Chapter 3

Discretisation and Solution Methods

We will begin this chapter by briefly introducing finite element methods before particularly focusing on least squares finite element approaches. Afterwards we will present a general overview of space–time solution methods, as well as discussing the approach applied in this thesis. Last but not least we will discuss different strategies to tackle the nonlinearity of the problem. While these concepts or methods are introduced separately here, we will use chapter 4 to tie them together in a comprehensive solver.

3.1 Finite Element Methods

In order to find a numerical approximation of the solution of a partial differential equation we need a way to approximate the operators involved. And while there are many different ideas of how to do so the one we have chosen to employ is a finite element approach, as they have shown to be one of the most powerful and versatile methodologies for the problem at hand [8]. The subsequent section introduces the tools required for the usage of finite elements as well as a suitable notation that we will rely on for the rest of this thesis. For a more comprehensive overview of the topic we refer to [17], [18] or [19], whereas this section as well as the next one are losely based on [8].

3.1.1 Variational Formulation

The foundation of every finite element formulation is finding an appropriate weak formulation which includes the choice of suitable trial and solution spaces. This is especially applicable in the case of a least squares approach and will be discussed in further detail in section (3.2).
Given Banach spaces X and Y, and a bounded linear operator $\mathcal{A} : X \to Y$, $f \in Y$, we consider the problem:

$$\text{Find } u \in X \text{ such that} \quad \mathcal{A}u = f \text{ in } Y. \tag{3.1}$$

We are interested in the case where $\mathcal{A}$ represents a partial differential operator. As mentioned before the process of discretisation begins with turning (3.1) into a suitable variational equation, which we assume to be defined in terms of a Hilbert space U, a continuous bilinear form $a(\cdot,\cdot) : U \times U \to \mathbb{R}$, and a bounded linear functional $L_f(\cdot) : U \to \mathbb{R}$ and can be written as follows

$$\text{Find } u \in U \text{ such that: } a(u,v) = L_f(v) \quad \forall v \in U \tag{3.2}$$

An operator equation such as (3.1) may be reformulated into several different variational equations. We can see that we were originally seeking for a solution u in the space X whereas in the weak formulation one attempts to find a solution in the space U, which generally does not lie in X, as U is less restrictive. However the relationship between the spaces of the strong and the weak form, and the operator $\mathcal{A}$ and the bilinear mapping $a(\cdot,\cdot)$ are of great importance, and while one generally wants the solution of the variational formulation (3.2) to be a "good" representation of the solution of the original problem (3.1), the definition of what that exactly means varies and usually depends on the nature of the problem, the overall objective, and often some

practicality issues. So let us assume for now that we have found a suitable weak formulation of the operator equation (3.1), and where for $a(\cdot,\cdot)$ it additionally holds that

$$a(u_1,u_2) = a(u_2,u_1), \quad \text{for all } u_1,u_2 \in U \quad \text{(symmetry)} \tag{3.3}$$
$$a(u_1,u_2) \leq \beta\|u_1\|_U \cdot \|u_2\|_U, \quad \text{for all } u_1,u_2 \in U \text{ and } \beta > 0 \quad \text{(boundedness)} \tag{3.4}$$
$$a(u_1,u_1) \geq \alpha\|u_1\|_U^2, \quad \text{for all } u_1 \in U \text{ and } \alpha > 0 \quad \text{(coercivity)} \tag{3.5}$$

as well as that $f \in U^*$, the dual space of U. Then by the *theorem of Lax-Milgram* we obtain the existence and uniqueness of a solution $u \in U$ of the variational formulation (3.2). We furthermore have that $a(\cdot,\cdot)$ induces a norm on U, as it fulfills all properties of being an innerproduct. Through *Riesz representation theorem*, which is used to prove the theorem of Lax-Milgram we additionally obtain the existence of an operator $\tilde{\mathcal{A}} : U \to U^*$ given by

$$a(u,v) = \langle \tilde{\mathcal{A}}u, v\rangle_{U^*,U} \quad \forall u,v \in U \tag{3.6}$$

where $\langle\cdot,\cdot\rangle_{U^*,U}$ denotes the duality pairing between U and its dual space $U*$ *should i explain more, or okay like that? in terms of inner product?*. Likewise we obtain for $L_f(\cdot)$ the existence of an unique element $\tilde{f}$ through the relation

$$L_f(v) = \langle \tilde{f}, v\rangle_{U^*,U} \quad \forall v \in U \tag{3.7}$$

The variational formulation is therefore equivalent to the problem

$$\text{Find } u \in U \text{ such that} \quad \tilde{\mathcal{A}}u = \tilde{f} \quad \text{in } U^* \tag{3.8}$$

In the special case that $X = U$ and $Y = U^*$ we have that $\mathcal{A} = \tilde{\mathcal{A}}$ and $f = \tilde{f}$ but this is generally not the case.
write more? dont really need this section here, but nice to refer back to it later in lsfem, or cut in both?

3.1.2 Galerkin Approach

In order to actually find a good approximation u^h of the unique solution u one chooses a suitable finite dimensional subspace U_h of U, where we search for a solution to the problem. Each subspace of a Hilbert space is again a Hilbert space itself and therefore the projected finite dimensional problem called Galerkin equation looks as follows

$$\text{Find } u_h \text{ in } U_h \text{ such that: } a(u_h,v_h) = L_f(v_h) \quad \forall v_h \in U_h. \tag{3.9}$$

By the theorem of Lax-Milgram it also has a unique solution. And since (3.2) holds for all $v \in U$, it also holds for all $v \in U_h$, and hence $a(u-u_h,v_h) = 0$, which we obtain by subtracting equation (3.9) from (3.2), a key property known as Galerkin orthogonality. With respect to the energy norm induced by $a(\cdot,\cdot)$, the finite dimensional solution u_h is a best approximation to u, in the sense that

$$\begin{aligned}\|u-u_h\|_a^2 &= a(u-u_h,u-u_h) = a(u-u_h,u) - a(u-u_h,u_h) - a(u-u_h,v_h)\\ &\leq \|u-u_h\|_a \cdot \|u-v_h\|_a \quad \forall v_h \in U_h.\end{aligned} \tag{3.10}$$

We derive the third term from the second by using the Galerkin orthogonality. If we now divide both sides by $\|u-u_h\|_a$, we obtain that $\|u-u_h\|_a \leq \|u-v_h\|_a$ for all $v_h \in U_h$. We also can

also have an estimate on $u - u_h$ in terms of the norm $\|\cdot\|_U$, by using the coercivity constant α and the bound from above β, we see that

$$\begin{aligned}\alpha\|u-u_h\|_U^2 &\leq a(u-u_h,u-u_h) = a(u-u_h,u-u_h) = a(u-u_h,u+v_h-v_h-u_h)\\ &= a(u-u_h,u-v_h) + a(u-u_h,v_h-u_h) = a(u-u_h,u-v_h)\\ &\leq \beta\|u-u_h\|_U\cdot\|u-v_h\|_U \quad \forall v_h \in U_h.\end{aligned} \tag{3.11}$$

Dividing by $\alpha\|u-u_h\|$ we have shown *Céa's lemma*, which states that

$$\|u-u_h\|_U \leq \inf_{v_h\in U_h}\frac{\beta}{\alpha}\|u-v_h\|_U, \quad u\in U, u_h\in U_h \tag{3.12}$$

Hence accuracy of our approximation depends in this case on the constants α and β.

If we assume that we have a discretisation Ω_h of our domain Ω, where $h > 0$ is a parameter depending on the mesh size and for which we have that as h tends to zero this implies that $dim(U_h) \Rightarrow \infty$. Additionally let $\{U_h : h > 0\}$ denote a family of finite dimensional subspaces of U, for which we assume that

$$\forall v \in U: \quad \inf_{v_h\in U_h}\|v-v_h\|_U \to 0 \text{ as } h\to 0. \tag{3.13}$$

That is with a mesh size tending to zero there exist increasingly precise approximations for every $v \in U$, whose infimum tends to zero as the mesh size does. Then we can also conclude by the beforementioned properties (3.11) and (3.12) that $\|u-u_h\|_U \to 0$ as $h \to 0$. Hence in this setting the approximate finite–dimensional solution u_h will converge to the weak solution u.

3.1.3 Matrix Formulation

After establishing the theoretical properties of the finite–dimensional problem our aim is now to recast it in a linear system of equations that can be solved efficiently. Since U_h is a finite dimensional Hilbert space, it has a finite basis $\{\phi_1, \phi_2, ..., \phi_n\}$ and we can write every element in U_h as a linear combination this basis, that is we have $u_h = \sum_{j=1}^n u_j\phi_j$, where $u_1, ..., u_n$ are constant coefficients. Writing (3.9) in terms of this basis we obtain by linearity

$$a(\sum_{j=1}^n u_j\phi_j, \phi_i) = \sum_{j=1}^n u_j a(\phi_j,\phi_i) = L_f(\phi_i) \quad \forall\ \phi_i,\ i = 1,2,...,n \tag{3.14}$$

If we now write this as a system of the form $Au_h = L_h$ with entries entries $A_{ij} = a(\phi_j,\phi_i)$, $(L_h)_i = L_f(\phi_i)$, then this becomes a linear system of equations which we can solve for the unknown vector u_h, where each matrix entry represents the evaluation of an integral expression. The question of how to choose favorable subspaces U_h, and a suitable basis for it has no trivial answer and depends on many factors and goes hand in hand with the question of how to best discretise the domain. Generally it seems like a sensible aim to opt for easily computable terms giving rise to a linear system that is in turn as easy as possible to solve. Hence one objective might be to choose the basis $\{\phi_1, ..., \phi_n\}$ such that $\text{supp}(\phi_i) \cap \text{supp}(\phi_j) = \emptyset$ for as many pairs (i, j) as possible, as this would give rise to a sparse system of equations. It is also worth noting that due to the symmetry of $a(\cdot,\cdot)$, we have that $a_{ij} = a_{ji}$.

Depending on the operator $\mathcal{A}$ from equation (3.1), there is not necessarily a straight forward way to translate the strong formulation into a symmetric variational formulation, that is a symmetric bilinear mapping $a(\cdot,\cdot)$, which will hence give rise to a symmetric linear system of equations. One strategy to enforce symmetry is through the formulation and subsequently the

differentiation of particular energy functionals, because we know by the theorem of Schwarz that order of differentiation with respect to partial derivatives is interchangeable and therefore leads to a symmetric Hessian. Hence if the operator $a(\cdot,\cdot)$ was somehow induced by a second order derivative it would have to be symmetric. How to construct these functionals such that they are related to the differential equation problem at hand will be discussed in the following section.

3.2 Least Squares Finite Element Methods

In this section which is based on [8], we introduce least squares finite element methods (LS-FEMs), a class of methods for finding the numerical solution of partial differential equations that is based on the minimisation of functionals which are constructed from residual equations. Historically finite element methods were first developed and analysed for problems like linear elasticity whose solutions describe minimisers for convex, quadratic functionals over infinite dimensional Hilbert spaces and therefore emerged in an optimisation setting. A Rayleigh-Ritz approximation of solutions of such problems is then found by minimising the functional over finite dimensional subspaces. For these classical problems the Rayleigh-Ritz setting gives rise to formulations that have a variety of favourable features and therefore have been and continue being highly successful. Among those are that:

1. domains and boundary conditions can be treated relatively easily in a systematic way
2. conforming finite element spaces are sufficient to guarantee stability and optimal accuracy of the approximate solutions
3. all variables can be approximated using the same finite element space, e.g. the space of degree n piecewise polynomials on a particular grid
4. the arising linear systems are
   - *(i)* sparse
   - *(ii)* symmetric
   - *(iii)* positive definite

Hence finite element methods originally emerged in the environment of an optimisation setting but have since then been extended to much broader classes of problems that are not necessarily associated to a minimisation problem anymore and generally lose the desirable features of the Rayleigh-Ritz setting except for 1 and 4 (i). Least squares finite element methods can be seen as a new attempt to re-establishing as many advantageous aspects of the Rayleigh-Ritz setting as possible, if not all, for more general classes of problems. In the following section we will have a look at a classical straightforward Rayleigh-Ritz setting to familiarise ourselves with the set up before extending it to the more complicated class of reaction diffusion equations introduced in the prologue.

We will consider a similar set up as in the finite element section (3.1) but with X and Y being Hilbert spaces, $f \in Y$ and a bounded, coercive linear operator $\mathcal{A} : X \to Y$, that is for some $\alpha, \beta > 0$:

$$\alpha\|u\|_X^2 \leq \|\mathcal{A}u\|_Y^2 \leq \beta\|u\|_X^2 \quad \forall u \in Y. \tag{3.15}$$

We consider the problem and the least squares functional:

$$\text{Find } u \in X \text{ such that } \mathcal{A}u = f \text{ in } Y \tag{3.16}$$

$$J(u; f) = \|\mathcal{A}u - f\|_Y^2 \tag{3.17}$$

which poses the minimisation problem:

$$\text{argmin}_{u \in X} J(u; f) \tag{3.18}$$

where we can see that the least squares functional (3.21) measures the residual of (3.20) in the norm of Y while seeking in for a solution in the space X. It follows that if a solution of the the problem (3.20) exists it will also be a solution of the minimisation problem. And a solution of the minimisation problem due to the definition of a norm will be a solution to (3.20) if the minimum is zero. If we consider $f = 0$, and using (3.19) we obtain that

$$\alpha^2 \|u\|_X^2 \leq J(u; 0)\|_Y^2 \leq \beta^2 \|v\|_X^2 \quad \forall u \in X \tag{3.19}$$

a property of $J(\cdot, \cdot)$ which we will call norm equivalence, which is an important property when defining least squares functionals. We can derive a candidate for a variational formulation of the following form

$$a(u, v) = \langle \mathcal{A}u, \mathcal{A}v \rangle_Y \text{ and } L_f(v) = \langle \mathcal{A}v, f \rangle_Y \quad \forall u, v \in X \tag{3.20}$$

where $\langle \cdot, \cdot \rangle_Y$ again denotes the innerproduct on Y, which will turn out to have all the desired properties. The operator form of (3.21) in the least squares setting is equivalent to the normal equations

$$\mathcal{A}^* \mathcal{A}u = \mathcal{A}^* f \quad \text{in } X \tag{3.21}$$

and corresponds to equation (3.9), with $\tilde{\mathcal{A}} = \mathcal{A}^* \mathcal{A}$, $\tilde{f} = \mathcal{A}^* f$ and $\mathcal{A}^*$ being the adjoint operator of $\mathcal{A}$. We can then move on to limiting our problem to a finite dimensional setting, where we choose a family of finite element subspaces $X^h \subset X$, parametrised by h tending to zero and restricting the minimisation problem to the subspaces. The LSFEM approximation $u^h \in X^h$ to the solution $x \in X$ of the infinite dimensional problem is the solution of the discrete minimisation problem

$$\min_{u^h \in X^h} J(u^h; f) \tag{3.22}$$

which is due to the fact that X^h is again a Hilbert space and therefore the same properties hold. Similarly to section (3.3.3) we can choose a basis $\{\phi_1, ..., \phi_n\}$ of X^h and will then obtain for the elements of $A^h \mathbb{R}^{n \times n}$, and $L_f^h \in \mathbb{R}^n$ that

$$A_{ij}^h = \langle \mathcal{A}\phi_j, \mathcal{A}\phi_i \rangle_Y \quad \text{and} \quad (L_f^h)_i = \langle \mathcal{A}\phi_i, f \rangle_Y \tag{3.23}$$

The following theorem establishes that this problem formulation actually gives rise to finite element set up.

Theorem 1. *Let $\alpha\|u\|_X^2 \leq \|\mathcal{A}u\|_Y^2 \leq \beta\|u\|_X^2$ for all $u \in X$ hold, under the same assumptions as established in this section and let $X^h \subset X$. Then,*

(i) the bilinear form $a(\cdot, \cdot)$ defined in (3.20) is continuous, symmetric and coercive

(ii) the linear functional $L_f(\cdot)$ defined in (3.20) is continuous

(iii) the variational formulation (3.20) is of the form (3.2) and has a unique solution $u \in X$ which is also the unique solution of the minimisation problem (3.18)

(iv) there exists a constant $c > 0$, such that u and u_h satisfy

$$\|u - u^h\|_X \leq c \inf_{v^h \in X^h} \|u - v^h\|_X \tag{3.24}$$

(v) the matrix A^h is symmetric positive definite

Idea of Proof: The properties *(i)* and *(ii)* directly follow from the boundedness and coercivity of $\mathcal{A}$ as well as the linearity of the inner product. Property *(iii)* follows from the theorem of Lax-Milgram while property *(iv)* is a consequence of Céa's lemma. The last property directly follows from the definition of A^h.

We therefore obtain that this least squares problem formulation has all the advantageous featurs of the Raleigh-Ritz setting without requiring $\mathcal{A}$ to be self-adjoint or symmetric which was our initial goal. However it is worth noting that the differential operator $\tilde{\mathcal{A}} = \mathcal{A}^*\mathcal{A}$ is of higher order than the one in the original formulation, which therefore requires higher regularity assumptions which might be unpreferrable as well as impractical. Potential ways to overcome this problem will be discussed in the following section as it is also an issue that arises in the problem formulation of the subsequent chapter.

3.3 Space-Time Approaches

Most solution methods for partial differential equations do not use the time direction for parallelisation. But with increasingly complex models, especially when many small steps in time are required and the rise of massively parallel computers, the idea of a parallelisation of the time axis has experienced a growing interest. Once parallelisation in space saturates it only seems natural to consider this remaining axis for parallelisation, after all, time is just another dimension [6]. However evolution over time behaves differently from the spatial dimensions, in the sense that it follows the causality principle. It means that the solution at later times is determined through earlier times whereas the opposite does not hold. This is not the case in the spatial domain.

The earliest papers on time parallelisation go back more than 50 years now to the 1960's, where it was mostly a theoretical consideration, before receiving an increasingly growing interest in the past two decades due to its computational need and feasibility. As mentioned in [6], on which this section is mainly based on and can be referred to for further details, time parallel methods can be classified into 4 different approaches, methods based on multiple shooting, domain decomposition and waveform relaxation, space-time multigrid and direct time parallel methods. Below a very brief overview of the main ideas behind these methods through some examples before taking a closer look at the strategy employed in this thesis.

Shooting type time parallel methods use a decomposition of the space-time domain Ω into time slabs Ω_j, i.e. $\Omega = \mathcal{S} \times [0, T]$ where $\mathcal{S}$ describes the spatial domain then $\Omega_j = \mathcal{S} \times [t_{j-1}, t_j]$ with $0 = t_0 < t_1 < .... < t_m = T$. Then there is usually an outer procedure that gives a coarse approximated solution y_j for all $x \in \mathcal{S}$ at t_j for all j, which are then used to compute solutions in the time subdomains Ω_j independently and in parallel and give rise to an overall solution. One important example of how this can be done was given by Lions, Maday and Turinici in 2001 [**?**], with an algorithm called parareal. A generalized version of it for a nonlinear problem of the form

$$y' = f(y), \quad y(t_0) = y_0 \tag{3.25}$$

can be formulated as follows using two propagation operators:

1. $G(t_j, t_{j-1}, y_{j-1})$ is a coarse approximation of $y(t_j)$ with initial condition $y(t_{j-1}) = y_{j-1}$