# Chapter 1

# Derivation of a Least Squares Finite Element Space-Time Discretisation

## 1.1 Construction of an Equivalent Minimisation Problem

In this chapter we will tie the beforementioned concepts together to derive a discretised problem formulation that can subsequently be turned into an algorithm capable of approximating reaction diffusion equations. In order to do so let us first set the ground for the overall framework we are looking at. We consider a space-time domain

$$\Omega = \mathcal{S} \times \mathcal{T}, \quad \mathcal{T} = (0, T),\ T > 0 \text{ and } \mathcal{S} \subset \mathbb{R}^N,\ N = 1, 2, 3 \tag{1.1}$$

where $\mathcal{T}$ represents the time domain and $\mathcal{S}$ the domain in space, which we require to be Lipschitz regular. We allow a mixture of Dirichlet and Neumann boundary conditions on $\partial\Omega$ which we denote as $\Gamma_D, \Gamma_N \subset \Gamma$ respectively. We further assume them to be such that the problem is well posed. *Is this enough?.* The class of partial differential equations introduced in the prologue that we would like to solve for then reads as the following:

$$\begin{aligned} \partial_t u - div(D(x)\nabla u) &= f(u) && \text{in } \Omega \\ u &= g_D && \text{on } \Gamma_D \\ \nabla u \cdot n &= g_N && \text{on } \Gamma_N \end{aligned} \tag{1.2}$$

It describes a parabolic partial differential equation with a potentially nonlinear right-hand side where the divergence is defined as

$$div(\sigma) = \frac{\partial \sigma}{\partial x_1} + ... + \frac{\partial \sigma}{\partial x_N} \text{ for } \sigma = \sigma(x, t) \text{ and } \nabla u = [\frac{\partial u}{\partial x_1}, ..., \frac{\partial u}{\partial x_N}]^T \text{ for } u = u(x, t).$$

We further assume that $D(x)$ is a bounded, symmetric positive definite matrix of size $N \times N$ with functions in $L^2(\mathcal{S})$ for almost all $x \in \bar{\mathcal{S}}$. *Are these sufficient assumptions on D(x)?* Typically we will have that $u(x, 0) = u_0 = g_D$ for all $x \in \mathcal{S}$ and Neumann boundary conditions on the boundary of $\mathcal{S}$ for $t \in (0, T)$.

The next step will be to derive an equivalent optimisation problem whose solution therefore then coincides with the solution of (1.2) at least in a weak sense. Since we are entirely working with finding solutions in Sobolev spaces in the least squares setting we can generally only require equivalence to a primal weak formulation of (1.2) or equivalence with respect to the solution space of the variational formulation, for a further discussion we refer to [**?**]. Generally when working with least squares finite element formulations choosing a suitable solution space $U$ and data space $Y$ is often non trivial as there are a number of difficulties that can arise. One usually faces a trade off between constructing a mathematically well-defined problem and allowing for a relatively simple, efficient, robust while still accurate implementation. Therefore further considerations need to be taken into account to make the methodology of LSFEMs competitive compared to other approaches like Galerkin approximations. We saw in the previous chapter that it is possible to derive least squares formulations that recover the properties of the Rayleigh–Ritz setting, however one hindrance one encounters in this setting as well as in many others

is the higher order operator arising in (3.25), that would require a solution space of higher regularity. When considering a simple Poisson equation with Dirichlet boundary conditions this would for example imply that we would require the solution $u$ to be from $H_0^2$, instead of $H_0^1$ [**?**] *more here or in previous LSFEM section ...?*. This does not only heavily limit the set of admissible solutions but it is additionally much harder to construct appropriate finite dimensional subspaces for and is therefore impractical to use. In order to succomb this obstacle we will recast (4.2) as a system of mixed equations only containing first order derivatives to apply the methodologies introduced in section (3.4) at the price of introducing an additional variable. Hence let

$$\begin{aligned} \partial_t u - div(\sigma) &= f(u) && \text{in } \Omega \\ \sigma &= D(x)\nabla u && \text{in } \Omega \\ u &= g_D && \text{on } \Gamma_D \\ \nabla u \cdot n &= g_N && \text{on } \Gamma_N. \end{aligned} \tag{1.3}$$

Rearranging this equation into a vector form we obtain in $\Omega$

$$\begin{pmatrix} I & -D(x)\nabla \\ -div & \frac{\partial}{\partial t} \end{pmatrix} \begin{pmatrix} \sigma \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ f(u) \end{pmatrix} \tag{1.4}$$

which we will refer to as the mixed strong form of our problem and can be shortened to $\mathcal{A}([\sigma, u]) = \tilde{f}(u)$, where $\mathcal{A}$ denotes the differential operator and $\tilde{f}(u)$ the right hand side.

**Remark.** It is not clear yet in which space we will be looking for a solution. In general when working with strong and weak formulations there is a a number of function spaces involved, problem formulations that are similar but not the same and it is difficult to represent them in clear but short notation. This can easily lead to some confusion. In the current case the situation is even more complicated since we are trying to derive a weak formulation in a least squares setting, while additionally there obviously exist the regular or so-called primal weak formulations of (1.2) or (1.3). Therefore we will first derive a least squares functional $J$ with a corresponding variational formulation, postponing the definition of appropriate spaces for now, but simply assuming them to already be well-defined, and later on show that they indeed exist.

So instead of looking for a solution of the strong formulation which is far more restrictive, let us now turn to the derivation of an optimisation problem that we would like to satisfy in a weak sense. The properties that we would like to be fulfilled are the following; we want a solution of (4.3) to be a global minimum of the optimisation problem, *independently* of the choice of the spaces that we are using and its associated norm, hence we additionally want the least squares functional $J$ to be zero for a solution of (4.3). On the other hand if $J([\sigma, u], f) = 0$ then the orginial problem (4.3) also has to be satisfied if only in a weak sense with respect to the spaces $U$ and $Y$. We can check that all three properties hold for the functional

$$\tilde{J}(\sigma, u) = ||u_t - div(\sigma) - f(u)||_Y^2 + ||\sigma - D(x)\nabla u||_Y^2 \tag{1.5}$$

Furthermore it has shown to be practical to be able to weight the two terms with coefficients [source] which does not affect the solution of the continuous problem but grants us the possibility to numerically give more importance to one term than the other, a further exploration of this will be in section 6.3.2. We also introduce additional scalars of $\frac{1}{2}$ to simplify further computations of the derivatives as we will see in the following section. Therefore the problem then reads

$$\min_{(\sigma,u)\in U} J([\sigma, u], f) = \frac{1}{2}c_1||u_t - div(\sigma) - f(u)||_Y^2 + \frac{1}{2}c_2||\sigma - D(x)\nabla u||_Y^2. \tag{1.6}$$

$J$ now defines an energy we can minimise. If we consider the functional for $f = 0$ we obtain

$$J([\sigma, u], 0) = \frac{1}{2}c_1\langle u_t - div(\sigma), u_t - div(\sigma)\rangle_Y + \frac{1}{2}c_2\langle \sigma - D(x)\nabla u, \sigma - D(x)\nabla u\rangle_Y. \tag{1.7}$$

which more specifically gives rise to the following bilinear form if we differentiate $J$ with respect to directional derivatives $\tau$ and $v$, and *subsequently sum over them*, see section 4.5, and Appendix A for further derivations.

$$\mathcal{B}([\sigma, u], [\tau, v]) = \langle \begin{pmatrix} I & -D\nabla \\ -div & \frac{\partial}{\partial t} \end{pmatrix} \begin{pmatrix} \sigma \\ u \end{pmatrix}, \begin{pmatrix} I & -D\nabla \\ -div & \frac{\partial}{\partial t} \end{pmatrix} \begin{pmatrix} \tau \\ v \end{pmatrix} \rangle_Y \tag{1.8}$$

The resulting candidate for a variational formulation can be derived as usual [source] that is as the *sum* of the directional derivatives of $J$, where the *directions* become the testfunctions from the space $W$.

$$\begin{aligned} &\text{Find } (\sigma, u) \in U \text{ such that } \mathcal{B}([\sigma, u], [\tau, v]) = \mathcal{L}_{\tilde{f}}(u)([\tau, v]) \quad \forall [\tau, v] \in W \\ &\text{with } \mathcal{B}([\sigma, u], [\tau, v]) = (\mathcal{A}([\sigma, u]), \mathcal{A}([\tau, v])) \text{ and } \mathcal{L}_{\tilde{f}}(u)([\tau, v]) = (\mathcal{A}[\tau, v], \tilde{f}(u))_Y \end{aligned} \tag{1.9}$$

**Remark:** $\mathcal{L}_{\tilde{f}}(u)([\tau, v])$ is a linear operator in $v$ but nonlinear in $u$. What remains to be determined are the function spaces $U$, $W$, and $Y$.

## 1.2 Function Spaces

In order to define $U$ and $Y$ let us consider the optimisation problem (1.6) and its variational formulation (1.9) again. We would like to allow for the broadest class of solutions possible while still ensuring that all terms are well-defined in $U$, and staying away from Sobolev spaces of negative or fractional powers due to the beforementioned practicality reasons. Additionally to guarantee the existence of all terms involved we have to make sure that the present weak derivatives of $\sigma$ and $u$ exist in the induced inner product of $Y$. Let us therefore consider the following spaces

$$H_{div}(\mathcal{S}) = \{\sigma \in (L^2(\mathcal{S}))^n : div(\sigma) \in L^2(\mathcal{S})\} \tag{1.10}$$

$$H_{DIV}(\Omega) = H^1_{div}(\mathcal{S}) \times L^2(\mathcal{T}) \tag{1.11}$$

*where and when to use $\times$, above it means something different from below ... still both direct sums?*

$$U = W = H^1_{DIV}(\Omega) \times H^1(\Omega) \tag{1.12}$$

$$Y = L^2(\Omega) \tag{1.13}$$

And thus we have that

$$\sigma \in H_{DIV}(\Omega) \text{ and } u \in H^1(\Omega) \text{ with} \tag{1.14}$$

$$||\sigma||^2_{H_{DIV}(\Omega)} = ||\sigma||^2_{L^2(\Omega)} + ||div(\sigma)||^2_{L^2(\Omega)}, \tag{1.15}$$

$$||u||^2_{H^1(\Omega)} = ||u||^2_{L^2(\Omega)} + ||\nabla u||^2_{L^2(\Omega)} + ||u_t||^2_{L^2(\Omega)} \text{ and hence} \tag{1.16}$$

$$||(\sigma, u)||^2_U = ||\sigma||^2_{H_{DIV}(\Omega)} + ||u||^2_{H^1(\Omega)} \tag{1.17}$$

$$CORRECT?! \tag{1.18}$$

The direct sum of two Hilbert spaces is again a Hilbert space using the inner product induced by the sum of the respective inner products [**?**]. We have that $L^2(\mathcal{T})$, $H^1(\Omega)$ and $H_{div}(\mathcal{S})$ are

Hilbert spaces [**?**], and hence we obtain that $U$, $W$, and $Y$ are as well. We can then check that all above terms in $J([\sigma, u], 0)$ are well defined. If we additionally assume $f(u) \in L^2(\Omega)$ for all $u$, the problem remains to be well-posed. Theoretically one could potentially only require $f \in H^{-1}(\Omega)$ but for the scope of this thesis, we will restrict ourselves to the former. Especially as we will later on have to require $f$ to be twice differentiable to actually solve the problem numerically, see chapter 6.

$\mathcal{B}$ induces an inner product and subsequently a norm on $U$, if we assume homogenous boundary conditions on $\sigma$ and $u$. We can check that it is a symmetric bilinear form. In addition we have that $\mathcal{B}([\sigma, u], [\sigma, u]) \geq 0$ for all $[\sigma, u] \in U$, since $\mathcal{B}([\sigma, u], [\sigma, u]) = J([\sigma, u], 0)$ is the sum of two squared $L^2$–norms. Therefore it only remains to show that

$$\mathcal{B}([\sigma, u], [\sigma, u]) = 0 \iff [\sigma, u] = 0. \tag{1.19}$$

If $[\sigma, u] = 0$ we immediately obtain that $\mathcal{B}([\sigma, u], [\sigma, u]) = 0$. Hence only the reverse direction remains to be shown. So let us assume that there exists $0 \neq [\sigma, u] \in U$
*to be filled*

In order to fully set the theoretical framework that guarantees us all the favourable attributes of the beforementioned Rayleigh–Ritz setting we would like to fulfill the assumptions of *Theorem 1* from section 3.3, which require the usage of conforming discrete subspaces $U^h \subset U$ and the following norm equivalence for some $\alpha, \beta > 0$

$$\alpha ||[\sigma, u]||_U^2 \leq J([\sigma, u], 0) \leq \beta ||[\sigma, u]||_U^2 \quad \forall\ [\sigma, u] \in U. \tag{1.20}$$

In order to show the upper bound on $J$ we use that $Y = L^2(\Omega)$, $D(x)$ is bounded, i.e. there exists $d_{\max} \in \mathbb{R}$ such that $\|D(x)\|_Y \leq d_{\max}$ for all $x \in \mathcal{S}$ and the parallelogram identity which holds in Hilbert spaces.

$$\begin{aligned}
J([\sigma, u], 0) &= \|u_t - div(\sigma)\|_{L^2(\Omega)}^2 + \|\sigma - D(x)\nabla u\|_{L^2(\Omega)}^2 \\
&\leq \|u_t - div(\sigma)\|_{L^2(\Omega)}^2 + \|u_t + div(\sigma)\|_{L^2(\Omega)}^2 + \|\sigma - D(x)\nabla u\|_{L^2(\Omega)}^2 + \|\sigma + D(x)\nabla u\|_{L^2(\Omega)}^2 \\
&= 2\|u_t\|_{L^2(\Omega)}^2 + 2\|div(\sigma)\|_{L^2(\Omega)}^2 + 2\|\sigma\|_{L^2(\Omega)}^2 + 2d_{\max}^2\|\nabla u\|_{L^2(\Omega)}^2 \\
&\leq 2\|u\|_{L^2(\Omega)}^2 + 2\|u_t\|_{L^2(\Omega)}^2 + 2d_{\max}^2\|\nabla u\|_{L^2(\Omega)}^2 + 2\|div(\sigma)\|_{L^2(\Omega)}^2 + 2\|\sigma\|_{L^2(\Omega)}^2 \\
&\leq \max(2, 2d_{\max}^2)(\|u\|_{H^1(\Omega)}^2 + \|\sigma\|_{H_{DIV}(\Omega)}^2) \\
&= \beta\|[\sigma, u]\|_U^2 \qquad \text{with } \beta := \max(2, 2d_{\max}^2).
\end{aligned} \tag{1.21}$$

The proof of the coercivity is not straight forward and potentially not even true. In the paper of Z. Cai et al. on "First-order system least squares for second-order partial differential equations: Part I", [**?**], they show norm equivalence for a similar class of problems that are however elliptic and therefore the terms and spaces involved differ. Nevertheless it might be possible to proceed similarly to their work to show that such an $\alpha$ exists, this is however beyond the scope of this thesis but could be an interesting extension to the topic.

A point that has not really been discussed so far but will have to be taken into account is the way of how to treat the boundary conditions in least–squares formulations. One possibility is to also include them in the functional as an additional term while another one would be to directly include them in the discretised system of the space. The former one entails the additional definition of an appropriate norm on the boundary while also requiring the treatment of the additional term. Since we have assumed it to be at least $L^2$-regular the appropriate conditions can directly be imposed as part of the discretised system which will be discussed in more detail in the implementation section.

## 1.3 A Finite Element Space-Time Formulation

After having derived a continuous least squares formulation let us turn towards deriving a finite element discretisation of the problem. We want to consider conforming subspaces of $U$ and $W$, hence we will be considering finite dimensional spaces $U_h \subset U$ and $W_h \subset W$ that are also defined on the entire space-time domain.

$U^h$ contains the solution space for $\sigma^h$ and $u^h$, that is $s^h = [\sigma^h, u^h] \in U^h$. However the subspaces for the variables do not have to be the same which can potentially be advantageous as the continuous spaces differ as well, due to their different properties. Hence let us assume that $\sigma^h \in \tilde{U}^h$ and $u^h \in \hat{U}^h$, where $U^h = \tilde{U}^h \times \hat{U}^h$.

Suppose we have $\tilde{n} = dim(\tilde{U}^h)$ and let $\{\tilde{\phi}_1, ..., \tilde{\phi}_{\tilde{n}}\}$ be a basis of $\tilde{U}^h$ and similarly $n = dim(\hat{U}^h)$ and $span\{\phi_1, ..., \phi_n\} = \hat{U}^h$. We furthermore assume $U^h$ to be constructed such that $\inf_{s^h \in U^h} ||s - s^h||_U \to 0$ as $h \to 0$ for all $s \in U$. This is a reasonable assumption, Raviart - Thomas something?! It is also worth noting that since we are in a space-time setting we have $\tilde{\phi} = \tilde{\phi}(x,t)$ and $\phi = \phi(x,t)$.

We can then represent $\sigma^h$ and $u^h$ as a linear combination of basis functions in $\tilde{U}^h$ or equivalently $\hat{U}^h$ that is

$$\sigma_h(x,t) = \sum_{i=1}^{\tilde{n}} \sigma_i \tilde{\phi}_i(x,t) \qquad u_h(x,t) = \sum_{i=1}^{n} u_i \phi_i(x,t) \tag{1.22}$$

The functional $J$ then looks as follows

$$J(\sigma_h, u_h) = || \sum_{i=1}^{n} u_i \ (\phi_i)_t - \sum_{i=1}^{\tilde{n}} \sigma_i \ div(\tilde{\phi}_i) - f(\sum_{i=1}^{n} u_i \phi_i)||_Y^2 + || \sum_{i=1}^{\tilde{n}} \sigma_i \tilde{\phi}_i - D(x) \nabla (\sum_{i=1}^{n} u_i \phi_i)||_Y^2 \tag{1.23}$$

which will give rise to a vector-valued system once we discretise the domain.

In the arising variational formulation that we will also have to consider finite dimensional subspaces of $W_h$ of $W$. In the scope of this thesis we restrict ourselves to the assumption that $W_h = U_h$. That is we introduce a set of test functions consisting of the basis vectors of $\tilde{U}^h$ and $U^h$. That is the discretised weak form then reads

$$\text{Find } (\sigma_h, u_h) \in U_h \text{ such that } B \cdot [\sigma_h, u_h]^T = L_{\tilde{f}}(u_h) \tag{1.24}$$

where $B \in \mathbb{R}^{m \times m}$, with $m = \tilde{n} + n$, be the matrix arising from the discretised bilinear operator, and $L_{\tilde{f}}(u_h) \in \mathbb{R}^m$ being a discretised right-hand side. Since we assume that the solution $s_h = [\sigma_h, u_h]$ first contains all values corresponding to $\sigma$ and then for $u$ we obtain a block structure for $B$ and $L_{\tilde{f}}$ of the following form.

$$B = \begin{bmatrix} B_{\sigma\sigma} & B_{\sigma u} \\ B_{u\sigma} & B_{uu} \end{bmatrix} \qquad L_{\tilde{f}}(u_h) = \begin{bmatrix} (L_{\tilde{f}}(u_h))_\sigma \\ (L_{\tilde{f}}(u_h))_u \end{bmatrix} \tag{1.25}$$

Each entry of each of the blocks of $B$ can be computed explicitly according to the subsequent

schemes.

$$\text{For } B_{\sigma\sigma}: \qquad B_{ij} = \langle \tilde{\phi}_j, \tilde{\phi}_i \rangle_Y + \langle div(\tilde{\phi}_j), div(\tilde{\phi}_i) \rangle_Y \qquad \forall i,j \in \{1,...,\tilde{n}\} \tag{1.26}$$

$$\text{For } B_{\sigma u}: \qquad B_{ij} = -\langle D(x)\nabla\phi_j, \tilde{\phi}_i \rangle_Y - \langle (\phi_j)_t, div(\tilde{\phi}_i) \rangle_Y \qquad \forall i \in \{1,...,\tilde{n}\}, j \in \{\tilde{n}+1,...,m\} \tag{1.27}$$

$$\text{For } B_{u\sigma}: \qquad B_{ij} = -\langle D(x)\nabla\phi_i, \tilde{\phi}_j \rangle_Y - \langle (\phi_i)_t, div(\tilde{\phi}_j) \rangle_Y \qquad i \in \{\tilde{n}+1,...,m\}, \forall j \in \{1,...,\tilde{n}\} \tag{1.28}$$

$$\text{For } B_{uu}: \qquad B_{ij} = \langle D(x)\nabla\phi_j, D(x)\nabla\phi_i \rangle_Y + \langle (\phi_j)_t, (\phi_i)_t \rangle_Y \qquad \forall i,j \in \{\tilde{n}+1,...,m\} \tag{1.29}$$

In the case that $f$ is independent of $u$, that is $f = 0$ or $f = f(x,t)$, the derivative of $f$ with respect to $u$, is clearly zero and therefore the right-hand side, which we will denote by $L_{\tilde{f}} = L_{\tilde{f}}(u_h)$ to underline its independence of $u$ only contains the following terms

$$(L_{\tilde{f}})_i = 0 \qquad i \in \{1,...,\tilde{n}\} \tag{1.30}$$

$$(L_{\tilde{f}})_i = \sum_{j=1}^{\tilde{n}} \langle div(\tilde{\phi}_j), f \rangle_Y - \langle (\phi_i)_t, f \rangle_Y \qquad i \in \{\tilde{n}+1,...,m\} \tag{1.31}$$

*where $P$ is some sort of projection in case $\tilde{n} \neq n$, how to really write this up?*
Thus we now know how compute each term we can assemble one large linear system of equations

$$B \begin{pmatrix} \sigma \\ u \end{pmatrix} = L_{\tilde{f}} \tag{1.32}$$

which can then be solved immediately for $[\sigma, u]^T$ using for example a multigrid method. In the case of $f = f(u)$ the situation is more complicated, since we also have to take the derivatives of $f$ with respect to $u$ into account, and then construct a nonlinear iteration scheme and consider linearisations of the problem to solve for $u$ unless we have that $f(u)$ is a linear function. In that case we would only obtain additional terms that can be added to $B_{uu}$.

## 1.4 Gradient and Hessian of the Objective

In order to formulate a nonlinear iteration scheme we will first derive the first and second order Gateaux derivatives of $J$ to be able to find concrete formulations for the gradient and hessian. We formulate them here now in their continuous form and will discuss the discretisation in the subsequent section. Furthermore for the remainder of this chapter we will assume that $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_Y$ in order to simplify the notation. For brevity and clarity we split the functional $J$ in three different terms that will be considered separately dividing it into the linear and nonlinear terms which is possible due to the linearity of the inner product.

$$\begin{aligned} J_1(\sigma,u) &= \frac{1}{2} c_1 \langle u_t - div(\sigma), u_t - div(\sigma) \rangle \\ J_2(\sigma,u) &= \frac{1}{2} c_1 \langle 2u_t - 2div(\sigma) - f(u), -f(u) \rangle \\ J_3(\sigma,u) &= \frac{1}{2} c_2 \langle \sigma - D(x)\nabla u, \sigma - D(x)\nabla u \rangle \end{aligned} \tag{1.33}$$

We can check that $J(\sigma,u) = J_1(\sigma,u) + J_2(\sigma,u) + J_3(\sigma,u)$.

Since $u \in H^1(\Omega)$ and $\sigma \in H_{DIV}(\Omega)$ they are not defined pointwise which we have to take into account for the subsequent computations.

*What changes, what doesn't? Weak derivatives.*

The partial directional derivatives of $J_1$ can be determined by once again taking the linearity of the inner product as well as its symmetry into account

$$\begin{aligned}
\frac{\partial J_1}{\partial \sigma} &= \lim_{\epsilon \to 0} \frac{J_1(\sigma + \epsilon\tau, u) - J_1(\sigma, u)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{c_1}{2\epsilon}(\langle u_t - div(\sigma + \epsilon\tau), u_t - div(\sigma + \epsilon\tau)\rangle - \langle u_t - div(\sigma), u_t - div(\sigma)\rangle) \\
&= \lim_{\epsilon \to 0} \frac{c_1}{2\epsilon}(\langle u_t, u_t\rangle - \langle u_t, div(\sigma)\rangle - \epsilon\langle u_t, div(\tau)\rangle - \langle div(\sigma), u_t\rangle + \langle div(\sigma), div(\sigma)\rangle \\
&+ \epsilon\langle div(\sigma), div(\tau)\rangle - \epsilon\langle div(\tau), u_t\rangle + \epsilon\langle div(\tau), div(\sigma)\rangle + \epsilon^2\langle div(\tau), div(\tau)\rangle \\
&- \langle u_t, u_t\rangle + \langle u_t, div(\sigma)\rangle + \langle div(\sigma), u_t\rangle - \langle div(\sigma), div(\sigma)\rangle) \\
&= \lim_{\epsilon \to 0} \frac{c_1}{2\epsilon}(-2\epsilon\langle u_t, div(\tau)\rangle + 2\epsilon\langle div(\sigma), div(\tau)\rangle + \epsilon^2\langle div(\tau), div(\tau)\rangle) \\
&= -c_1\langle u_t, div(\tau)\rangle + c_1\langle div(\sigma), div(\tau)\rangle
\end{aligned} \tag{1.34}$$

By proceeding analogously for equation $J_2$ and $J_3$ we obtain in these cases:

$$\frac{\partial J_2}{\partial \sigma} = c_1\langle div(\tau), f(u)\rangle \tag{1.35}$$

$$\frac{\partial J_3}{\partial \sigma} = c_2\langle \sigma, \tau\rangle - c_2\beta\langle \tau, \nabla u\rangle \tag{1.36}$$

Let us now turn to the partial derivatives with respect to $u$. Here we obtain the following for $J_1$ and $J_3$:

$$\begin{aligned}
\frac{\partial J_1}{\partial u} &= \lim_{\epsilon \to 0} \frac{J_1(\sigma, u + \epsilon v) - J_1(\sigma, u)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{c_1}{2\epsilon}(\langle (u + \epsilon v)_t - div(\sigma), (u + \epsilon v)_t - div(\sigma)\rangle - \langle u_t - div(\sigma), u_t - div(\sigma)\rangle) \\
&= c_1\langle u_t, v_t\rangle - c_1\langle v_t, div(\sigma)\rangle
\end{aligned} \tag{1.37}$$

$$\frac{\partial J_3}{\partial u} = -c_2\langle \sigma, D(x)\nabla v\rangle + c_2\langle D(x)\nabla u, D(x)\nabla v\rangle$$

*Repeat assumptions on D(x). But can leave it where it is for now?*

In the case of $J_2$, we have to take the non-linearity of $f$ into account. If we assume that $f$ sufficiently smooth (what do we need exactly?!) that is

$$\lim_{\epsilon \to 0} f(u + \epsilon v) = f(u) \text{ and} \tag{1.38}$$

$$\lim_{\epsilon \to 0} \langle f(u + \epsilon v), f(u + \epsilon v)\rangle - \langle f(u), f(u)\rangle = \langle f'(u) \cdot v, f(u)\rangle + \langle f(u), f'(u) \cdot v\rangle \tag{1.39}$$

which can be added due to symmetry.

$$\begin{aligned}
\frac{\partial J_2}{\partial u} &= \lim_{\epsilon\to 0}\frac{J_2(\sigma,u+\epsilon v)-J_2(\sigma,u)}{\epsilon}\\
&= \lim_{\epsilon\to 0}\frac{c_1}{2\epsilon}(\langle 2(u+\epsilon v)_t - 2div(\sigma) - f(u+\epsilon v), -f(u+\epsilon v)\rangle - \langle 2u_t - 2div(\sigma) - f(u), -f(u)\rangle)\\
&= \lim_{\epsilon\to 0}\frac{c_1}{2\epsilon}(-2\langle u_t, f(u+\epsilon v)\rangle + 2\langle u_t, f(u)\rangle\\
&- 2\epsilon\langle v_t, f(u+\epsilon v)\rangle\\
&+ 2\langle div(\sigma), f(u+\epsilon v\rangle - 2\langle div(\sigma), f(u)\rangle\\
&+ \langle f(u+\epsilon v), f(u+\epsilon v)\rangle - \langle f(u), f(u)\rangle)\\
&= -c_1\langle u_t, f'(u)\cdot v\rangle - c_1\langle v_t, f(u)\rangle + c_1\langle div(\sigma), f'(u)\cdot v\rangle + c_1\langle f(u), f'(u)\cdot v\rangle
\end{aligned} \tag{1.40}$$

Hence we obtain the following partial first order directional derivatives.

$$J_\sigma[\tau] = \frac{\partial}{\partial\sigma}J(\sigma,u)[\tau] = c_2\langle\sigma,\tau\rangle + c_1\langle div(\sigma), div(\tau)\rangle - c_2\langle D(x)\nabla u,\tau\rangle - c_1\langle u_t, div(\tau)\rangle - c_1\langle f(u), div(\tau)\rangle \tag{1.41}$$

$$\begin{aligned}
J_u[v] = \frac{\partial}{\partial u}J(\sigma,u)[v] =& c_1\langle u_t, v_t\rangle - c_1\langle v_t, div(\sigma)\rangle - c_2\langle\sigma, D(x)\nabla v\rangle + c_2\langle D(x)\nabla u, D(x)\nabla v\rangle\\
&- c_1\langle u_t, f'(u)\cdot v\rangle - c_1\langle v_t, f(u)\rangle - c_1\langle div(\sigma), f'(u)\cdot v\rangle + c_1\langle f(u), f'(u)\cdot v\rangle
\end{aligned} \tag{1.42}$$

Following the same principles one can determine the second order partial derivatives whose derivation will only be briefly outlined here for the most difficult terms which are those including $f$.

$$\frac{\partial^2}{\partial\sigma^2}J[\tau][\rho] = c_2\langle\rho,\tau\rangle + c_1\langle div(\rho), div(\tau)\rangle \tag{1.43}$$

$$\frac{\partial^2}{\partial\sigma\partial u}[v][\tau] = \frac{\partial^2}{\partial u\partial\sigma}[\tau][v] = -\langle\tau,\nabla v\rangle - \langle v_t, div(\tau)\rangle - \langle div(\tau), f'(u)v\rangle \tag{1.44}$$

$$\begin{aligned}
\frac{\partial^2 J}{\partial u^2}[v][w] &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(J_u(\sigma,u+\epsilon w)[v] - J_u(\sigma,u)[v])\\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(c_1\langle(u+\epsilon w)_t, v_t\rangle + c_2\langle\nabla(u+\epsilon w),\nabla v\rangle - c_1\langle(u+\epsilon w)_t, f'(u+\epsilon w)\cdot v\rangle\\
&- c_1\langle v_t, f(u+\epsilon w)\rangle - c_1\langle div(\sigma), f'(u+\epsilon w)\cdot v\rangle + c_1\langle f(u+\epsilon w), f'(u+\epsilon w)\cdot v\rangle\\
&- J_u(\sigma,u)[v])\\
&= c_1\langle w_t, v_t\rangle + c_2\langle\nabla w,\nabla v\rangle\\
&+ \lim_{\epsilon\to 0}\frac{1}{\epsilon}(c_1\langle u_t, f'(u+\epsilon w)\cdot v\rangle - c_1\langle u_t, f'(u)\cdot v\rangle\\
&- \epsilon\cdot c_1\langle w_t, f'(u+\epsilon w)\cdot v\rangle\\
&- c_1\langle v_t, f(u+\epsilon w)\rangle + c_1\langle v_t, f(u)\rangle\\
&- c_1\langle div(\sigma), f'(u+\epsilon w)\cdot v\rangle + c_1\langle div(\sigma), f'(u)\cdot v\rangle\\
&+ c_1\langle f(u+\epsilon w), f'(u+\epsilon w)\cdot v\rangle - c_1\langle f(u), f'(u)\cdot v\rangle)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 J}{\partial u^2}[v][w] &= c_1\langle w_t, v_t\rangle + c_2\langle\nabla w,\nabla v\rangle + c_1\langle u_t, w^T f''(u)v\rangle - c_1\langle w_t, f'(u)\cdot v\rangle - c_1\langle v_t, f'(u)\cdot w\rangle\\
&- c_1\langle div(\sigma), w^T f''(u)v\rangle + c_1\langle f(u), w^T f''(u)v\rangle + \langle f'(u)\cdot w, f'(u)\cdot v\rangle
\end{aligned} \tag{1.45}$$

*Yet some more on weak derivatives etc.*
Thus we have now computed the necessary terms needed for the gradient as well as the Hessian of $J$. So let us now construct the nonlinear iteration scheme in order to minimse $J$.

## 1.5 Nonlinear Iteration Scheme

*We only require the whole thing to be differentiable in a weak sense .... not like mentioned before.*

As discussed in the previous chapter [see section 3.4] the general approach to finding a minimiser of the functional $J$ is to search for a tuple $[\sigma, u]$ for which $\nabla J([\sigma, u]) = 0$, while $\nabla^2 J([\sigma, u])$ is positive definite. We start with an initial guess $s_0 = s_{\text{init}} = [\sigma_{\text{init}}, u_{\text{init}}]$ and then successively try to decrease energy. In case of a Newton step by finding a quadratic approximation of $J$ at the value of the current iterate $s_k$, where the updated solution $s_{k+1}$ is the minimiser of the quadratic approximation. However if $J$ is not convex, and hence $\nabla^2 J([\sigma, u])$ not positive definite, the extremum of the quadratic approximation can actually lead us to a maximum of $J$. Therefore if we want to use a Newton iteration and maintain a decrease or at least no increase of energy in every step, that is $J(s_0) \geq J(s_1) \geq ... \geq J(s_k) \geq ...$ we need to be checking for convexity. In order to perform a Newton step we have to compute the Hessian of $J$, as well as its gradient in each iterate. The other option that was discussed previously to obtain a reduction in energy is to use a gradient descent method where we simply take an iteration step into the steepest descent direction. Hence we only need to evaluate the gradient in this case, which is computationally much cheaper but only leads to a linear convergence rate, instead of a quadratic one in the convex Newton case.

There are different ways to linearise and discretise in $f$, that is we have to decide how to represent $f$ in our finite dimensional subspace formulation. For simplicity we assumed that $f \in C^2$, and denote the first and second derivative with respect to $u$ by $f'$ and $f''$. This is in line with the forcing term stemming from the FitzHugh-Nagumo model as well as many other physical applications [source]. Hence for each degree of freedom in $u$ we can determine coefficients for $f, f', f''$ and represent them in the basis of $\hat{U}^h$, that is

$$(f_h)_i = \sum_{i=1}^{n} f_i \phi_i, \qquad (f'_h)_i = \sum_{i=1}^{n} f'_i \phi_i, \qquad (f''_h)_i = \sum_{i=1}^{n} f''_i \phi_i \tag{1.46}$$

Therefore $f_h, f'_h$ and $f''_h$ now represent a finite dimensional approximation of $f$ for one particular approximation of $u$. With this representation we can now proceed to formulate a discretisation of the gradient of $J$, linearised in $s_k = [\sigma_k, u_k]$ where we know that inner products are still well-defined since we are working in conforming subspaces.
We want $\nabla J = 0$

$$(L_{\tilde{f}}(u_h))_\sigma = \langle div(\sigma_h), f'_h \rangle_Y \tag{1.47}$$

$$(L_{\tilde{f}}(u_h))_u = -\langle u_t, f'_h \rangle_Y - \langle v_t, f_h \rangle_Y + \langle f_h, f'_h \rangle_Y \tag{1.48}$$

$$\nabla J_k = \nabla J(s_k) = B \begin{pmatrix} \sigma_k \\ u_k \end{pmatrix} - L_{\tilde{f}}(u_k) \tag{1.49}$$

For one step of gradient descent we therefore compute the update by

$$s_{k+1} = s_k + \alpha(-\nabla J_k) \tag{1.50}$$

where $\alpha > 0$ is a scaling parameter that can be chosen in different ways, for example a line search algorithm and which will be discussed in more detail in the implementation chapter.

From the gradient and previous section we can determine the discretised Hessian which is needed for a Newton step. We obtain

$$H_k = \nabla^2 J(s_k) = B_{\text{lin}} + Q \tag{1.51}$$

where $B_{\text{lin}} = B$ from before and $Q$ contains the nonlinear part, that is

$$Q = \begin{bmatrix} 0 & Q_{\sigma u} \\ Q_{u\sigma} & Q_{uu} \end{bmatrix} \text{ where} \tag{1.52}$$

$$(Q_{\sigma u})_{ij} = (Q_{u\sigma})_{ji} = -\langle div(\tilde{\phi}), f_h' \rangle \tag{1.53}$$

$$s_{k+1} = s_k - H_k^{-1}(\nabla J_k) \tag{1.54}$$

that is we would like to solve the linear system of equations

$$e_{k+1} = -H_k^{-1}(\nabla J_k) \quad \text{with} \quad s_{k+1} = s_k + e_{k+1} \tag{1.55}$$

But since it is generally expensive to compute the inverse of a large matrix, even if $H_k$ is sparse and symmetric, because this property does in general not translate to the inverse we apply a multigrid method to solve (4.33). That is we need to solve a linear system of equations in each Newton step.
put derivative to zero therefore if we assume local convexity must be an extremum, positive definite, means minimum, therefore variational formulation makes sense

As mentioned in the prologue we consider an iterative approach to solve this problem and will therefore consider linearisations of the problem. But before going into more detail about that, let us consider the variational formulation of the coupled reaction-diffusion system (4.3)
A solution to (4.3) will always be a minimiser of the functional $J$ independent on the choice of $Y$, however the sequence of iterates is dependent on the choice of the norm.

After having derived this discrete least squares space-time finite element formulation we will in the next chapter be looking at a concrete implementation which contains numerical results for some sample problems.