathbf{J}_{41} & & \mathbf{J}_{43} & \mathbf{J}_{44} & \\ \mathbf{J}^*_{56} & & & & \mathbf{J}_{54} & \mathbf{J}_{55} \end{bmatrix} \quad \text{and} \quad \mathbf{P}_2 = \begin{bmatrix} \mathbf{J}_{55} & & & & & \\ \mathbf{J}^*_{65} & \mathbf{J}^*_{66} & & & & \\ & \mathbf{J}^*_{16} & \mathbf{J}_{11} & & & \\ & & \mathbf{J}_{21} & \mathbf{J}_{22} & & \\ & \mathbf{J}^*_{36} & \mathbf{J}_{31} & & \mathbf{J}_{33} & \\ & \mathbf{J}^*_{46} & \mathbf{J}_{41} & & \mathbf{J}_{43} & \mathbf{J}_{44} \end{bmatrix},$$

which allow for a resolution of the linear iteration by BFS. The two choices are characterized by the fact that mechanics is solved before the electrical model.

We perform a numerical simulation for $t \in (0, T)$, $T = 750$ [ms] with the parameter settings defined in Section 5.1. We use $\Delta t = 1$ [ms] and the Crank-Nicolson method for time integration. We measure the total number of Newton and linear iterations used by solving the problem with the different preconditioners, for different values of the Newton's tolerance ϵ_N. The results are reported in Tables 5.2, 5.3 and 5.4. We find that by using $\mathbf{P}_{GS}$ or $\mathbf{P}_2$ as preconditioner, for $\epsilon_N = 10^{-6}$ we do one linear iteration for each Newton's iteration. This result suggests that the lower blocks of the Jacobian have more influence on the decrease of the non-linear and linear residuals, than the upper ones. An additional indication which leads to this conclusion is the fact that, although $\mathbf{P}_1$ possess more blocks than $\mathbf{P}_{GS}$ and $\mathbf{P}_2$, and the magnitude of the spectral radius of its iteration matrix is comparable to the ones observed when $\mathbf{P}_{GS}$ or $\mathbf{P}_2$ are used, using it as preconditioner causes slower convergence. In Figure 5.6(a) we plot $\rho(\mathbf{B}_{GS})$, $\rho(\mathbf{B}_1)$ and $\rho(\mathbf{B}_2)$ as function of time, where $\rho(\mathbf{B}_{GS})$ is the spectral radius of the iteration matrix associated to BGS and so on. They are computed numerically at each time iteration during the first Newton's iteration. Although the magnitude of $\rho(\mathbf{B}_1)$ is equal to the ones of $\rho(\mathbf{B}_2)$ (in Figure 5.6(a) the



(a)

(b) $\mathbf{P}_{GS}$.

(c) $\mathbf{P}_1$.

(d) $\mathbf{P}_2$.

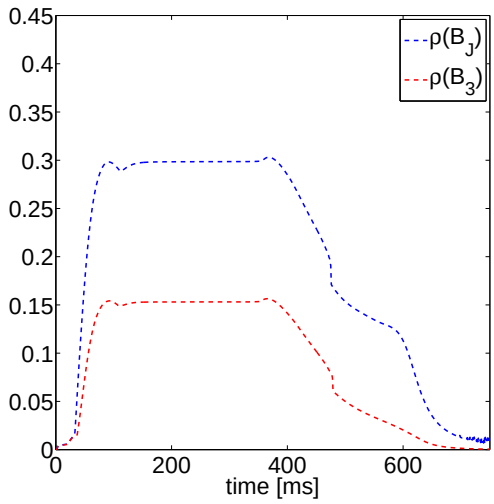
Figure 5.6: (a): Spectral radii of $\mathbf{B}_{GS}$, $\mathbf{B}_1$ and $\mathbf{B}_2$ as function of time for $\Delta t = 1$ [ms]. (b): Decrease of non-linear and linear residual when $\mathbf{P}_{GS}$ is used. (c): Decrease of non-linear and linear residual when $\mathbf{P}_1$ is used. (d): Decrease of non-linear and linear residual when $\mathbf{P}_2$ is used.

two lines overlap perfectly), using $\mathbf{P}_1$ as preconditioner gives as result a higher number of Newton's iterations, and therefore a higher number of linear iterations. In Figure 5.6 is showed the decrease of the non-linear residual and the linear residual when $\mathbf{P}_{GS}$, $\mathbf{P}_1$ and $\mathbf{P}_2$ are used. The plot is relative to one time step of the numerical simulation ($t = 400$ [ms]). From the figure we can observe that when $\mathbf{P}_1$ is used the non-linear residual decreases slower. Then, one more Newton's iteration could imply more than one linear iteration, especially when the non-linear residual is relatively small. We guess that this result is due to the fact that $\mathbf{P}_1$ does not include the block $\mathbf{J}^*_{65}$, relative to the coupling between the active stress and the mechanical equations, which results to have more influence on the decrease of the residuals than the other blocks placed over the block diagonal of $\mathbf{J}_{\mathscr{F}}$.

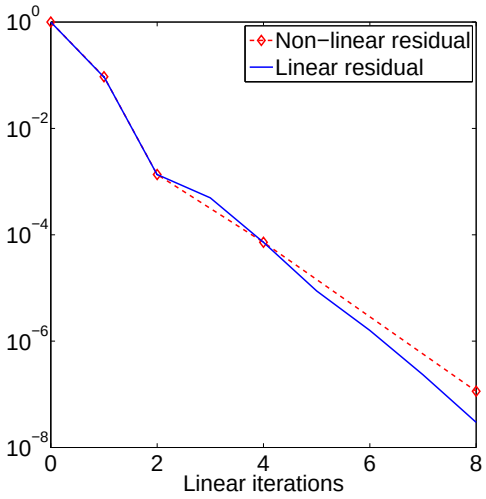
To verify our guess we perform additional simulations by considering the preconditioner

$$\mathbf{P}_3 = \begin{bmatrix} \mathbf{J}_{11} & & & & & & \\ & \mathbf{J}_{22} & & & & & \\ & & \mathbf{J}_{33} & & & & \\ & & & \mathbf{J}_{44} & & & \\ & & & & \mathbf{J}_{55} & & \\ & & & & \mathbf{J}^*_{65} & \mathbf{J}^*_{66} & \end{bmatrix},$$

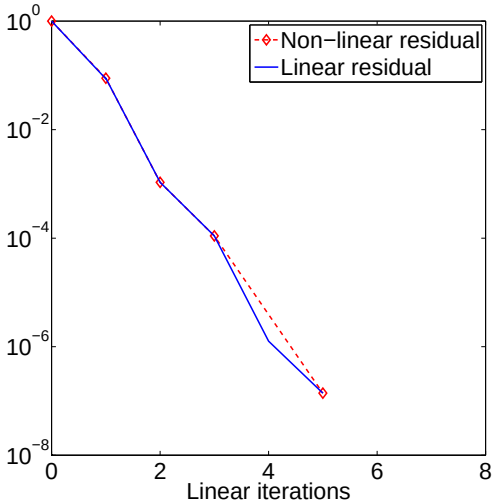
which possess only the block $\mathbf{J}^*_{65}$ more than $\mathbf{P}_J$. We find that by adding only this block reduces significantly the number of Newton's and linear iterations (these about by a factor 2) performed by BJ. In Figure 5.7(a) we compare $\rho(\mathbf{B}_J)$ and $\rho(\mathbf{B}_3)$, while in Figure 5.7(b) and (c) is plotted the decrease of the non-linear and linear residuals for one time step of the numerical simulation ($t = 500$ [ms]).



(a)



(b) $\mathbf{P}_J$.



(c) $\mathbf{P}_3$.

Figure 5.7: (a): Spectral radius of $\mathbf{B}_J$ and $\mathbf{B}_3$ as function of time for $\Delta t = 1$ [ms]. (b): Decrease of non-linear and linear residuals when $\mathbf{P}_J$ is used. (c): Decrease of non-linear and linear residuals when $\mathbf{P}_3$ is used.

On the other hand, it is important to underline that, although BJ employs more Newton's and linear iterations than the other choices, it still manages to converge. Thus, by considering the fact that BJ allows for simultaneous updating of the different state variables, using BJ as iterative solver could also represent a very efficient way of solving the system. Furthermore, decreasing Δt reduces $\rho(\mathbf{B}_J)$ and BJ converges faster. In Table 5.5 are reported the total number of Newton's iterations and the total number of linear iterations used by BJ to converge for different choices of Δt. The tolerance for the non-linear residual is set to 10^{-6}. In Figure 5.8(a) is plotted the spectral radius $\rho(\mathbf{B}_J)$ as function of time for different choices of Δt, while in Figure 5.8(b), (c) and (d) is shown the decrease of non-linear and linear residuals for BJ for different choices of Δt (the plot is relative to one time iteration: $t = 500$ [ms]).

Prec.	# New. its.	# lin. its.	# New. its./# time its.	# lin. its. /# New. its.
$\mathbf{P}_J$	3256	8201	4.34	2.52
$\mathbf{P}_{GS}$	2155	2186	2.87	1.01
$\mathbf{P}_1$	2777	4629	3.70	1.67
$\mathbf{P}_2$	2144	2170	2.86	1.01
$\mathbf{P}_3$	2683	3693	3.58	1.38

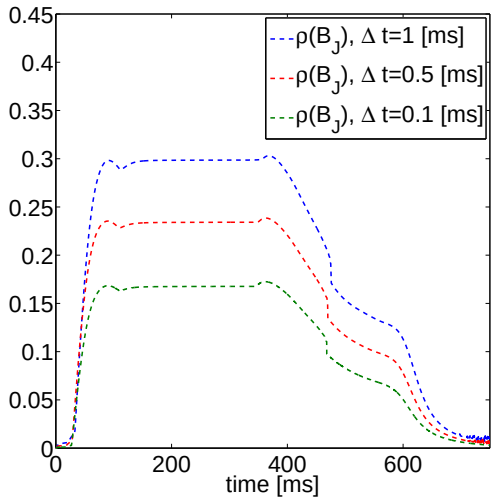
Table 5.2: Comparison between different preconditioners in terms of total number of Newton's iterations (# New. its.) and total number of linear iterations (# lin. its.) performed during the numerical simulation when $\epsilon_N = 10^{-6}$ ($\Delta t = 1$ [ms]).

Prec.	# New. its.	# lin. its.	# New. its./# time its.	# lin. its. /# New. its.
$\mathbf{P}_J$	3408	9001	4.54	2.64
$\mathbf{P}_{GS}$	2698	3311	3.60	1.23
$\mathbf{P}_1$	2999	5293	4.00	1.76
$\mathbf{P}_2$	2278	2393	3.07	1.05
$\mathbf{P}_3$	2852	4078	3.80	1.43

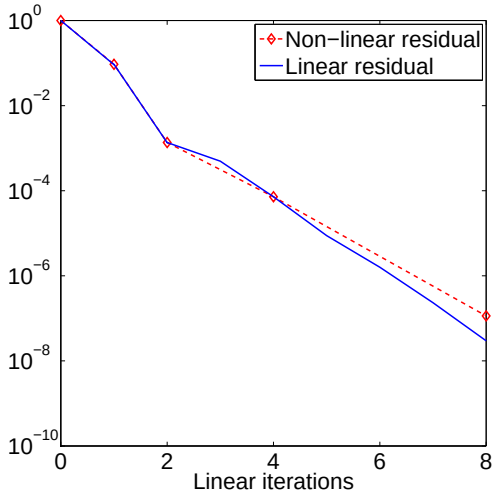
Table 5.3: Comparison between different preconditioners in terms of total number of Newton's iterations (# New. its.) and total number of linear iterations (# lin. its.) performed during the numerical simulation when $\epsilon_N = 10^{-7}$ ($\Delta t = 1$ [ms]).

Prec.	# New. its.	# lin. its.	# New. its./# time its.	# lin. its. /# New. its.
$\mathbf{P}_J$	3836	12347	5.11	3.22
$\mathbf{P}_{GS}$	2784	3440	3.71	1.24
$\mathbf{P}_1$	3357	6607	4.48	1.97
$\mathbf{P}_2$	2707	3237	3.61	1.20
$\mathbf{P}_3$	3302	6173	4.40	1.87

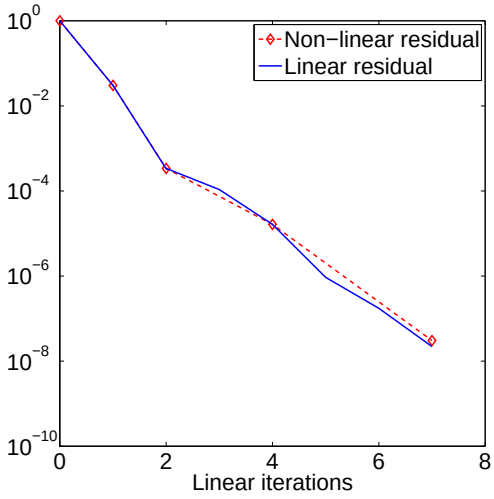
Table 5.4: Comparison between different preconditioners in terms of total number of Newton's iterations (# New. its.) and total number of linear iterations (# lin. its.) performed during the numerical simulation when $\epsilon_N = 10^{-8}$ ($\Delta t = 1$ [ms]).



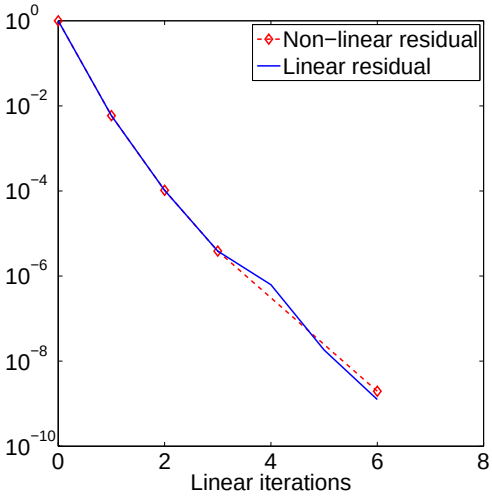
(a)



(b) $\Delta t = 1$ [ms].



(c) $\Delta t = 0.5$ [ms].



(d) $\Delta t = 0.1$ [ms].

Figure 5.8: (a): Spectral radius of $\mathbf{B}_J$ as function of time for different choices of Δt. (b): Decrease of non-linear and linear residuals for BJ with $\Delta t = 1$ [ms]. (c): Decrease of non-linear and linear residuals for BJ with $\Delta t = 0.5$ [ms]. (d): Decrease of non-linear and linear residuals for BJ with $\Delta t = 0.1$ [ms].

Time step	# New. its.	# lin. its.	# New. its./# time its.	# lin. its. /# New. its.
$\Delta t = 1$ [ms]	3256	8201	4.34	2.52
$\Delta t = 0.5$ [ms]	5990	11126	3.99	1.86
$\Delta t = 0.1$ [ms]	27585	47349	3.68	1.72

Table 5.5: Total number of Newton's iterations (# New. its.) and linear iterations (# lin. its.) performed by using BJ for different time steps. The stopping threshold for the non-linear residual is $\epsilon_N = 10^{-6}$.

5.4 Simplified Description of Termination of an Arrhytmias

This last section is devoted to showing and discussing a numerical experiment involving the mechanism of the stretch-activated currents. In particular we aim at showing how the stretch-activated currents can induce termination of re-entrant waves.

We model a simplified version of a re-entrant arrythmia as follows. We take model (4.3.1) and we impose periodic boundary conditions in the AP model, so that $v(0,t) = v(1,t)\ \forall t \in (0,T)$. Thanks to this modification, it is now possible to observe the front wave of the electrical signal propagating in both directions of the physical domain. When the two fronts meet, the potential goes back to its resting value, and it stays there until a new external stimulus is applied (Figure 5.9).

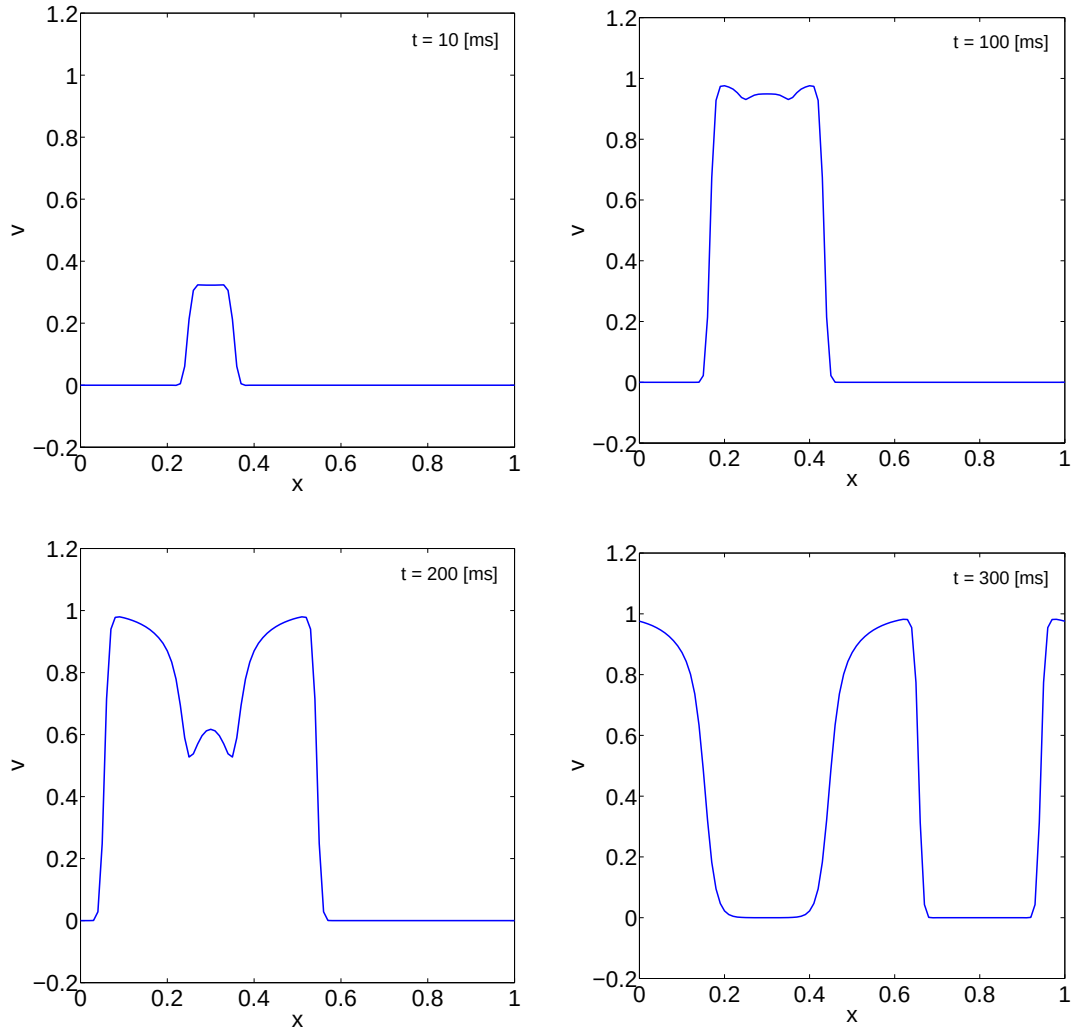


Figure 5.9: Time evolution of the electrical signal along the physical domain with periodic boundary conditions. ($\mu = 10^{-4}$).

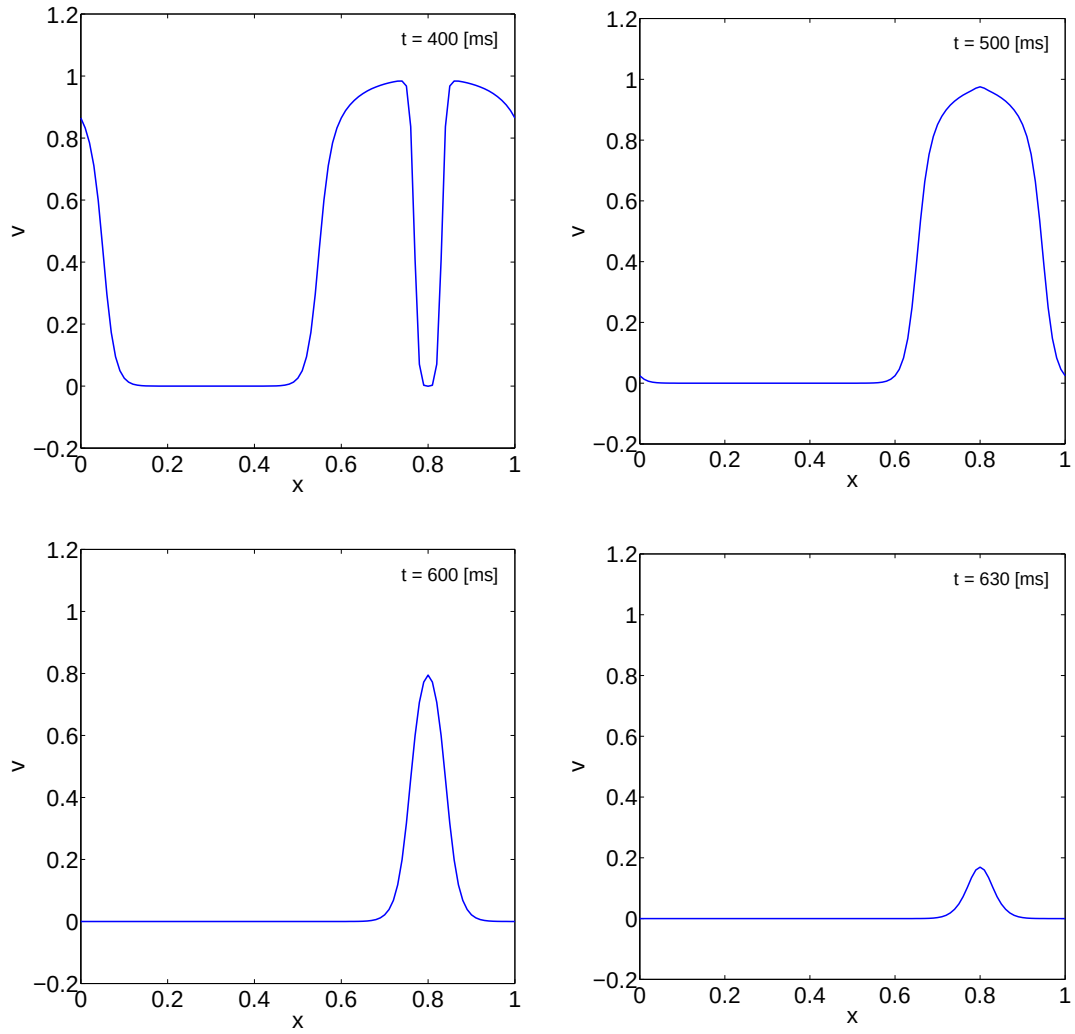


Figure 5.9: Time evolution of the electrical signal along the physical domain with periodic boundary conditions. ($\mu = 10^{-4}$).

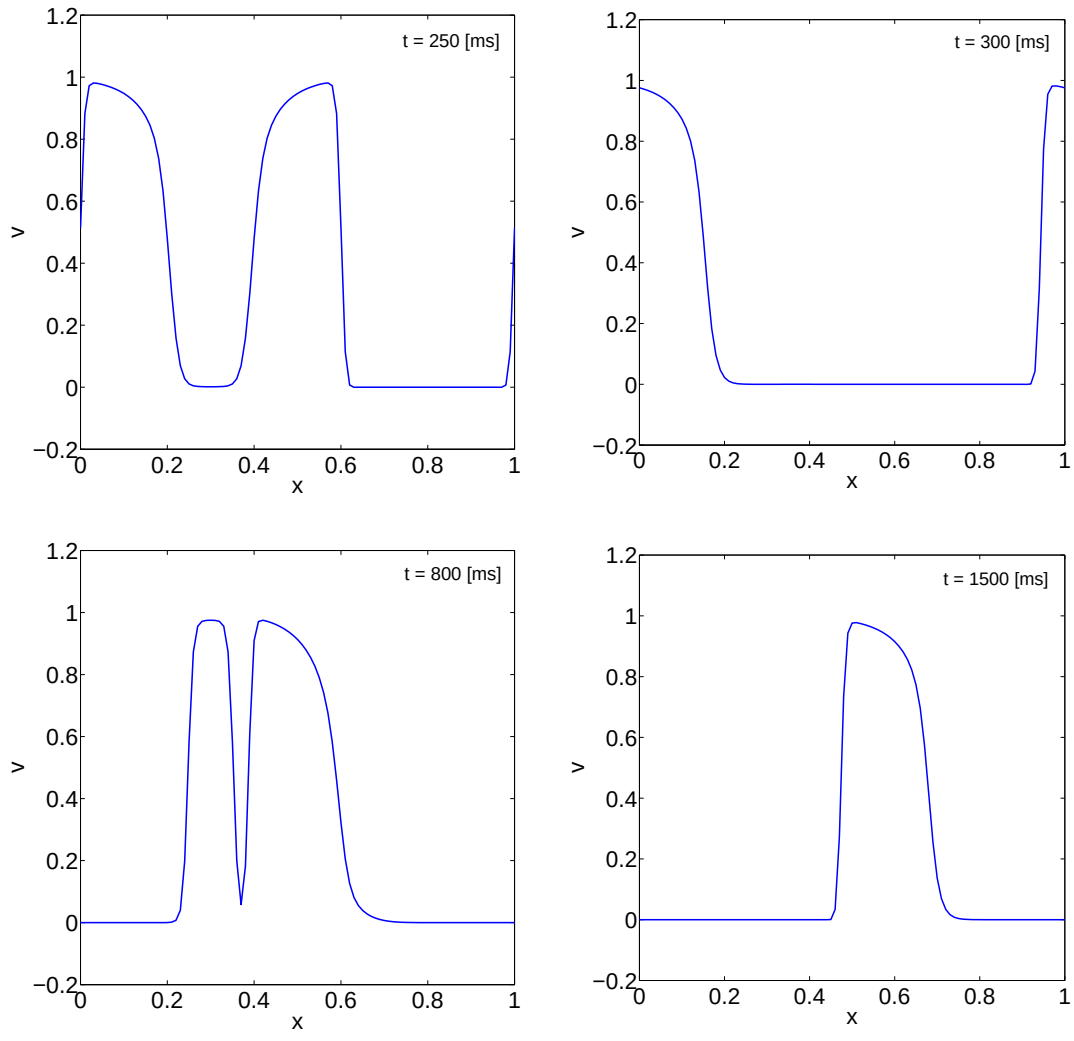
This normal behaviour can be corrupted artificially by enforcing one of the two fronts to go back to its resting value, before it meets the opposite front. Now the electrical signal travels along only one direction, and the normal behaviour can not be recovered by changing, for example, the periodicity or the magnitude of the external input. However we find that by applying an external tension to the free node of the mechanical model, and therefore causing the activation of the current I_s along all the physical domain, we manage to bring the action potential back to its resting value, and then the normal excitation behaviour is recovered.

In our simulation, once the two front waves are completely developed, we enforce the one moving from left to right to be equal to zero. By applying periodically the external stimulus, we do not manage to recover the bidirectional excitation behaviour. At $t = 1800$ [ms] we apply an external tension in the free node $x = 1$, modelled as

$$T(t) = \begin{cases} T_i \left(\sin\left(\dfrac{2\pi}{\omega} t + \varphi \right) + 1 \right), & t \in (t_1, t_2), \\ 0 & \text{otherwise.}, \end{cases}$$

and we set $T_i = 0.01$ [g cm^{-1}ms^{-2}], $\omega = 600$, $\varphi = 1.5\pi$, $t_1 = 1800$ [ms] and $t_2 = 2400$ [ms]. In Figure 5.10 we show the time evolution of the action potential along the physical domain. The system is solved fully implicitly and $\Delta = 1$ [ms] and $\theta = 0.5$ are used for time integration. The external tension is active at $t > 1800$ [ms] and slightly increases up to 0.02 [g cm^{-1}ms^{-2}] at time $t = 2100$ [ms]. It can be observed how the dilatation of the domain generates a new electrical potential which acts to bring the electrical signal to rest. After some time, when a new external stimulus arrives, the action potential is newly capable of propagation along both directions of the domain, and the normal excitation-contraction mechanism is recovered.

We perform the same simulation by treating explicitly the coupling term responsible of the stretch-activated currents, and by using the same values of Δt and θ. No numerical instabilities occur and no significant differences in the numerical solutions are observed. Therefore we can conclude that for these value the coupling term between v and u is not so strong to cause numerical instabilities when treated explicitly.



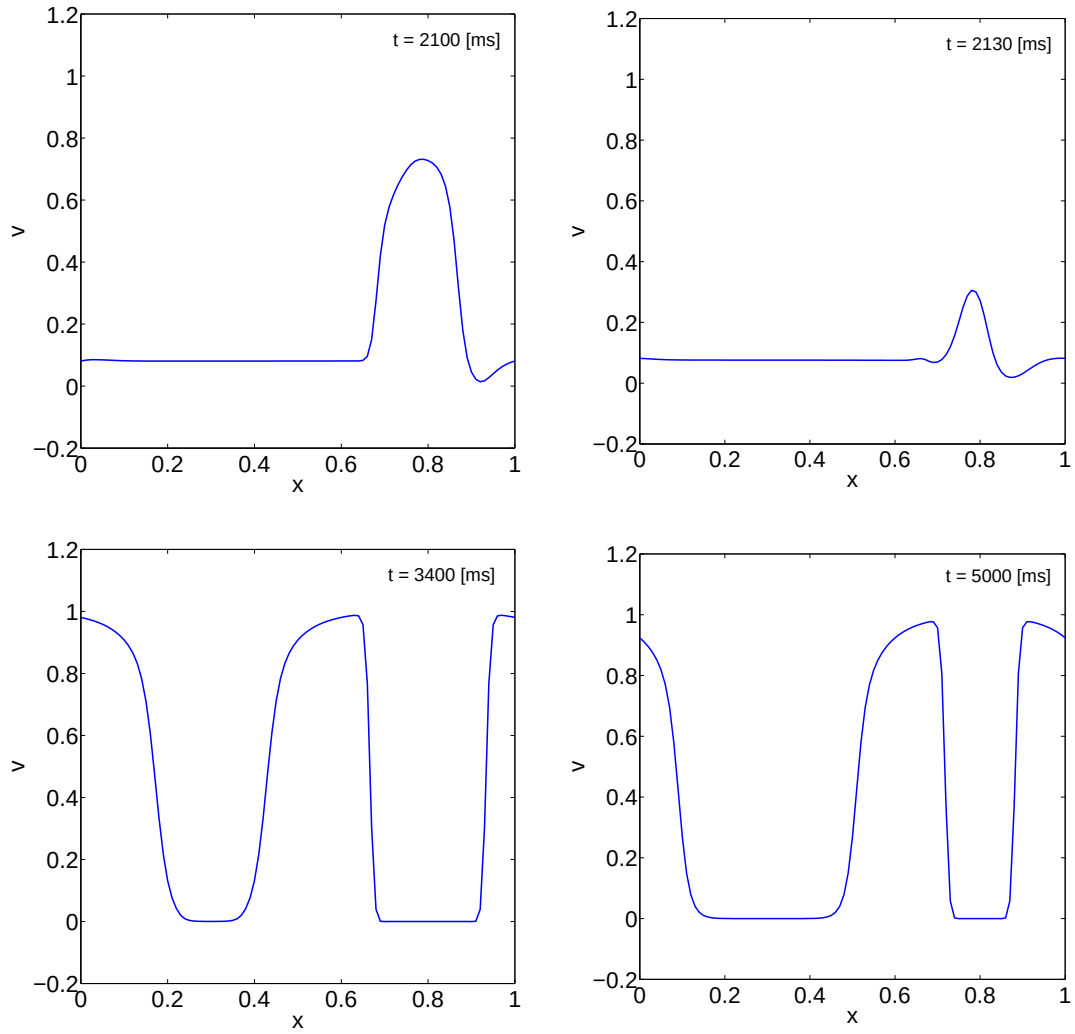


Figure 5.10: Time evolution of the electrical signal along the physical domain. The external additional stretch is activated at $t = 1800$ [ms] and reaches its maximum value at $t = 2100$ [ms]. The stretch-activated current makes the potential to go back to its resting state and the whole system to recover the normal excitation behaviour. ($\mu = 10^{-4}$).

Conclusions

From the numerical experiments performed some significant insights on the numerical approximation of the Aliev-Panfilov equations (AP model) and of the fully coupled model (4.3.1) can be drawn.

As widely known in literature, we have shown that different approaches can be considered to approximate the elementwise integrals when solving the AP model by means of the finite element method. However, depending on the approach used, we can experience significant differences in the approximated solution to the model. The SVI approach leads to an overestimation of the front velocity. The ICI and L-ICI approaches, which can reduce significantly the computational cost spent during the assembling part, produce an underestimation of the front velocity. The AP equations, as many reaction-diffusion problems, suffer from severe time step restriction when time integration is performed by using numerical schemes such as explicit Euler's method, and usually implicit methods are preferred. However implicit methods require at each time step the solution of a non-linear system. Therefore we showed two valid alternatives (OS and RKC) to fully implicit methods, which allow to solve the problem explicitly without imposing severe restriction on the time step.

Regarding the fully coupled model (4.3.1), it is important to remark that the choice one makes for the parameters can lead to significant differences in the numerical approximations and in the numerical stability of the whole system. With our parameter settings we showed that using the Crank-Nicolson method for time integration can lead to numerical instabilities or inaccurate solutions if the time step Δt is not small enough. The coupling between active stress and the mechanical equations is more prone to giving rise to numerical instabilities. If the evolution law of the internal variable H is characterized by a small time constant and Δt is not small enough, numerical instabilities can occur. We found that, with our parameter settings, to reach sufficient accuracy $\Delta t = 0.1$ [ms] is a suitable choice.

By treating some of the couplings explicitly, it is possible to solve more efficiently the system. The coupling terms to be treated explicitly have to be chosen carefully, otherwise numerical instabilities can occur if the numerical scheme chosen for time integration is not dissipative or the time step Δt is not small enough, as showed in Section 5.2. Critical situations are observed when the coupling terms between microscopic stress, macroscopic active stress and mechanical equations are treated explicitly. On the other hand, using implicit Euler's method for the parts of the problem which are solved implicitly, assures convergence of the numerical scheme and absence of oscillations in the approximated solution.

Another way of solving the problem reducing the computational effort is represented by not solving the linearised system (4.3.4) exactly, but by using an iterative solver. Of course, using an iterative solver implies the choice of a stopping criterion, which is usually based on the current non-linear residual to avoid unnecessary computations. We verified that the stopping criterion proposed in [20, 32] can reduce significantly the computational effort, avoiding unnecessary computations. Finally we compared different possible choices for the preconditioner in terms of total number of Newton's iterations and linear iterations. Block Gauss-Seidel (BGS) shows very fast convergence. On the other hand, even if employing a larger number of Newton and linear iterations, also Block Jacobi (BJ) converges. Since the

latter is embarrassingly parallel, it can result even more efficient than BGS, especially for small Δt. We tried to reduce the total number of Newton and linear iterations performed by the iterative solver, by looking for other preconditioners, which posses more blocks than the one used in BGS, and which can be placed into a block lower/upper triangular matrix by simultaneous row/column permutations. In particular the other two choices we made ($\mathbf{P}_1$ and $\mathbf{P}_2$) are characterized by the fact that the mechanics is solved before the electrical model. Both choices reach convergence. However while $\mathbf{P}_2$ reduces (only slightly) the number of iterations observed in BGS, using $\mathbf{P}_1$ causes slower convergence of the non-linear residual. We hypothesized that this behaviour is due to the fact that $\mathbf{P}_1$ does not include the block containing the linearised active stress. We then performed other simulations by using as precondioner the matrix given by the block diagonal of the Jacobian plus the block containing the active stress. We observed that this choice can reduce significantly the number of Newton and linear iterations observed when using BJ.

Lastly, in Section 5.4 we showed a brief experiment to investigate how the effect of stretch-activated currents can induce termination of re-entrant waves. Moreover we found that the coupling between the action potential and the displacement (mechanoelectrical feedback) is not so strong to induce numerical instabilities when treated explicitly.

Bibliography

[1] R. R. Aliev, A. V. Panfilov. (1996) A simple two-variable model of cardiac excitation. *Chaos, Solitons & Fractals*. 7: 3: 293-301.

[2] J. Bestel, F. Clément, M. Sorine. (2001) A Biomechanical Model of Muscle Contraction. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 1159-1161. Springer, Berlin, Heidelberg.

[3] C. J. Brokaw. (1976) Computer simulation of movement-generating cross-bridges. *Biophysical Journal*. 16: 1013-1027.

[4] M. Courtemanche, L. Glass, J. P. Keener. (1993) Instabilities of a propagating pulse in a ring of excitable media. *Physical Review Letters*. 70: 2182-2185.

[5] S. C. Eisenstat, H. F. Walker. (1996) Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*. 22(12): 1155-1162.

[6] R. FitzHugh. (1960) Thresholds and plateaus in the Hodgkin-Huxley nerve equations. *Journal of General Physiology*. 43: 867-896.

[7] R. FitzHugh. (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*. 1: 445-465.

[8] R. FitzHugh. (1969) *Mathematical models of excitation and propagation in nerve*. In: Biological Engineering, Ed: H. P. Schwan. McGraw-Hill, New York.

[9] A. C. Guyton, J. E. Hall. (2001) *Textbook of Medical Physiology*. 12th Edition. Elsevier.

[10] E. Hairer, G. Wanner. (1996) *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Second Edition, Springer Series in Computational Mathematics, Vol. 14. Springer, Berlin.

[11] M. Hanslien, R. Artebrant, A. Tveito, G. T. Lines, X. Cai. (2011) Stability of two time-integrators for the Aliev-Panfilov system. *International Journal of Numerical Analysis and Modeling*. 8: 3: 427-442.

[12] A. V. Hill. (1938) The heat of shortening and the dynamic constants of muscle. *Proceedings of the Royal Society of London B*. 126: 136-195.

[13] A. L. Hodgkin, A. F. Huxley. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*. 117: 500-544.

[14] P. J. van der Houwen, B. P. Sommeijer. (1980) *On the internal stability of explicit, m-stage Runge-Kutta methods for large m values. JAMM - Journal of Applied Mathematics and Mechanics*. 60: 479-485.

[15] T. J. R. Hughes (2000) *The Finite Element Method: Linear Static and Dynamic Finite Elements Analysis*. Dover Publications, Mineola, New York.

[16] W. Hundsdorfer, J. G. Verwer. (2003) *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer Series in Computational Mathematics, Vol. 33. Springer, Amsterdam.

[17] A. F. Huxley. (1957) Muscle structure and theories of contraction. *Progress in Biophysics*. 7: 255-318.

[18] A. F. Huxley, R. M. Simmons. (1971) Proposed mechanism of force generation in striated muscle. *Nature*. 233: 533-538.

[19] J. Keener, J. Sneyd. (2008) *Mathematical Physiology: Systems Physiology*. Springer, New York.

[20] C. T. Kelley. (2003) *Solve nonlinear equations with Newton's method*. Philadelphia, PA: SIAM.

[21] B. M. Koeppen, B. A. Stanton. (2010) *Berne & Levy Physiology*. 6th Edition. Mosby, Elsevier.

[22] S. Krishnamoorthi, M. Sarkar, W. S. Klug. (2013) Numerical quadrature and operator splitting in finite element methods for cardiac electrophysiology. *International Journal for Numerical Methods in Biomedical Engineering*. 29(11): 1243-1266.

[23] S. Marchessau, H. Delingette, M. Sermesant, M. Sorine, K. Rhode, S. G. Duckett, C. A. Rinaldi, R. Razavi, N. Ayache. (2012) Preliminary specificity study of the Bestel-Clément-Sorine electromechanical model of the heart using parameter calibration from medical images. *Journal of the Mechanical Behaviour of Biomedical Materials*. 20: 259-271.

[24] M. P. Nash, A. V. Panfilov. (2004) Electromechanical model of excitable tissue to study reentrant cardiac arrhytmias. *Progress in Biophysics & Molecular Biology*. 85: 501-522.

[25] A. V. Panfilov, R. H. Keldermann, M. P. Nash. (2005) Self-organized pacemakers in a coupled reaction-diffusion-mechanics system. *Physical Review Letters*. 95: 258104.

[26] A. Quarteroni, A. Valli. (1994) *Numerical Approximation of Partial Differential Equations*. Springer Series in Computational Mathematics, Vol. 23. Springer, Berlin.

[27] A. Quarteroni, R. Sacco, F. Saleri. (2007) *Numerical Mathematics*. 2nd edition. Texts in Applied mathematics, Vol. 37. Springer, Berlin.

[28] J. Rinzel, J. B. Keller. (1973) Traveling wave solutions of a nerve conduction equation. *Biophysical Journal*. 13: 12: 1313-1337.

[29] A. M. Robertson. (2009) Supplementary Notes for: Nonlinear elasticity with applications to the arterial wall. University of Pittsburgh.

[30] S. Rossi, T. Lassila, R. Ruiz-Baier, A. Sequeira, A. Quarteroni. (2013) Termodynamically consistent orthotropic activation model capturing ventricular systolic wall thickening in cardiac electromechanics. *European Journal of Mechanics/A Solids*. In press, DOI: 10.1016/j.euromechsol.2013.10.009.

[31] A. M. Stuart, A. R. Humphries (1996) *Dynamical systems and Numerical Analysis*. Cambridge University Press.

[32] N. Tardieu, E. Cheignon. (2012) A Newton-Krylov method for solid mechanics. *European Journal of Computational Mechanics*. 21: 3-6, 374-384.

[33] J. G. Verwer. (1996) Explicit Runge-Kutta methods for parabolic partial differential equations. *Applied Numerical Mathematics*. 22: 359-379.

[34] G. I. Zahalak. (1981) A distribution-moment approximation for kinetic theories of muscular contraction. *Mathematical Biosciences*. 114: 55: 89-114.