Multi-parameter regularization arising in optimal control of fluid flows

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Markus Klein
aus Reutlingen

Tübingen
2014

Tag der mündlichen Prüfung:	26.01.2015
Dekan:	Prof. Dr. Wolfgangs Rosenstiel
1. Berichterstatter:	Prof. Dr. Andreas Prohl
2. Berichterstatter:	Prof. Dr. Michael Hinze

Abstract

The objective of this thesis is to study optimal control problems subject to equations arising in the field of fluid dynamics. This thesis is split into two essential parts. Each of them deals with an important partial differential equation, that are of interest in various applications and are widely considered in current research: The density dependent Navier–Stokes equation and the thin-film equation.

These optimal control problems are motivated in many ways: First, the equations are mathematically interesting due to strong nonlinear effects occurring additionally as coupling effects in the context of optimization. Also, it is not immediate that properties (such as convergence of numerical approximations) are inherited by the optimal control problem. The literature on optimal control subject to nonlinear partial differential equation is rare, while the knowledge on those problems subject to the mentioned equations is even more rare: Only very few works are known, and the content of this thesis is a big contribution to this topic. Finally, for both control problems, there are industrial applications requiring the optimal control of fluid flows (which will also be addressed within the this thesis) such as the control of the interface in aluminum production, or the control of thin liquid layer on a silicon wafer.

In both parts, the use of regularization parameters is vital in order to overcome analytical issues. The coupling of these parameters is specified, and (in the second part) a limiting problem is solved for these parameters tending to zero.

In the first part of this thesis, we consider an optimal control problem for the interface in a two-dimensional two-phase fluid problem. The minimization functional consists of two parts: The L^2-distance to a given density profile and the interfacial length. We show existence of an optimal control and derive necessary first order optimality conditions for a corresponding phase field approximation. An unconditionally stable fully discrete scheme which is based on low order finite element discretization is proposed, and convergence of corresponding iterates to solutions of the continuous optimality conditions for vanishing discretization parameters is shown.

The second part consists of an optimal control problem subject to the thin-film equation which is deduced from the Navier–Stokes equation. The thin-film equation lacks well-posedness for general controls due to possible degeneracies; state constraints are used to circumvent this problematic issue, and ensure well-posedness of the optimal control problem as well as the rigorous derivation of necessary first order optimality conditions for the optimal control problem. A multi-parameter regularization addressing both, the possibly degenerate term in the equation and the state constraint, is considered, and convergence is shown for vanishing regularization parameters by decoupling both effects.

Both parts are concluded by corresponding numerical experiments, validating the models, comparing parameters (of the regularization and of the numerical algorithms) and their scaling to each other, and including academic examples of industrial applications.

Zusammenfassung in deutscher Sprache

Ziel dieser Arbeit ist es, optimale Steuerungsprobleme mit Gleichungen, die auf dem Gebiet der Fluiddynamik auftauchen, zu studieren. Die vorliegende Doktorarbeit ist in zwei wesentliche Teile aufgespalten. Jeder Teil behandelt eine wichtige partielle Differentialgleichung, die von Interesse in weitreichenden Anwendungen sind und noch immer aktuell erforscht werden: Die dichteabhängige Navier-Stokes Gleichung und die dünne Filme Gleichung.

Diese optimalen Steuerungsprobleme sind vielseitig motiviert: Zunächst sind die Gleichungen aufgrund starker nichtlinearer Effekte, die im Rahmen der Optimierung zusätzlich zu Kopplungseffekten führen, mathematisch interessant. Weiter ist es nicht unmittelbar klar, ob sich Eigenschaften (wie etwa die Konvergenz numerischer Approximationen) innerhalb der Optimalsteuerungsproblems vererben. Die Erkenntnisse in der Literatur über Optimalsteuerungsprobleme bezüglich nichtlinearen partiellen Differentialgleichungen sind rar, während der Kenntnisstand über jene Optimalsteuerungsprobleme, die sich mit den benannten Gleichungen beschäftigten, noch viel unvollständiger ist: Es sind bisher nur sehr wenige Arbeiten darüber bekannt und der Inhalt der vorliegenden Doktorarbeit ist ein großer Beitrag zu diesem Thema. Schließlich gibt es für beide Probleme industrielle Anwendungen, welche die Steuerung von Flüssigkeitsströmungen erforderlich machen (diese werden innerhalb der vorliegenden Arbeit auch thematisiert), wie etwa die Kontrolle von Grenzflächen in der Aluminiumproduktion oder die Kontrolle von dünnen Flüssigkeitsschicht auf Siliziumwafern.

In beiden Teilen dieser Arbeit ist die Verwendung von Regularisierungsparametern unerlässlich, um analytische Probleme zu überwinden. Die Kopplung dieser Parameter wird spezifiziert und (im zweiten Teil) ist ein Grenzproblem für den Fall gelöst, dass die Parameter gegen Null konvergieren.

Im ersten Teil der vorliegenden Doktorarbeit betrachten wir ein Optimierungsproblem, um die Grenzfläche eines zweidimensionalen Zwei-Phasen Problems zu steuern. Das zu minimierende Funktional besteht dabei aus zwei Teilen: Dem Abstand zu einem gegebenen gewünschten Dichteprofil (gemessen in der L^2-Norm) sowie der Länge der Grenzfläche. Wir zeigen Existenz einer optimalen Steuerung und leiten notwendige Optimalitätsbedingungen erster Ordnung für eine zugehörige Phasenfeld-Approximation her. Wir schlagen ein vorbehaltlos stabiles volldiskretes Schema vor, welches auf einer Finite Elemente Diskretisierung niedriger Ordnung beruht, und wir zeigen für verschwindende Diskretisierungsparameter die Konvergenz zugehöriger optimaler Steuerungen gegen Lösungen der kontinuierlichen Optimalitätsbedingungen.

Der zweite Teil besteht aus einem Optimalsteuerungsproblem bezüglich der dünnen Filme Gleichung, welche aus der Navier-Stokes Gleichung hergeleitet wird. Der dünnen Filme

Gleichung fehlt für allgemeine Kontrolle die Wohlgestelltheit aufgrund möglicher Degeneriertheit; Zustandsbeschränkungen werden benutzt, um dieses Problem in den Griff zu umgehen und Wohlgestelltheit des Optimalsteuerungsproblems sicherzustellen sowie notwendige Optimalitätsbedingungen erster Ordnung rigoros herzuleiten. Ein Mehrparameteransatz zur Regularisierung, der beide Probleme – den möglicherweise degenerierten Term in der Gleichung und die Zustandsbeschränkungen – anspricht, wird betrachtet, und für diesen Ansatz wird Konvergenz für verschwindende Regularisierungsparameter gezeigt, der beide Effekte entkoppelt.

Beide Teile werden werden von entsprechenden numerischen Experimenten abgeschlossen, welche die Modelle prüfen, Parameter und deren Skalierungen miteinander vergleichen (solche, die für die Regularisierung nötig sind, aber auch welche, die in den numerischen Algorithmen vorkommen) und Beispiel der industriellen Anwendungen auf akademischen Niveau beinhalten.

Acknowledgement

I thank everyone who supported me during the preparation of this thesis in whatever way.

There are some people deserving a special note of thanks.

I would like to thank my advisor, Prof. Dr. Andreas Prohl, for the opportunity to write my thesis in the interesting and important field of optimal control, as well as the constant support over the years. I also think him and Prof. Dr. Michael Hinze for being the two evaluators of this thesis.

I thank my recent and former colleagues of the numerical analysis group: Bernd Brumm, Erich Carelli, Thomas Dunst, Jonas Hähnle, Heiko Kröner, Dhia Mansour, Mikhail Neklyudov, Christian Power, Anton Prochel, Christian Schellnegger, Jonathan Seyrich, Hanna Wallach, and Daniel Weiß. In particular, I am grateful to my colleagues and friends Bernd Brumm, Thomas Dunst, Dhia Mansour, Christopher Nerz, Ailyn Schäfer, Jonathan Seyrich, and Pirmin Vollert, who were involved in many helpful mathematical discussions, and who gave me a constant moral support over the last four years and even longer. I thank Kristina Kohls and Christopher Nerz for reading parts of this thesis and giving me feedback.

I would also like to thank all my friends for helping me to relax and having enough energy to complete my research and this thesis. In particular my deep gratitude goes to my best friend Verena who was and who is always the greatest support for me in all parts of my life.

Finally, I would like to thank my parents. Without their help it would have been impossible for me to study, and great interesting field of mathematics would have remained unknown to me.

The work during the preparation of this thesis was supported by a DFG grant within the Priority Program SPP 1253 (Optimization with Partial Differential Equations). The simulations in chapter 16 were performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC; cf. [25].

Contents

List of Figures

Introduction

The Navier–Stokes equation describes the motion of viscous incompressible fluid in a given domain $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) over a time interval $[0, T]$, and is a subject of recent research and has many physical and engineering applications. In order to solve the equation, we search for the velocity $\boldsymbol{y} : \Omega_T \to \mathbb{R}^d$, the density $\rho : \Omega_T \to \mathbb{R}$, and the pressure $p : \Omega_T \to \mathbb{R}^d$ with

$$\begin{aligned} \rho \boldsymbol{y}_t + \rho[\boldsymbol{y} \cdot \nabla]\boldsymbol{y} - \operatorname{div}(\mu(\rho)\nabla \boldsymbol{y}) + \nabla p = \rho \boldsymbol{u}, \\ \rho_t + [\boldsymbol{y} \cdot \nabla]\rho = 0, \\ \operatorname{div} \boldsymbol{y} = 0, \end{aligned} \tag{NSE}$$

together with initial conditions $y(0, .) = y_0 : \Omega \to \mathbb{R}^d$ and $\rho(0, .) = \rho_0 : \Omega \to \mathbb{R}$; here, $\boldsymbol{u} : \Omega_T \to \mathbb{R}^d$ is a given right-hand side acting as an external force, and $\mu(\rho) : \Omega_T \to \mathbb{R}$ is the (density-dependent) viscosity. The set of equations is completed with suitable boundary conditions which will be detailed later. On a physical level, the first equation of (NSE) is conservation of momentum, while the second equation of (NSE) models conservation of mass, and the last one in (NSE) is the incompressible condition. If we set ρ to be constant, we arrive at the classical (or, one-fluid) Navier–Stokes equation, which has still some unsolved properties regarding existence and regularity, cf. [31]. Of course, the situation is more complicated if ρ is present in the equation: First existence results were provided in the middle of the 1970s for special cases (e.g., if μ does not depend on ρ). A more general existence and regularity result containing the above formulation (NSE) of the equation with more quantitative properties of solutions was shown in [54] with techniques based on [29]. For an overview about history, physical derivation of the equations and related models, we refer the reader to [54].

The range of applications of the Navier–Stokes equation is very wide, as it can be used for engineering applications related to fluid flows, such as the production of aluminum (cf. [32]), the description of blood flow (cf. [60]), manufacturing semiconductor devices (cf. [17]), the transport of human cells through a fluid channel (cf. [8])—to name only a few concrete examples. For the Navier–Stokes equation and for most related models, existence is shown and converging numerical algorithms are developed; see, e.g., [33, 63] for the one-fluid Navier–Stokes equation, and, e.g., [12, 38, 55] for the density-dependent case.

In many applications, it is desirable not only to calculate how the fluid behaves for a given driving force $\boldsymbol{u}$, but we want to control the behavior of the fluid(s): To use the given examples above, we want to control the aluminum production with such forces that the aluminum part of the fluid does not to touch the electrodes in order to prevent it from a short circuit; we want to introduce forces in the manufacturing of Si wafers such that the height profile of a layer forms a specific pattern; we want to transport human cells with the help of such forces that the shape of the cells is similar during the movement, etc.

As noted above, in optimal control problems the force $\boldsymbol{u}$ is not given, but to be found, whereas a desired profile for other variables (in our context, for $\boldsymbol{y}$ or ρ) is also given. In optimal control problems a standard approach is to minimize the following functional,

$$J(\boldsymbol{y},\rho,\boldsymbol{u}) := \frac{\alpha_{\boldsymbol{y}}}{2}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|^2_{\boldsymbol{L}^2(\Omega_T)} + \frac{\alpha_\rho}{2}\|\rho-\tilde{\rho}\|^2_{L^2(\Omega_T)} + \frac{\alpha_{\boldsymbol{u}}}{2}\|\boldsymbol{u}\|^2_{\boldsymbol{L}^2(\Omega_T)}$$

subject to the Navier–Stokes equation, where $\alpha_{\boldsymbol{y}}, \alpha_\rho, \alpha_{\boldsymbol{u}}$ are nonnegative numbers and $\tilde{\boldsymbol{y}}$ and $\tilde{\rho}$ are given target functions. The first two terms are called tracking type functions, where the third part is called cost part of the function: If $(\boldsymbol{y}^*, \rho^*, \boldsymbol{u}^*)$ is a minimum of J, the velocity $\boldsymbol{y}^*$ is near the desired velocity $\tilde{\boldsymbol{y}}$, the density distribution ρ^* is near the desired density distribution $\tilde{\rho}$ (both measured in the L^2-norm), and the $\boldsymbol{L}^2$-norm of $\boldsymbol{u}^*$ (which reflects the cost of this particular control) is not too big. From a mathematical viewpoint, it is important to have $\alpha_{\boldsymbol{u}} > 0$ as this is one key—together with well-posedness and a-priori estimates of the equation—to imply existence of optimal solutions, which is detailed in, e.g., Theorem 4.2 and Theorem 13.2.

Optimal control of the one-fluid Navier–Stokes equation has been studied since more than twenty years, where one of the first works was [2]. Existence of minima are proven, necessary and sufficient optimality conditions are derived, and convergence of time and space discretizations are proven. For a summary of many results, we refer the reader to, e.g., [39, 40] and the references therein.

In contrast, optimal control of the density dependent Navier–Stokes equation is a very recent subject and the literature is very rare, in particular for works where rigorous derivation of optimality conditions and convergence proofs for numerical discretizations are provided. The author is aware of the article [52], where an optimal control problem with an L^2 tracking-type functional subject to a regularized density dependent Stokes equation in $\Omega \subset \mathbb{R}^2$ is studied, and optimality conditions are derived. Moreover, by some assumptions on a nonregularized solution, it is shown that minima (with a regularization parameter $\varepsilon > 0$) converge to a minimum of the limiting problem for $\varepsilon = 0$. The first part of this thesis provides a notable contribution on the optimal control of the density dependent Navier–Stokes equation.

Before we go into further detail of the new results, we want to introduce an other equation, which will be the equation of interest in the second part of this thesis: The thin-film equation, which can be derived from the Navier–Stokes equation, and which describes the motion of a (thin) fluid film on a plane. Assume that $\Omega = \{(x,y) \in (a,b)\times\mathbb{R} : \ 0 < y < g(x)\}$ is filled with a fluid, where $g : [a,b] \to \mathbb{R}$ is a given continuous function with $g > 0$. There are two main parts of its boundary: A lower boundary $\underline{\partial}\Omega := [a,b] \times \{0\}$ and the upper boundary $\overline{\partial}\Omega := \operatorname{graph} g$.

In order to model the motion of the fluid, we consider the one-fluid Navier–Stokes equation with an conservative force of the type $\boldsymbol{u}(x,y) = \nabla g_0(y)$ (where $g_0 : \mathbb{R} \to \mathbb{R}$ is a scalar potential) within the domain Ω together with suitable boundary conditions: To model the solid-liquid interface on $\underline{\partial}\Omega$, we suppose so-called no slip conditions, which model friction; to model the liquid-gas interface on $\overline{\partial}\Omega$, we suppose a surface tension condition; to model the fact that the fluid film is thin, we suppose that $\tau := \frac{\max u}{b-a} \ll 1$. A nondimensional transformation (based on τ) of the involved terms and of the boundary conditions leads to an asymptotic expansion in τ. Neglecting higher order terms of τ then leads to the so-called thin-film equation,

$$y_t = -(f(y)y_{xxx})_x - (g_0(y)y_x)_x, \tag{TF}$$

where y is the height of the film on at a given position and $f(y) = |y|^3$. The equation is completed with initial data, and with homogeneous boundary data for y_x and y_{xxx}. A self-contained derivation of this equation is done in chapter B, where we follow [16]. By virtue of the leading spatial term in the thin-film equation, (TF) is degenerate: If a potential solution is zero in a given region, no a-priori estimates in this region would be accessible, which questions existence results—or, at least, complicates their proof.

In [19], the authors considered the thin-film for $g_0 \equiv 0$ and showed existence for nonnegative initial data, regularity results and showed a so-called entropy estimates which ensures the solution to be positive for all times, provided the initial data is positive. This kind of maximum principle is surprising since linear equation of fourth order do not meet a maximum principle. We also refer the reader to [21], where an overview about the equation, its solution, and properties of the solution are provided; and we refer the reader to [20], where the existence result and the entropy estimates are extended to $g_0 \equiv -1$. In the meantime, other boundary conditions for $\partial\Omega$ are used, other potential functions are used, and convergence of property conserved discretization schemes are development; see, e.g., [15, 16, 34, 35, 36, 37, 48, 57].

In constrast, the literature on optimal control of thin films is more than rare: The author is not aware of any literature dealing with optima control problems subject to (TF). In the second part of this thesis, we consider an optimization problem subject to the thin-film equation, which is the first time such a problem is presented in the literature. However, there is one example in the literature [67], where the authors deal with a much more easy constraint, which is related to the thin-film equation (TF). We will go into more detail about the setup in [67] after presenting our approach.

The problem in directly addressing (TF) is that the control functions would be potential functions g_0, and the coupling between the control and state y would be too strong—unless the problem is transformed into a problem with finite controls parameters by assuming g_0 to be a polynomial or similar. The author is not aware of any literature dealing with an optimization problem, where controls appear in the form of potential functions instead of L^2 forces (or similar). Since this is not even conceptionally clear, we consider a modified equation,

$$y_t = -(f(y)y_{xxx})_x - (g_0(y)y_x)_x + u_x, \tag{TF-mod}$$

where g_0 is now a given potential function, and $u : \Omega_T \to \mathbb{R}$ is considered as control. Since every term in (TF-mod) appears in divergence-form, the mass (which is $\int_\Omega y(t,.)\,\mathrm{d}x$) is conserved as long as we consider all controls u to vanish at the boundary. If we now consider the following standard tracking-type functional,

$$J_{TF}(y,u) = \frac{1}{2}\|y - \tilde{y}\|^2 + \frac{\alpha}{2}\|u\|^2,$$

it is unclear if an optimal control problem of minimizing J subject to (TF-mod), is well-posed: For a given control u, it is unclear if a corresponding solution y of (TF-mod) exists due to the degenerate leading part of the equation. For a given "good" potential function g_0, entropy estimates are valid (which ensure positivity; see above). If we add state constraints $y \geq C_0 > 0$ (for a suitable constant $C_0 > 0$) and assume g_0 to be of this type in order to have at least one feasible control, we can prove existence of an optimum. With the main result from [4] and an appropriate choice of function spaces, necessary optimality conditions are derived containing measure-valued Lagrange multipliers. Other authors have dealt with degeneracies in a similar way, cf. [26, 27, 28]. The results in the second part of this thesis seems to be—except for these mentioned results—one of the first works for optimal control problems subject to a degenerate partial differential equation.

In order to circumvent the measure-valued Lagrange multipliers and to make the control problem implementable, we want to use a relaxation method, e.g., to remove the state constraint and add a penalty type functional,

$$\frac{1}{2\gamma} \left\| (C_0 - y)^+ \right\|^2,$$

to the original functional J, and let the regularization parameter $\gamma > 0$ tend to zero. But doing this at this point would lead to the same problems as above since it could lead to non-feasible solutions (i.e., $y < C_0$ in a region), which could lead to ill-posedness of the equation. This vicious circle is broken by introducing a regularization parameter $\varepsilon > 0$ into the equation, leading to

$$y_t = -\left([f(y) + \varepsilon] y_{xxx}\right)_x - (g_0(y) y_x)_x + u_x. \tag{TF-eps}$$

We can now consider an intermediate optimal control problem where state constraints $y \geq C_0$ are present, and $\varepsilon > 0$ is active. Clearly, this intermediate problem is solvable, and necessary optimality conditions can be derived. The main result of the second part of this thesis is the convergence of optimal state-control pairs $\{(y_\varepsilon, u_\varepsilon)\}$ of the intermediate problem towards a minimum of the original problem for $\varepsilon \to 0$ (for a subsequence); see Theorem 14.4. To do this, several additional technical difficulties have to be overcome—details are provided in the second part of this thesis.

After establishing the intermediate problem, a relaxation of the state constraint via a penalty type approach leads to a well-posed optimal control problem, for which existence is shown. As a second key result, convergence to a solution of the intermediate problem (i.e., state constraint being present and $\varepsilon > 0$) is shown for $\gamma \to 0$, and necessary optimality conditions are derived—now, they contain only regular Lagrange multiplier, which is a good starting point for the numerical implementation.

We note that the convergence results are only proven for $g_0 \equiv 0$ or $g_0 \equiv -1$. For other fixed potential functions, some particular results do not hold, which is why the theory for more general potential functions is left open.

In [67], the authors study (TF-eps) with $\varepsilon = 1$ as the governing equation, and then they consider an optimal control problem where J_{TF} is to minimized with respect to (TF-eps) (for $\varepsilon = 1$) with minor obvious differences (which are not crucial for their analysis). This problem coincides with the intermediate optimal control problem in our context without state constraints. The main theorem in [67] is to proof existence of solutions for the described optimal control problem. In contrast to our work, the main difficulty of paying attention to the degenerate governing equation is omitted, as well as a convergence study. Hence the contribution to this topic by the author of this thesis seems significant since main difficulties are addressed and solved in the second part of this thesis.

Coming back to the optimal control problem subject to (NSE), there are also technical problem calling for the introduction of regularization parameters. Let us first specify the problem: We are interested in minimizing the following functional, leading to

$$J_{NSE}(\rho, \boldsymbol{u}) := \int_0^T \left\{ \beta \mathcal{H}^1(S_\rho) + \frac{\lambda}{2} \int_\Omega |\rho - \tilde{\rho}|^2 \,\mathrm{d}\boldsymbol{x} + \frac{\alpha}{2} \int_\Omega |\boldsymbol{u}|^2 \,\mathrm{d}\boldsymbol{x} \right\} \mathrm{d}t$$

for $\beta, \lambda, \alpha > 0$, subject to (NSE) ($\mu$ is assumed not to depend on ρ) with homogeneous Dirichlet boundary conditions for the velocity and homogeneous Neumann boundary conditions for the density. We assume ρ_0 to be consisting of two values standing for the two fluids, respectively. The set S_ρ is the so called jump-set containing of those points touching both fluids, and $\mathcal{H}^1$ is the one-dimensional Hausdorff-measure. This first term in J is motivated in order to minimize the interfacial length, which is important in several applications; see chapter 1.

The problem here is that J_{NSE} is not well-defined since the interfacial length may not be bounded; and at the same time, the mass equation in (NSE) lacks regularity properties which impedes the derivation of necessary optimality conditions—the latter issue was also a problem in [52]. The second problematic issue can be overcome by introducing artificial diffusion into the mass equation by adding $-\varepsilon\Delta\rho$ for a small $\varepsilon > 0$,

$$\begin{aligned} \rho \boldsymbol{y}_t + \rho[\boldsymbol{y} \cdot \nabla]\boldsymbol{y} - \operatorname{div}(\mu \nabla \boldsymbol{y}) + \nabla p &= \rho \boldsymbol{u}, \\ \rho_t + [\boldsymbol{y} \cdot \nabla]\rho - \varepsilon\Delta\rho &= 0, \\ \operatorname{div} \boldsymbol{y} &= 0, \end{aligned} \tag{NSE-eps}$$

This method was once used to show existence of (NSE) and therefore, it is also known that for $\varepsilon \to 0$, the sequence of regularized states $\{(y_\varepsilon)\}$ converges to the solution of (NSE) in suitable function spaces, cf. [54]. It is now possible to show all required regularity properties and to derive necessary optimality conditions, but the regularity of ρ is so good that it implies $S_\rho = \varnothing$, hence the first term in the functional J_{NSE} has no impact. This issue can be overcome by replacing the term $\mathcal{H}^1(S_\rho)$ by the so-called perimeter approximation (cf. [22]),

$$F_\delta(\rho) := \int_\Omega \left\{ \delta |\nabla\rho|^2 + \frac{1}{4\delta} W(\rho) \right\} \mathrm{d}\boldsymbol{x},$$

and regularizing the initial value. It is known that the perimeter approximation, F_δ, Γ-convergence to $\mathcal{H}^1(S_\rho)$, cf. [22]. An important consequence of Γ-convergence is that if $f_\delta \xrightarrow{\Gamma} f$, then the sequence of minima x_δ of f_δ converges to the minimum of f, if they exist.

An ingenuous strategy would be the following: Solve the optimal control problem for $\delta, \varepsilon > 0$ and consider the sequence of the minima $(\rho_{\delta,\varepsilon}, \boldsymbol{u}_{\delta,\varepsilon})$. Since the side constraint converges for $\varepsilon \to 0$ and the functional Γ-converges, this would be a method to construct a minimum of the original problem, which was not clear if it is solvable. However, this is not true. The author has constructed an easy example for functions $f_i, f : \mathbb{R} \to \mathbb{R}$ which Γ-converge, and invertable side constraints $g_i = 0$ and $g = 0$, respectively, for $g_i, g : \mathbb{R} \to \mathbb{R}$, but the combined functions do not Γ-converge, although “something seems to converge”. We refer to chapter A for the definition of Γ-convergence and the discussed example. The author is not aware of a concept where Γ-convergence is considered with additional constraints.

It does not seem to be possible that one of the two regularization parameters for this problem can tend to zero (or even set to zero), but there is a hint on how to scale $\delta > 0$ and $\varepsilon > 0$ against each other: By a-priori estimates on (NSE-eps), the gradient of ρ is bounded by means of ε. In order to bound the first term in the functional F_δ uniformly (with respect to $\delta > 0$ or $\varepsilon > 0$), we can insert the a-prior estimate into the functional F_δ, and derive $\varepsilon \approx \delta$ in order to have at least this term uniformly bounded. We note that this is only necessary for convergence, but far away from being sufficient. However, numerical experiments in chapter 9 confirm that this choice is quite reasonable. More details are provided in chapter 1 and chapter 9.

For both parameters being positive, existence of an optimum can be shown, and necessary optimality conditions can be derived, which are the starting point for the numerical analysis. It is not immediate that convergence results of the equation (for discretization parameters tending to zero) are inherit by the discretization of the optimal control problem since the adjoint equation appearing in necessary optimality conditions (which is linear) is highly coupled with the state equation: The state variables (whose regularity properties are restricted) appear as coefficients in the adjoint equation; in this particular case we have two adjoint variables which couple with each other as they both appear in both adjoint equations. We want to discretize the equation by the method introduced in [12], use the “first discretize, then optimize”-strategy (which implies directly the solvability of the coupled state-adjoint equation system), and then we want to bound all emerging variables uniformly with respect to discretization parameters, which implies the existence of weak limits ensuring the convergence towards the continuous optimality conditions (up to a subsequence). In order to derive uniform bounds, the coupling effects in the adjoint equation have to be understood leading to a sophisticated analysis; details are provided in the chapter 5 and chapter 7. The main theorem of the first part of this thesis states that the sequence of minima of the discretized problem converges to functions (up to subsequences) which solve the continuous necessary optimality condition; see Theorem 8.3. The same strategy was also successfully used in [30].

In the second part of this thesis, a convergence study of the numerical analysis is left open and experiments are based directly on the discretization of the equation.

To sum up, both parts of this thesis deal with optimal control problems subject to equations arising in fluid dynamics; both equations come with their own intrinsic difficulties, but both require the introduction of regularization parameters in a natural fashion and therefore illuminate different aspects of optimal control problems: The need to introduce fixed regularization parameters in order to well defined functional, and to perform numerical analysis (for discretization parameters tending to zero) in the first part of this thesis; the need to introduce state constraints which then will be regularized and the need to regularize the equation in order to perform the convergence of the overall convergence in the second part of this thesis. Also, in both parts, the interface is controlled: In the first part we control the interface between the fluids and address it with the term modeling interfacial length, while in the second part the interface between the liquid-gas interface is given by the graph of the state variable which is directly addressed in the optimal control problem.

In this sense, the use of the regularization parameters is somehow complementary: In the first part of this thesis, they are coupled as described above, and both have to appear simultaneously; in the second part of this thesis, the effect of the regularization parameters is decoupled and one variable after the other can tend to zero.

The analysis of both parts leads to open questions. There are two important questions about general concepts, which emerged within this work: However, not only the answers, but even the concept behind it is unclear. The author thinks that it could be worth to investigate on the development of concepts.

- How would an optimal control problem work, if the control is not an L^2 function, but a potential function $g_0 : \mathbb{R} \to \mathbb{R}$ occurring as $g_0(y)$ in the equation (where y is the state)? How can existence of optimal solutions be shown without restricting the class of controls g_0 too much (e.g., if g_0 is a polynomial of a fixed degree, the problem would transform into one with a finite dimensional control space)? How can necessary optimality conditions be derived?
- How can one define Γ-convergence with respect to constraints? Should the value of the functional be set to $+\infty$, if the constraint does not hold, or is there an other possibility? Is it possible to prove a corresponding result to the fundamental result of Γ-convergence (i.e., Γ convergence implies the convergence of corresponding minima)?

Both parts of this thesis are concluded by numerical experiments which show evidence of the models: In the first part in chapter 9, it is shown that presence of the term $\mathcal{H}^1(S_\rho)$ leads to fundamental different results than its absence. For the second part in chapter 16, it is shown that a fixed given right-hand side u can lead to negative solutions of the equation, which confirms the necessity of state constraints. In both parts, comparison of various parameters were done, scaling effects of various parameters are studied, and examples leading to applications (on an academic level) are studied.

The first part of this work relies on the published version [9], while the second part of this thesis relies on [49]. The respective introductions of the two parts of this thesis contain a more detailed introduction to the specific topic, and they give an overview of the structure of the respective part.

Part I.

Control of interface evolution in multiphase fluid flows

1. Introduction

We plan to control the motion of a multi-phase fluid flow in a bounded domain $\Omega \subset \mathbb{R}^2$. A typical application includes the production of aluminum via electrolysis, where liquid aluminum oxide flows on top of liquid aluminum in some container, and aluminum may not come into contact with the electrolytes in some region at the top of the container [32]; as a consequence, oscillatory effects of the interface between the fluids should be avoided. Other applications for the control of interfacial regions between different fluids can be found in microfluidics for material processing, chemistry, biology, and medicine; see, e.g., [8]. Below, the motion of the fluid is described by the incompressible multi-phase fluid Navier–Stokes equations (cf., e.g., [54]), and the objective functional consists of a L^2 tracking-type term, together with the perimeter functional. The optimal control problem then reads as follows.

Problem 1.1
Let $\Omega \subset \mathbb{R}^2$ be open and bounded, and let $\tilde{\rho} : \Omega_T := (0,T) \times \Omega \to \mathbb{R}$ be given. Let $\alpha, \beta, \lambda > 0$, and $T > 0$ be fixed. Find $\boldsymbol{y}, \boldsymbol{u} : \Omega_T \to \mathbb{R}^2$, and $\rho : \Omega_T \to \mathbb{R}$, such that

$$G(\rho, \boldsymbol{u}) := \int_0^T \left\{ \beta \mathcal{H}^1(S_\rho) + \frac{\lambda}{2} \int_\Omega |\rho - \tilde{\rho}|^2 \, \mathrm{d}\boldsymbol{x} + \frac{\alpha}{2} \int_\Omega |\boldsymbol{u}|^2 \, \mathrm{d}\boldsymbol{x} \right\} \mathrm{d}t \tag{1.1}$$

is minimized subject to the density dependent Navier–Stokes equations,

$$\rho \boldsymbol{y}_t + \rho[\boldsymbol{y} \cdot \nabla]\boldsymbol{y} - \operatorname{div}(\mu(\rho)\nabla \boldsymbol{y}) + \nabla p = \rho \boldsymbol{u}, \tag{1.2a}$$

$$\rho_t + [\boldsymbol{y} \cdot \nabla]\rho = 0, \tag{1.2b}$$

$$\operatorname{div} \boldsymbol{y} = 0, \tag{1.2c}$$

together with $\rho(0,.) = \rho_0$, $\boldsymbol{y}(0,.) = \boldsymbol{y}_0$, and $\boldsymbol{y} = \boldsymbol{0}$ on $(0,T] \times \partial\Omega$.

This model tracks given profiles $\tilde{\rho} \in L^2(\Omega_T)$ while simultaneously minimizing the interface length; see Figure 9.12 for the regularizing role of the perimeter functional in the optimal control problem.

In Problem 1.1, $S_\rho \subseteq \Omega$ denotes the jump set of the function ρ, and $\mathcal{H}^1$ the one-dimensional Hausdorff measure, see, e.g., [5]. The variable $\boldsymbol{y}$ denotes the velocity, and ρ the density of the fluid, with $\mu > 0$ its local viscosity. A typical situation for the initial density is $\rho_0 = \rho_1 \chi_{\Omega_1} + \rho_2 \chi_{\Omega_2}$ with $\Omega_1 \cap \Omega_2 = \varnothing$ and $0 < \rho_1 \leq \rho_2 < \infty$.

Minimizing the functional G without additional constraints has been studied in former works [5, 22], where the perimeter functional is approximated by the following regularization (cf. [22, Chapter 4]),

$$F_\delta(\rho) := \int_0^T \int_\Omega \left\{ \delta |\nabla \rho|^2 + \frac{1}{4\delta} W(\rho) \right\} \mathrm{d}\boldsymbol{x} \, \mathrm{d}t, \tag{1.3}$$

where $\delta > 0$ and $W(\rho) = (\rho - \rho_1)^2(\rho - \rho_2)^2$. It is well known (for the time independent case) that F_δ Γ-converges on $L^2(\Omega)$ to the perimeter functional $\rho \mapsto \mathcal{H}^1(S_\rho)$, which takes finite values if ρ is in $GSBV(\Omega)$; see [5, 22]. An important consequence of Γ-convergence is that every sequence of minima of the approximative problem converges to a minimum of the original problem.

The authors are not aware of any work in the literature where Γ-convergence of (sequences of) problems with analytical constraints like differential equations has been studied. A naive way to deal with additional constraints is to set the corresponding functional equal to infinity at points where the constraint does not hold. It is easy to construct examples (even for functionals $f_n, f : \mathbb{R} \to \mathbb{R}$ with $f_n \xrightarrow{\Gamma} f$), such that a corresponding Γ-convergence result will not be inherited in the presence of such additional constraints; see Lemma A.2 The problem here is that by the well-posedness of the equation, the modified functional would have to be changed in "too many points" to infinity.

In order to show existence and derive optimality conditions for a regularized version of Problem 1.1, we have to overcome several difficulties.

1. Since $\rho \in L^\infty(\Omega_T)$ in (1.2b), and not in $SBV(\Omega)$ for almost all times in general, neither is the jump set S_ρ well-defined nor the mapping $\rho \mapsto S_\rho$ weakly lower semicontinuous for almost all times. Hence, it is not clear how to construct solutions for Problem 1.1. Moreover, it is not obvious how to derive optimality conditions due to the lack of differentiability of the perimeter functional. We note that in [65] it is shown that a solution of (1.2b) is in $BV(\Omega)$ for almost every time, provided initial data contained in $BV(\Omega)$ as long as $\Omega \subset \mathbb{R}$. Even though this does not imply that ρ is in $SBV(\Omega)$, it motivates hope that this can be shown in future—maybe under some additional assumptions.
2. The derivation of optimality conditions via the Lagrange multiplier theorem is non-trivial due to the limited regularity of the density in (1.2b), and it is not clear if the minimum is a regular point; as a consequence, the associated Lagrange multiplier lacks regularity properties; see, e.g., [47, Chapter 1] or [56, Chapter 9]. In particular, the Lagrange multiplier to (1.2b) would not be a function that is defined on Ω_T.

In order to handle the first problematic issue, we use the perimeter approximation (1.3) and replace the objective function G in (1.1) by

$$J_\delta(\rho, \boldsymbol{u}) := \frac{\lambda}{2}\int_0^T \int_\Omega |\rho - \tilde{\rho}|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \frac{\beta}{2}\int_0^T \int_\Omega \left\{ \delta |\nabla \rho|^2 + \frac{1}{4\delta} W(\rho) \right\} \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \frac{\alpha}{2}\int_0^T \int_\Omega |\boldsymbol{u}|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t,$$

where $\delta > 0$. This phase-field approximation is done in order to construct a well-defined weakly lower semicontinuous functional $J_\delta : L^2(H^1) \times L^2(\boldsymbol{L}^2) \to \mathbb{R}$.

To handle the second problematic issue, we regularize (1.2b) by adding artificial diffusion at a scale $\varepsilon > 0$. This modification improves regularity properties of both, the density ρ and the related Lagrange multipliers, and allows us to construct solutions of the modified Problem 4.1. The idea of adding artificial diffusion is used to solve (1.2) (cf. [54]) and was also used in [52] to derive optimality conditions for the density dependent Stokes equations.

It can then be shown that solutions of the modified equation (3.2) converge to those of (1.2). We also refer the reader to a corresponding discussion based on computational studies in chapter 9.

Theorem 4.2 asserts solvability of the regularized optimality problem for finite $\varepsilon, \delta > 0$. An open question remains as to how to choose pairings ε and δ in such a way that the functional is bounded uniformly with respect to ε and δ. A heuristic choice which is due to standard parabolic a priori estimates is to set $\delta = \mathcal{O}(\varepsilon)$ in order to have at least $\mathcal{H}^1(S_\rho) < \infty$ for a limiting density ρ $(\varepsilon, \delta \to 0)$. This scaling is supported by computational evidence reported in chapter 9: Choosing $\varepsilon \gg \delta$ causes highly diffuse interfaces (see Figure 9.3), as opposed to "parasitic velocities" in the opposited scenario where $\varepsilon \ll \delta$ (see Figure 9.2).

There are other multiphase fluid flow models which include surface tension terms, and thus avoid highly oscillatory behavior of the surface on physical grounds; see, e.g., [1]. In those cases, accordingly, regular interfaces occur due to combined effects of surface tension and the perimeter functional. Hence, we choose the present setup of the optimization problem to decouple effects of the equation from those of the functional by addressing fluids with negligible surface tension in this work. Moreover, this identification of effects allows for future works with extended functionals of, e.g., Willmore energy type, which penalizes areas with large mean curvature on the surface.

However, the literature on optimal control problems subject to the density dependent Navier–Stokes equations is rare: A main difficulty in Problem 1.1 is the strong coupling between the mass equation and the momentum equation, which leads to a strong coupling in the adjoint equations; another problem comes from the lack of regularity of the solution of the mass equation. We mention the work of Kunisch and Lu [52], where an optimal control problem with an L^2 tracking-type functional subject to the regularized density dependent Stokes equation in $\mathbb{R}^2$ is studied, and optimality conditions are derived. Moreover, by some assumptions on a nonregularized solution, it is shown that minima $(\varepsilon > 0)$ converge to a minimum of the limiting problem for $\varepsilon = 0$. We note that a corresponding result seems unclear in the present setting, where $\delta, \varepsilon > 0$; furthermore, as already discussed above, solvability of the limiting problem, Problem 1.1, has to remain an open problem.

The construction of convergent numerical discretizations of (1.2) is a very recent subject. The first work which accomplished this goal is [55], where a discontinuous Galerkin scheme is studied for (1.2), which uses piecewise constant functions for the pressure, in particular. In view of optimal control, corresponding (discrete) optimality conditions couple primal and dual variables, which requires bounding primal variables in stronger norms to show stability of the overall scheme. For this reason, we consider instead the continuous Galerkin scheme [12], where continuous functions for the pressure space are admitted. The discretization for (1.2) in [12] introduces numerical stabilization terms in order to conclude convergence against a weak solution, which fits into the discussion above of a regularized version of Problem 1.1 as well. In the present case, since an artificial diffusion term is introduced in (1.2a), we do not need most of the regularization terms as suggested in [12] in our scheme; see (5.5). An alternative strategy for discretizing (1.2) is given in [38], where (1.2) is solved numerically using artificial diffusion as well; however, the convergence analysis uses higher order finite elements for the velocity space, which is more expensive on a computational level than the schemes in [12] and [55].

In order to construct necessary optimality conditions by a fully practical discrete scheme for the regularized optimization problem, we use the "first discretize, then optimize" ansatz: We propose a corresponding implementable discrete optimization problem, which involves the discretized equations as discussed above in [12], derive related discrete optimality conditions, and show convergence of the iterates to a solution of the continuous optimality conditions. This method benefits from the available stable, convergent finite element based fully practical discretization of the state equation, and leads to a construction of solutions of the continuous optimality system (4.1). The solvability of discrete optimality conditions follows directly via the Lagrange multiplier theorem. This approach to set up discrete optimality conditions has the advantage of being a natural and structure preserving discretization of the adjoint equation. The main challenging part is the strong coupling of both primal variables ρ and $\boldsymbol{y}$, their strong coupling with the two adjoint variables, and the coupling between both adjoint variables themselves. In order to address this issue, we first have to derive strong stability properties for the discrete primal variables. The second step is to derive standard parabolic regularity properties for the discrete adjoint variables. Here, we need the regularity of the primal variables and a combined argument: Since derivatives of both adjoint variables are present in both adjoint equations (6.2a) and (6.2c), we have to multiply the adjoint equations simultaneously with different test functions and to consider a proper weighted sum of the resulting inequalities.

From a practical point of view, other ways of controls include a finite-dimensional control space, where amplitudes of given forces are unknown, or boundary control. However, the distributed control provided here gives hints to regions in which region an optimal control should act. This is relevant in certain engineering applications, such as magnetohydrodynamics, or the control of ferrofluids; cf. [61]. We note that all proofs can also be performed directly in the same manner (even more easy in some points) for a finite-dimensional control space, where only amplitudes of given forces are unknown. It is likely that under some assumptions even a boundary control could be possible.

This part is organized as follows. In chapter 3 we study the regularized equation (3.2) and prove regularity results; see Theorem 3.2. In chapter 4, Problem 1.1 is restated in a proper form: We assert solvability in Theorem 4.2 and derive first order necessary optimality conditions (4.1). In chapter 5, we describe the numerical setup, state the numerical scheme for the primal equation (5.5), and show solvability and standard parabolic bounds of the discrete density and velocity in Lemma 5.1. Moreover, we prove boundedness of the discrete density and velocity in stronger norms in Lemmas 5.2 and 5.3. In chapter 6, we study the discrete optimization problem and derive discrete optimality conditions (6.2). In chapter 7, we derive bounds for the discrete adjoint variables in standard parabolic norms. The proof has to cope with the subtle coupling between both dual variables, i.e., the Lagrange multipliers related to (1.2b) and (1.2a); We also motivate why we needed that strong bounds on the primal variables. Finally, in chapter 8 we show the main result of this part in Theorem 8.3: For numerical parameters $h, k \to 0$, a subsequence of solutions of the discrete optimality conditions (5.5)–(6.2) converges to a solution of the continuous optimality conditions (4.1) for fixed $\delta, \varepsilon > 0$. Moreover, the discrete optimal control function $\{\boldsymbol{U}^n\}$ converges to the continuous optimal control function $\boldsymbol{u}$ strongly in $L^2(\boldsymbol{L}^2)$, which will be proven in Theorem 8.4. Lastly, we present several numerical experiments in chapter 9. Here, we propose a variable step-size gradient type algorithm for the solution of the discrete problem and study the relative effect of the phase-field formulation of the functional, and the stabilization of the PDE constraint. We also demonstrate qualitatively different behaviors of the fluids for $\beta = 0$ and $\beta > 0$ (see Figure 9.12) to show evidence of the regularizing effect of the perimeter functional onto the initial interface. Those experiments are motivated from corresponding behaviors of solutions of the L^2-gradient flow of the perimeter functional; cf. [7].

2. Preliminaries

2.1. General notation

Let $W^{k,p}$ and $H^k := W^{k,2}$ denote standard Sobolev spaces. By

$$W^{k,p}(W^{m,q}) := W^{k,p}(0,T;W^{m,q})$$

we refer the reader to standard Bochner spaces. The space $\mathcal{C}(X)$ denotes the space of continuous functions taking values in X. Vector-valued functions and spaces containing such functions are written in bold-face notation. We define

$$L_0^2(\Omega) := \left\{ u \in L^2(\Omega) : \int_\Omega u \,\mathrm{d}\boldsymbol{x} = 0 \right\}.$$

The space $\boldsymbol{V}$ (respectively, $\boldsymbol{H}$) denotes the closure of $\{\boldsymbol{v} \in \boldsymbol{\mathcal{C}}_0^\infty(\Omega) : \operatorname{div} \boldsymbol{v} = 0\}$ in the $\boldsymbol{H}^1$-norm (respectively, $\boldsymbol{L}^2$-norm). For the scalar products in L^2 and $L^2(L^2)$, respectively, of f and g, we write (f,g) in cases where no confusion arises; otherwise, we add the corresponding space as index to the scalar product. The notation $\|.\|$ stands for the L^2- or the $L^2(L^2)$-norm, which will be clear from the context.

The dual pairing of X and its dual space X^* is written as $\langle .,. \rangle_{X,X^*}$. The space of linear functionals from X to Y is denoted by $L(X,Y)$.

We use C as a generic nonnegative constant; to indicate dependencies, we write $C(.)$.

2.2. Known results

We recall well-known results for the state equation (1.2). For details, we refer the reader to [54].

Theorem 2.1
Let $0 < T < \infty$. Assume $\boldsymbol{u} \in L^2(\boldsymbol{L}^2)$, $\Omega \subset \mathbb{R}^d$ $(d = 2,3)$ bounded, open, and assume as well as $\boldsymbol{y}_0 \in \boldsymbol{H}$ and $\rho_0 \in L^\infty(\Omega)$. Then there exists a global weak solution $(\boldsymbol{y}, \rho) \in L^2(\boldsymbol{V}) \times L^\infty(L^\infty)$ of (1.2). Every global weak solution $(\boldsymbol{y}, \rho)$ has the following property: For every $0 \le \alpha \le \beta < \infty$ the measure of $\{\boldsymbol{x} \in \Omega : \alpha \le \rho(\boldsymbol{x}, t) \le \beta\}$ is independent of $t \ge 0$. In particular, we have for almost all $(t, \boldsymbol{x}) \in \Omega_T$

$$0 < \inf_\Omega \rho_0 \le \rho(t, \boldsymbol{x}) \le \|\rho_0\|_{L^\infty(\Omega)}. \tag{2.1}$$

In addition, the following estimate holds:

$$\|\boldsymbol{y}\|_{L^\infty(\boldsymbol{H})} + \|\boldsymbol{y}\|_{L^2(\boldsymbol{V})} \le C\Big(\Omega, T, \|\rho_0\|_{L^\infty}, \|\boldsymbol{u}\|_{L^2(\boldsymbol{L}^2)}, \|\boldsymbol{u}_0\|_{\boldsymbol{H}}\Big).$$

3. The regularized state equation

The following hypotheses are valid for the rest of this part. In particular, smoother initial data are required for the following analysis.

Hypothesis 3.1
We assume that

1. $0 < T < \infty$.
2. $\Omega \subset \mathbb{R}^2$ is bounded and open, with $\partial\Omega \in \mathcal{C}^{1,1}$, or Ω is polyhedral and convex.
3. $\rho_0 \in H^2(\Omega)$ with $0 < \rho_1 \leq \rho_0(.) \leq \rho_2 < \infty$.
4. $\boldsymbol{y}_0 \in \boldsymbol{V}$.
5. $\mu(\rho) = \mu > 0$ is constant.

We now want to regularize (1.2). Before doing so, we write the equation in a different way be using the identity (as long as $\rho_t + [\boldsymbol{y} \cdot \nabla]\rho = 0$; cf. [55])

$$\rho\Big(\boldsymbol{y}_t + [\boldsymbol{y} \cdot \nabla]\boldsymbol{y}\Big) = \frac{1}{2}\Big(\rho\boldsymbol{y}_t + (\rho\boldsymbol{y})_t + \rho[\boldsymbol{y} \cdot \nabla]\boldsymbol{y} + \operatorname{div}(\rho\boldsymbol{y} \otimes \boldsymbol{y})\Big). \tag{3.1}$$

This modification is used in order to prove solvability for the numerical scheme (5.5); cf. Lemma 5.1. We note that this modification has no effect in the continuous setting and is only used here to make the continuous optimality system (4.1) and the discrete one (6.2) comparable.

After applying (3.1), we regularize the state equation (1.2) to improve regularity properties of the density $\rho : \Omega_T \to \mathbb{R}$, and of the corresponding Lagrange multiplier $\eta : \Omega_T \to \mathbb{R}$ of the mass equation (3.2b) below. The system then reads (for $\varepsilon > 0$)

$$\frac{1}{2}\rho\boldsymbol{y}_t + \frac{1}{2}(\rho\boldsymbol{y})_t + \frac{1}{2}\rho[\boldsymbol{y} \cdot \nabla]\boldsymbol{y} + \frac{1}{2}\operatorname{div}(\rho\boldsymbol{y} \otimes \boldsymbol{y}) - \operatorname{div}(\mu(\rho)\nabla\boldsymbol{y}) + \nabla p = \rho\boldsymbol{u}, \tag{3.2a}$$

$$\rho_t + [\boldsymbol{y} \cdot \nabla]\rho - \varepsilon\Delta\rho = 0, \tag{3.2b}$$

$$\operatorname{div}\boldsymbol{y} = 0, \tag{3.2c}$$

together with boundary conditions $\boldsymbol{y} = \boldsymbol{0}$ and $\partial_{\boldsymbol{n}}\rho = 0$ on $(0, T] \times \partial\Omega$, as well as the initial conditions $\boldsymbol{y}(0, .) = \boldsymbol{y}_0$, and $\rho(0, .) = \rho_0$.

Theorem 3.2
Let $\varepsilon > 0$ and $\boldsymbol{u} \in L^2(\boldsymbol{L}^2)$. Then there exists a global weak solution of (3.2) such that

$$\boldsymbol{y} \in \boldsymbol{Y} := L^2(\boldsymbol{H}^2) \cap H^1(\boldsymbol{H}) \subset \mathcal{C}(\boldsymbol{V}),$$
$$\rho \in R := H^1(H^1) \cap L^\infty(H^2) \subset \mathcal{C}(\Omega_T).$$

In particular, there exists a constant $C = C(\varepsilon, T, \boldsymbol{u}, \boldsymbol{y}_0, \rho_0) \geq 0$ such that

$$\|\boldsymbol{y}\|_{\boldsymbol{Y}} + \|\rho\|_R \leq C. \tag{3.3}$$

Bounds on the solution, which are uniform in $\varepsilon > 0$ may be obtained in the following norms,

$$\boldsymbol{y} \in L^2(\boldsymbol{V}) \cap L^\infty(\boldsymbol{H}) \cap H^1(\boldsymbol{V}^*), \qquad \rho \in L^\infty(L^\infty) \cap H^1(H^{-1}). \tag{3.4}$$

In particular, we have $\rho_1 \leq \rho \leq \rho_2$ a.e. in Ω_T.

Proof
The existence of a weak solution (ρ, u) follows by Schauder's fixed point theorem and standard parabolic theory, similarly to [54].

The improved regularity of $\boldsymbol{y}$ follows from [54, pp. 32ff], while the regularity of ρ follows from formally testing (3.2b) with $-\Delta\rho$ and testing the time derivative of (3.2b) with $-\Delta\rho$, respectively. The estimates on $\boldsymbol{y}$ and ρ in (3.3) are both based on Hypothesis 3.1.

To get uniform bounds with respect to $\varepsilon > 0$ in the norms which are indicated in (3.4), we test (3.2a) with $\boldsymbol{y}$ and (3.2b) with ρ, and neglect (nonnegative) ε-terms. Finally, the uniform lower and upper pointwise bounds for ρ follow from the maximum principle for parabolic equations, together with the bounds of ρ_0; cf. Hypothesis 3.1. □

A consequence of Theorem 3.2 is that $p \in L^2(H^1 \cap L^2_0)$.

4. Optimal Control of the regularized system

Problem 4.1
Let $0 < T < \infty$, and $\varepsilon, \delta > 0$. Minimize J_δ subject to (3.2).

4.1. Existence

Theorem 4.2
There exists at least one solution $(\bar{\boldsymbol{y}}, \bar{\rho}, \bar{\boldsymbol{u}}) \in \boldsymbol{Y} \times R \times L^2(\boldsymbol{L}^2)$ of Problem 4.1.

Proof
For every $\boldsymbol{u} \in L^2(\boldsymbol{L}^2)$, Theorem 3.2 ensures the existence of a solution to (3.2); hence the set of feasible points is not empty, and there exists $\bar{J} := \inf J_\delta(\boldsymbol{y}, \rho, \boldsymbol{u}) \geq 0$, where the infimum is taken over all feasible $(\boldsymbol{y}, \rho, \boldsymbol{u}) \in \boldsymbol{Y} \times R \times L^2(\boldsymbol{L}^2)$. Thus, we may consider a minimizing sequence $\{(\boldsymbol{y}_n, \rho_n, \boldsymbol{u}_n)\} \in \boldsymbol{Y} \times R \times L^2(\boldsymbol{L}^2)$, such that for $n \to \infty$

$$J_\delta(\boldsymbol{y}_n, \rho_n, \boldsymbol{u}_n) \searrow \bar{J}.$$

By the definition of the cost functional J_δ, the sequence $\{\boldsymbol{u}_n\}$ is bounded in $L^2(\boldsymbol{L}^2)$ and—thanks to Theorem 3.2—the sequences $\{\boldsymbol{y}_n\}$ and $\{\rho_n\}$ are bounded in $\boldsymbol{Y}$ and R respectively. Then there exist $\boldsymbol{y}^* \in \boldsymbol{Y}$, $\rho^* \in R$, and $\boldsymbol{u}^* \in L^2(\boldsymbol{L}^2)$ such that for corresponding subsequences (not relabeled) and $n \to \infty$,

$$\begin{aligned} \boldsymbol{y}_n &\rightharpoonup \boldsymbol{y}^* && \text{weakly in } \boldsymbol{Y}, \\ \rho_n &\rightharpoonup \rho^* && \text{weakly in } R, \\ \boldsymbol{u}_n &\rightharpoonup \boldsymbol{u}^* && \text{weakly in } L^2(\boldsymbol{L}^2). \end{aligned}$$

We have to show that $(\boldsymbol{y}^*, \rho^*, \boldsymbol{u}^*)$ is a solution of (3.2), and $J_\delta(\boldsymbol{y}^*, \rho^*, \boldsymbol{u}^*) = \bar{J}$.

1. Following the argumentation in [54, Section 2.4], we can pass to the limit in the momentum equation and in the mass equation (except for the much easier term $\varepsilon\Delta\rho^n$) because of the bounds we have deduced in Theorem 3.2. By the definition of R, the distributional limit of the term $-\varepsilon\Delta(\rho^n)$ is $-\varepsilon\Delta\rho^*$.

2. All terms of J_δ are continuous and convex (and therefore weakly lower semicontinuous), except for $\int_{\Omega_T} W(\rho)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$. Since $\rho_n \rightharpoonup \rho$ in R, we conclude by Sobolev embeddings that $\rho_n \to \rho$ in $L^4(L^4)$, in particular (up to a subsequence) $\rho_n \to \rho$ a.e. in Ω_T. Fatou's lemma then guarantees weakly lower semi-continuity. Hence, we have $J_\delta(\boldsymbol{y}^*, \rho^*, \boldsymbol{u}^*) = \bar{J}$. □

4.2. Optimality conditions

Next, we show that the Frechet derivative of the side constraints (3.2) is surjective on appropriate spaces, and then we derive optimality conditions. These will be compared in chapter 8 with the discrete ones from chapter 6.

For the next theorem, we use the mapping $\boldsymbol{e} : (\boldsymbol{Y} \times L^2(H^1 \cap L_0^2) \times R \times L^2(\boldsymbol{L}^2)) \to (L^2(\boldsymbol{L}^2) \times L^2(L^2) \times L^2(L^2) \times \boldsymbol{V} \times H^2(\Omega))$, which is defined by

$$\boldsymbol{e}(\boldsymbol{y}, p, \rho, \boldsymbol{u}) := \begin{pmatrix} e_1(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \\ e_2(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \\ e_3(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \\ a_1(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \\ a_2(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2}\rho\boldsymbol{y}_t + \frac{1}{2}(\rho\boldsymbol{y})_t + \frac{1}{2}\rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{y} + \frac{1}{2}\operatorname{div}(\rho\boldsymbol{y}\otimes\boldsymbol{y}) - \mu\Delta\boldsymbol{y} - \rho\boldsymbol{u} + \nabla p \\ \rho_t + [\boldsymbol{y}\cdot\nabla]\rho + \frac{1}{2}\rho\operatorname{div}\boldsymbol{y} - \varepsilon\Delta\rho \\ \operatorname{div}\boldsymbol{y} \\ \boldsymbol{y}(0,.) - \boldsymbol{y}_0 \\ \rho(0,.) - \rho_0 \end{pmatrix}.$$

We omit boundary conditions in $\boldsymbol{e}$, which may be treated by standard methods; see, e.g., [40, Section 2.6]. Initial conditions are treated in the same manner as there. Note that we have added $\frac{1}{2}\rho\operatorname{div}\boldsymbol{y} = 0$ to the mass equation e_2. Using this term is a standard way to treat the incompressible condition in the context of a finite element approximation, and it stabilizes the discrete operator in chapter 5. Adding this term here allows us to compare continuous and discrete effects in later chapters.

Theorem 4.3
Let $0 < T < \infty$ and $\varepsilon, \delta > 0$. The mapping $\boldsymbol{e}$ is well-defined and Frechet differentiable. Moreover, for each $(\boldsymbol{y}, p, \rho, \boldsymbol{u}) \in \boldsymbol{Y} \times L^2(H^1 \cap L_0^2) \times R \times L^2(\boldsymbol{L}^2)$, the derivative $\boldsymbol{e}'(\boldsymbol{y}, p, \rho, \boldsymbol{u})$ is surjective.

Proof

1. With the estimates in Theorem 3.2, we see that the mapping $\boldsymbol{e}$ is well-defined.

2. The candidate for the derivative of $\boldsymbol{e}$ is

$$\begin{aligned}\left\langle e_1'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle &= \rho(\boldsymbol{\delta y})_t + \frac{1}{2}\rho_t\boldsymbol{\delta y} + \frac{1}{2}\rho[\boldsymbol{\delta y}\cdot\nabla]\boldsymbol{y} + \frac{1}{2}\rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{\delta y} \\ &\quad + \frac{1}{2}\operatorname{div}(\rho\boldsymbol{\delta y}\otimes\boldsymbol{y}) + \frac{1}{2}\operatorname{div}(\rho\boldsymbol{y}\otimes\boldsymbol{\delta y}) - \mu\Delta\boldsymbol{\delta y} \\ &\quad + \delta\rho\boldsymbol{y}_t + \frac{1}{2}(\delta\rho)_t\boldsymbol{y} + \frac{1}{2}\delta\rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{y} - \frac{1}{2}\delta\rho\boldsymbol{u} \\ &\quad + \frac{1}{2}\operatorname{div}(\delta\rho\boldsymbol{y}\otimes\boldsymbol{y}) - \rho\boldsymbol{\delta u} + \nabla\delta p, \\ \left\langle e_2'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle &= (\delta\rho)_t + [\boldsymbol{y}\cdot\nabla]\delta\rho + \frac{1}{2}\delta\rho\operatorname{div}\boldsymbol{y} - \varepsilon\Delta\delta\rho\end{aligned}$$

$$+ [\boldsymbol{\delta y} \cdot \nabla]\rho + \frac{1}{2}\rho \operatorname{div} \boldsymbol{\delta y},$$

$$\left\langle e_3'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle = \operatorname{div} \boldsymbol{\delta y},$$

$$\left\langle a_1'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle = \boldsymbol{\delta y}(0, .),$$

$$\left\langle a_2'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle = \delta\rho(0, .),$$

which is obtained by direct calculation as in [40, Section 2.6].

3. Let $(\boldsymbol{f}, g, h, \boldsymbol{\varphi}, \psi) \in L^2(\boldsymbol{L}^2)\times L^2(L^2)\times L^2(L^2)\times \boldsymbol{V}\times H^2(\Omega)$ and $(\boldsymbol{\varphi}, \psi) \in \boldsymbol{L}^2(\Omega)\times L^2(\Omega)$. The existence of solutions $(\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u}) \in \boldsymbol{Y} \times L^2(H^1) \times R \times L^2(\boldsymbol{L}^2)$ for

$$\left\langle \boldsymbol{e}'(\boldsymbol{y}, p, \rho, \boldsymbol{u}), (\boldsymbol{\delta y}, \delta p, \delta\rho, \boldsymbol{\delta u})\right\rangle = (\boldsymbol{f}, g, h),$$
$$\boldsymbol{\delta y}(0, .) = \boldsymbol{\varphi},$$
$$\delta\rho(0, .) = \psi,$$

together with suitable boundary conditions, follows from standard linear parabolic theory; cf. [53] and Theorem 3.2. □

Theorem 4.3 allows us to apply the Lagrange multiplier theorem (cf. [56, Section 9.3]), and thus to deduce necessary optimality conditions for Problem 6.1 below. We define the Lagrange functional $\mathcal{L} : \boldsymbol{Y} \times L^2(H^1 \cap L^2_0) \times R \times L^2(\boldsymbol{L}^2) \times L^2(\boldsymbol{H}) \times L^2(L^2) \times L^2(L^2) \to \mathbb{R}$ via

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}, p, \rho, \boldsymbol{u}; \boldsymbol{z}, q, \eta) := J(\rho, \boldsymbol{u}) &+ \left\langle \eta, \rho_t + [\boldsymbol{y} \cdot \nabla]\rho + \frac{1}{2}\rho \operatorname{div} \boldsymbol{y} - \varepsilon\Delta\rho\right\rangle_{L^2(L^2),L^2(L^2)}\\
&+ \left\langle \boldsymbol{z}, \frac{1}{2}\rho\boldsymbol{y}_t + \frac{1}{2}(\rho\boldsymbol{y})_t + \frac{1}{2}\rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{y} - \mu\Delta\boldsymbol{y} + \nabla p - \rho\boldsymbol{u}\right\rangle_{L^2(\boldsymbol{L}^2),L^2(\boldsymbol{L}^2)}\\
&- \frac{1}{2}\left\langle \rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{z}, \boldsymbol{y}\right\rangle_{L^2(\boldsymbol{H}^{-1}),L^2(\boldsymbol{H}^1_0)} + \langle q, \operatorname{div} \boldsymbol{y}\rangle_{L^2(L^2),L^2(L^2)}.\end{aligned}$$

By using the directional derivatives of $\boldsymbol{e}$ from the proof of Theorem 4.3, together with integration by parts, and setting the derivatives of $\mathcal{L}$ equal to zero, then a straightforward calculation, together with methods from [40, Section 2.6] leads to the following optimality conditions:

$$\begin{aligned}\boldsymbol{0} = &\frac{1}{2}\eta\nabla\rho - \frac{1}{2}\rho\nabla\eta - \frac{1}{2}\rho_t\boldsymbol{z} - \rho\boldsymbol{z}_t + \frac{1}{2}\rho\nabla\boldsymbol{y}\boldsymbol{z} - \frac{1}{2}[\nabla\rho\cdot\boldsymbol{y}]\boldsymbol{z}\\
&- \rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{z} - \frac{1}{2}\rho\nabla\boldsymbol{z}\boldsymbol{y} - \mu\Delta\boldsymbol{z} - \nabla q,\end{aligned} \tag{4.1a}$$

$$0 = \operatorname{div} \boldsymbol{z}, \tag{4.1b}$$

$$\begin{aligned}0 = &\lambda(\rho - \tilde{\rho}) - \beta\delta\Delta\rho + \frac{\beta}{8\delta}W'(\rho) - \eta_t - [\boldsymbol{y}\cdot\nabla]\eta - \varepsilon\Delta\eta\\
&+ \frac{1}{2}\boldsymbol{z}\cdot\boldsymbol{y}_t - \frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{z}_t + \frac{1}{2}[\boldsymbol{y}\cdot\nabla]\boldsymbol{y}\cdot\boldsymbol{z} - \boldsymbol{u}\cdot\boldsymbol{z} - \frac{1}{2}[\boldsymbol{y}\cdot\nabla]\boldsymbol{z}\cdot\boldsymbol{y},\end{aligned} \tag{4.1c}$$

$$\boldsymbol{0} = \alpha\boldsymbol{u} - \rho\boldsymbol{z}, \tag{4.1d}$$

$$\boldsymbol{0} = \rho\boldsymbol{y}_t - \rho[\boldsymbol{y}\cdot\nabla]\boldsymbol{y} - \mu\Delta\boldsymbol{y} - \rho\boldsymbol{u} + \nabla p, \tag{4.1e}$$

$$0 = \rho_t + [\boldsymbol{y} \cdot \nabla]\rho - \varepsilon\Delta\rho, \tag{4.1f}$$
$$0 = \operatorname{div} \boldsymbol{y}, \tag{4.1g}$$

together with the initial conditions

$$\boldsymbol{y}(0,.) = \boldsymbol{y}_0, \quad \rho(0,.) = \rho_0, \quad \boldsymbol{z}(T,.) = \mathbf{0}, \quad \eta(T,.) = 0$$

and the homogeneous boundary conditions

$$\boldsymbol{y} = 0, \quad \partial_{\boldsymbol{n}}\rho = 0, \quad \boldsymbol{z} = \mathbf{0}, \quad \partial_{\boldsymbol{n}}\eta = 0 \qquad \text{on } (0,T] \times \partial\Omega.$$

The Lagrange multiplier theorem assures that this system has at least one solution.

The derivation of initial and boundary conditions is done by a standard argument; cf. [40, Section 2.6].

Equations (4.1a) and (4.1c) are the adjoint equations, and equation (4.1d) is the optimality condition, while equations (4.1e) and (4.1f) are are identical to the state equations (3.2a) and (3.2b).

5. Discretization of the state equation

We now consider a discrete version of (3.2), a modification of which is studied in [12].

5.1. Numerical setup and notation

Let $\mathcal{T}_h$ be a quasi-uniform triangulation of Ω with $h := \max_{T\in\mathcal{T}_h} \operatorname{diam} T$ and

$$R_h := \left\{ X_h \in \mathcal{C}(\bar{\Omega}) : \ X_h\big|_T \in P_\ell(T) \quad \forall T \in \mathcal{T}_h \right\}.$$

We assume that the triangulation is strongly acute; see, e.g., [51]. For the finite element approximation, we define the following spaces.

- R_h for the approximation of the density ρ;
- $\boldsymbol{V}_h$ and M_h as an inf-sup stable conforming pair (e.g. Taylor–Hood or MINI elements) for velocity $\boldsymbol{y}$ and pressure p, involving zero Dirichlet boundary conditions for $\boldsymbol{y}$;
- the space of discrete divergence-free functions

$$\boldsymbol{J}_h := \{\boldsymbol{v}_h \in \boldsymbol{V}_h : \ (\operatorname{div} \boldsymbol{v}_h, \chi_h) = 0 \quad \text{for all } \chi_h \in M_h\}.$$

Recall the discrete Laplace operator $\Delta_h : R_h \to R_h$, where

$$-(\Delta_h V, \Phi) = (\nabla V, \nabla \Phi) \quad \forall V, \Phi \in R_h.$$

Analogously, we define the vector-valued discrete Laplacian for the space $\boldsymbol{V}_h$ by $\tilde{\Delta}_h : \boldsymbol{V}_h \to \boldsymbol{V}_h$. The discrete Stokes operator $\boldsymbol{A}_h$ is defined by $\boldsymbol{A}_h := -\boldsymbol{P}_h\tilde{\Delta}_h$, where $\boldsymbol{P}_h : \boldsymbol{L}^2 \to \boldsymbol{J}_h$ denotes the $\boldsymbol{L}^2$-projection. For details, we refer the reader to [42, Section 4]. The subset consisting of finite element functions $V \in R_h$ such that $\|\Delta_h V\|_{L^2} \le C < \infty$ with $C > 0$ independent of h will be denoted by $H^2_{\text{disc}} \subseteq R_h$. The subset $\boldsymbol{H}^2_{\text{disc}} \subseteq \boldsymbol{V}_h$ is defined in the same way.

We will often use the following discrete embedding and Gagliardo–Nirenberg type inequalities (cf. [42, Lemma 4.4.]):

$$\|\nabla V\|_{L^4} \le C(\|\Delta_h V\| + \|\nabla V\|), \tag{5.1}$$

$$\|\nabla V\|_{L^4} \le C\|\nabla V\|^{\frac{1}{2}}(\|\Delta_h V\| + \|\nabla V\|)^{\frac{1}{2}}. \tag{5.2}$$

$$\|\nabla \boldsymbol{V}\|_{\boldsymbol{L}^4} \le C\|\tilde{\Delta}_h \boldsymbol{V}\|, \tag{5.3}$$

$$\|\nabla \boldsymbol{V}\|_{\boldsymbol{L}^4} \le C\|\nabla \boldsymbol{V}\|^{\frac{1}{2}}\|\tilde{\Delta}_h \boldsymbol{V}\|^{\frac{1}{2}}. \tag{5.4}$$

Let $t_n := nk$ (for $n = 0, \dots, N$), for $k = \frac{T}{N}$. Let $(\boldsymbol{Y}^0, R^0) \in \boldsymbol{V}_h \times R_h$ be the projection of $(\boldsymbol{y}_0, \rho_0)$, with $\rho_1 \leq R^0 \leq \rho_2$.

We will use the following notation for discrete functions: The notation $\{V^n\} \subseteq X_h$ describes a family of finite element functions (in a finite element space X_h) evaluated at subsequent times t_n, while $\mathcal{V} : \Omega_T \to \mathbb{R}$ stands for the piecewise affine, globally continuous time interpolant of $\{V^n\}$. Moreover, we define the following piecewise constant in time interpolants of $\{V^n\}$ for $t \in [t_j, t_{j+1})$,

$$V^+(t) := V^{j+1}, \qquad V^\bullet(t) := V^j, \qquad V^-(t) := V^{j-1}.$$

For vector-valued functions $\boldsymbol{v} : \Omega_T \to \mathbb{R}^n$, we shall write all quantities in boldface, i.e., $\boldsymbol{V}^n$ for the discrete iterates and $\boldsymbol{\mathcal{V}} : \Omega_T \to \mathbb{R}^n$ for its time interpolant. For the variables ρ and η, we use the capital letters $\mathcal{R}$ and $\mathcal{E}$ for the time interpolant, while there should be no confusion with the space R from Theorem 3.2. The discrete time derivative of the function $\mathcal{V}$ will be denoted as

$$d_t V^n := \frac{V^n - V^{n-1}}{k}.$$

The discrete version of (3.2) reads as follows: For $1 \leq n \leq N$ find $(\boldsymbol{Y}^n, P^n, R^n) \in \boldsymbol{V}_h \times M_h \times R_h$ such that for all $(\boldsymbol{Z}, \Pi, E) \in \boldsymbol{V}_h \times M_h \times R_h$

$$0 = (d_t R^n, E) + \varepsilon(\nabla R^n, \nabla E) + ([\boldsymbol{Y}^n \cdot \nabla] R^n, E) + \frac{1}{2}(R^n \operatorname{div} \boldsymbol{Y}^n, E), \tag{5.5a}$$

$$\begin{aligned}(R^{n-1}\boldsymbol{U}^n, \boldsymbol{Z}) = {}& \frac{1}{2}(R^{n-1} d_t \boldsymbol{Y}^n, \boldsymbol{Z}) + \frac{1}{2}(d_t(R^n \boldsymbol{Y}^n), \boldsymbol{Z}) + \mu(\nabla \boldsymbol{Y}^n, \nabla \boldsymbol{Z}) \\ & + \frac{1}{2}([R^{n-1}\boldsymbol{Y}^{n-1} \cdot \nabla]\boldsymbol{Y}^n, \boldsymbol{Z}) - \frac{1}{2}([R^{n-1}\boldsymbol{Y}^{n-1} \cdot \nabla]\boldsymbol{Z}, \boldsymbol{Y}^n) + (\nabla P^n, \boldsymbol{Z}),\end{aligned} \tag{5.5b}$$

$$0 = (\operatorname{div} \boldsymbol{Y}^n, \Pi). \tag{5.5c}$$

The assumptions on the strongly acute triangulation imply a lower bound for the discrete density. The additional term $\frac{1}{2} R^n \operatorname{div} \boldsymbol{Y}^n$ in (5.5a), together with the reformulation in (5.5b) via (3.1) again make the convective operators in (5.5a) and (5.5b) skew-symmetric. For details, we refer the reader to [12]. Just as in [12], we can establish ρ_2 as an upper bound of $\{R^n\}$, which is due to the discrete maximum principle.

In the remainder of this chapter, we derive bounds for the iterates in (5.5) and verify that solutions of (5.5) converge to those of (3.2) for vanishing numerical parameters $k, h \to 0$. We proceed as follows:

1. Derive uniform bounds for the fully discrete scheme in standard parabolic norms; see Lemma 5.1. By stability of the interpolation, all interpolants inherit these bounds.
2. Derive uniform bounds in higher norms for the fully discrete scheme for $\{R^n\}$; i.e., bound $\mathcal{R}$ in $L^2(H^2_{\text{disc}}) \cap H^1(L^2) \cap L^\infty(H^1)$ uniformly with respect to k and h (see Lemma 5.2). To do this, we have to test (5.5a) with $-\Delta_h R^n$ and with $d_t R^n$.
3. In Lemma 5.3, we want to bound $\mathcal{R}$ in $H^1(H^1)$ and $\boldsymbol{\mathcal{Y}}$ in $L^2(\boldsymbol{H}^2_{\text{disc}}) \cap H^1(\boldsymbol{L}^2)$, which requires a bit more regularity on $\boldsymbol{\mathcal{Y}}$ and $\mathcal{R}$, respectively, than is available from Lemma 5.1. In order to conclude, we have to simultaneously test (5.5b) and (5.5a) with different test functions and combine all inequalities. We highlight more details of this strategy at the beginning of the proof of Lemma 5.3. This approach partly mimics ideas from [54, Section 2.2].

4. With these bounds, and a discrete version of the Aubin–Lions compactness theorem (see Lemma 5.4), it is possible to derive strong convergence for the affine interpolants.
5. By arguments from [62], strong convergence also holds for constant in time interpolants. This will lead to the convergence of the scheme (5.5) to (3.2) up to subsequences.

5.2. Stability of the scheme

Lemma 5.1
Let $0 < T < \infty$ and $\varepsilon > 0$. For every $1 \le n \le N$ there exists a solution $(\boldsymbol{Y}^n, P^n, R^n) \in \boldsymbol{V}_h \times M_h \times R_h$ to (5.5) which satisfies

$$\frac{1}{2} d_t \left[\left\| \sqrt{R^n} \boldsymbol{Y}^n \right\|^2 \right] + \mu \left\| \sqrt{R^{n-1}} \nabla \boldsymbol{Y}^n \right\|^2 + \frac{k}{2} \left[\left\| \sqrt{R^{n-1}} d_t \boldsymbol{Y}^n \right\|^2 \right] = \int_\Omega R^{n-1} \boldsymbol{U}^n \boldsymbol{Y}^n \, \mathrm{d}\boldsymbol{x},$$
$$\frac{1}{2} d_t \|R^n\|^2 + \frac{k}{2} \left[\|d_t R^n\|^2 \right] + \varepsilon \|\nabla R^n\|^2 = 0,$$

as long as $k \le k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$ is sufficiently small. For small enough $h, k > 0$, and every $1 \le n \le N$, there holds

$$0 < \rho_1 \le R^n \le \rho_2 < \infty. \tag{5.6}$$

In particular, we have the following uniform bounds for $\boldsymbol{\mathcal{Y}}$ and $\mathcal{R}$:

$$\|\boldsymbol{\mathcal{Y}}\|_{L^\infty(\boldsymbol{L}^2)} + \|\boldsymbol{\mathcal{Y}}\|_{L^2(\boldsymbol{H}^1)} + \|\mathcal{R}\|_{L^\infty(\boldsymbol{L}^2)} + \|\mathcal{R}\|_{L^2(H^1)} \le C(\varepsilon, \boldsymbol{u}, T).$$

The bounds also hold for $\boldsymbol{\mathcal{Y}}^{\bullet/-}$ and $\mathcal{R}^{\bullet/-}$ respectively.

Proof
This lemma relies on [12, Lemma 3.1] and can be proven with small modifications. □

Property (5.6) is a weaker form of the property (2.1), which is hold by the continuous solution of (1.2). The property (5.6) is very important for the following estimates.

Lemma 5.2
There holds uniformly with respect to $k, h > 0$

$$\|\Delta_h \mathcal{R}\|^2_{L^2(L^2)} + \|\nabla \mathcal{R}\|^2_{L^\infty(L^2)} + \|d_t \mathcal{R}\|^2_{L^2(L^2)} \le C(\varepsilon, T),$$

as long as $k \le k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$ is sufficiently small.

Proof
Step 1. Test (5.5a) with $-\Delta_h R^n \in R_h$,

$$(d_t \nabla R^n, \nabla R^n) + \varepsilon \|\Delta_h R^n\|^2 = ([\boldsymbol{Y}^n \cdot \nabla] R^n, \Delta_h R^n) + \frac{1}{2} (R^n \operatorname{div} \boldsymbol{Y}^n, \Delta_h R^n) =: I_1 + I_2.$$

For $\sigma > 0$, we estimate both terms by

$$I_1 \le \sigma \|\Delta_h R^n\|^2 + C(\sigma) \|\nabla \boldsymbol{Y}^n\|^2,$$

$$\begin{aligned} I_2 &\le \sigma\|\Delta_h R^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^4}\|\nabla R^n\|^2_{\boldsymbol{L}^4} \\ &\le \sigma\|\Delta_h R^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|\|\nabla \boldsymbol{Y}^n\|\|\nabla R^n\|\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big) \\ &\le \sigma\|\Delta_h R^n\|^2 + C(\sigma)\Big(\|\boldsymbol{Y}^n\|^2\|\nabla \boldsymbol{Y}^n\|^2 + \|\boldsymbol{Y}^n\|\|\nabla \boldsymbol{Y}^n\|\Big)\|\nabla R^n\|^2, \end{aligned}$$

where we used (5.6). By Lemma 5.1 and an appropriate choice of σ, we may conclude by Gronwall's inequality to bound $\Delta_h \mathcal{R}$ in $L^2(L^2)$ and $\nabla\mathcal{R}$ in $L^\infty(L^2)$.

Step 2. In order to show bounds for $d_t\mathcal{R}$, we test (5.5a) with $d_t R^n \in R_h$ and get

$$\|d_t R^n\|^2 + \varepsilon(d_t\nabla R^n, \nabla R^n) = -([\boldsymbol{Y}^n\cdot\nabla]R^n, d_t R^n) - \frac{1}{2}(R^n \operatorname{div}\boldsymbol{Y}^n, d_t R^n) =: II_1 + II_2.$$

Both terms can be estimated as follows for $\sigma > 0$:

$$\begin{aligned} II_1 &\le \sigma\|d_t R^n\|^2 + C(\sigma)\Big(\|\boldsymbol{Y}^n\|^2\|\nabla \boldsymbol{Y}^n\|^2 + \|\boldsymbol{Y}^n\|\|\nabla \boldsymbol{Y}^n\|\Big)\|\nabla R^n\|^2 + C(\sigma)\|\Delta_h R^n\|^2, \\ II_2 &\le \sigma\|d_t R^n\|^2 + C(\sigma)\|\nabla \boldsymbol{Y}^n\|^2. \end{aligned}$$

Again, we conclude by Gronwall's inequality, Lemma 5.5, and the first part of this proof. □

Lemma 5.3

There holds uniformly in $k, h > 0$

$$\|\boldsymbol{\mathcal{Y}}\|_{H^1(\boldsymbol{L}^2)} + \|\boldsymbol{\mathcal{Y}}\|_{L^\infty(\boldsymbol{H}^1_0)} + \|\Delta_h\boldsymbol{\mathcal{Y}}\|_{L^2(\boldsymbol{L}^2)} + \|\nabla d_t\mathcal{R}\|^2_{L^2(L^2)} \le C(\varepsilon, T),$$

as long as $k \le k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$ is sufficiently small.

Proof

Before starting with the technical part, let us mention the main difficulties to overcome in the proof: We will first test (5.5b) with $d_t\boldsymbol{Y}^n$ and $\boldsymbol{A}_h\boldsymbol{Y}^n$ in order to get positive terms to obtain the desired norms for $\boldsymbol{\mathcal{Y}}$. In the following calculation, the terms I_1 and K_2 are responsible for additional terms with no corresponding positive term, and which are not accessible to a Gronwall type argument. These new bad terms are $\|\Delta_h\boldsymbol{Y}^n\|$ and $\|\nabla d_t R^n\|$, respectively. Luckily, some of the bad terms are obtained with an arbitrary small constant, which allows us to complete the proof in its entirety. In the first three steps, we will deduce independently three inequalities; in the last step, we will combine them in a proper manner and deduce the desired bounds. Throughout the proof, we denote by $\tilde{J}, \tilde{L}, \tilde{N}$ functions which are summable in time by Lemma 5.1 and Lemma 5.2, i.e., $k\sum\tilde{J}, \ldots < \infty$ uniformly with respect to $k, h > 0$.

Step 1. Choose $\boldsymbol{Z} = d_t\boldsymbol{Y}^n$ in (5.5b),

$$\begin{aligned} &\|\sqrt{R^{n-1}}d_t\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}d_t\|\nabla\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}k\|d_t\nabla\boldsymbol{Y}^n\|^2 \\ &\quad\le \frac{1}{2}\left|(d_t R^n\boldsymbol{Y}^n, d_t\boldsymbol{Y}^n)\right| + \frac{1}{2}\left|\Big([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Y}^n, d_t\boldsymbol{Y}^n\Big)\right| \\ &\qquad + \frac{1}{2}\left|([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]d_t\boldsymbol{Y}^n, \boldsymbol{Y}^n)\right| + \left|(R^{n-1}\boldsymbol{U}^n, d_t\boldsymbol{Y}^n)\right| =: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

We derive estimates for each term $I_1, \dots, I_4$ separately. Let $\sigma, \tau, \theta > 0$. We calculate

$$\begin{aligned}
I_1 &\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|d_tR^n\|_{L^4}^2\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^4}^2 \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|d_tR^n\|\Big(\|d_t\nabla R^n\| + \|d_tR^n\|\Big)\|\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^n\| \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + \tau\|\nabla d_tR^n\|^2 + C(\sigma,\tau)\|d_tR^n\|^2\|\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^n\|^2 + C(\sigma,\tau)\|d_tR^n\|^2 \\
&=: \sigma\|d_t\boldsymbol{Y}^n\|^2 + \tau\|\nabla d_tR^n\|^2 + C(\sigma,\tau)J_1\|\nabla\boldsymbol{Y}^n\|^2 + C(\sigma,\tau)\tilde{J}_1 \\
I_2 &\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|\|\nabla\boldsymbol{Y}^{n-1}\|\|\nabla\boldsymbol{Y}^n\|\|\tilde{\Delta}_h\boldsymbol{Y}^n\| \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + \theta\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma,\theta)\|\boldsymbol{Y}^{n-1}\|^2\|\nabla\boldsymbol{Y}^{n-1}\|^2\|\nabla\boldsymbol{Y}^n\|^2 \\
&=: \sigma\|d_t\boldsymbol{Y}^n\|^2 + \theta\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma,\theta)J_2\|\nabla\boldsymbol{Y}^n\|^2,
\end{aligned}$$

where we used Sobolev embeddings and the Gagliardo–Nirenberg inequalities. By Lemma 5.2, we have $k\sum J_1 + J_2 + \tilde{J}_1 \le C < \infty$ uniformly with respect to $h, k > 0$, but depending on $\varepsilon, T > 0$. Integration by parts yields

$$\begin{aligned}
I_3 &\le \Big|\Big((\nabla R^{n-1}\cdot\boldsymbol{Y}^{n-1})d_t\boldsymbol{Y}^n, \boldsymbol{Y}^n\Big)\Big| + \Big|\Big((R^{n-1}\operatorname{div}\boldsymbol{Y}^{n-1})d_t\boldsymbol{Y}^n, \boldsymbol{Y}^n\Big)\Big| + I_2 \\
&=: I_{3a} + I_{3b} + I_2.
\end{aligned}$$

We estimate with (5.2), (5.4), and Sobolev embeddings,

$$\begin{aligned}
I_{3a} &\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|_{L^4}^2\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^8}^2\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^8}^2 \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|\Big(\|\Delta_hR^{n-1}\| + \|\nabla R^{n-1}\|\Big) \\
&\qquad \times \|\nabla\boldsymbol{Y}^{n-1}\|^{\frac{3}{2}}\|\boldsymbol{Y}^{n-1}\|^{\frac{1}{2}}\|\nabla\boldsymbol{Y}^n\|^{\frac{3}{2}}\|\boldsymbol{Y}^n\|^{\frac{1}{2}} \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|\Big(\|\Delta_hR^{n-1}\| + \|\nabla R^{n-1}\|\Big) \\
&\qquad \times \|\nabla\boldsymbol{Y}^{n-1}\|\|\boldsymbol{Y}^{n-1}\|^{\frac{1}{2}}\|\boldsymbol{Y}^n\|^{\frac{1}{2}}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^{\frac{1}{2}}\|\nabla\boldsymbol{Y}^n\|^{\frac{3}{2}}\Big) \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|\Big(\|\Delta_hR^{n-1}\| + \|\nabla R^{n-1}\|\Big) \\
&\qquad \times \|\nabla\boldsymbol{Y}^{n-1}\|\|\boldsymbol{Y}^{n-1}\|^{\frac{1}{2}}\|\boldsymbol{Y}^n\|^{\frac{1}{2}}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big) \\
&=: \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)J_{3a}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big), \\
I_{3b} &\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^4}^2\|\nabla\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^4}^2 \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^{n-1}\|\|\tilde{\Delta}_h\boldsymbol{Y}^{n-1}\| \\
&\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + \theta\|\tilde{\Delta}_h\boldsymbol{Y}^{n-1}\|^2 + C(\sigma,\theta)\|\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^{n-1}\|^2 \\
&=: \sigma\|d_t\boldsymbol{Y}^n\|^2 + \theta\|\tilde{\Delta}_h\boldsymbol{Y}^{n-1}\|^2 + C(\sigma,\theta)J_{3b}\|\nabla\boldsymbol{Y}^n\|^2, \\
I_4 &\le \sigma\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{U}^n\|^2,
\end{aligned}$$

where we have used the inequality valid in space dimension $d = 2$,

$$\|v\|_{L^8} \le C\|\nabla v\|^{\frac{3}{4}}\|v\|^{\frac{1}{4}} \quad \forall v \in H_0^1(\Omega).$$

Like above, J_{3a} and J_{3b} are uniformly summable by Lemmas 5.1 and 5.2.

By choosing an approximate $\sigma > 0$, we deduce for a summable function $\tilde{J}$ that

$$\frac{1}{2}\|\sqrt{R^{n-1}}d_t\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}d_t\|\nabla\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}k\|d_t\nabla\boldsymbol{Y}^n\|^2$$
$$\leq \theta\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + \tau\|\nabla d_t R^n\|^2 + C(\tau,\theta)\tilde{J}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big) + \tilde{J}. \quad (5.7)$$

Step 2. By the definition of the projection $\boldsymbol{P}_h$, we have

$$\mu(\tilde{\Delta}_h\boldsymbol{Y}^n, \boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n) = \mu\|\boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 = \mu\|\boldsymbol{A}_h\boldsymbol{Y}^n\|^2.$$

By [42, Corollary 4.4], every $\boldsymbol{V}\in\boldsymbol{J}_h$ satisfies

$$\|\tilde{\Delta}_h\boldsymbol{V}\| \leq C\|\boldsymbol{A}_h\boldsymbol{V}\|.$$

We may now test (5.5b) with $\boldsymbol{A}_h\boldsymbol{Y}^n$ to conclude

$$\begin{aligned}
C\mu\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 &\leq \mu\|\boldsymbol{A}_h\boldsymbol{Y}^n\|^2 \\
&\leq \frac{1}{2}\Big|(R^{n-1}d_t\boldsymbol{Y}^n, \boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n)\Big| + \frac{1}{2}\Big|(d_t(R^n\boldsymbol{Y}^n), \boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n)\Big| \\
&\quad + \frac{1}{2}\Big|([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Y}^n, \boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n)\Big| + \frac{1}{2}\Big|([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n, \boldsymbol{Y}^n)\Big| \\
&\quad + \Big|(R^{n-1}\boldsymbol{U}^n, \boldsymbol{P}_h\tilde{\Delta}_h\boldsymbol{Y}^n)\Big| =: K_1 + K_2 + K_3 + K_4 + K_5.
\end{aligned}$$

Exactly as in the first step, we estimate all terms $K_1,\ldots,K_5$ and use an argument similar to the first step, and use as well as $\|\boldsymbol{A}_h\boldsymbol{Y}^n\| \leq \|\tilde{\Delta}_h\boldsymbol{Y}^n\|$ (which holds by definition and by the continuity of the projection $\boldsymbol{P}_h$); then we get for $\sigma,\tau>0$

$$\begin{aligned}
K_1 &\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C_{\boldsymbol{Y}}(\sigma)\|d_t\boldsymbol{Y}^n\|^2, \\
K_2 &\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C_{\boldsymbol{Y}}(\sigma)\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|d_tR^n\|^2_{L^4}\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^4} \\
&\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C_{\boldsymbol{Y}}(\sigma)\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)\|d_tR^n\|^2_{L^4}\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^4} \\
&\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + \tau\|\nabla d_tR^n\|^2 + C_{\boldsymbol{Y}}(\sigma)\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma,\tau)\|d_tR^n\|^2\|\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^n\|^2, \\
&=: \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + \tau\|\nabla d_tR^n\|^2 + C_{\boldsymbol{Y}}(\sigma)\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma)L_2\|\nabla\boldsymbol{Y}^n\|^2, \\
K_3 &\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^4}\|\nabla\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^4} \\
&\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|\|\nabla\boldsymbol{Y}^{n-1}\|\|\nabla\boldsymbol{Y}^n\|\|\tilde{\Delta}_h\boldsymbol{Y}^n\| \\
&\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|^2\|\nabla\boldsymbol{Y}^{n-1}\|^2\|\nabla\boldsymbol{Y}^n\|^2 \\
&=: \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)L_3\|\nabla\boldsymbol{Y}^n\|^2, \\
K_4 &\leq K_3 + \Big|([\nabla R^n\cdot\boldsymbol{Y}^{n-1}]\tilde{\Delta}_h\boldsymbol{Y}^n, \boldsymbol{Y}^n)\Big| + \Big|(R^{n-1}\operatorname{div}\boldsymbol{Y}^{n-1}\tilde{\Delta}_h\boldsymbol{Y}^n, \boldsymbol{Y}^n)\Big| \\
&=: K_3 + K_{4a} + K_{4b}, \\
K_{4a} &\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|^2_{\boldsymbol{L}^4}\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^8}\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^8} \\
&\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\nabla R^{n-1}\|\Big(\|\Delta_hR^{n-1}\| + \|\nabla R^{n-1}\|\Big) \\
&\qquad \times\|\nabla\boldsymbol{Y}^{n-1}\|\|\boldsymbol{Y}^{n-1}\|^{\frac{1}{2}}\|\boldsymbol{Y}^n\|^{\frac{1}{2}}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big) \\
&=: \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)L_{4a}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big), \\
K_{4b} &\leq \sigma\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^4}\|\nabla\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^4}
\end{aligned}$$

$$
\begin{aligned}
&\le \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|\|\nabla \boldsymbol{Y}^n\|\|\nabla \boldsymbol{Y}^{n-1}\|\|\tilde{\Delta}_h \boldsymbol{Y}^{n-1}\| \\
&\le \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^n\|^2 + \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^{n-1}\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|^2\|\nabla \boldsymbol{Y}^n\|^2\|\nabla \boldsymbol{Y}^{n-1}\|^2 \\
&=: \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^n\|^2 + \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^{n-1}\|^2 + C(\sigma) L_{4b}\|\nabla \boldsymbol{Y}^n\|^2, \\
K_5 &\le \sigma\|\tilde{\Delta}_h \boldsymbol{Y}^n\|^2 + C(\sigma)\|\boldsymbol{U}^n\|^2,
\end{aligned}
$$

where we used integration by parts for term K_4. For each L_i we get uniform bounds $k\sum L_i \le C < \infty$ by Lemmas 5.2 and 5.1. Choosing $\sigma > 0$ small enough then leads, for a summable function $\tilde{L}$ and a constant $C_{\boldsymbol{Y}} \in \mathbb{R}$, to the following estimate

$$
\|\tilde{\Delta}_h \boldsymbol{Y}^n\|^2 \le \tau\|\nabla d_t R^n\| + C_{\boldsymbol{Y}}\|d_t \boldsymbol{Y}^n\|^2 + \tilde{L} + C(\tau)\tilde{L}\Big(\|\nabla \boldsymbol{Y}^{n-1}\|^2 + \|\nabla \boldsymbol{Y}^n\|^2\Big). \tag{5.8}
$$

Step 3. We take the time derivative of (5.5a), which reads as

$$
\begin{aligned}
(d_t)^2 R^n - \varepsilon\Delta_h d_t R^n = &- [d_t \boldsymbol{Y}^n \cdot \nabla] R^n - [\boldsymbol{Y}^{n-1}\cdot\nabla] d_t R^n \\
&- \frac{1}{2} d_t R^n \operatorname{div} \boldsymbol{Y}^{n-1} - \frac{1}{2} R^n \operatorname{div} d_t \boldsymbol{Y}^n.
\end{aligned} \tag{5.9}
$$

Multiply (5.9) with $d_t R^n$; then integration in space leads to

$$
\begin{aligned}
&\Big((d_t)^2 R^n, d_t R^n\Big) + \varepsilon\|\nabla d_t R^n\|^2 \\
&\quad = -\big([d_t \boldsymbol{Y}^n\cdot\nabla]R^n, d_t R^n\big) - \big([\boldsymbol{Y}^{n-1}\cdot\nabla] d_t R^n, d_t R^n\big) - \frac{1}{2}\big(d_t R^n \operatorname{div} \boldsymbol{Y}^{n-1}, d_t R^n\big) \\
&\qquad - \frac{1}{2}\big(R^n \operatorname{div} d_t \boldsymbol{Y}^n, d_t R^n\big) =: M_1 + M_2 + M_3 + M_4.
\end{aligned}
$$

We calculate for $\sigma, \lambda > 0$

$$
\begin{aligned}
M_1 &\le \lambda\|d_t \boldsymbol{Y}^n\|^2 + C(\lambda)\|\nabla R^n\|_{L^4}^2\|d_t R^n\|_{L^4}^2 \\
&\le \lambda\|d_t \boldsymbol{Y}^n\|^2 + C(\lambda)\|\nabla R^n\|\|\Delta_h R^n\|\|d_t R^n\|\|\nabla d_t R^n\| \\
&\le \lambda\|d_t \boldsymbol{Y}^n\|^2 + \sigma\|\nabla d_t R^n\|^2 + C(\lambda,\sigma)\|\nabla R^n\|^2\|\Delta_h R^n\|^2\|d_t R^n\|^2 \\
&=: \lambda\|d_t \boldsymbol{Y}^n\|^2 + \sigma\|\nabla d_t R^n\|^2 + C(\lambda,\sigma) N_1\|d_t R^n\|^2, \\
M_2 &\le \sigma\|\nabla d_t R^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^4}^2\|d_t R^n\|_{L^4}^2 \\
&\le \sigma\|\nabla d_t R^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|\|\nabla \boldsymbol{Y}^{n-1}\|\|d_t R^n\|\Big(\|\nabla d_t R^n\| + \|\nabla R^n\|\Big) \\
&\le \sigma\|\nabla d_t R^n\|^2 + C(\sigma)\Big(\|\boldsymbol{Y}^{n-1}\|^2\|\nabla \boldsymbol{Y}^{n-1}\|^2 + \|\boldsymbol{Y}^{n-1}\|\|\nabla \boldsymbol{Y}^{n-1}\|\Big)\|d_t R^n\|^2 \\
&=: \sigma\|\nabla d_t R^n\|^2 + C(\sigma) N_2\|d_t R^n\|^2, \\
M_3 &\le \|\nabla \boldsymbol{Y}^{n-1}\|\|d_t R^n\|_{L^4}^2 \le \|\nabla \boldsymbol{Y}^{n-1}\|\|d_t R^n\|\Big(\|\nabla d_t R^n\| + \|\nabla R^n\|\Big) \\
&\le \sigma\|\nabla d_t R^n\|^2 + C(\sigma)\Big(\|\nabla \boldsymbol{Y}^{n-1}\|^2 + \|\nabla \boldsymbol{Y}^{n-1}\|\Big)\|d_t R^n\|^2 \\
&=: \sigma\|\nabla d_t R^n\|^2 + C(\sigma) N_3\|d_t R^n\|^2, \\
M_4 &= \frac{1}{2}\big(d_t \boldsymbol{Y}^n, \nabla(R^n d_t R^n)\big) = \frac{1}{2}\big(d_t \boldsymbol{Y}^n, \nabla R^n d_t R^n\big) + \frac{1}{2}\big(d_t \boldsymbol{Y}^n, R^n \nabla d_t R^n\big) \\
&=: M_{4a} + M_{4b}, \\
M_{4a} &\le \lambda\|d_t \boldsymbol{Y}^n\|^2 + C(\lambda)\|\nabla R^n\|_{L^4}^2\|d_t R^n\|_{L^4}^2
\end{aligned}
$$

$$
\begin{aligned}
&\le \lambda\|d_t\boldsymbol{Y}^n\|^2 + C(\lambda)\|\nabla R^n\|\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)\|d_t R^n\|\Big(\|\nabla d_t R^n\| + \|d_t R^n\|\Big)\\
&\le \sigma\|\nabla d_t R^n\|^2 + \lambda\|d_t\boldsymbol{Y}^n\|^2 + C(\lambda)\bigg(\|\nabla R^n\|^2\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)^2\\
&\quad + \|\nabla R^n\|\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)\|d_t R^n\|\bigg)\|d_t R^n\|^2\\
&=: \sigma\|\nabla d_t R^n\|^2 + \lambda\|d_t\boldsymbol{Y}^n\|^2 + C(\sigma,\lambda)N_{4a}\|d_t R^n\|^2,\\
M_{4b} &\le \sigma\|\nabla d_t R^n\|^2 + C(\sigma)\|d_t\boldsymbol{Y}^n\|^2.
\end{aligned}
$$

All functions N_i are summable in time, i.e., we have $k\sum N_i < \infty$ uniformly in $k, h > 0$. For an appropriate choice of σ and $\tilde{N}$ being a summable function and $C_{\boldsymbol{Y},2} \in \mathbb{R}$, we arrive at

$$
d_t\|d_t R^n\|^2 + \varepsilon\|\nabla d_t R^n\|^2 \le C_{\boldsymbol{Y},2}\|d_t\boldsymbol{Y}^n\|^2 + C\tilde{N}\|d_t R^n\|^2. \tag{5.10}
$$

Step 4: We insert (5.8) into (5.7), choose $\theta > 0$ small enough, and arrive at

$$
\begin{aligned}
&\frac{1}{4}\|\sqrt{R^{n-1}}d_t\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}d_t\|\nabla\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}k\|d_t\nabla\boldsymbol{Y}^n\|^2\\
&\qquad\le \tau\|\nabla d_t R^n\|^2 + C(\tau)\tilde{F}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big) + \tilde{F},
\end{aligned} \tag{5.11}
$$

where here and below $\tilde{F}$ is generic summable function $k\sum\tilde{F} < \infty$ uniformly in $k, h > 0$ consisting of all above functions $\tilde{J}, \tilde{L}, \tilde{N}$. Inserting (5.11) into (5.8), we get

$$
\|\tilde{\Delta}_h\boldsymbol{Y}^n\|^2 \le \tau\|\nabla d_t R^n\| + \tilde{F} + C(\tau)\tilde{F}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big). \tag{5.12}
$$

We add $(4C_{\boldsymbol{Y},2}\rho_{\min}^{-2} + 1)(5.11) + (5.10)$, choose τ small enough and arrive at

$$
\begin{aligned}
&\frac{1}{4}\|\sqrt{R^{n-1}}d_t\boldsymbol{Y}^n\|^2 + \frac{\mu}{2}d_t\|\nabla\boldsymbol{Y}^n\|^2 + d_t\|d_t R^n\|^2 + \varepsilon\|\nabla d_t R^n\|^2\\
&\qquad\le \tilde{F} + C\tilde{F}\|d_t R^n\|^2 + C\tilde{F}\Big(\|\nabla\boldsymbol{Y}^{n-1}\|^2 + \|\nabla\boldsymbol{Y}^n\|^2\Big).
\end{aligned}
$$

We conclude with the discrete version of Gronwall's lemma to obtain for sufficiently small $k \le k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$ all bounds expect that for $\tilde{\Delta}_h\boldsymbol{\mathcal{Y}}$. This bound of $\|\tilde{\Delta}_h\boldsymbol{\mathcal{Y}}\|_{L^2(\boldsymbol{L}^2)}$ can be derived by inserting all existing bounds into (5.12). □

The lemma above is the reason why we need $\mu > 0$ to be constant—otherwise there is no way to get a lower bound for the terms involving $\boldsymbol{\mathcal{Y}}$ with the methods used in its proof.

5.3. Convergence of the scheme

In order to pass to the limit, we need a discrete version of the Aubin–Lions compactness theorem.

Lemma 5.4
Let $\mathcal{T}_h$ be a quasi-uniform triangulation of Ω, let $\{\boldsymbol{\Phi}_h^n\}_{n=0}^N \subset \boldsymbol{V}_h$, with $\boldsymbol{\mathcal{P}}$ its time interpolation, such that

$$\|\boldsymbol{\mathcal{P}}\|_{H^1(\boldsymbol{L}^2)} + \|\boldsymbol{\mathcal{P}}\|_{L^2(\boldsymbol{H}^1)} + \|\tilde{\Delta}_h \boldsymbol{\mathcal{P}}\|_{L^2(\boldsymbol{L}^2)} \leq C.$$

Then, there exist a subsequence $\{\mathcal{P}\}_{k,h} \subset L^2(\boldsymbol{H}^1)$ (not relabeled), and $\boldsymbol{P} \in L^2(\boldsymbol{H}^1)$ such that $\boldsymbol{\mathcal{P}} \to \boldsymbol{P}$ in $L^2(\boldsymbol{H}^1)$ for $h, k \to 0$.

Proof
See [18, Lemma 2.4] or [66, Lemma 4.9]. □

The same result holds for the discrete Laplacian Δ_h, which is defined on R_h.

Lemma 5.5
There exists a subsequence (not relabeled) such that for $h, k \to 0$ the following hold:

1. $\boldsymbol{\mathcal{Y}}^{\bullet/-}, \boldsymbol{\mathcal{Y}} \overset{*}{\rightharpoonup} \boldsymbol{y}$ in $L^\infty(\boldsymbol{H}_0^1)$ and $\mathcal{R}^{\bullet/-}, \mathcal{R} \overset{*}{\rightharpoonup} \rho$ in $L^\infty(H^1)$,
2. $\boldsymbol{\mathcal{Y}} \to \boldsymbol{y}$ in $L^2(\boldsymbol{H}_0^1)$ and $\mathcal{R} \to \rho$ in $L^2(\boldsymbol{H}^1)$,
3. $\operatorname{div} \boldsymbol{\mathcal{Y}} \to \operatorname{div} \boldsymbol{y} = 0$ in $L^2(\boldsymbol{L}^2)$,
4. $\mathcal{R} \to \rho$ in $L^q(L^q)$ for $1 \leq q < \infty$.

Proof
1. This is a direct consequence of Lemmas 5.3 and 5.2.
2. This follows from the estimates in Lemmas 5.3 and 5.2, respectively, as well as Lemma 5.4.
3. This follows from the second part of this lemma and (5.5c).
4. Since $\operatorname{div} \boldsymbol{\mathcal{Y}} \to 0$ in $L^2(\boldsymbol{L}^2)$, we conclude by [12, Lemma 3.2] that $\mathcal{R} \to \rho$ in $L^2(L^2)$. By the L^∞-bound of $\mathcal{R}$ and this information, it follows easily by Hölder's inequality that $\mathcal{R} \to \rho$ in $L^q(L^q)$ for every $1 \leq q < \infty$. □

The bounds from Lemmas 5.1, 5.2, and 5.3 yield the convergence of all linear terms in (5.5), and the convergence results from Lemma 5.5 are sufficient for the convergence of all nonlinear terms in (5.5); we refer the reader to [12] for details of the concluding convergence of the nonlinear terms.

In the next chapter, we consider a discretization of Problem 4.1 and derive first order optimality conditions in chapter 6. The main goal in chapter 7 is to verify stability estimates for a solution of the adjoint equation. A main problem in achieving this is that the adjoint equation couples adjoint variables with primal variables. The stability of the adjoint equation, together with the bounds from this chapter, are used in chapter 8 to practically construct weak solutions of the continuous optimality conditions (4.1).

6. Discrete optimization problem

We define a discretization of Problem 4.1, and show existence of a minimum. The finite-dimensional version of the Lagrange multiplier theorem directly yields existence of a solution to the related discrete optimality system (6.2), which corresponds to (4.1). The "first discretize, then optimize" ansatz allows us to benefit from the results in the stability analysis for (6.2).

Problem 6.1
Let $\varepsilon, \delta > 0$, and $h, k > 0$. Minimize $J_{\delta,h,k} : \{R_h\}_{n=1}^N \times \{\boldsymbol{L}^2(\Omega)\}_{n=1}^N$ via

$$\begin{aligned} J_{\delta,h,k}(\mathcal{R},\boldsymbol{\mathcal{U}}) := \frac{\lambda}{2} k \sum_{n=1}^N \int_\Omega |R^n - \tilde{\rho}(t_n)|^2 \,\mathrm{d}\boldsymbol{x} + \frac{\alpha}{2} k \sum_{n=1}^N \int_\Omega |\boldsymbol{U}^n|^2 \,\mathrm{d}\boldsymbol{x} \\ + \frac{\beta}{2} k \sum_{n=1}^N \int_\Omega \left\{ \delta |\nabla R^n|^2 + \frac{1}{4\delta} W(R^n) \right\} \mathrm{d}\boldsymbol{x} \end{aligned} \tag{6.1}$$

subject to (5.5).

A variational discretization for $\{\boldsymbol{U}^n\}_n$ is used, i.e., every iterate $\boldsymbol{U}^n$ is in $\boldsymbol{L}^2(\Omega)$ a priori; however, $\{\boldsymbol{U}^n\}$ is discretized implicitly through (6.2d) by means of $\mathcal{R}$ and $\boldsymbol{\mathcal{Z}}$, and its time interpolant is bounded uniformly in $L^2(\boldsymbol{L}^2)$. The advantages are an easier analysis and a natural discretization; i.e., we do not have to consider projections in (6.2d). We refer the reader to [45] for details. At the end of chapter 8, we show that $\boldsymbol{\mathcal{U}} \to \boldsymbol{u}$ in $L^2(\boldsymbol{L}^2)$.

Similarly to the proof of Theorem 4.2, we can prove the following theorem.

Theorem 6.2
There exists at least one solution of Problem 6.1.

We state the Lagrange functional and consider derivatives of it with respect to all unknowns. As in chapter 4, we can use the Lagrange multiplier theorem in order to derive optimality conditions. We define the Lagrange functional via

$$\mathcal{L}_{h,k}(\boldsymbol{\mathcal{Y}}, \mathcal{P}, \mathcal{R}, \boldsymbol{\mathcal{U}}; \boldsymbol{\mathcal{Z}}, \mathcal{Q}, \mathcal{E})$$

$$
\begin{aligned}
&:= J_{\delta,h,k}(\mathcal{R},\boldsymbol{\mathcal{U}}) + k\sum_{n=1}^{N}\Big(d_t R^n + [\boldsymbol{Y}^n\cdot\nabla]R^n + \frac{1}{2}R^n \operatorname{div}\boldsymbol{Y}^n - \varepsilon\Delta_h R^n, E^n\Big)\\
&\quad + k\sum_{n=1}^{N}\Big(\frac{1}{2}R^{n-1}d_t\boldsymbol{Y}^n + \frac{1}{2}d_t(R^n\boldsymbol{Y}^n) + \frac{1}{2}[R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Y}^n, \boldsymbol{Z}^n\Big)\\
&\quad + k\sum_{n=1}^{N}\Big(-\mu\tilde{\Delta}_h\boldsymbol{Y}^n + \nabla P^n - R^{n-1}\boldsymbol{U}^n, \boldsymbol{Z}^n\Big)\\
&\quad - k\sum_{n=1}^{N}\frac{1}{2}\Big([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Z}^n, \boldsymbol{Y}^n\Big) + k\sum_{n=1}^{N}(\operatorname{div}\boldsymbol{Y}^n, Q^n).
\end{aligned}
$$

The first line stands for (5.5a), and the following lines for (5.5b) and (5.5c), respectively. For notational simplicity, let $\mathcal{L} = \mathcal{L}_{h,k}$ below.

The derivatives of $\mathcal{L}$ with respect to the Lagrange multipliers $\{\boldsymbol{Z}^n\}_n, \{P^n\}_n$, and $\{E^n\}_n$ lead to (5.5). Setting all derivatives of $\mathcal{L}$ equal to zero, we may infer by the Lagrange multiplier theorem and integration by parts in the same manner as in chapter 4, that the coupled system of (5.5) and the following system has at least one weak solution:

$$
\begin{aligned}
0 = {}& \frac{1}{2}E^n\nabla R^n - \frac{1}{2}R^n\nabla E^n - \frac{1}{2}d_tR^n\boldsymbol{Z}^n - R^n d_t\boldsymbol{Z}^{n+1} + \frac{1}{2}R^n\nabla\boldsymbol{Y}^{n+1}\cdot\boldsymbol{Z}^{n+1} &\text{(6.2a)}\\
& - \frac{1}{2}R^n\nabla\boldsymbol{Z}^{n+1}\cdot\boldsymbol{Y}^{n+1} - \frac{1}{2}(\nabla R^{n-1}\cdot\boldsymbol{Y}^{n-1})\boldsymbol{Z}^n - \frac{1}{2}R^{n-1}\operatorname{div}\boldsymbol{Y}^{n-1}\boldsymbol{Z}^n\\
& - [R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Z}^n - \mu\tilde{\Delta}_h\boldsymbol{Z}^n - \nabla Q^n,\\
0 = {}& -\operatorname{div}\boldsymbol{Z}^n, &\text{(6.2b)}\\
0 = {}& -d_tE^{n+1} - [\boldsymbol{Y}^n\cdot\nabla]E^n + \frac{1}{2}(\operatorname{div}\boldsymbol{Y}^n)E^n - \varepsilon\Delta_hE^{n+1} + \frac{1}{2}d_t\boldsymbol{Y}^{n+1}\cdot\boldsymbol{Z}^{n+1} &\text{(6.2c)}\\
& - \frac{1}{2}\boldsymbol{Y}^n\cdot d_t\boldsymbol{Z}^{n+1} + \frac{1}{2}[\boldsymbol{Y}^n\cdot\nabla]\boldsymbol{Y}^{n+1}\cdot\boldsymbol{Z}^{n+1} - \boldsymbol{U}^{n+1}\cdot\boldsymbol{Z}^{n+1} - \frac{1}{2}[\boldsymbol{Y}^n\cdot\nabla]\boldsymbol{Z}^{n+1}\cdot\boldsymbol{Y}^{n+1}\\
& + \lambda(R^n - \tilde{\rho}(t_n)) - \beta\delta\Delta_hR^n + \frac{\beta}{8\delta}W'(R^n),\\
0 = {}& \alpha\boldsymbol{U}^n - R^{n-1}\boldsymbol{Z}^n, &\text{(6.2d)}
\end{aligned}
$$

together with the final conditions $E^{N+1} = 0$, $\boldsymbol{Z}^{N+1} = \boldsymbol{0}$, and $Q^{N+1} = 0$ and homogeneous Dirichlet boundary conditions for $\{\boldsymbol{Z}^n\}_n$.

7. Stability of the discrete adjoint equation

We derive uniform bounds for existing solutions of (6.2). These results are used in chapter 8 in order to identify the limit of (6.2) for $h, k \to 0$.

Lemma 7.1
There holds uniformly in $k, h > 0$

$$\|\boldsymbol{\mathcal{Z}}_t\|^2_{L^2(\boldsymbol{L}^2)} + \|\boldsymbol{\mathcal{Z}}\|_{L^\infty(\boldsymbol{L}^2)} + \|\boldsymbol{\mathcal{Z}}\|_{L^2(\boldsymbol{H}^1_0)} + \varepsilon\|\mathcal{E}\|^2_{L^2(H^1)} + \|\mathcal{E}\|_{L^\infty(L^2)} + \|\mathcal{E}_t\|_{L^2((H^1)^*)} \leq C(\varepsilon, T),$$

as long as $k \leq k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$ is sufficiently small.

Proof
The proof consists of four steps. We test (6.2a) with $\boldsymbol{Z}^n$ and $d_t\boldsymbol{Z}^n$, and we test (6.2c) with E^n. The crucial terms arising in the analysis below are in particular I_2, II_4, and III_2. To properly deal with them, we have to consider a weighted sum of the single inequalities similarly to that used in the proof of Lemma 5.3.

Step 1. We test (6.2a) with $\boldsymbol{Z}^n$. Thanks to $(\operatorname{div}\boldsymbol{Z}^n, Q^n) = 0$, we arrive at

$$\begin{aligned}
-(R^n d_t\boldsymbol{Z}^{n+1}, \boldsymbol{Z}^n) + \mu\|\nabla\boldsymbol{Z}^n\|^2 &= \frac{1}{2}(d_tR^n\boldsymbol{Z}^n, \boldsymbol{Z}^n) - \frac{1}{2}(E^n\nabla R^n, \boldsymbol{Z}^n) \\
&+ \frac{1}{2}(R^n\nabla E^n, \boldsymbol{Z}^n) - \frac{1}{2}(R^n\nabla\boldsymbol{Y}^{n+1}\cdot\boldsymbol{Z}^{n+1}, \boldsymbol{Z}^n) \\
&+ \frac{1}{2}(R^n\nabla\boldsymbol{Z}^{n+1}\cdot\boldsymbol{Y}^{n+1}, \boldsymbol{Z}^n) + \frac{1}{2}\Big([\nabla R^{n-1}\cdot\boldsymbol{Y}^{n-1}]\boldsymbol{Z}^n, \boldsymbol{Z}^n\Big) \\
&+ \frac{1}{2}(R^{n-1}\operatorname{div}\boldsymbol{Y}^{n-1}\boldsymbol{Z}^n, \boldsymbol{Z}^n) + \Big([R^{n-1}\boldsymbol{Y}^{n-1}\cdot\nabla]\boldsymbol{Z}^n, \boldsymbol{Z}^n\Big).
\end{aligned}$$

By elementary algebraic calculations, we find

$$-(R^n d_t\boldsymbol{Z}^{n+1}, \boldsymbol{Z}^n) = -\frac{1}{2}d_t\|\sqrt{R^n}\boldsymbol{Z}^{n+1}\|^2 - \frac{1}{2}(d_tR^n\boldsymbol{Z}^n, \boldsymbol{Z}^n) + \frac{1}{2}k\|\sqrt{R^n}d_t\boldsymbol{Z}^n\|^2.$$

Hence, we get

$$\begin{aligned}
&-\frac{1}{2}d_t\|\sqrt{R^n}\boldsymbol{Z}^{n+1}\|^2 + \frac{k}{2}\|\sqrt{R^n}d_t\boldsymbol{Z}^n\|^2 + \mu\|\nabla\boldsymbol{Z}^n\|^2 \\
&\qquad\leq \|d_tR^n\|\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} + \frac{1}{2}\|\nabla R^n\|_{L^4}\|E^n\|_{L^4}\|\boldsymbol{Z}^n\| + C\|\nabla E^n\|\|\boldsymbol{Z}^n\| \\
&\qquad\quad + C\|\nabla\boldsymbol{Y}^{n+1}\|\|\boldsymbol{Z}^{n+1}\|_{\boldsymbol{L}^4}\|\boldsymbol{Z}^n\|_{\boldsymbol{L}^4} + C\|\nabla\boldsymbol{Z}^{n+1}\|\|\boldsymbol{Y}^{n+1}\|_{\boldsymbol{L}^\infty}\|\boldsymbol{Z}^n\| \\
&\qquad\quad + \frac{1}{2}\|\nabla R^{n-1}\|_{L^4}\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^4}\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} + C\|\operatorname{div}\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^2}\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} \\
&\qquad\quad + C\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^\infty}\|\nabla\boldsymbol{Z}^n\|_{\boldsymbol{L}^2}\|\boldsymbol{Z}^n\| =: I_1 + I_2 + \ldots + I_8.
\end{aligned}$$

For a small $\sigma, \theta > 0$, we calculate with the same tools as in the proofs of Lemmas 5.2 and 5.3,

$$
\begin{aligned}
I_1 &\le \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)\|d_t R^n\|^2\|\boldsymbol{Z}^n\|^2 =: \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)J_1\|\boldsymbol{Z}^n\|^2,\\
I_2 &\le C\|\nabla R^n\|_{L^4}\|(\|E^n\| + \|\nabla E^n\|)\|\boldsymbol{Z}^n\| \le \theta\|E^n\|^2 + \theta\|\nabla E^n\|^2\\
&\qquad + C(\theta)\|\nabla R^n\|_{L^4}^2\|\boldsymbol{Z}^n\|^2 =: \theta\|E^n\|^2 + \theta\|\nabla E^n\|^2 + C(\theta)J_2\|\boldsymbol{Z}^n\|^2,\\
I_3 &\le \theta\|\nabla E^n\|^2 + C(\theta)\|\boldsymbol{Z}^n\|^2,\\
I_4 &\le C\|\nabla \boldsymbol{Y}^{n+1}\|^2\|\boldsymbol{Z}^{n+1}\|\|\nabla \boldsymbol{Z}^{n+1}\| + C\|\nabla \boldsymbol{Y}^{n+1}\|^2\|\boldsymbol{Z}^n\|\|\nabla \boldsymbol{Z}^n\|\\
&\le \sigma\|\nabla \boldsymbol{Z}^n\|^2 + \sigma\|\nabla \boldsymbol{Z}^{n+1}\|^2 + C(\sigma)\|\nabla \boldsymbol{Y}^{n+1}\|^4\Big(\|\boldsymbol{Z}^{n+1}\|^2 + \|\boldsymbol{Z}^n\|^2\Big)\\
&=: \sigma\|\nabla \boldsymbol{Z}^n\|^2 + \sigma\|\nabla \boldsymbol{Z}^{n+1}\|^2 + C(\sigma)J_4\Big(\|\boldsymbol{Z}^{n+1}\|^2 + \|\boldsymbol{Z}^n\|^2\Big),\\
I_5 &\le \sigma\|\nabla \boldsymbol{Z}^{n+1}\|^2 + C(\sigma)\|\boldsymbol{Y}^{n+1}\|_{L^\infty}^2\|\boldsymbol{Z}^n\|^2 =: \sigma\|\nabla \boldsymbol{Z}^{n+1}\|^2 + C(\sigma)J_5\|\boldsymbol{Z}^n\|^2,\\
I_6 &\le C\|\nabla R^n\|_{L^4}\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^4}\|\boldsymbol{Z}^n\|\|\nabla \boldsymbol{Z}^n\|\\
&\le \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)\|\nabla R^n\|_{L^4}\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^4}\|\boldsymbol{Z}^n\|^2\\
&=: \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)J_6\|\boldsymbol{Z}^n\|^2,\\
I_7 &\le \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)\|\nabla \boldsymbol{Y}^{n-1}\|^2\|\boldsymbol{Z}^n\|^2 =: \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)J_7\|\boldsymbol{Z}^n\|^2,\\
I_8 &\le \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)\|\boldsymbol{Y}^{n-1}\|_{\boldsymbol{L}^\infty}^2\|\boldsymbol{Z}^n\|^2 =: \sigma\|\nabla \boldsymbol{Z}^n\|^2 + C(\sigma)J_8\|\boldsymbol{Z}^n\|^2,
\end{aligned}
$$

where $k\sum J_i \le C < \infty$ uniformly with respect to $h, k > 0$, by Lemmas 5.2 and 5.3. Summing up and choosing $\sigma > 0$, we get for a summable function $\tilde{J}$ and a $\theta > 0$ (which will be chosen later)

$$
\begin{aligned}
-\frac{1}{2}d_t\|\sqrt{R^n}\boldsymbol{Z}^{n+1}\|^2 + \mu\|\nabla \boldsymbol{Z}^n\|^2 + \frac{k}{2}\|\sqrt{R^n}d_t\boldsymbol{Z}^n\|^2&\\
\le \theta\|E^n\|^2 + \theta\|\nabla E^n\|^2 + C(\theta)\tilde{J} + C(\theta)\tilde{J}\Big(\|\boldsymbol{Z}^{n+1}\|^2 + \|\boldsymbol{Z}^n\|^2\Big).& \qquad (7.1)
\end{aligned}
$$

Step 2. We test (6.2c) with E^n and get

$$
\begin{aligned}
&\varepsilon\|\nabla E^{n+1}\|^2 - \frac{1}{2}d_t\|E^{n+1}\|^2 + \frac{k}{2}\|d_tE^{n+1}\|^2\\
&\le \|\boldsymbol{Y}^n\|_{\boldsymbol{L}^\infty}\|\nabla E^n\|\|E^n\| + \frac{1}{2}\|\operatorname{div}\boldsymbol{Y}^n\|\|E^n\|_{L^4}^2 + \frac{1}{2}\|d_t\boldsymbol{Y}^{n+1}\|\|\boldsymbol{Z}^{n+1}\|_{\boldsymbol{L}^4}\|E^n\|_{L^4}\\
&\quad + \frac{1}{2}\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^\infty}\|d_t\boldsymbol{Z}^{n+1}\|\|E^n\| + \frac{1}{2}\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^4}\|\nabla \boldsymbol{Y}^{n+1}\|_{\boldsymbol{L}^4}\|\boldsymbol{Z}^{n+1}\|_{\boldsymbol{L}^4}\|E^n\|_{L^4}\\
&\quad + \|\boldsymbol{U}^{n+1}\|\|\boldsymbol{Z}^{n+1}\|_{\boldsymbol{L}^4}\|E^n\|_{L^4} + \frac{1}{2}\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^8}\|\nabla \boldsymbol{Z}^{n+1}\|\|\boldsymbol{Y}^{n+1}\|_{\boldsymbol{L}^8}\|E^n\|_{L^4}\\
&\quad + \lambda\Big|(R^n - \tilde{\rho}(t_n), E^n)\Big| + \beta\delta\|\Delta_h R^n\|\|E^n\| + \frac{\beta}{8\delta}|(W'(R^n), E^n)|\\
&=: II_1 + II_2 + \ldots + II_{10}.
\end{aligned}
$$

For $\sigma, \tau, \lambda > 0$ small enough, we derive the following estimates using the techniques from the first part:

$$
II_1 \le \sigma\|\nabla E^n\|^2 + C(\sigma)\|\boldsymbol{Y}^n\|_{\boldsymbol{L}^\infty}^2\|E^n\|^2 =: \sigma\|\nabla E^n\|^2 + C(\sigma)K_1\|E^n\|^2,
$$

$$
\begin{aligned}
II_2 &\le C\|\nabla \boldsymbol{Y}^n\|\|E^n\|\Big(\|\nabla E^n\| + \|E^n\|\Big)\\
&\le \sigma\|\nabla E^n\|^2 + C\Big(\|\nabla \boldsymbol{Y}^n\|^2 + \|\nabla \boldsymbol{Y}^n\|\Big)\|E^n\|^2,\\
II_3 &\le C\|d_t\boldsymbol{Y}^{n+1}\|_{\boldsymbol{L}^2}\|\boldsymbol{Z}^{n+1}\|^{\frac12}\|\nabla\boldsymbol{Z}^{n+1}\|^{\frac12}\|E^n\|^{\frac12}\Big(\|\nabla E^n\|^{\frac12} + \|E^n\|^{\frac12}\Big)\\
&\le \sigma\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\sigma)\|d_t\boldsymbol{Y}^{n+1}\|^2\|\boldsymbol{Z}^{n+1}\|^2\\
&\quad + C(\sigma)\Big(\|d_t\boldsymbol{Y}^{n+1}\|^2 + \|d_t\boldsymbol{Y}^{n+1}\|\Big)\|E^n\|^2\\
&=: \sigma\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\sigma)K_3\|\boldsymbol{Z}^{n+1}\|^2 + C(\sigma)K_3\|E^n\|^2,\\
II_4 &\le \tau\|d_t\boldsymbol{Z}^{n+1}\|^2 + C(\tau)\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^\infty}\|E^n\|^2 =: \tau\|d_t\boldsymbol{Z}^{n+1}\|^2 + C(\tau)K_4\|E^n\|^2\\
II_5 &\le C\|\boldsymbol{Z}^{n+1}\|\|\nabla\boldsymbol{Z}^{n+1}\|\\
&\quad + C\|\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^{n+1}\|\|\tilde\Delta_h\boldsymbol{Y}^{n+1}\|\|E^n\|\Big(\|\nabla E^n\| + \|E^n\|\Big)\\
&\le \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\sigma)\Big(\|\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^n\|^2\|\|\nabla\boldsymbol{Y}^{n+1}\|^2\|\tilde\Delta_h\boldsymbol{Y}^{n+1}\|^2\\
&\quad + \|\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^n\|\|\nabla\boldsymbol{Y}^{n+1}\|\|\tilde\Delta_h\boldsymbol{Y}^{n+1}\|\Big)\|E^n\|^2 + C(\lambda)\|\boldsymbol{Z}^{n+1}\|^2\\
&=: \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\lambda)\|\boldsymbol{Z}^{n+1}\|^2 + C(\sigma)K_5\|E^n\|^2,\\
II_6 &\le C\|\boldsymbol{U}^{n+1}\|\|\boldsymbol{Z}^{n+1}\|\|\nabla\boldsymbol{Z}^{n+1}\| + \|\boldsymbol{U}^{n+1}\|\|E^n\|\Big(\|\nabla E^n\| + \|E^n\|\Big)\\
&\le \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\lambda)\|\boldsymbol{U}^{n+1}\|^2\|\boldsymbol{Z}^{n+1}\|^2\\
&\quad + C(\sigma)\Big(\|\boldsymbol{U}^{n+1}\|^2 + \|\boldsymbol{U}^{n+1}\|\Big)\|E^n\|^2\\
&=: \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\lambda)K_6\|\boldsymbol{Z}^{n+1}\|^2 + C(\sigma)K_6\|E^n\|^2,\\
II_7 &\le \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + C(\lambda)\|\boldsymbol{Y}^n\|^2_{\boldsymbol{L}^8}\|\boldsymbol{Y}^{n+1}\|^2_{\boldsymbol{L}^8}\|E^n\|\Big(\|\nabla E^n\| + \|E^n\|\Big)\\
&\le \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2\\
&\quad + C(\lambda,\sigma)\Big(\|\nabla\boldsymbol{Y}^n\|^4\|\nabla\boldsymbol{Y}^{n+1}\|^4 + \|\nabla\boldsymbol{Y}^n\|^2\|\nabla\boldsymbol{Y}^{n+1}\|^2\Big)\|E^n\|^2\\
&=: \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \sigma\|\nabla E^n\|^2 + C(\lambda,\sigma)K_7\|E^n\|^2,\\
II_8 &\le C(\lambda) + \|E^n\|^2,\\
II_9 &\le C(\beta,\delta)\|\Delta_h R^n\|^2 + \|E^n\|^2 =: K_9 + \|E^n\|^2,\\
II_{10} &\le C(\beta,\delta) + \|E^n\|^2.
\end{aligned}
$$

Again, we have $k\sum K_i < \infty$ uniformly in k and h for each i by Lemmas 5.2 and 5.3. We have derived the following estimate:

$$
\begin{aligned}
&\varepsilon\|\nabla E^{n+1}\|^2 - d_t\|E^{n+1}\|^2 + k\|d_tE^{n+1}\|^2\\
&\qquad \le \lambda\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \tau\|d_t\boldsymbol{Z}^{n+1}\|^2 + C(\lambda,\tau)\tilde K\|\boldsymbol{Z}^{n+1}\|^2 + C(\lambda,\tau)\tilde K\|E^n\|^2. \quad (7.2)
\end{aligned}
$$

Choosing $\lambda, \theta > 0$ small enough and adding (7.1) and (7.2) together, we get

$$
\begin{aligned}
&- d_t\|\sqrt{R^n}\boldsymbol{Z}^{n+1}\|^2 + \frac{\mu}{2}\|\nabla\boldsymbol{Z}^n\|^2 - d_t\|E^{n+1}\|^2 + \frac{\varepsilon}{2}\|\nabla E^n\|^2\\
&\qquad \le \tau\|d_t\boldsymbol{Z}^{n+1}\|^2 + C(\tau)\tilde J\|\boldsymbol{Z}^{n+1}\|^2 + C(\tau)\tilde K\|E^n\|^2, \quad (7.3)
\end{aligned}
$$

where $\tilde F$ is a generic summable function consisting of $\tilde J$ and $\tilde K$ from above.

Step 3. We test (6.2a) with $-d_t\boldsymbol{Z}^{n+1}$ and get

$$\begin{aligned}
&\|\sqrt{R^n}d_t\boldsymbol{Z}^{n+1}\|^2 - \frac{\mu}{2}d_t\|\nabla\boldsymbol{Z}^{n+1}\|^2 + \frac{\mu}{2}k\|d_t\nabla\boldsymbol{Z}^{n+1}\|^2 \\
&\quad\le \frac{1}{2}\|\sqrt{R^n}d_t\boldsymbol{Z}^{n+1}\|^2 + C\|E^n\|^2_{L^4}\|\nabla R^n\|^2_{L^4} + C\|\nabla E^n\|^2 + C\|d_tR^n\|^2_{L^4}\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} \\
&\qquad + C\|\nabla\boldsymbol{Y}^{n+1}\|^2_{\boldsymbol{L}^4}\|\boldsymbol{Z}^{n+1}\|^2_{\boldsymbol{L}^4} + C\|\nabla\boldsymbol{Z}^{n+1}\|^2\|\boldsymbol{Y}^{n+1}\|^2_{\boldsymbol{L}^\infty} \\
&\qquad + C\|\nabla R^{n-1}\|^2_{L^4}\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^\infty}\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} + C\|\nabla\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^4}\|\boldsymbol{Z}^n\|^2_{\boldsymbol{L}^4} \\
&\qquad + C\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^\infty}\|\nabla\boldsymbol{Z}^n\|^2 =: \frac{1}{2}\|\sqrt{R^n}d_t\boldsymbol{Z}^{n+1}\|^2 + III_1 + III_2 + \ldots + III_8.
\end{aligned}$$

Again, every III_i can be estimated as follows, using some positive constants $\theta > 0$:

$$\begin{aligned}
III_1 &\le C\|\nabla R^n\|\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)\|E^n\|\Big(\|\nabla E^n\| + \|E^n\|\Big) \\
&\le \theta\|\nabla E^n\|^2 + C(\theta)\|\nabla R^n\|^2\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)^2\|E^n\|^2 \\
&\quad + C(\theta)\|\nabla R^n\|\Big(\|\Delta_h R^n\| + \|\nabla R^n\|\Big)\|E^n\|^2 \\
&=: \theta\|\nabla E^n\|^2 + C(\theta)L_1\|E^n\|^2, \\
III_2 &\le C_E\|\nabla E^n\|^2, \\
III_3 &\le C\|d_tR^n\|\Big(\|\nabla d_tR^n\| + \|d_tR^n\|\Big)\|\boldsymbol{Z}^n\|\|\nabla\boldsymbol{Z}^n\| \\
&\le \|\nabla\boldsymbol{Z}^n\|^2 + C\|d_tR^n\|^2\Big(\|\nabla d_tR^n\| + \|d_tR^n\|\Big)^2\|\boldsymbol{Z}^n\|^2 =: \|\nabla\boldsymbol{Z}^n\|^2 + CL_3\|\boldsymbol{Z}^n\|^2, \\
III_4 &\le \|\nabla\boldsymbol{Z}^{n+1}\|^2 + C\|\nabla\boldsymbol{Y}^{n+1}\|^2\|\tilde{\Delta}_h\boldsymbol{Y}^{n+1}\|^2\|\boldsymbol{Z}^{n+1}\|^2 =: \|\nabla\boldsymbol{Z}^{n+1}\|^2 + CL_4\|\boldsymbol{Z}^{n+1}\|^2, \\
III_5 &\le \|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^\infty}\|\nabla\boldsymbol{Z}^{n+1}\|^2 =: L_5\|\nabla\boldsymbol{Z}^{n+1}\|^2, \\
III_6 &\le C\|\nabla R^{n-1}\|\Big(\|\Delta_h R^{n-1}\| + \|\nabla R^{n-1}\|\Big)\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^\infty}\|\nabla\boldsymbol{Z}^n\|^2 =: CL_6\|\nabla\boldsymbol{Z}^n\|^2, \\
III_7 &\le C\|\nabla\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^4}\|\nabla\boldsymbol{Z}^n\|^2 =: CL_7\|\nabla\boldsymbol{Z}^n\|^2, \\
III_8 &\le C\|\boldsymbol{Y}^{n-1}\|^2_{\boldsymbol{L}^\infty}\|\nabla\boldsymbol{Z}^n\|^2 =: CL_8\|\nabla\boldsymbol{Z}^n\|^2.
\end{aligned}$$

As above, all L_i are such that $k\sum L_i \le C$. We now have shown the following estimate:

$$\begin{aligned}
&\|\sqrt{R^n}d_t\boldsymbol{Z}^{n+1}\|^2 - d_t\|\nabla\boldsymbol{Z}^{n+1}\|^2 \\
&\qquad\le C_E\|\nabla E^n\|^2 + \tilde{L}\|E^n\|^2 + \tilde{L}\Big(\|\boldsymbol{Z}^n\|^2 + \|\nabla\boldsymbol{Z}^n\|^2 + \|\boldsymbol{Z}^{n+1}\|^2\Big) \qquad (7.4)
\end{aligned}$$

Step 4. We consider now the sum $(2\varepsilon^{-1}C_E + 1)(7.3) + (7.4)$ and get for a appropriate choice of τ that

$$\begin{aligned}
&-d_t\|\sqrt{R^n}\boldsymbol{Z}^{n+1}\|^2 - d_t\|E^{n+1}\|^2 - d_t\|\nabla\boldsymbol{Z}^{n+1}\|^2 \\
&\quad + C\|\nabla\boldsymbol{Z}^n\|^2 + C\|\nabla E^n\|^2 + \|\sqrt{R^n}d_t\boldsymbol{Z}^{n+1}\|^2 \\
&\qquad\qquad \le \tilde{F}\|\boldsymbol{Z}^{n+1}\|^2 + \tilde{F}\|E^n\|^2 + \tilde{F}\|\nabla\boldsymbol{Z}^n\|^2
\end{aligned}$$

with some generic summable functions $\tilde{F}$ consisting of $\tilde{J}, \tilde{K}$, and $\tilde{L}$ from above. The discrete version of Gronwall's lemma leads then to the assertion of the lemma for a sufficiently small $k \le k_0(\Omega, \rho_{\min}, \rho_{\max}, T, \mu, \varepsilon)$. □

8. Convergence of the Scheme

In this chapter, we pass to the limit for the adjoint variables, thanks to the bounds from the previous chapter, and identify the limit with a weak solution of the corresponding continuous optimality system (4.1).

Lemma 8.1
There exist functions $\rho^* \in H^1(H^1) \cap L^2(H^2) \cap L^\infty(H^1)$, $\boldsymbol{y}^* \in L^\infty(\boldsymbol{V}) \cap L^2(\boldsymbol{H}^2) \cap H^1(\boldsymbol{H})$, $\eta^* \in H^1(H^{-1}) \cap L^2(H^1) \cap L^\infty(L^2)$, $\boldsymbol{z}^* \in H^1(\boldsymbol{H}) \cap L^\infty(\boldsymbol{H}) \cap L^2(\boldsymbol{V})$, $\boldsymbol{u}^* \in L^2(\boldsymbol{L}^2)$ such that a subsequence (not relabeled) of $\{R^n\}, \{\boldsymbol{Y}^n\}, \{E^n\}, \{\boldsymbol{Z}^n\}, \{\boldsymbol{U}^n\}$ converges to their counterparts in the following sense (for $h, k \to 0$):

$$
\begin{aligned}
\mathcal{R}^{\bullet/-}, \mathcal{R} &\overset{*}{\rightharpoonup} \rho^* && \text{weakly star in } L^\infty(L^\infty),\\
\mathcal{R}^{\bullet/-}, \mathcal{R} &\to \rho^* && \text{strongly in } L^2(H^1),\\
\mathcal{R}^{\bullet/-}, \mathcal{R} &\rightharpoonup \rho^* && \text{weakly in } H^1(L^2),\\
\operatorname{div} \boldsymbol{\mathcal{Y}}^{\bullet/-}, \operatorname{div} \boldsymbol{\mathcal{Y}} &\to 0 && \text{strongly in } L^2(\boldsymbol{L}^2),\\
\boldsymbol{\mathcal{Y}}^{\bullet/-}, \boldsymbol{\mathcal{Y}} &\rightharpoonup \boldsymbol{y}^* && \text{weakly in } H^1(\boldsymbol{L}^2),\\
\boldsymbol{\mathcal{Y}}^{\bullet/-}, \boldsymbol{\mathcal{Y}} &\overset{*}{\rightharpoonup} \boldsymbol{y}^* && \text{weakly star in } L^\infty(\boldsymbol{V}),\\
\boldsymbol{\mathcal{Y}}^{\bullet/-}, \boldsymbol{\mathcal{Y}} &\to \boldsymbol{y}^* && \text{strongly in } L^2(\boldsymbol{H}^1),\\
\mathcal{E}^{+/\bullet}, \mathcal{E} &\rightharpoonup \eta^* && \text{weakly in } L^2(H^1),\\
\mathcal{E}^{+/\bullet}, \mathcal{E} &\to \eta^* && \text{strongly in } L^2(L^2),\\
\boldsymbol{\mathcal{Z}}^{+/\bullet}, \boldsymbol{\mathcal{Z}} &\rightharpoonup \boldsymbol{z}^* && \text{weakly in } H^1(\boldsymbol{L}^2),\\
\boldsymbol{\mathcal{Z}}^{+/\bullet}, \boldsymbol{\mathcal{Z}} &\rightharpoonup \boldsymbol{z}^* && \text{weakly in } L^2(\boldsymbol{H}^1),\\
\boldsymbol{\mathcal{Z}}^{+/\bullet}, \boldsymbol{\mathcal{Z}} &\to \boldsymbol{z}^* && \text{strongly in } L^2(\boldsymbol{L}^2),\\
\operatorname{div} \boldsymbol{\mathcal{Z}}^{+/\bullet}, \operatorname{div} \boldsymbol{\mathcal{Z}} &\rightharpoonup 0 && \text{weakly in } L^2(\boldsymbol{L}^2),\\
\boldsymbol{\mathcal{U}}^{+/\bullet}, \boldsymbol{\mathcal{U}} &\rightharpoonup \boldsymbol{u}^* && \text{weakly in } L^2(\boldsymbol{L}^2).
\end{aligned}
$$

Proof
The weak convergence and weak-star convergence, respectively, follows from Lemmas 5.1, 5.2, 5.3, 5.5, and 7.1. The strong convergence for the affine time interpolants are obtained by the bounds and Lemma 5.4. Since increments are bounded (see Lemmas 5.2, 5.3, and 7.1), all constant in time interpolants inherit the strong convergence as already mentioned at the end of section 5.1. □

These convergence properties are sufficient, such that all linear terms in (6.2) will directly converge to their continuous counterparts in a weak sense. It remains to identify weak convergence of all nonlinear parts, where we use in particular the strong convergence results from above. For a simpler notation, we drop the stars from the limits derived above. By a standard density argument, it is enough to identify the limit for used smooth test functions.

Lemma 8.2
Let $\boldsymbol{\varphi} \in \mathcal{C}_0^\infty([0,T), \boldsymbol{\mathcal{C}}_0^\infty)$ or $\varphi \in \mathcal{C}_0^\infty([0,T), \mathcal{C}^\infty)$ respectively. We have

1. $(\nabla\mathcal{R}^\bullet\mathcal{E}^\bullet - \nabla\rho\eta, \boldsymbol{\varphi}) \to 0.$
2. $(\nabla\mathcal{E}^\bullet\mathcal{R}^\bullet - \nabla\eta\rho, \boldsymbol{\varphi}) \to 0.$
3. $(\mathcal{R}^\bullet\boldsymbol{\mathcal{Z}}_t - \rho\boldsymbol{z}_t, \boldsymbol{\varphi}) \to 0.$
4. $(\mathcal{R}_t\boldsymbol{\mathcal{Z}}^+ - \rho_t\boldsymbol{z}, \boldsymbol{\varphi}) \to 0.$
5. $(\nabla\mathcal{R}^+ \cdot \boldsymbol{\mathcal{Y}}^+\boldsymbol{\mathcal{Z}}^\bullet - \nabla\rho \cdot \boldsymbol{y}\boldsymbol{z}, \boldsymbol{\varphi}) \to 0.$
6. $(\mathcal{R}^- \operatorname{div}\boldsymbol{\mathcal{Y}}^-\boldsymbol{\mathcal{Z}}^\bullet - \rho\operatorname{div}\boldsymbol{y}\boldsymbol{z}, \boldsymbol{\varphi}) = (\mathcal{R}^-\operatorname{div}\boldsymbol{\mathcal{Y}}^-\boldsymbol{\mathcal{Z}}^\bullet, \boldsymbol{\varphi}) \to 0.$
7. $([\mathcal{R}^-\boldsymbol{\mathcal{Y}}^- \cdot \nabla]\boldsymbol{\mathcal{Z}}^\bullet - [\rho\boldsymbol{y} \cdot \nabla]\boldsymbol{z}, \boldsymbol{\varphi}) \to 0.$
8. $(\mathcal{R}^\bullet\nabla\boldsymbol{\mathcal{Z}}^+\boldsymbol{\mathcal{Y}}^+ - \rho\nabla\boldsymbol{z}\boldsymbol{y}, \boldsymbol{\varphi}) \to 0.$
9. $(\mathcal{R}^\bullet\nabla\boldsymbol{\mathcal{Y}}^+\boldsymbol{\mathcal{Z}}^+ - \rho\nabla\boldsymbol{y}\boldsymbol{z}, \boldsymbol{\varphi}) \to 0.$
10. $([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\mathcal{E}^\bullet - [\boldsymbol{y} \cdot \nabla]\eta, \varphi) \to 0.$
11. $(\operatorname{div}\boldsymbol{\mathcal{Y}}^\bullet\mathcal{E}^\bullet - \operatorname{div}\boldsymbol{y}\eta, \varphi) = (\operatorname{div}\boldsymbol{\mathcal{Y}}^\bullet\mathcal{E}^\bullet, \varphi) \to 0.$
12. $(\boldsymbol{\mathcal{Y}}_t \cdot \boldsymbol{\mathcal{Z}}^+ - \boldsymbol{y}_t \cdot \boldsymbol{z}, \varphi) \to 0.$
13. $(\boldsymbol{\mathcal{Y}}^\bullet \cdot \boldsymbol{\mathcal{Z}}_t - \boldsymbol{y} \cdot \boldsymbol{z}_t, \varphi) \to 0.$
14. $([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\boldsymbol{\mathcal{Y}}^+ \cdot \boldsymbol{\mathcal{Z}}^+ - [\boldsymbol{y} \cdot \nabla]\boldsymbol{y} \cdot \boldsymbol{z}, \varphi) \to 0.$
15. $([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Y}}^+ - [\boldsymbol{y} \cdot \nabla]\boldsymbol{z} \cdot \boldsymbol{y}, \varphi) \to 0.$
16. $(\boldsymbol{\mathcal{U}}^+ \cdot \boldsymbol{\mathcal{Z}}^+ - \boldsymbol{u} \cdot \boldsymbol{z}, \varphi) \to 0.$

PROOF

1. We write

$$(\nabla\mathcal{R}^\bullet\mathcal{E}^\bullet - \nabla\rho\eta, \boldsymbol{\varphi}) = (\nabla\mathcal{R}^\bullet(\mathcal{E}^\bullet - \eta), \boldsymbol{\varphi}) + (\nabla(\mathcal{R}^\bullet - \rho)\eta, \boldsymbol{\varphi}) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\nabla\mathcal{R}^\bullet\|_{L^2(L^2)}\|\mathcal{E}^\bullet - \eta\|_{L^2(L^2)}\|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0, \\ II &= (\nabla(\mathcal{R}^\bullet - \rho), \eta\boldsymbol{\varphi}) \to 0. \end{aligned}$$

2. We write

$$(\nabla\mathcal{E}^\bullet\mathcal{R}^\bullet - \nabla\eta\rho, \boldsymbol{\varphi}) = (\nabla\mathcal{E}^\bullet(\mathcal{R}^\bullet - \rho), \boldsymbol{\varphi}) + (\nabla(\mathcal{E}^\bullet - \eta)\rho, \boldsymbol{\varphi}) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\nabla\mathcal{E}^\bullet\|_{L^2(L^2)}\|\mathcal{R}^\bullet - \rho\|_{L^2(L^2)}\|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0, \\ II &= (\nabla(\mathcal{E}^\bullet - \eta), \rho\boldsymbol{\varphi}) \to 0. \end{aligned}$$

3. We write

$$(\mathcal{R}^\bullet \boldsymbol{\mathcal{Z}}_t - \rho \boldsymbol{z}_t, \boldsymbol{\varphi}) = ((\mathcal{R}^\bullet - \rho)\boldsymbol{\mathcal{Z}}_t, \boldsymbol{\varphi}) + (\rho(\boldsymbol{\mathcal{Z}} - \boldsymbol{z})_t, \boldsymbol{\varphi}) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\boldsymbol{\varphi}(\mathcal{R}^\bullet - \rho)\|_{L^2(\boldsymbol{H}^1)} \|\boldsymbol{\mathcal{Z}}_t\|_{L^2(\boldsymbol{H}^{-1})} \le \|\mathcal{R}^\bullet - \rho\|_{L^2(H^1)} \|\boldsymbol{\mathcal{Z}}_t\|_{L^2(\boldsymbol{H}^{-1})} \|\nabla \boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0, \\ II &= ((\boldsymbol{\mathcal{Z}} - \boldsymbol{z})_t, \rho \boldsymbol{\varphi}) \to 0. \end{aligned}$$

4. We write

$$(\mathcal{R}_t \boldsymbol{\mathcal{Z}}^+ - \rho_t \boldsymbol{z}, \boldsymbol{\varphi}) = (\mathcal{R}_t(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}), \boldsymbol{\varphi}) + ((\mathcal{R} - \rho)_t \boldsymbol{z}, \boldsymbol{\varphi}) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\mathcal{R}_t\|_{L^2(L^2)} \|\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0, \\ II &= ((\mathcal{R} - \rho)_t, \boldsymbol{z} \cdot \boldsymbol{\varphi}) \to 0. \end{aligned}$$

5. We write

$$\begin{aligned} &(\nabla \mathcal{R}^- \cdot \boldsymbol{\mathcal{Y}}^- \boldsymbol{\mathcal{Z}}^\bullet - \nabla \rho \cdot \boldsymbol{y} \boldsymbol{z}, \boldsymbol{\varphi}) \\ &= (\nabla(\mathcal{R}^- - \rho) \cdot \boldsymbol{y} \boldsymbol{z}, \boldsymbol{\varphi}) + (\nabla \mathcal{R}^- (\boldsymbol{\mathcal{Y}}^- - \boldsymbol{y}) \boldsymbol{z}, \boldsymbol{\varphi}) + (\nabla \mathcal{R}^- \cdot \boldsymbol{\mathcal{Y}}^- (\boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z}), \boldsymbol{\varphi}) =: I + II + III \end{aligned}$$

and calculate

$$\begin{aligned} I &= (\nabla(\mathcal{R}^- - \rho), \boldsymbol{y} \cdot \boldsymbol{z} \boldsymbol{\varphi}) \to 0, \\ II &\le \|\nabla \mathcal{R}^-\|_{L^\infty(L^2)} \|\boldsymbol{\mathcal{Y}}^- - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^4)} \|\boldsymbol{z}\|_{L^2(\boldsymbol{L}^4)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0, \\ III &\le \|\nabla \mathcal{R}^-\|_{L^2(L^4)} \|\boldsymbol{\mathcal{Y}}^-\|_{L^\infty(L^4)} \|\boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \\ &\le \|\nabla \mathcal{R}^-\|_{L^2(L^2)} \Big(\|\Delta_h \mathcal{R}^-\|_{L^2(L^2)} + \|\nabla \mathcal{R}^-\|_{L^2(L^2)} \Big) \\ &\quad \times \|\boldsymbol{\mathcal{Y}}^-\|_{L^\infty(L^4)} \|\boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0. \end{aligned}$$

6. We calculate

$$(\mathcal{R}^- \operatorname{div} \boldsymbol{\mathcal{Y}}^- \boldsymbol{\mathcal{Z}}^\bullet, \boldsymbol{\varphi}) \le \|\mathcal{R}^-\|_{L^\infty(L^\infty)} \|\operatorname{div} \boldsymbol{\mathcal{Y}}^-\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}^\bullet\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0.$$

7. We write

$$\begin{aligned} &([\mathcal{R}^- \boldsymbol{\mathcal{Y}}^- \cdot \nabla] \boldsymbol{\mathcal{Z}}^\bullet - [\rho \boldsymbol{y} \cdot \nabla] \boldsymbol{z}, \boldsymbol{\varphi}) \\ &\qquad = -(\nabla \mathcal{R}^- \cdot \boldsymbol{\mathcal{Y}}^- \boldsymbol{\mathcal{Z}}^\bullet - \nabla \rho \cdot \boldsymbol{y} \boldsymbol{z}, \boldsymbol{\varphi}) - (\mathcal{R}^- \operatorname{div} \boldsymbol{\mathcal{Y}}^- \boldsymbol{\mathcal{Z}}^\bullet - \rho \operatorname{div} \boldsymbol{y} \boldsymbol{z}, \boldsymbol{\varphi}) \\ &\qquad\quad - \Big(([\mathcal{R}^- \boldsymbol{\mathcal{Y}}^- \cdot \nabla] \boldsymbol{\varphi}, \boldsymbol{\mathcal{Z}}^\bullet) - ([\rho \boldsymbol{y} \cdot \nabla] \boldsymbol{\varphi}, \boldsymbol{z}) \Big) =: -I - II - III. \end{aligned}$$

The terms I and II are estimated in the last two parts. We rewrite III in the following way:

$$III = ([\mathcal{R}^- (\boldsymbol{\mathcal{Y}}^- - \boldsymbol{y}) \cdot \nabla] \boldsymbol{\varphi}, \boldsymbol{\mathcal{Z}}^\bullet) + ([(\mathcal{R}^- - \rho) \boldsymbol{y} \cdot \nabla] \boldsymbol{\varphi}, \boldsymbol{\mathcal{Z}}^\bullet) + ([\rho \boldsymbol{y} \cdot \nabla] \boldsymbol{\varphi}, \boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z})$$

$$=: III_a + III_b + III_c.$$

The three terms can be estimated as follows:

$$\begin{aligned}
III_a &\le \|\mathcal{R}^-\|_{L^\infty(L^\infty)} \|\boldsymbol{\mathcal{Y}}^- - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\nabla\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \|\boldsymbol{\mathcal{Z}}^\bullet\|_{L^2(\boldsymbol{L}^2)} \to 0,\\
III_b &\le \|\mathcal{R}^- - \rho\|_{L^2(L^4)} \|\boldsymbol{y}\|_{L^\infty(\boldsymbol{L}^4)} \|\nabla\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \|\boldsymbol{\mathcal{Z}}^\bullet\|_{L^\infty(\boldsymbol{L}^2)}\\
&\le C\|\nabla(\mathcal{R}^- - \rho)\|_{L^2(L^2)} \|\nabla\boldsymbol{y}\|_{L^\infty(\boldsymbol{L}^2)} \|\nabla\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \|\boldsymbol{\mathcal{Z}}^\bullet\|_{L^\infty(\boldsymbol{L}^2)} \to 0,\\
III_c &\le \|\mathcal{R}^-\|_{L^\infty(L^\infty)} \|\boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\nabla\boldsymbol{\varphi}\|_{L^\infty(L^\infty)} \|\boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \to 0.
\end{aligned}$$

8. We write

$$\begin{aligned}
&(\mathcal{R}^\bullet \nabla\boldsymbol{\mathcal{Z}}^+\boldsymbol{\mathcal{Y}}^+ - \rho\nabla\boldsymbol{z}\boldsymbol{y}, \boldsymbol{\varphi})\\
&= \Big((\mathcal{R}^\bullet - \rho)\nabla\boldsymbol{\mathcal{Z}}^+\boldsymbol{\mathcal{Y}}^+, \boldsymbol{\varphi}\Big) + \Big(\rho\nabla(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z})\boldsymbol{y}, \boldsymbol{\varphi}\Big) + \Big(\rho\nabla\boldsymbol{\mathcal{Z}}^+(\boldsymbol{\mathcal{Y}}^+ - \boldsymbol{y}), \boldsymbol{\varphi}\Big) =: I + II + III,
\end{aligned}$$

and by integration by parts,

$$II = -(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}, [\nabla\rho\cdot\boldsymbol{\varphi}]\boldsymbol{y}) - (\rho\nabla\boldsymbol{y}(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}), \boldsymbol{\varphi}) - (\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}, \rho \operatorname{div}\boldsymbol{\varphi}\boldsymbol{y}) =: II_a + II_b + II_c.$$

We calculate

$$\begin{aligned}
I &\le \|\mathcal{R}^\bullet - \rho\|_{L^2(L^4)} \|\nabla\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Y}}^+\|_{L^\infty(\boldsymbol{L}^4)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0,\\
II_a &\to 0,\\
II_b &\le \|\rho\|_{L^\infty(L^\infty)} \|\nabla\boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0,\\
II_c &\to 0,\\
III &\le \|\rho\|_{L^\infty(L^\infty)} \|\nabla\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Y}}^+ - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(\boldsymbol{L}^\infty)} \to 0.
\end{aligned}$$

9. This estimate is just the same like the last one except that $\boldsymbol{\mathcal{Z}}^+$ and $\boldsymbol{\mathcal{Y}}^+$ are interchanged and the derivative affects $\boldsymbol{\mathcal{Y}}^+$, which has an improved regularity in contrast to $\boldsymbol{\mathcal{Z}}^+$.

10. We write

$$([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\mathcal{E}^\bullet - [\boldsymbol{y}\cdot\nabla]\eta, \varphi) = ([(\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y})\cdot\nabla]\mathcal{E}^\bullet, \varphi) + ([\boldsymbol{y}\cdot\nabla](\mathcal{E}^\bullet - \eta), \varphi) =: I + II$$

and calculate

$$\begin{aligned}
I &\le \|\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\nabla\mathcal{E}^\bullet\|_{L^2(L^2)} \|\varphi\|_{L^\infty(L^\infty)} \to 0,\\
II &= -([\boldsymbol{y}\cdot\nabla]\varphi, \mathcal{E}^\bullet - \eta) \to 0,
\end{aligned}$$

where we used $\operatorname{div}\boldsymbol{y} = 0$.

11. We calculate

$$(\operatorname{div}\boldsymbol{\mathcal{Y}}^\bullet\mathcal{E}^\bullet, \varphi) \le \|\operatorname{div}\boldsymbol{\mathcal{Y}}^\bullet\|_{L^2(\boldsymbol{L}^2)} \|\mathcal{E}^\bullet\|_{L^2(L^2)} \|\varphi\|_{L^\infty(L^\infty)} \to 0.$$

12. We write

$$(\boldsymbol{\mathcal{Y}}_t \cdot \boldsymbol{\mathcal{Z}}^+ - \boldsymbol{y}_t \cdot \boldsymbol{z}, \varphi) = (\boldsymbol{\mathcal{Y}}_t \cdot (\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}), \varphi) + ((\boldsymbol{\mathcal{Y}} - \boldsymbol{y})_t \cdot \boldsymbol{z}, \varphi) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\boldsymbol{\mathcal{Y}}_t\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(L^\infty)} \to 0, \\ II &= ((\boldsymbol{\mathcal{Y}} - \boldsymbol{y})_t, \varphi \boldsymbol{z}) \to 0. \end{aligned}$$

13. We write

$$(\boldsymbol{\mathcal{Y}}^\bullet \cdot \boldsymbol{\mathcal{Z}}_t - \boldsymbol{y} \cdot \boldsymbol{z}_t, \varphi) = \Big((\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y}) \boldsymbol{\mathcal{Z}}_t, \varphi\Big) + \Big(\boldsymbol{y} \cdot (\boldsymbol{\mathcal{Z}} - \boldsymbol{z})_t, \varphi\Big) =: I + II$$

and calculate

$$\begin{aligned} I &\le \|\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}_t\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\varphi}\|_{L^\infty(L^\infty)} \to 0, \\ II &= ((\boldsymbol{\mathcal{Z}} - \boldsymbol{z})_t, \varphi \boldsymbol{y}) \to 0. \end{aligned}$$

14. We write

$$\begin{aligned} &([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\boldsymbol{\mathcal{Y}}^+ \cdot \boldsymbol{\mathcal{Z}}^+ - [\boldsymbol{y} \cdot \nabla]\boldsymbol{y} \cdot \boldsymbol{z}, \varphi) \\ &\qquad = \Big([(\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y}) \cdot \nabla]\boldsymbol{\mathcal{Y}}^+ \cdot \boldsymbol{\mathcal{Z}}^+, \varphi\Big) + \Big([\boldsymbol{y} \cdot \nabla](\boldsymbol{\mathcal{Y}}^+ - \boldsymbol{y}) \cdot \boldsymbol{\mathcal{Z}}^+, \varphi\Big) \\ &\qquad\quad + \Big([\boldsymbol{y} \cdot \nabla]\boldsymbol{y} \cdot (\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}), \varphi\Big) =: I + II + III \end{aligned}$$

and calculate

$$\begin{aligned} I &\le \|\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y}\|_{L^2(\boldsymbol{L}^4)} \|\nabla \boldsymbol{\mathcal{Y}}^+\|_{L^\infty(L^2)} \|\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^4)} \|\varphi\|_{L^\infty(L^\infty)} \\ &\le C\|\nabla(\boldsymbol{\mathcal{Y}}^\bullet - \boldsymbol{y})\|_{L^2(\boldsymbol{L}^2)} \|\nabla \boldsymbol{\mathcal{Y}}^+\|_{L^\infty(L^2)} \|\nabla \boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)} \|\varphi\|_{L^\infty(L^\infty)} \to 0, \\ II &\le \|\boldsymbol{y}\|_{L^\infty(\boldsymbol{L}^4)} \|\nabla(\boldsymbol{\mathcal{Y}}^+ - \boldsymbol{y})\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^4)} \|\varphi\|_{L^\infty(L^\infty)} \\ &\le C\|\nabla \boldsymbol{y}\|_{L^\infty(\boldsymbol{L}^2)} \|\nabla(\boldsymbol{\mathcal{Y}}^+ - \boldsymbol{y})\|_{L^2(\boldsymbol{L}^2)} \|\nabla \boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)} \|\varphi\|_{L^\infty(L^\infty)} \to 0, \\ III &= \Big(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}, \varphi \boldsymbol{y} \cdot \nabla]\boldsymbol{y}\Big) \to 0. \end{aligned}$$

15. Since we have

$$([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Y}}^+, \varphi) = (\operatorname{div} \boldsymbol{\mathcal{Y}}^\bullet \boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Y}}^+, \varphi) + ([\boldsymbol{\mathcal{Y}}^\bullet \cdot \nabla]\boldsymbol{\mathcal{Y}}^+ \cdot \boldsymbol{\mathcal{Z}}^+, \varphi) \qquad \square$$

and the same for the continuous counterpart, we have only to show $(\operatorname{div} \boldsymbol{\mathcal{Y}}^\bullet \boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Y}}^+, \varphi) \to 0$. The second term has been dealt with in step 14 of the proof. We calculate

$$\begin{aligned} (\operatorname{div} \boldsymbol{\mathcal{Y}}^\bullet \boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Y}}^+, \varphi) &\le \|\operatorname{div} \boldsymbol{\mathcal{Y}}^\bullet\|_{L^2(\boldsymbol{L}^2)} \|\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^4)} \|\boldsymbol{\mathcal{Y}}^+\|_{L^\infty(\boldsymbol{L}^4)} \|\varphi\|_{L^\infty(L^\infty)} \\ &\le C\|\operatorname{div} \boldsymbol{\mathcal{Y}}^\bullet\|_{L^2(\boldsymbol{L}^2)} \|\nabla \boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)} \|\nabla \boldsymbol{\mathcal{Y}}^+\|_{L^\infty(\boldsymbol{L}^2)} \|\varphi\|_{L^\infty(L^\infty)} \to 0. \end{aligned}$$

16. With (6.2d) and (4.1d) respectively, we rewrite the terms to

$$\alpha(\boldsymbol{\mathcal{U}}^+ \cdot \boldsymbol{\mathcal{Z}}^+ - \boldsymbol{u}\cdot\boldsymbol{z},\varphi) = (\mathcal{R}^\bullet \boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Z}}^+ - \rho\boldsymbol{z}\cdot\boldsymbol{z},\varphi)$$
$$= \Big((\mathcal{R}^\bullet - \rho)\boldsymbol{\mathcal{Z}}^+ \cdot \boldsymbol{\mathcal{Z}}^+,\varphi\Big) + (\rho(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z})\cdot\boldsymbol{\mathcal{Z}}^+,\varphi) + (\rho\boldsymbol{z}\cdot(\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}),\varphi) =: I + II + III.$$

We calculate finally

$$\begin{aligned} I &\leq \|\mathcal{R}^\bullet - \rho\|_{L^2(L^4)}\|\boldsymbol{\mathcal{Z}}^+\|_{L^\infty(\boldsymbol{L}^2)}\|\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^4)}\|\varphi\|_{L^\infty(L^\infty)} \\ &\leq C\|\nabla(\mathcal{R}^\bullet - \rho)\|_{L^2(L^2)}\|\boldsymbol{\mathcal{Z}}^+\|_{L^\infty(\boldsymbol{L}^2)}\|\nabla\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)}\|\varphi\|_{L^\infty(L^\infty)} \to 0, \\ II &\leq \|\rho\|_{L^\infty(L^\infty)}\|\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)}\|\boldsymbol{\mathcal{Z}}^+\|_{L^2(\boldsymbol{L}^2)}\|\varphi\|_{L^\infty(L^\infty)} \to 0, \\ III &\leq \|\rho\|_{L^\infty(L^\infty)}\|\boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)}\|\boldsymbol{\mathcal{Z}}^+ - \boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)}\|\varphi\|_{L^\infty(L^\infty)} \to 0. \end{aligned}$$

Theorem 8.3
The functions $\rho^* \in H^1(H^1) \cap L^\infty(H^2)$, $\boldsymbol{y}^* \in L^\infty(\boldsymbol{V}) \cap L^2(\boldsymbol{H}^2) \cap H^1(\boldsymbol{H})$, $\eta^* \in H^1(H^{-1}) \cap L^\infty(H^1)$, $\boldsymbol{z}^* \in H^1(\boldsymbol{H}) \cap L^\infty(\boldsymbol{H}) \cap L^2(\boldsymbol{V})$, $\boldsymbol{u}^* \in L^2(\boldsymbol{L}^2)$ obtained in Lemma 8.1 solve (4.1).

Proof
With the first nine parts of Lemma 8.2, we deduce that solutions of (6.2a) converge to solutions of (4.1a) (up to subsequences). With the remaining parts of Lemma 8.2, we deduce that solutions of (6.2c) converge to solutions of (4.1c). All other parts in those equations are linear and we can pass to their limits by the weak convergence results obtained in Lemma 8.1. □

Theorem 8.4
For the function $\boldsymbol{u}^* \in L^2(\boldsymbol{L}^2)$ obtained in Lemma 8.1 we have $\boldsymbol{\mathcal{U}}^\bullet, \boldsymbol{\mathcal{U}}^+ \to \boldsymbol{u}^*$ strongly in $L^2(\boldsymbol{L}^2)$ (up to a subsequence) for $h, k \to 0$.

Proof
We use (4.1d) and (6.2d) to rewrite $\boldsymbol{\mathcal{U}}^\bullet$ and u, respectively. We have then to show that

$$0 \leftarrow \|u - \mathcal{U}^+\|_{L^2(\boldsymbol{L}^2)} = \|\mathcal{R}^-\boldsymbol{\mathcal{Z}}^\bullet - \rho\boldsymbol{z}\|_{L^2(\boldsymbol{L}^2)} = \|(\boldsymbol{\mathcal{Z}}^\bullet - \boldsymbol{z})\rho\|_{L^2(\boldsymbol{L}^2)} + \|(\mathcal{R}^-\rho)\boldsymbol{\mathcal{Z}}^\bullet\|_{L^2(\boldsymbol{L}^2)}.$$

For the first term, we use $L^\infty(L^\infty)$ bounds on ρ and the strong convergence of $\boldsymbol{Z}^\bullet \to \boldsymbol{z}$ in $L^2(\boldsymbol{L}^2)$.

For the second term, we use the Aubin–Lions lemma for the spaces $H^2(\Omega) \hookrightarrow\hookrightarrow W^{1,4}(\Omega) \hookrightarrow H^1(\Omega)$ and the bounds on $\mathcal{R}^-$ in order to have strong convergence $\mathcal{R}^- \to \rho$ in $L^2(W^{1,4}) \subseteq L^2(L^\infty)$. This leads to convergence in the second term together with the bound of $\boldsymbol{\mathcal{Z}}^\bullet$ in $L^\infty(\boldsymbol{L}^2)$. □

9. Computational studies

In this chapter we present a numerical algorithm and perform computational studies for the system (5.5)–(6.2). The numerical experiments in this chapter were implemented by one coauthor in [9].

9.1. Implementation details and minimization algorithm

We use the Taylor–Hood finite element pair with vector-valued quadratic finite elements for $\boldsymbol{V}_h$, standard continuous piecewise linear elements for M_h, and standard linear finite elements for R_h as discussed in chapter 5. It then follows from (6.2d) that for $n = 1, \ldots, N$ the control variable $\boldsymbol{U}^n$ belongs to the space of vector-valued conforming piecewise continuous finite elements of degree three.

Due to the advective character of the mass equation, without proper stabilization the numerical solution may exhibit oscillations near the interface. Our experience suggests that one should choose ε_h in the stabilization term $-\varepsilon_h \Delta_h R^n$ to be dependent on the mesh size and time step values. However, an optimal choice of ε_h for this type of stabilization is not clear. We choose $\varepsilon_h|_K = \xi h_K$, where h_K is the diameter of the mesh element K. The constant $\xi > 0$ is chosen a priori to minimize the oscillations. For fixed values of $\xi > 0$ we still observe oscillations in the values of the computed density in some experiments. In the majority of cases the oscillations exceeded only $2\% - 5\%$ of the initial densities ρ_1, ρ_2.

Remark 9.1
The oscillations in the density values can always be eliminated by choosing a sufficiently large parameter $\xi = \|\boldsymbol{\mathcal{Y}}\|_{L^\infty(\boldsymbol{L}^\infty)}$ in $-\varepsilon_h \Delta_h R^n$. A more sophisticated alternative is to use an upwind scheme for the mass equation. This corresponds to a space-time adaptive choice of the artificial diffusion parameter $\varepsilon_h|_K = \frac{h_K}{2} \|\boldsymbol{Y}\|_{\boldsymbol{L}^\infty(K)}$ (along with an appropriate modification of the dual problem); cf. [12, Section 5.1.3]. The upwind discretization preserves monotonicity of the computed density $\mathcal{R}$ and introduces less artificial diffusion in the numerical solution than a constant diffusion parameter. However, we experienced difficulties with the convergence of the minimization algorithm for the upwind discretization. In addition, an adaptive choice of the diffusion parameter may introduce an artificial control mechanism into the system.

The numerical experiments have been performed for different initial conditions for the density; the initial condition for the velocity was always set to zero. We prescribe homogeneous Dirichlet boundary conditions for (5.5b)–(6.2a); for (5.5a), (6.2c) we prescribe homogeneous Neumann boundary conditions. We use a fixed time step size and employ a simple mesh refinement strategy. In the regions of the computational domain where

$\rho_1 + 10^{-3} < R^n \leq \rho_2 - 10^{-3}$, for some $n = 1, \ldots, N$ (i.e., along the interfacial region) we use a fine mesh size $h = h_{\min}$ (in most experiments $h_{\min} = 1/64$), and elsewhere we use a coarse mesh size $h_{\max} = 1/16$. We do not perform any mesh coarsening. Note, that in general the minimum mesh size $h_{\min}$ should be chosen according to the thickness of the diffuse interface which is determined by the parameter δ; cf. [10, 11, 13].

A standard approach (see, e.g., [46]) is to express $\mathcal{R} \equiv \mathcal{R}(\boldsymbol{\mathcal{U}})$. Given $h > 0$, we note that solutions are unique by a contraction argument for sufficiently small time step sizes $k > 0$. The dependence of the functional (6.1) on the density $\mathcal{R}$ can therefore be eliminated, and the functional can be rewritten as

$$\hat{J}_{\delta,h,k}(\boldsymbol{\mathcal{U}}) = J_{\delta,h,k}(\mathcal{R}(\boldsymbol{\mathcal{U}}), \boldsymbol{\mathcal{U}})\,.$$

The gradient of $\hat{J}_{\delta,h,k}$ is equivalent to (6.2d), i.e.,

$$\nabla_{\boldsymbol{\mathcal{U}}} \hat{J}_{\delta,h,k} = \alpha \boldsymbol{\mathcal{U}} - \boldsymbol{\mathcal{R}}^{-} \boldsymbol{\mathcal{Z}}.$$

We then look for the minimum of the reduced functional $\hat{J}_{\delta,h,k}$, which is equivalent to finding a solution of the system (5.5)–(6.2).

To minimize the functional $\hat{J}_{\delta,h,k}$ we employ a modified gradient descent algorithm with adaptive step size according to the Barzilai and Borwein criterion [14].

1. Choose the constants σ_{init}, $\sigma_{\min}$, $\sigma_{\max}$, σ_*, δ_{TOL}, set $\boldsymbol{\mathcal{U}}_0 = \mathbf{0}$.
2. Iterate for $k = 0, \ldots$.
   a) If $k < 1$ set $\sigma_k = \sigma_{\text{init}}$; for $k \geq 1$ choose step size
   $$\sigma_k = \sigma_* \frac{\int_0^T \int_\Omega (\boldsymbol{\mathcal{S}}_k, \boldsymbol{\mathcal{W}}_k)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t}{\|\boldsymbol{\mathcal{W}}_k\|^2_{L^2(\boldsymbol{L}^2)}},$$
   where $\boldsymbol{\mathcal{S}}_k = \boldsymbol{\mathcal{U}}_k - \boldsymbol{\mathcal{U}}_{k-1}$ and $\boldsymbol{\mathcal{W}}_k = \boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_{k-1}$, $\boldsymbol{\mathcal{G}}_k = \nabla_{\boldsymbol{\mathcal{U}}} \hat{J}(\boldsymbol{\mathcal{U}}_k)$.

   b) If $\sigma_k > \sigma_{\max}$ or $\sigma_k < 0$, set $\sigma_k = \sigma_{\min}$.

   c) Compute
   $$\boldsymbol{\mathcal{U}}_{k+1} = \boldsymbol{\mathcal{U}}_k - \sigma_k \nabla_{\boldsymbol{\mathcal{U}}} \hat{J}(\boldsymbol{\mathcal{U}}_k)$$
3. Stop when $\|\nabla_{\boldsymbol{\mathcal{U}}} \hat{J}\|^2_{L^2(\boldsymbol{L}^2)} < \delta_{TOL}$.

In the computational experiments below we choose $\sigma_{\text{init}} = 10^{-4}$, $\sigma_{\min} = 100$, $\sigma_{\max} = 20\sigma_{\min}$, $\sigma_* = 0.2$, $\delta_{TOL} = 10^{-9}$. For the given tolerance δ_{TOL} the gradient algorithm converges after 200–800 steps in all numerical experiments.

We found that the above algorithm with suitably chosen parameters σ_{init}, $\sigma_{\min}$, $\sigma_{\max}$, σ_* was more efficient for the given problem of this chapter than a corresponding exact line search algorithm, where the optimal step size was determined by bisection.

Note, that we use the steepest descent method for its simplicity. More advanced algorithms, such as SQP, could provide better performance for our problem; cf. [41] for an overview of different algorithms. However, the problem with the SQP algorithm is that a coupled system for solving (5.5) and (6.2) has to be solved for all times steps simultaneously due to the parabolic nature of these equations, which leads to very large system matrices, which is a difficult computational task.

Remark 9.2
To evaluate the gradient $\nabla_{\boldsymbol{\mathcal{U}}} \hat{J}(\boldsymbol{\mathcal{U}}_k) = \alpha \boldsymbol{\mathcal{U}}_k - \boldsymbol{\mathcal{R}}_k^- \boldsymbol{\mathcal{Z}}_k$ requires the values of $\boldsymbol{\mathcal{U}}_k$, $\boldsymbol{\mathcal{R}}_k^-$, $\boldsymbol{\mathcal{Z}}_k$. The control $\boldsymbol{\mathcal{U}}_k$ is known from the previous iteration of the gradient algorithm. The value of $\boldsymbol{\mathcal{R}}_k^-$ (along with the value of $\boldsymbol{\mathcal{Y}}_k$) is obtained by solving the forward equations (5.5) in space and time with the control $\boldsymbol{\mathcal{U}} \equiv \boldsymbol{\mathcal{U}}_k$ in (5.5b). Then, the dual variable $\boldsymbol{\mathcal{Z}}_k$ is obtained by solving the backward problem (6.2) with $\boldsymbol{\mathcal{Y}} \equiv \boldsymbol{\mathcal{Y}}_k$, $\boldsymbol{\mathcal{R}} \equiv \boldsymbol{\mathcal{R}}_k$.

9.2. Effect of the phase-field term in the energy functional

The numerical experiments in this section demonstrate the effects of the phase-field approximation of the perimeter functional (1.3) in the energy functional $J_{\delta,h,k}$. If not mentioned otherwise, we set $\lambda = 0$ in the experiments in this whole section.

The optimal solution for the phase-field approximation for a fixed parameter δ can be characterized as follows: The domain Ω can be split into two disjoint regions Ω_1, Ω_2, where the density attains distinct values $\Omega_1 = \{\boldsymbol{x} : \ \rho(\boldsymbol{x}) = \rho_1\}$ and $\Omega_2 = \{\boldsymbol{x} : \ \rho(\boldsymbol{x}) = \rho_2\}$. The two regions are separated by a thin region $\Omega_I = \{\boldsymbol{x} : \ \rho_1 < \rho(\boldsymbol{x}) < \rho_2\}$, with a fixed width that depends on the parameter δ; cf. [6]. Within this diffuse interface region the density varies continuously between the two density values ρ_1, ρ_2, and the interface between the two fluids is represented by a hypersurface defined by the mean level-set

$$\Gamma_I = \{\boldsymbol{x} : \ \rho(\boldsymbol{x}) = \frac{1}{2}(\rho_1 + \rho_2)\}.$$

The first set experiment examines the effects of the parameters δ and ε. We set $T = 0.5$, $k = 0.05$, $\rho_1 = 1$, $\rho_2 = 2$, $\xi = 10^{-3}$, $\alpha = 10^{-5}$, $\beta = 5$, $\delta = (3.68\pi)^{-1}$.

To eliminate the effect of diffusion on the shape of the interfacial region we first take $\xi = 0.001$. We start with an initial condition with discontinuous density. To minimize the phase-field energy, the discontinuous solution will evolve towards the optimal solution with a diffuse interface; see Figure 9.1, where the solution on the last time level is displayed, together with the level sets of the density corresponding to distinct values $\rho_1 + i \cdot 0.1$ for $i = 0, 1, \ldots, 10$. The value of the phase-field energy functional is reduced from the initial value 0.468 to the value 0.274 for the computed optimal solution. Note that rather large overshoots can occur in the solution for the density due to the value of ξ which is small. The computed optimal velocity is displayed in Figure 9.2. We observe a nontrivial "parasitic" velocity profile in the diffuse interface area: A diffuse interface layer is produced which leads to an overall decrease of the energy, but the influence on the shape of the interface is undesirable.

We continue with two experiments that demonstrate the effects of different values of the diffusion parameter ξ. We start with a discontinuous density and use a large value for the diffusion coefficient $\xi = 0.1$. The energy of the initial iterate of the gradient algorithm is 0.2519 and the energy of the computed optimal solution is 0.2516. As in the previous case, there is very little control involved in the computation, since the interface assumes a diffuse shape due to the large value of the diffusion constant $\xi > 0$. Note however, that due to the excessive diffusion, the interface is smeared out, and hence the thickness of the diffuse interface region does not correspond to the optimal thickness for the considered value of the parameter $\delta > 0$; see Figure 9.3. There is no control mechanism included in the model that would allow us to counteract the diffusion effects. Consequently, the energy of the optimal solution is larger than in the previous experiment. For $\xi = 0.5$ the initial energy is 0.409753, the final energy is 0.409684, and the interface Γ_I (represented by a white line) has disappeared; see Figure 9.3.

Remark 9.3

The behavior that can be observed in Figure 9.1 is not desirable in practice and can be eliminated by choosing $\xi > 0$ sufficiently large. Our experience indicates that $\xi \approx C\delta$ produces satisfactory results. For numerical reasons the constant C can be chosen according to Remark 9.1. Note that from the modeling point of view, the stabilization parameter ε in (3.2b) can be interpreted as the diffusion coefficient of the fluids. For diffusion dominated flows physical diffusivity may provide enough stabilization so that no additional numerical stabilization is necessary.

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 0.25)$ **(c)** $\mathcal{R}(t = 0.5)$

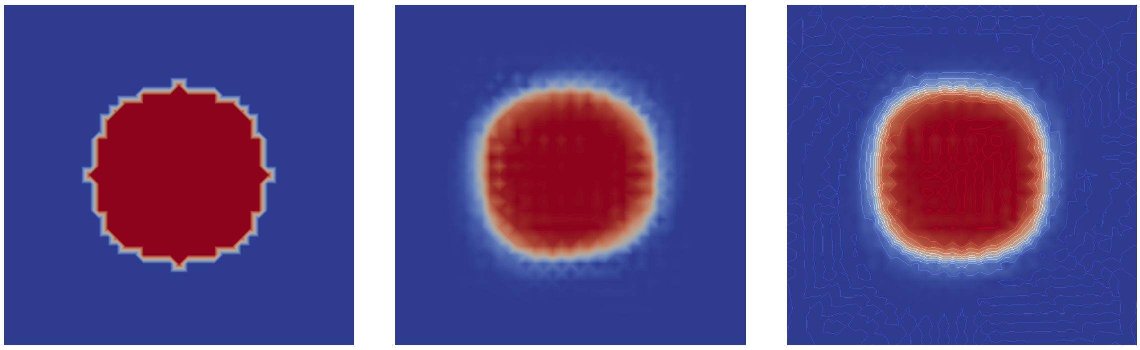


Figure 9.1. Optimal solution (density) with discontinuous initial condition at different times with $\xi = 0.001$; corresponding velocities are shown in Figure 9.2.

The initial condition for the next experiment has the shape of a square with a diffuse transition region across the interface. We set $T = 1$, $\alpha = 10^{-4}$, $\xi = 0.03$; the remaining parameters are the same as in previous experiments. The phase-field functional minimizes the perimeter of the interface, and the square initial condition evolves into a circle. In Figure 9.4 we display the optimal solution at different times. The interface Γ_I is represented by a white line. The optimal velocity field is displayed in Figure 9.5.

For comparison, we include the tracking-type part in the computation. We set $\lambda = 10$ and set the desired state equal to the initial condition for the density $\tilde{\mathcal{R}}(\boldsymbol{x}) = \rho(0, \boldsymbol{x})$. The snapshot of the solution at final time in Figure 9.6 reveals that for $\lambda > 0$ the tracking-type part of the functional forces the interface towards the square shape.

In the next experiment, the initial condition consists of two circles; the density is discontinuous. We set $\alpha = 10^{-5}$, $\lambda = 0$, $T = 0.5$, and the remaining parameters have the same

(a) $\boldsymbol{\mathcal{Y}}(t = 0.05)$ (scaled by 0.02) **(b)** $\boldsymbol{\mathcal{Y}}(t = 0.15)$ (scaled by 0.06) **(c)** $\boldsymbol{\mathcal{Y}}(t = 0.35)$ (scaled by 0.12)

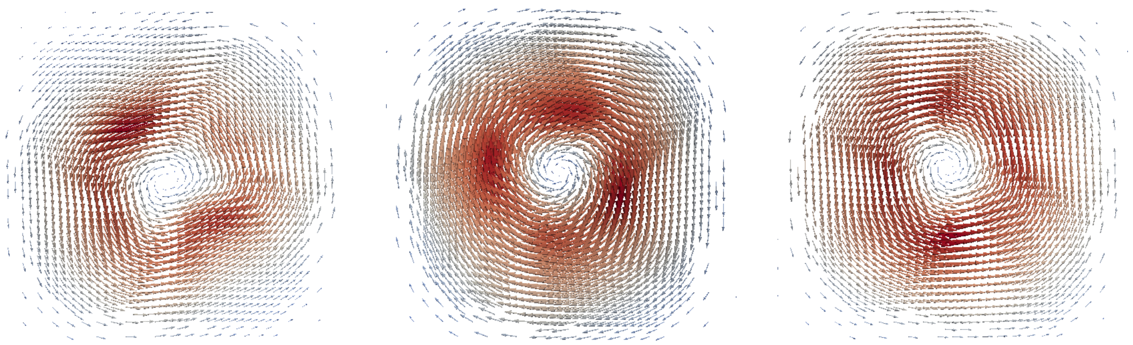


Figure 9.2. Optimal solution (velocity $\boldsymbol{\mathcal{Y}}$) at different times with $\xi = 0.001$; the vectors are scaled by different factors; corresponding densities are shown in Figure 9.1.

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 0.5)$ with $\xi = 0.1$ **(c)** $\mathcal{R}(t = 0.5)$ with $\xi = 0.5$

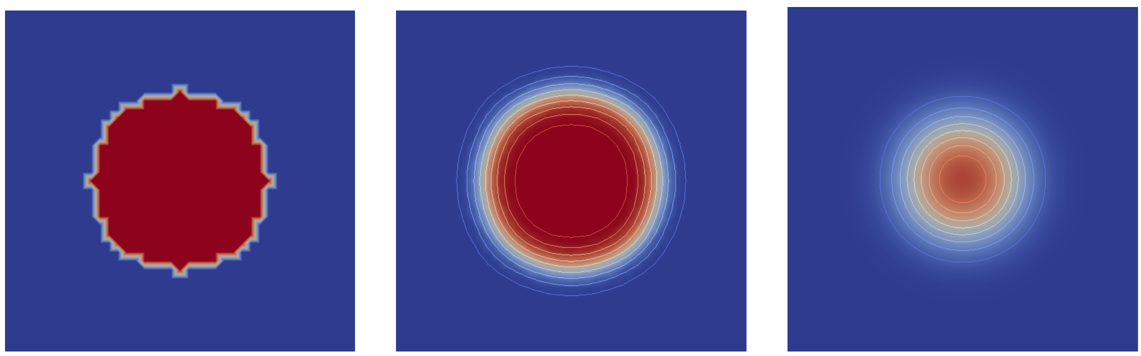


Figure 9.3. Optimal solution (density $\mathcal{R}$) with different large values of diffusion constant ξ.

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 0.25)$ **(c)** $\mathcal{R}(t = 0.5)$ **(d)** $\mathcal{R}(t = 1.0)$

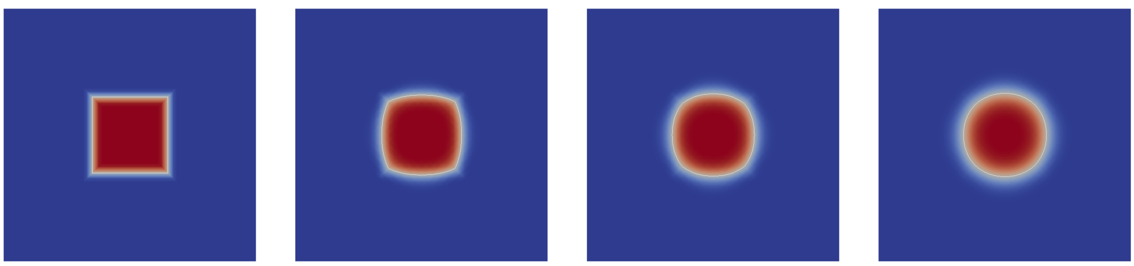


Figure 9.4. Optimal solution (density $\mathcal{R}$) at different times; corresponding velocities are shown in Figure 9.5.

(a) $\mathcal{Y}(t = 0.05)$ (scaled by 0.3) **(b)** $\mathcal{Y}(t = 0.25)$ (scaled by 1.0) **(c)** $\mathcal{Y}(t = 0.5)$ (scaled by 30)

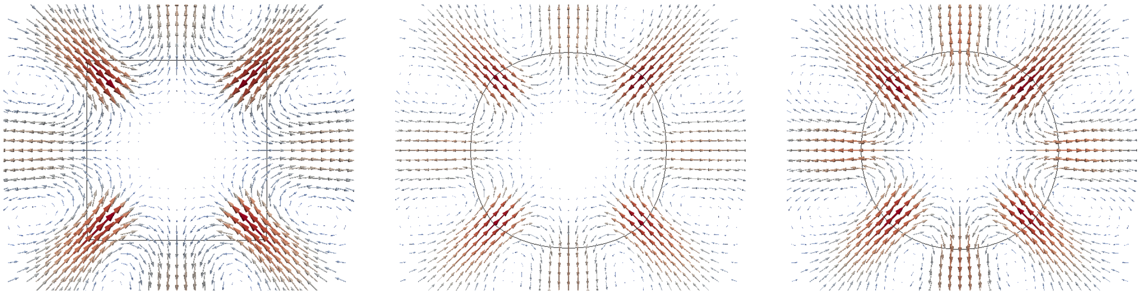


Figure 9.5. Optimal solution (velocity $\mathcal{Y}$) at different times; the vectors are scaled by different factors; corresponding densities are shown in Figure 9.4.

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 1.0)$ for $\lambda = 0$ **(c)** $\mathcal{R}(t = 1.0)$ for $\lambda = 10$

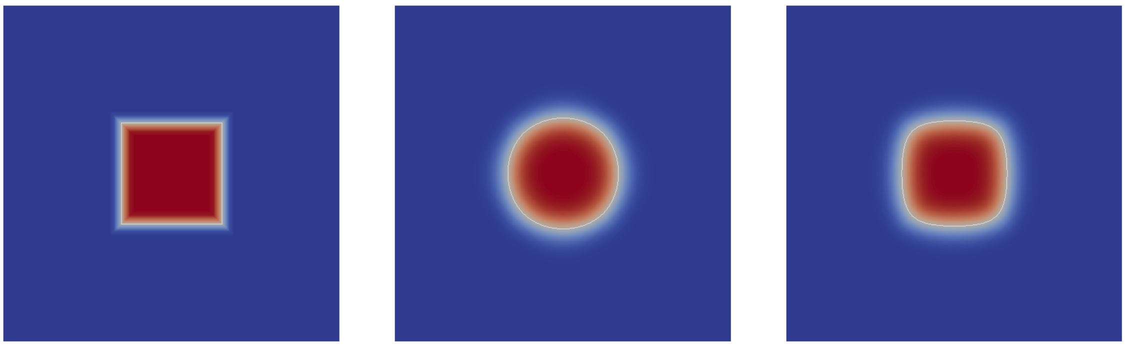


Figure 9.6. Comparison of the optimal solution (density $\mathcal{R}$) for different values for λ.

values as in the previous experiment. To minimize the phase-field energy, the two circles are forced to join together, and the interface then evolves towards a circle; see Figure 9.7. The velocity field is displayed in Figure 9.8; the asymmetry of the results can be attributed to the round-off errors in the numerical approximation. Note that for $\alpha = 10^{-4}$ the two circles remain separated for $T = 0.5$. To further verify the effects of round-off errors, we repeated the experiment with a different linear solver. The results were almost identical with the exception that they were antisymmetric along the vertical direction. This phenomenon can be explained by a possible nonuniqueness of the solution: Due to the round-off errors the gradient algorithm may converge towards a different solution with similar energy.

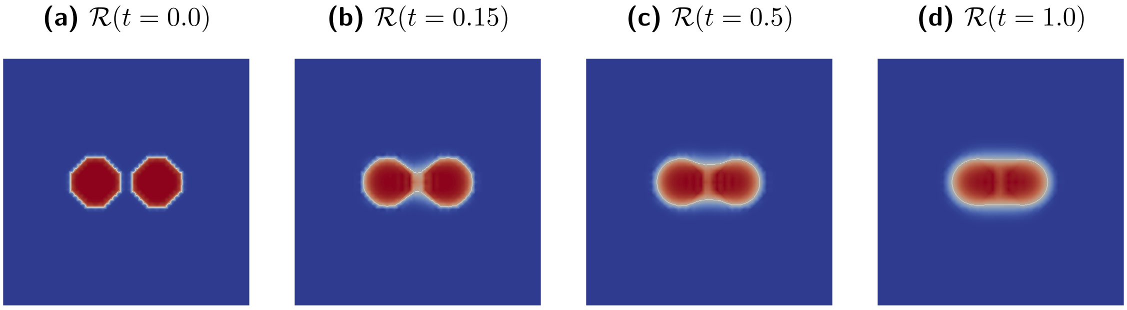


Figure 9.7. Optimal solution (density $\mathcal{R}$) at different times with $\lambda = 0$; the corresponding velocities are shown in Figure 9.8.

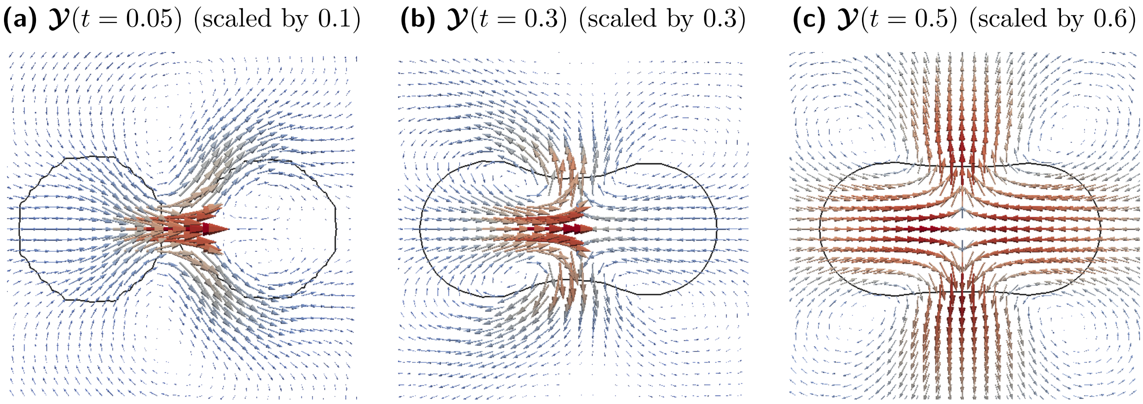


Figure 9.8. Optimal solution (velocity $\boldsymbol{\mathcal{Y}}$) at different times; the vectors are scaled by different factors; the corresponding densities are shown in Figure 9.7.

In the next experiment, we start with an initial condition in the shape of a dumbbell; set $\alpha = 10^{-4}$, $T = 0.5$, $\lambda = 0$, and take the remaining parameters as in the previous experiments. The evolution of the computed density is depicted in Figure 9.9—the dumbbell shape remains connected and the interface evolves towards an ellipsoidal shape. To illustrate the adaptive algorithm we also display the finite element mesh obtained by the adaptive algorithm described in Section 9.1; see Figure 9.9d.

In the final experiment from this section we start with a dumbbell-shaped initial condition, with rectangular ends that are further apart from each other than in the previous example, and a thinner connecting neck. We choose $\alpha = 10^{-5}$, $\lambda = 0$, $T = 0.5$, and the other parameters remain unchanged. The evolution of the optimal solution for this setting leads

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 0.15)$ **(c)** $\mathcal{R}(t = 1.0)$ **(d)** Used mesh

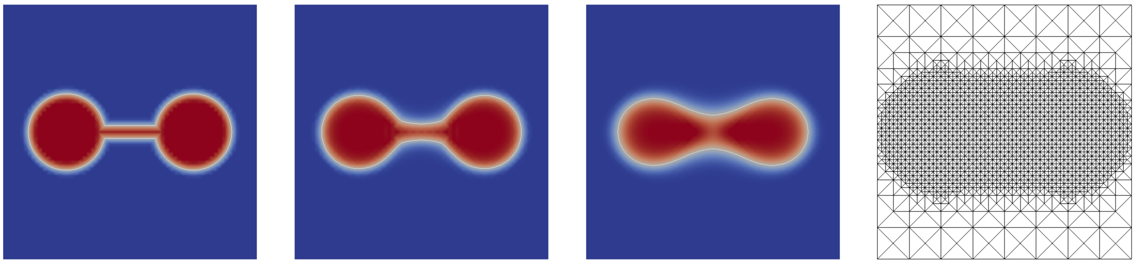


Figure 9.9. Optimal solution (density $\mathcal{R}$) at different times with $\lambda = 0$, and the finite element mesh used in the computation.

to a pinch-off, the squares become circular, and the initially connected dumbbell shape eventually splits up into two separate bubbles, see Figure 9.10. In Figure 9.10d we display the solution without control, i.e., with zero velocity, which shows that the connecting region becomes thinner due to the effects of diffusion, but the dumbbell remains connected. The optimal velocity and the interface are displayed in Figure 9.11.

(a) $\mathcal{R}(t = 0.0)$ **(b)** $\mathcal{R}(t = 0.15)$ **(c)** $\mathcal{R}(t = 0.5)$ **(d)** $\mathcal{R}(t = 0.5)$, without control

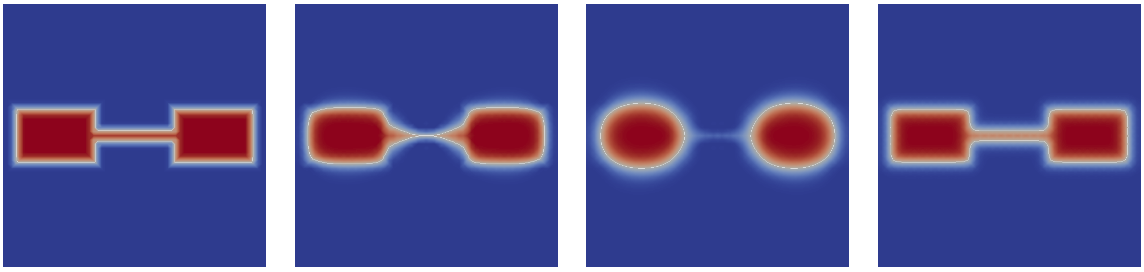


Figure 9.10. Optimal solution (density $\mathcal{R}$) at different times, with and without control; $\lambda = 0$; corresponding velocities are shown in Figure 9.11.

9.3. Effects of the distance and phase-field functionals

In the next experiment we demonstrate the differences between the optimal solutions for the functional (1.3) and the L^2 tracking-type part, i.e., (1.3) with $\beta = 0$.

In the first experiment we study a problem with a non-convex initial condition (see Figure 9.12b) and the convex hull of it being the target (see Figure 9.12a), and we set $T = 0.5$, $k = 0.05$, $\xi = 0.03$, $\alpha = 10^{-5}$, $\beta = 5$, $\lambda = 5$, $\delta = (3.68\pi)^{-1}$. The remaining parameters were as in the previous experiments from the last section.

In Figure 9.12 we compare the solution at the final time for $\beta = 0$ and $\beta = 5$. The solution at the final time for $\beta = 5$ becomes convex (see Figure 9.12c), while for the pure L^2-distance energy the solution remains non-convex (see Figure 9.12d).

(a) $\boldsymbol{\mathcal{Y}}(t = 0.1)$ (scaled by 0.1) **(b)** $\boldsymbol{\mathcal{Y}}(t = 0.15)$ (scaled by 0.15) **(c)** $\boldsymbol{\mathcal{Y}}(t = 0.5)$ (scaled by 0.4)

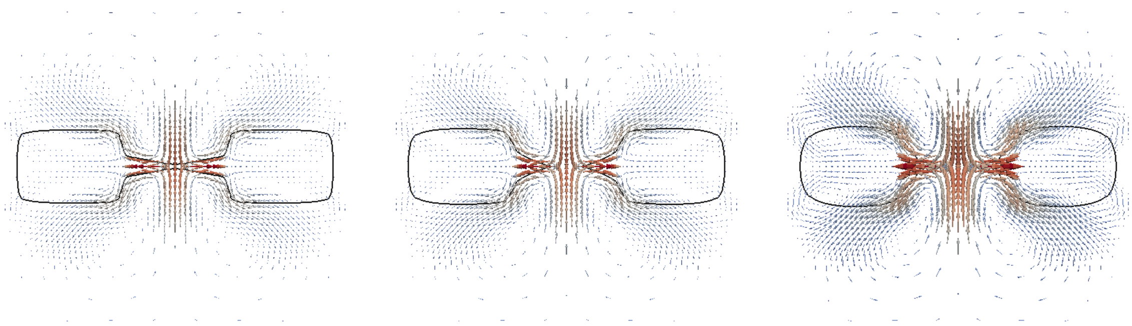


Figure 9.11. Optimal solution (velocity $\boldsymbol{\mathcal{Y}}$) at different times; the vectors are scaled by different factors; $\lambda = 0$; corresponding densities are shown in Figure 9.10.

(a) Target $\tilde{\mathcal{R}}$ **(b)** $\mathcal{R}(t = 0.0)$ **(c)** $\mathcal{R}(t = T)$ for $\beta = 5$ **(d)** $\mathcal{R}(t = T)$ for $\beta = 0$

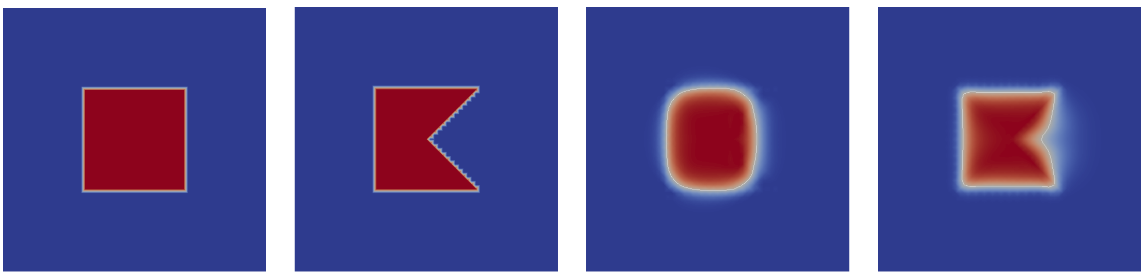


Figure 9.12. Optimal solution at $t = T$ for different values of β; $\lambda = 5$.

In the second experiment we set $T = 1$, $k = 0.05$, $\rho_1 = 1$, $\rho_2 = 2$, $\varepsilon = 0.03$, $\alpha = 10^{-4}$, $\beta = 5$, $\delta = (3.68\pi)^{-1}$. The initial condition consists of two squares, while the desired target is a rectangular region that covers the two squares; see Figure 9.13). In Figure 9.14, we compare the evolution for $\lambda = 0$ and $\lambda = 10$, respectively. The interface is considerably smoother for the latter case.

(a) Initial condition $\mathcal{R}(t = 0.0)$ **(b)** Target $\tilde{\mathcal{R}}$

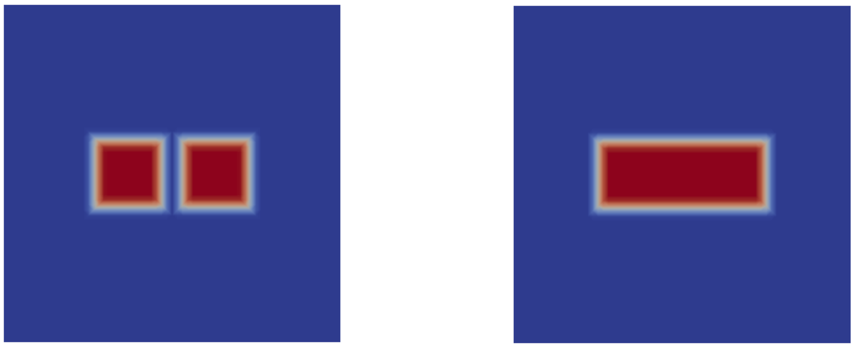


Figure 9.13. Initial condition $\mathcal{R}(t = 0.0)$ and target $\tilde{\mathcal{R}}$

9.4. Convergence for vanishing discretization parameters

We take a closer look at the last example from the last section with initial condition and target from 9.13, whose evolution in time is shown in Figure 9.14. We compute the example with $\lambda = 10$ for different mesh size $h_{min} = 2^{-l}/64$ and time step $l = 2^{-l}/20$ for $l = 0, 1, 2$, and we fix $h_{max} = 1/16$.

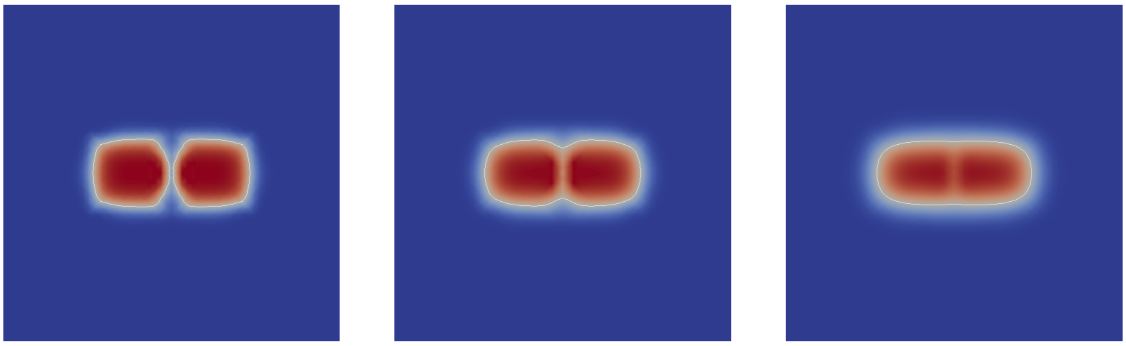
The value of the functional $J_{\delta,h,k}$ at the optimal solution for decreasing discretization parameters is displayed in Table 9.1; we observe that the energy is decreases with growing l.

h	k	$J_{\delta,h,k}$
1/64	0.05	0.719421
1/128	0.025	0.659382
1/256	0.0125	0.634922

Table 9.1. Behavior of the cost functional $J_{\delta,h,k}$ for decreasing discretization parameters.

A zoom at the optimal solution and the interface at the final time for $l = 0, 1, 2$ is displayed in Figure 9.15. The shape of the interface is very similar for different discretization parameters (see Figure 9.16). The artificial diffusion effects in the mass equation decreases with increasing l as the diffusion parameter $\varepsilon_h = 0.03h$ depends on the mesh size.

(a) $\mathcal{R}(t = 0.25)$ for $\lambda = 0$ **(b)** $\mathcal{R}(t = 0.5)$ for $\lambda = 0$ **(c)** $\mathcal{R}(t = 1.0)$ for $\lambda = 0$



(d) $\mathcal{R}(t = 0.25)$ for $\lambda = 10$ **(e)** $\mathcal{R}(t = 0.5)$ for $\lambda = 10$ **(f)** $\mathcal{R}(t = 1.0)$ for $\lambda = 10$

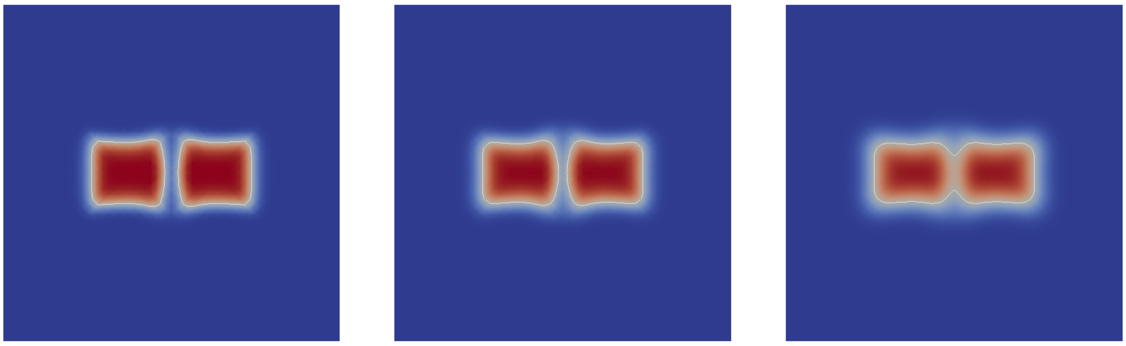


Figure 9.14. Optimal solution (density $\mathcal{R}$) at different times for $\lambda = 0$ (top) and $\lambda = 10$ (bottom).

(a) $\mathcal{R}(t = 1.0)$ for $l = 0$ **(b)** $\mathcal{R}(t = 1.0)$ for $l = 1$ **(c)** $\mathcal{R}(t = 1.0)$ for $l = 2$

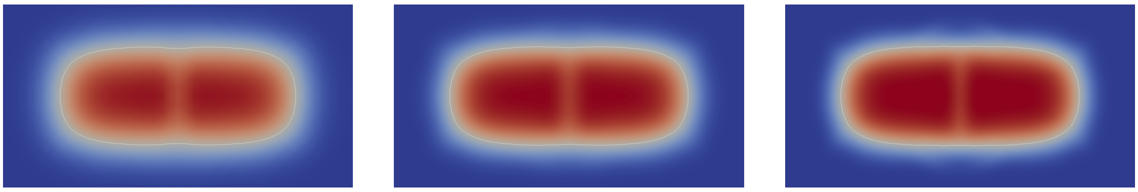


Figure 9.15. Optimal solution (density $\mathcal{R}$) at time $t = 1$ for mesh size $h = 2^{-l}/64$ and time step $l = 2^{-l}/20$ for different values of l.

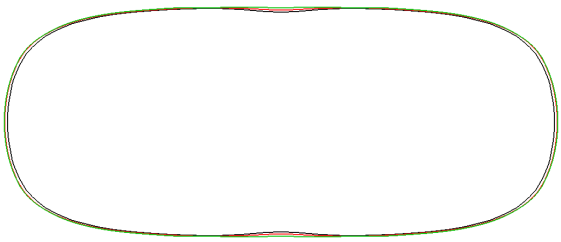


Figure 9.16. Interface at $t = 1$ for $h = 1/64$ (black), $h = 1/124$ (red), $h = 1/256$ (green).

Part II.

Optimal control the thin-film equation

10. Introduction

Let $\Omega = (a,b) \subset \mathbb{R}$, $0 < T < \infty$, $g_0 : (0,\infty) \to \mathbb{R}$ be a given function, $u \in L^2(0,T;\Omega)$ and $y_0 \geq C_0 > 0$ be smooth enough. The one-dimensional thin-film equation (weak slip) reads as follows: Find $y : \Omega_T \to \mathbb{R}$ such that

$$y_t = -(f(y)y_{xxx})_x - (g_0(y)y_x)_x + u_x, \tag{10.1}$$

together with the initial condition $y(0,.) = y_0$ and boundary conditions $y_x = y_{xxx} = 0$ in a, b.

The variable y describes the height of a thin fluid film on some (flat) surface. Here and below, we assume $f(y) = \lambda|y|^3$ for some $\lambda > 0$. The potential function $g_0 : \mathbb{R} \to \mathbb{R}$ will be specified in the following. The potential function g_0 models forces which are present in the evolution of the film. Roughly speaking, g_0 has to be in such a way that the solution of (10.1) for $u \equiv 0$ is strictly positive provided initial data y_0 are strictly positive. Relevant potentials g_0 are $g_0 \equiv 0$ (cf. [19]), $g_0 \equiv -1$ (cf. [20]), or $g_0(y) = -y^{-\kappa}$ for some $\kappa > 0$ (cf. [15, 21]). Typically, the potential g_0 becomes singular for $y \to 0$, resulting in strong forces in the equation (10.1). A survey addressing general issues of the equation (10.1) is given in [21] and the references therein.

In this work, we study the following constrained optimization problem related to (10.1).

Problem 10.1
Let $\tilde{y} \in L^2(\Omega_T)$ be a given, $\alpha > 0$, and $C_0 > 0$. Find a minimum (y^*, u^*) of

$$J(y,u) := \frac{1}{2}\int_0^T \int_\Omega |y - \tilde{y}|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{\alpha}{2}\int_0^T \int_\Omega |u|^2 \,\mathrm{d}x\,\mathrm{d}t$$

subject to (10.1) and $y \geq C_0$ in Ω_T.

The aim of the problem is to control the evolution height y of a fluid film which is driven by an external control $u_x : \Omega_T \to \mathbb{R}$. The governing equation (10.1) is in divergence form, avoiding evaporation or wetting effects. The first term on the right-hand side of (10.1) models the dynamics of the fluid film coming from the Navier–Stokes equations. The term g_0 models intrinsic forces such as gravity, van der Waals forces between molecules, etc. Some further properties of g_0 are detailed in Hypothesis 12.3 and Hypothesis 12.9. The last contribution on the right-hand side u_x is also in divergence form and models external forces. It is known that in the absence of external forces and for some potential functions g_0, the solution y of (10.1) converges to its spatial mean value in the limit $t \to \infty$ as long as a global solution can be provided; cf. [20]. In the context of the optimization, we want to force the optimal solution not to form a flat profile in time, i.e., not to be a flat profile for big values of t, but to approximate the spatial shape $\tilde{y}$.

We refer the reader to [21, 35] for the equation (10.1), and to [57] for the derivation of the equation and corresponding models (e.g., the strong slip case). The fundamental work for the equation with $g_0, u \equiv 0$ is [19]. Since our goal is to show existence and derive optimality conditions for Problem 10.1, we need to recapitulate and modify the proofs given in [19] for (10.1) with $g_0, u \neq 0$. Typically, a solution of the leading equation (10.1) is endowed with an energy equation and an entropy inequality, from which we may deduce non-negativity of solutions. It is due to the presence of the u-term, that an entropy inequality is not clear to hold any more; see Figures 16.1 and 16.2, where the solution to a given non-trivial right-hand side u is displayed. For a general given target profile $\tilde{y}$, the external control u is not expected to have a sign, and u should force the state y to take almost zero values if $\tilde{y}$ is of this kind. This is the reason why we do not expect that in such a case an entropy inequality holds for equation (10.1).

A possible application of the above optimal control Problem 10.1 is in the fabrication of electronic chips, where thin layers of different material are deposited on a Si wafer. For an efficient electronical circuit, each layer has to constitute a specific profile, which defines where there is material and where no material is allowed. The problem is to find external forces such that the solution of (10.1) is near the desired profile $\tilde{y}$. Typically, the initial condition in this application is constant and the goal is to form the profile by so-called dewetting; see [17] and section 16.7 below. This goal can either be accomplished by background engineering knowledge or by solving Problem 10.1.

Equation (10.1) is derived in, e.g., [16, 59]: We consider the fluid to be thin, i.e., $\tau :=$ height/length $\ll 1$. A nondimensional transformation from the classical Navier–Stokes equation which is based on the small ratio τ and a Taylor expansion of the terms, together with the assumption of a so-called no slip boundary conditions (cf. [59, p. 936]) leads to an asymptotic expansion in τ. Neglecting higher order terms of τ, and the proper use of boundary conditions then leads to (10.1). A detailed derivation of the equation will be revisited in chapter B following the named sources.

It is important to note that through the transformation process, a conservative force on the right-hand side of the Navier–Stokes equation transforms into a potential function g_0 in (10.1). Hence, a control problem for the Navier–Stokes equation where a distributed conservative force is to be found (cf. [2]) transforms “naturally” into an optimal control problem of the thin-film equation, where a potential function is to be found: Instead of searching a L^2 control function u, one would like to find a potential function like g_0 in (10.1) for minimizing the functional J. However, we do not know how to accomplish this goal since nothing about the potential is known. Up to our knowledge there is no work in the context of optimal control where a potential as control variable is to be found. The coupling of such a control and state variables is much stronger as in the case where they are only coupled via the right-hand side of a partial differential equation or similar. In particular, it is not clear if this coupling allows to derive continuous solution operators to deduce boundedness of the state variable to prove existence of optimal controls. Also, it seems that there is no literature on deriving necessary optimality conditions for such a situation. Only in cases where a specific algebraic form of the potential is given (i.e., if the potential is polynomial or a sum of given potentials), there is hope that the corresponding optimal control problem is well-posed—and in these cases the space of controls becomes finite-dimensional since only a few real numbers are to be found. The authors are not aware of any literature dealing with optimal control problems, where potential functions are the subject of interest.

The problem in solving Problem 10.1 is that the nonlinear term of the equation (10.1) may degenerate, and therefore the equation could not be well-posed. To avoid this deficiency, we can only take into account those exterior forces u, where the corresponding solution y exists. Unfortunately, we cannot give a good characterization to it, and we do not know topological properties of this set. There are a few recent articles dealing with degenerate optimal control problems for different equations than (10.1); see e.g., [26, 27, 28].

There are two possible ways to overcome this problematic issue:

1. Restrict to a rich enough class of external forces $u : \Omega_T \to \mathbb{R}$ such that solvability of (10.1) is ensured and solutions y are strictly positive. From an optimization viewpoint this strategy is convenient since only control constraints appear. Unfortunately, there is no such result for equation (10.1), and the possibility of too severely restricted controls sets in order to ensure well-posedness of (10.1) has to be encountered.

2. Force the solution to be strictly positive by state constraints as indicated in Problem 10.1. In this case, we only aim for strict positivity of a solution y of (10.1), but have no further restriction regarding controls u, i.e., solutions y near an almost degenerate target function $\tilde{y}$ are possible and can be reached by the optimization procedure. As a drawback, we have to overcome several mathematical difficulties.

We refer the reader to [27] where the authors compare both strategies for a different equation, and conclude that the second scenario is more suitable in order to cope with possible degeneracies arising in the governing equation since the set of external forces in the first scenario may not be rich enough, and therefore possible target profiles $\tilde{y}$ may not be reached.

We are able to show existence for the optimal control Problem 10.1. With the help of an abstract result for state constrained optimization problems from [4], we are able to derive necessary optimality conditions for Problem 10.1. In order to overcome technical difficulties arising later in the convergence proofs, we want to make the optimal control problem at this point compatible and need to aim for right-hand sides u from $L^2(H_0^1)$. This can be accomplished in two different ways: The first possibility is to consider a different cost functional J including the $L^2(H^1)$-norm of the control u, whereby the issue is solved directly. This approach is used by the author in [50]. Instead of this idea, we consider $\left(\Delta^{-1}(u_x)\right)_x$ as external control in (10.1) instead of u_x, where Δ^{-1} is the inverse Poisson operator with homogeneous Dirichlet boundary conditions. This idea was used in [49]. As a result, the driving term in (10.1) takes values in $L^2(L^2)$ rather than in $L^2(H^{-1})$. We emphasize that this is only needed in order to pass to proper limiting functions in chapter 14. Later in chapter 16, we consider a finite dimensional version of the optimal control Problem 10.1, where no convergence analysis will be done, hence there is no need to modify the right-hand side of (10.1), and we will use there u_x as external control.

The optimality conditions (13.13) involve non-regular Lagrange multipliers in the dual space of $L^2(H^4) \cap H^1(L^2)$, where it is not clear how to handle them in a numerical simulation: Typical strategies for solving such problems use some sort of relaxation of the state constraint $y \geq C_0$ such as penalty approximation [24], the Moreau-Yosida approximation [44], or mixed control-state constraints (Lavrentiev regularization) [64]. The problem in our case is that the state constraint is not additional, but essential in order to ensure well-posedness of the equation (10.1).

In order to circumvent the problematic issue of possibly loosing the well-posedness property of the state equation in the context of relaxation methods, our strategy is as follows: First, we regularize the state equation (10.1) by adding εy_{xxxx}, which introduces a regularization to the equation and ensures well-posedness for general exterior forces u. We consider the optimal control problem subject to the regularized equation (12.1) and state constraints (see Problem 14.1). Similarly to the original Problem 10.1, we show existence of an optimum, and derive necessary optimality conditions. We show that the sequence of solutions of the optimal control problem is uniformly bounded with respect to ε which allows to construct limiting functions and Lagrange multipliers. In order to show the bounds, it is crucial that we modified the external control term in (10.1) to $\left(\Delta^{-1}(u_x)\right)_x$, which helps us at this particular point to bound all corresponding Lagrange multipliers in their particular spaces. We are able to show that these derived limiting functions and Lagrange multipliers solve the necessary optimality conditions (13.13) of the original Problem 13.1. Here, the additional term plays an important role.

Finally, we consider the optimal control problem subject to the regularized equation without state constraints, but with a modified cost functional J_γ which additionally contains a penalization term to account for the state constraint with a parameter $\gamma > 0$; see Problem 15.1. Since the equality constraint is well-posed for every $\varepsilon > 0$, we may use a standard numerical approach to solve the corresponding optimality conditions (15.4) in chapter 16. We can show that the sequence of minimizers of the fully regularized Problem 15.1 converges to functions solving the intermediate optimization Problem 14.1 for $\gamma \to 0$. We use here the penalty approach because of its simple implementation and flexibility. However, the drawback is that the condition number of the underlying problem grows for $\gamma \to 0$. This leads to ill-posed problems on the level of numerical linear algebra, which can also be observed in the numerical experiments in chapter 16. An alternative way would be to use more sophisticated regularization approaches as mentioned above, where some further details are provided in chapter 15.

We emphasize that it is necessary to study the intermediate optimization Problem 14.1 since it is not possible to simultaneously let both regularization parameters tend to zero. It is important that the parameter $\gamma > 0$ dealing with the regularization of the state constraint is the first which tends to zero: Here, we benefit from the well-posedness of the involved equality constraint for every $\varepsilon > 0$ to construct a solution of the intermediate optimization Problem 14.1. Vice versa, a direct approximation with the penalty method and $\gamma > 0$ only (or any other relaxation method) could lead to an optimal control problem with possibly non-invertable state equation due to the non-feasibility of the iterates—which was the issue why we introduced the state constraint in the first place.

To sum up, our main result is to show the existence of a solution of Problem 10.1, and to construct it by a multi-parameter regularization of the optimal control problem, where the limit of a corresponding subsequence solves the original Problem 10.1.

This part is organized as follows:

- In chapter 12, we study a regularization of the state equation (12.1) and derive results for it, with either the regularization parameter $\varepsilon > 0$ or state constraints being present.

In the context of this chapter, we consider rather general potential functions g_0 pertaining to improved regularity results for solutions of (13.1). A few results are only shown for $g_0 \equiv 0$ or $g_0 \equiv -1$.

- In chapter 13, we study a modification (i.e., Problem 13.1) of Problem 10.1 which uses the external control $\left(\Delta^{-1} u_x\right)_x$ with a non-regularized equation and the state constraint.
- In chapter 14, we study an optimization problem (i.e., Problem 14.1) which is connected to the one in the previous chapter, but with a regularization term scaled by $\varepsilon > 0$ in the equation to make it well-posed. This problem can be understood as an intermediate one between the original Problem 13.1 in chapter 13 and Problem 15.1 in chapter 15 which is suitable as starting point for numerical studies. We show solvability and construct a sequence of minimizers which converges to a function (up to a subsequence) which solves the optimality conditions from chapter 13 for $\varepsilon \to 0$.
- In chapter 15, we study the penalty approximation with parameter $\gamma > 0$ of Problem 14.1 introduced in the previous chapter, which allows to get rid of the non-regular Lagrange multiplier associated to the state constraint. We will show that the sequence of solutions converges to a minimum of Problem 14.1 from chapter 14 for $\gamma \to 0$.
- In chapter 16, we present computational studies, using a mixed first order finite element method. We detail how to implement the optimal control problem and compare different parameters. In particular, we show that a goal like optimal dewetting mentioned above can be accomplished; see Figure 16.9; we show that the state constraints are necessary to provide solutions being positive; we give hints concerning how other parameters like α, ε, should be chosen in order to perform reasonable experiments.

We do not include convergence studies for the involved parameters, hence the change of u_x to $\left(\Delta^{-1}(u_x)\right)_x$ in (10.1) is unnecessary, and this replacement may be considered as a scaling of the control for a fixed spatial discretization parameter $h > 0$. Therefore, we do not include this modification into our computational experiments in chapter 16.

11. Preliminaries

We write $\|.\|$ for the $L^2(\Omega)$ or $L^2(\Omega_T)$-norm, when it is clear if we only integrate in space or both, in space and time. Let $W^{k,p}$ and $H^k := W^{k,2}$ denote standard Sobolev spaces. By

$$W^{k,p}(W^{m,q}) := W^{k,p}(0,T;W^{m,q})$$

we refer the reader to standard Bochner spaces. The space $\mathcal{C}$ denotes the space of continuous functions, while $\mathcal{C}^{0,\alpha}$ denotes corresponding Hölder spaces.

The dual pairing of X and its dual space X^* is written as $\langle .,. \rangle_{X,X^*}$. For the scalar products in L^2 and $L^2(L^2)$, respectively, of f and g, we write (f,g) in cases where no confusion arises; otherwise, we add the corresponding space as index to the scalar product.

We use C as a generic nonnegative constant; to indicate dependencies, we write $C(.)$.

12. The regularized state equation

In this chapter, we will show properties of solutions of a regularization of the equation (10.1). At the end of this chapter, we discuss aspects of its solvability, which relies on the proven results and is crucially depending on the chosen potential function g_0.

Problem 12.1
Let $\varepsilon > 0$. Find $y : \Omega_T \to \mathbb{R}$ such that

$$y_t = -\left([f(y) + \varepsilon]y_{xxx}\right)_x - (g_0(y)y_x)_x + u_x, \tag{12.1}$$

together with initial condition $y(0) = y_0$ and boundary conditions $y_x = y_{xxx} = 0$ in a, b, where $f(y) = \lambda|y|^3$ for a given $\lambda > 0$. We will also use the abbreviation $f_\varepsilon(y) := f(y) + \varepsilon$.

12.1. Regularity and properties of solutions

Lemma 12.2
Let $\varepsilon > 0$, $u \in L^2(\Omega_T)$ and let y be a solution of (12.1). Then, the mass is conserved, i.e.,

$$\int_\Omega y(t,.)\,\mathrm{d}x = \int_\Omega y_0\,\mathrm{d}x \quad \forall\, 0 \le t \le T. \tag{12.2}$$

Proof
Integrate (10.1) over Ω and use the divergence theorem to proof (12.2). □

The following hypothesis gathers minimum requirements concerning the potential function g_0 to prove some regularity properties of possible solutions y of (12.1).

Hypothesis 12.3
Assume that one of the following hypothesis is true.

(A1) The potential g_0 is continuous and uniformly bounded in $L^2(0,T;L^\infty([C_0,\infty)))$.

(A2) The potential g_0 is smooth, uniformly bounded in $L^2(0,T;L^2([C_0,\infty)))$, and $g_0 \le 0$ as well as $g_0'' \ge 0$ in Ω_T.

In both cases, a term like $\|g_0\|_X$ (where X is one of the spaces above) means that the range of g_0 in the particular domain is bounded for all possible arguments in g_0.

Lemma 12.4
Let $\varepsilon > 0$, let Hypothesis 12.3 be true and let $y : \Omega_T \to \mathbb{R}$ be a solution of (12.1) with $y \geq C_0$ a.e. Then there exists a constant $C > 0$ such that the following energy inequality holds

$$\|y_x\|^2_{L^\infty(L^2)} + (\lambda C_0^3 + \varepsilon)\|y_{xxx}\|^2_{L^2(L^2)} \leq C\left(T, C_0, \|y_0\|_{H^1}, \|u\|_{L^2(L^2)}\right). \tag{12.3}$$

In particular, y is Hölder continuous in space, i.e., there exists a constant $H_{\text{space}} > 0$ such that

$$|y(t,x_1) - y(t,x_2)| \leq H_{\text{space}}|x_1 - x_2|^{\frac{1}{2}} \quad \forall\, 0 \leq t \leq T,\ x_1, x_2 \in \Omega.$$

Proof
We multiply (12.1) with $-y_{xx}$, integrate over Ω, and arrive for almost all $t \in [0,T]$ at

$$\frac{1}{2}\frac{d}{dt}\|y_x\|^2 + \int_\Omega f_\varepsilon(y) y_{xxx}^2 \,\mathrm{d}x = -\int_\Omega g_0(y) y_x y_{xxx} \,\mathrm{d}x - \int_\Omega u_x y_{xx} \,\mathrm{d}x =: I + II. \tag{12.4}$$

We estimate now the terms I and II, depending on whether g_0 satisfies (A1) or (A2).

If (A1) is true, we estimate I as follows

$$I \leq \|g_0\|_{L^\infty[C_0,\infty)}\|y_x\|\|y_{xxx}\| \leq \sigma\|y_{xxx}\|^2 + C(\sigma)\|g_0\|^2_{L^\infty[C_0,\infty)}\|y_x\|^2,$$

where $\sigma > 0$. In the case of (A2), we calculate

$$I = \int_\Omega g_0(y) y_{xx}^2 \,\mathrm{d}x + \int_\Omega g_0'(y) y_x^2 y_{xx} \,\mathrm{d}x = \int_\Omega g_0(y) y_{xx}^2 \,\mathrm{d}x - \frac{1}{3}\int_\Omega g_0''(y) y_x^4 \,\mathrm{d}x \leq 0,$$

since by integration by parts, there holds

$$\int_\Omega g_0'(y) y_x^2 y_{xx} \,\mathrm{d}x = -\int_\Omega g_0''(y) y_x^4 \,\mathrm{d}x - 2\int_\Omega g_0'(y) y_x^2 y_{xx} \,\mathrm{d}x.$$

The term II can be estimated by

$$II = \int_\Omega u y_{xxx} \,\mathrm{d}x \leq \sigma\|y_{xxx}\|^2 + C(\sigma)\|u\|^2.$$

Putting things together, using that $f_\varepsilon(y) \geq \lambda C_0^3 + \varepsilon$, and Gronwall's inequality, we have proven the lemma. The Hölder continuity follows by one-dimensional Sobolev embeddings. □

Lemma 12.5
Let $\varepsilon > 0$, let Hypothesis 12.3 be true, $u \in L^2(\Omega_T)$, and let $y : \Omega_T \to \mathbb{R}$ be a solution of (12.1) with $y \geq C_0$ a.e. Then there exists a constant $H_{\text{time}} \equiv H_{\text{time}}(T, C_0, y_0, u) > 0$ such that

$$|y(t_2,x) - y(t_1,x)| \leq H_{\text{time}}|t_2 - t_1|^{\frac{1}{8}} \quad \forall\, 0 \leq t_1, t_2 \leq T,\ x \in \Omega.$$

Proof
The proof uses arguments similar (for $g_0 \equiv 0$) to those given in [19, Lemma 2.1].

Step 1. Assume the statement is not correct. Then for every $M > 0$ there exist $x_0 \in \Omega$ and $0 \leq t_1, t_2 \leq T$ such that

$$|y(t_2, x_0) - y(t_1, x_0)| > M|t_2 - t_1|^{\beta} \tag{12.5}$$

for $\beta = \frac{1}{8}$. Without restriction let us assume that $t_1 < t_2$ and $y(t_2) > y(t_1)$. Then (12.5) reads as

$$y(t_2, x_0) - y(t_1, x_0) > M(t_2 - t_1)^{\beta}. \tag{12.6}$$

In the proof, we will show that M can be uniformly bounded with respect to x_0, t_1 and t_2, which contradicts (12.6).

We construct an appropriate test function of the equation (12.1). Let

$$\xi(x) := \xi_0 \left(\frac{x - x_0}{\frac{M^2}{16 H_{\text{space}}^2}(t_2 - t_1)^{2\beta}} \right),$$

where M is from (12.6), H_{space} from Lemma 12.4. The function $\xi_0 \in \mathcal{C}_0^{\infty}$ has the properties $\xi_0(x) = \xi_0(-x)$, $\xi_0(x) \equiv 1$ for $0 \leq x < \frac{1}{2L}$ for some $L > 0$ (L will be chosen later and will only depend on $H_{\text{space}} > 0$ from Lemma 12.4 and on Ω), $\xi_0(x) \equiv 0$ for $x \geq 1$ and $\xi_0'(x) \leq 0$ for $x \geq 0$. In particular, we have

$$\xi(x) = \begin{cases} 0, & |x - x_0| \geq \frac{M^2}{16 H_{\text{space}}^2}(t_2 - t_1)^{2\beta}, \\ 1, & |x - x_0| \leq \frac{1}{2L} \frac{M^2}{16 H_{\text{space}}^2}(t_2 - t_1)^{2\beta}. \end{cases}$$

We define the function θ_δ by

$$\theta_\delta(t) := \int_{-\infty}^{t} \theta_\delta'(s) \, \mathrm{d}s,$$

where

$$\theta_\delta'(t) = \begin{cases} \frac{1}{\delta}, & |t - t_2| < \delta, \\ -\frac{1}{\delta}, & |t - t_1| < \delta, \\ 0, & \text{else} \end{cases}$$

for $0 < \delta < \min\{\frac{1}{2}(t_2 - t_1), t_1, T - t_2\}$ small enough.

We consider the function $\phi(t, x) := \xi(x)\theta_\delta(t)$, multiply (12.1) with ϕ, integrate over Ω_T and get

$$\int_0^T y\phi_t \, \mathrm{d}x \, \mathrm{d}t = -\int_0^T \int_\Omega f_\varepsilon(y) y_{xxx} \phi_x \, \mathrm{d}x \, \mathrm{d}t - \int_0^T \int_\Omega g_0(y) y_x \phi_x \, \mathrm{d}x \, \mathrm{d}t + \int_0^T \int_\Omega u\phi_x \, \mathrm{d}x \, \mathrm{d}t. \tag{12.7}$$

Step 2. We derive a lower bound for the left-hand side of (12.7). By the construction of θ_δ, its time derivative approximates like a Dirac function evaluated at t_1 and t_2, respectively. More precisely, we have for $\delta \to 0$

$$\int_0^T \int_\Omega y(t, x)\xi(x)\theta_\delta'(t) \, \mathrm{d}x \, \mathrm{d}t \to \int_\Omega \xi(x) [y(t_2, x) - y(t_1, x)] \, \mathrm{d}x. \tag{12.8}$$

We consider points x such that

$$|x - x_0| \le \frac{M^2}{16 H_{\text{space}}^2}(t_2 - t_1)^{2\beta} \tag{12.9}$$

since outside this ball, the corresponding integral in (12.7) vanishes. For such x, there holds by (12.6) and Lemma 12.4

$$\begin{aligned} y(t_2,x) - y(t_1,x) &= [y(t_2,x) - y(t_2,x_0)] + [y(t_2,x_0) - y(t_1,x_0)] + [y(t_1,x_0) - y(t_1,x)] \\ &\ge -2H_{\text{space}}|x - x_0|^{\frac{1}{2}} + M(t_2 - t_1)^{\beta} \ge \frac{M}{2}(t_2 - t_1)^{\beta}, \end{aligned}$$

where we also used (12.9). For $L = L(\Omega, H_{\text{space}}) > 0$ appropriate, we have $\{\xi = 1\} \subset \Omega$. We may estimate the term in (12.8) from below as follows,

$$\int_\Omega \xi(x)[y(t_2,x) - y(t_1,x)]\,\mathrm{d}x \ge \frac{M}{2}(t_2 - t_1)^{\beta}\frac{1}{2L}\frac{M^2}{16H_{\text{space}}^2}(t_2 - t_1)^{2\beta} = CM^3(t_2 - t_1)^{3\beta}. \tag{12.10}$$

Step 3. We derive an upper bound for the right-hand side of (12.7). The first term can be estimated as follows

$$\begin{aligned} &\int_0^T \int_\Omega f_\varepsilon(y) y_{xxx}\phi_x \,\mathrm{d}x\,\mathrm{d}t \\ &\le \|f_\varepsilon(y)\|_{L^\infty(\Omega_T)}\|y_{xxx}\|_{L^2(L^2)}\left(\int_0^T\int_\Omega [\xi'(x)]^2[\theta_\delta(t)]^2\,\mathrm{d}x\,\mathrm{d}t\right)^{\frac{1}{2}} \\ &\le \|f_\varepsilon(y)\|_{L^\infty(\Omega_T)}\|y_{xxx}\|_{L^2(L^2)}\overbrace{\left(\int_\Omega [\xi'(x)]^2\,\mathrm{d}x\right)^{\frac{1}{2}}}\underbrace{\left(\int_0^T [\theta_\delta(t)]^2\,\mathrm{d}t\right)^{\frac{1}{2}}} \\ &\le C(H_{\text{space}})\overbrace{\frac{1}{\frac{M^2}{16H_{\text{space}}^2}(t_2 - t_1)^{2\beta}}\|\xi_0'\|_{L^\infty(\Omega)}\frac{M}{4H_{\text{space}}}(t_2 - t_1)^{\beta}}\underbrace{2(t_2 - t_1 + 2\delta)^{\frac{1}{2}}}, \end{aligned}$$

where we used that the first two norms are uniformly bounded via Lemma 12.4 by $C(H_{\text{space}})$. The factor $\frac{M}{4H_{\text{space}}}(t_2 - t_1)^{\beta}$ is the integral of 1 over $\operatorname{supp}\xi$, while $(t_2 - t_1 + 2\delta)^{\frac{1}{2}}$ is the Lebesgue measure of the support of θ_δ, where we use that θ_δ is uniformly bounded by 2 (We highlight the affiliation of each term in the last estimate). We emphasize that the constant C does depends on H_{space} from Lemma 12.4 (i.e., on T, C_0, y_0, and u), but it does not depend on ε, M or δ.

We estimate the remaining terms in (12.7),

$$\begin{aligned} \int_0^T \int_\Omega g_0(y) y_x \phi_x \,\mathrm{d}x\,\mathrm{d}t &\le \|g_0(y)\|_{L^\infty[C_0,\infty)}\|y_x\|\|\phi_x\| \\ &\le C(H_{\text{space}})\frac{1}{M}(t_2 - t_1)^{-\beta}(t_2 - t_1 + 2\delta)^{\frac{1}{2}}, \end{aligned}$$

where we used the same calculation as above for ϕ_x as well as Lemma 12.4 and we considered (A1) from Hypothesis 12.3. In case of (A2) from Hypothesis 12.3, we estimate as follows.

$$\int_0^T \int_\Omega g_0(y) y_x \phi_x \,\mathrm{d}x\,\mathrm{d}t \le \|g_0(y)\|_{L^2[C_0,\infty)} \|y_x\|_{L^\infty(\Omega_T)} \|\phi_x\| \\ \le C(H_{\text{space}}) \frac{1}{M} (t_2 - t_1)^{-\beta} (t_2 - t_1 + 2\delta)^{\frac{1}{2}},$$

where we used Lemma 12.4 and Sobolev embeddings. The last term is easy to estimate,

$$\int_0^T \int_\Omega u \phi_x \,\mathrm{d}x\,\mathrm{d}t \le \|u\| \|\phi_x\| \le C(H_{\text{space}}) \frac{1}{M} (t_2 - t_1)^{-\beta} (t_2 - t_1 + 2\delta)^{\frac{1}{2}}.$$

Step 4. For $\delta \to 0$, we get at the end

$$M^3 (t_2 - t_1)^{3\beta} \le C \frac{1}{M} (t_2 - t_1)^{\frac{1}{2} - \beta},$$

where the constant C is independent of x_0, t_1, t_2 and M. This leads to $M \le \sqrt[4]{C}$, which contradicts (12.6), and the lemma follows. □

The following lemma does not hold for each potential function g_0 which satisfies Hypothesis 12.3 since derivatives of g_0 come into play. We restrict ourselves to the two main examples.

Lemma 12.6
Let $\varepsilon > 0$, let $g_0 \equiv 0$ or $g_0 \equiv -1$. Moreover, let $u \in L^2(H^1)$, and let $y : \Omega_T \to \mathbb{R}$ be a solution of (12.1) with $y \ge C_0$ a.e. Then, for every $\sigma > 0$, there holds

$$\|y_{xx}\|^2_{L^\infty(L^2)} + (C_0^3 + \varepsilon) \|y_{xxxx}\|^2_{L^2(L^2)} \le \sigma \|y_{xxxx}\|^2_{L^2(L^2)} + C(\sigma)(\|u_x\|^2_{L^2(L^2)} + 1), \quad (12.11)$$

where $C(\sigma)$ denotes a positive constant depending on $\sigma > 0$.

Proof
We rewrite the main part of the equation (12.1) in non-divergence form,

$$\left([f(y) + \varepsilon] y_{xxx}\right)_x = [f(y)]_x y_{xxx} + [f(y) + \varepsilon] y_{xxxx}, \qquad \left(g_0(y) y_x\right)_x = g_0(y) y_{xx},$$

where we already know that $[f(y)]_x = 3y^2 y_x$. We multiply (12.1) with y_{xxxx}, integrate over Ω and arrive for $\sigma > 0$ at

$$\frac{1}{2}\frac{d}{dt}\|y_{xx}\|^2 + \underbrace{\int_\Omega [f(y) + \varepsilon] y_{xxxx}^2 \,\mathrm{d}x}_{=:I_1} \le \underbrace{-\int_\Omega [f(y)]_x y_{xxx} y_{xxxx} \,\mathrm{d}x}_{=:I_2} \quad (12.12) \\ + \underbrace{\int_\Omega |g_0(y)||y_{xx}||y_{xxxx}| \,\mathrm{d}x}_{=:I_3} + \sigma \|y_{xxxx}\|^2 + C(\sigma) \|u_x\|^2.$$

We calculate for $\sigma > 0$

$$
\begin{aligned}
I_1 &\geq (\lambda C_0^3 + \varepsilon)\|y_{xxxx}\|^2, \\
I_2 &\leq C\|y\|_{L^\infty}^2 \underbrace{\|y_x\|_{L^\infty}}_{=:I_4} \underbrace{\|y_{xxx}\|}_{=:I_5} \|y_{xxxx}\|, \\
I_3 &\leq \sigma\|y_{xxxx}\|^2 + C(\sigma)\|g_0(y)\|_{L^\infty(\Omega)}^2\|y_{xx}\|^2, \\
I_4 &\leq C\|y_x\|^{\frac{1}{2}}\|y_{xx}\|^{\frac{1}{2}} \leq C\|y\| + C\|y_x\|, \\
I_5 &\leq \sigma\|y_{xxxx}\| + C(\sigma)\|y\|,
\end{aligned}
$$

where we used that $\Omega \subset \mathbb{R}$ (for I_4) and we used [3, Theorem 5.2(1)] (for I_5). With the estimates of I_4 and I_5, we arrive at

$$
\begin{aligned}
I_2 + I_3 = \sigma C_1\|y_{xxxx}\|^2 &+ \sigma\|y_{xxxx}\|^2 \\
&+ C(\sigma)\|y\|_{L^\infty(\Omega)}^2\big(\|y_x\| + \|y\|\big)^2\|y\|^2 + C(\sigma)\|g_0(y)\|_{L^\infty(\Omega)}^2\|y_{xx}\|^2,
\end{aligned} \tag{12.13}
$$

where C_1 depends on T, C_0, y_0, u and comes from Lemma 12.4, but is independent of σ. The constant $C(\sigma)$ is also justified from Lemma 12.4 and depends on T, C_0, y_0, u, and on σ.

We absorb the first two terms (with a leading $\sigma > 0$) in the first row of (12.13) into the lower bound of I_1. Since the remaining two terms in the last row of (12.13) which are led by $C(\sigma)$ are integrable in time by (12.3), we deduce (12.11) with Gronwall's lemma. □

12.2. Existence

For every $\varepsilon > 0$, the regularized equation (12.1) has at least one weak solution.

Lemma 12.7
Let $\varepsilon > 0$, let Hypothesis 12.3 be true, and let $u \in L^2(\Omega_T)$. Then (12.1) has at least one solution $y \in L^2(H^3) \cap H^1(H^{-1})$.

Proof
This follows from standard parabolic theory since the leading part of the equation is uniformly parabolic. □

In some cases, it is also possible to prove uniqueness. We note that the subsequent analysis does not require uniqueness of (12.1).

Lemma 12.8
Let $\varepsilon > 0$, let $g_0 \equiv 0$ or $g_0 \equiv -1$, and let $u \in L^2(H^1)$. Then, the solution obtained in Lemma 12.7 is unique.

Proof
Let $y, z \in L^2(H^3) \cap H^1(H^{-1})$ be solutions of (12.1).

Step 1. We first note that due to Lemma 12.6, we have $y, z \in L^2(H^4) \cap H^1(L^2)$. The difference $e := y - z$ fulfills $e(0, .) = 0$ and

$$
e_t + \varepsilon e_{xxxx} = -(f(y)e_{xxx})_x - \big([f(y) - f(z)]z_{xxx}\big)_x - (g_0(y)e_x)_x. \tag{12.14}
$$

Step 2. We multiply (12.14) with $-e_{xx}$, integrate over Ω and arrive at

$$\frac{1}{2}\frac{d}{dt}\|e_x\|^2 + \varepsilon\|e_{xxx}\|^2 = -(f(y)e_{xxx}, e_{xxx}) - ([f(y)-f(z)]z_{xxx}, e_{xxx}) - (g_0(y)e_x, e_{xxx}) =: I + II + III. \tag{12.15}$$

We estimate the right-hand side of (12.15),

$$\begin{aligned} I &\leq 0, \\ II &\leq C\|e\|\|z_{xxx}\|_{L^\infty(\Omega)}\|e_{xxx}\| \leq \frac{\varepsilon}{2}\|e_{xxx}\|^2 + C(\varepsilon)\|z_{xxxx}\|^2\|e\|^2, \\ III &= (g_0(y)e_{xx}, e_{xx}) \leq 0, \end{aligned}$$

where we used that $f(y) \geq 0$ (in term I), the fact that f is Lipschitz and one-dimensional Sobolev embeddings for the z-term (in term II), and the fact that g_0 is either zero or constant negative (in term III). Absorbing $\frac{\varepsilon}{2}\|e_{xxx}\|^2$ to the left-hand side, we arrive at

$$\frac{1}{2}\frac{d}{dt}\|e_x\|^2 + \frac{\varepsilon}{2}\|e_{xxx}\|^2 \leq C(\varepsilon)\|z_{xxxx}\|^2\|e\|^2.$$

By Gronwall's lemma, we deduce that $e_x \equiv 0$.

Step 3. We multiply (12.14) with e, integrate over Ω and arrive at

$$(e_t, e) + \varepsilon(e_{xxxx}, e) = (f(y)e_{xxx}, e_x) + ([f(y)-f(z)]z_{xxx}, e_x) + (g_0(y)e_x, e_x) = 0, \tag{12.16}$$

since $e_x \equiv 0$ by the previous part of the proof. We conclude that $e \equiv 0$. □

In general, there is no solution of (10.1) for $\varepsilon = 0$, g_0, and u arbitrary. For the next chapter, we have to restrict ourselves to special cases of g_0 which provides a solution for at least one right-hand side u for $\varepsilon = 0$ in order to have a non-empty feasible set for the optimization problem.

Hypothesis 12.9
We assume that the potential function g_0 is in such a way that there exists at least one function $u \in L^2(H_0^1)$ such that there exists a global solution y of (12.1) and a constant $C_0 > 0$ with $y \geq 2C_0 > 0$ in Ω_T for the case $\varepsilon = 0$.

Remark 12.10
Hypothesis 12.9 is valid for, e.g., $g_0 \equiv 0$ (see [19]), or $g_0 \equiv -1$ (see [20]). In both cases, Hypothesis 12.9 holds for $u \equiv 0$.

Even for $g_0 \equiv 0$, equation (10.1) might be too degenerate to have a solution for a general right-hand side $u \in L^2(H^1)$. There are two ways to construct a solution of (10.1) by a sequence (y_ε) solving (12.1) for a sequence $\varepsilon \to 0$: Either, we restrict ourselves to more regular right-hand sides $u \in L^2(H^2)$ which allows uniform estimates as in Lemma 12.6 with respect to $\varepsilon > 0$. Another possibility, which we will use in the optimization problem is the following: If all iterates y_ε have a pointwise lower bound which is uniformly bounded away from zero with respect to $\varepsilon > 0$, then it is also possible to pass to the limit, even without the use of more regular right-hand sides u. The uniform lower bound is obtained, e.g., when the sequence (y_ε) ensembles from solutions of an optimization problem with suitable state constraints. The following two lemmas reflect both situations separately.

Lemma 12.11
Let $g_0 \equiv 0$ or $g_0 \equiv -1$, $u \in L^2(H^2)$ and let y_ε be the solution of (12.1). Then, there exist a function $y \in H^1(L^2) \cap L^2(H^4)$ and a subsequence (still denoted by ε) such that $y_\varepsilon \to y$ uniformly in Ω_T. The limit function y solves (10.1).

Proof
In the case $u \in L^2(H^2)$, the proof of Lemma 12.4 can be modified such that $y_{\varepsilon,x} \in L^\infty(L^2)$ without the need of $y_\varepsilon \geq C_0$ (we have to perform integration by parts on the term $-(u_x, y_{xx}) = (u_{xx}, y_x) \leq \|u_{xx}\|^2 + \|y_x\|^2$ in (12.4), which can then be treated by Gronwall's lemma. In the case of $g_0 \equiv -1$, the first term on the right-hand side of (12.4) is just $-\|y_{xx}\|^2$ after integration by parts.), i.e., we have $y_\varepsilon \in \mathcal{C}(\mathcal{C}^{0,\frac{1}{2}})$ bounded uniformly with respect to $\varepsilon > 0$. Together with Lemma 12.5, we can deduce that the sequence (y_ε) is bounded uniformly in $\mathcal{C}^{0,\frac{1}{8}}(\mathcal{C}^{0,\frac{1}{2}})$, i.e., y_ε is equicontinuous and uniformly bounded and there exist a subsequence and y such that $y_\varepsilon \to y$ uniformly.

The fact that y solves (10.1) follows from [19, Theorem 3.1]. □

Lemma 12.12
Let $g_0 \equiv 0$ or $g_0 \equiv -1$, $u \in L^2(H^1)$ and let y_ε be the solution of (12.1) with $y_\varepsilon \geq C_0$ independent of $\varepsilon > 0$. Then, there exist a $y \in H^1(L^2) \cap L^2(H^4)$ and a subsequence (still denoted by ε) such that $y_\varepsilon \to y$ uniformly in Ω_T. The limiting function y solves (10.1).

Proof
If $y_\varepsilon \geq C_0$ uniformly in $\varepsilon > 0$, then it is possible to absorb all the terms to the second term on the left-hand side of (12.4) in the proof of Lemma 12.4, i.e., we get uniform (with respect of $\varepsilon > 0$) bounds for y_ε in the $L^2(H^3) \cap H^1(H^{-1})$ norm. We follow the proof of Lemma 12.6 to show that (y_ε) is uniformly bounded in $L^2(H^4) \cap H^1(L^2)$. By the uniform bounds, there exists a limiting function $y \in L^2(H^4) \cap H^1(L^2)$ such that $y_\varepsilon \rightharpoonup y$ weakly in $L^2(H^4) \cap H^1(L^2)$ (up to a subsequence). It remains to show that y solves (10.1). We can now either use the second part of the proof of Lemma 12.11 to conclude, or we verify it by hand: For the linear terms, this is clear. For the nonlinear terms, we calculate for $\varphi \in \mathcal{C}^\infty(\Omega_T)$ and the subsequence mentioned

$$\big(f(y_\varepsilon)y_{\varepsilon,xxx} - f(y)y_{xxx}, \varphi_x\big) = \big([f(y_\varepsilon) - f(y)]y_{\varepsilon,xxx}, \varphi_x\big) + \big(f(y)[y_{\varepsilon,xxx} - y_{xxx}], \varphi_x\big) \to 0.$$

For the second nonlinear term, we calculate

$$\big(g_0(y_\varepsilon)y_{\varepsilon,x} - g_0(y)y_x, \varphi_x\big) = \big([g_0(y_\varepsilon) - g_0(y)]y_{\varepsilon,x}, \varphi_x\big) + \big(g_0(y)[y_{\varepsilon,x} - y_x], \varphi_x\big).$$

This concludes the proof. □

In this section, we discussed different cases for which solvability of (10.1) and (12.1) may be established.

1. For $\varepsilon > 0$, g_0 satisfying Hypothesis 12.3, and an arbitrary $u \in L^2(\Omega_T)$, a solution y_ε of (12.1) exists; see Lemma 12.7.

2. For $\varepsilon = 0$, g_0 satisfying Hypothesis 12.9, $u \equiv 0$, and state constraints (i.e., $y \geq C_0$) being absent, equation (10.1) is solvable; cf. Hypothesis 12.9 and Remark 12.10.

3. For $\varepsilon > 0$, $g_0 \equiv 0$, and regular right-hand sides $u \in L^2(H^2)$, the sequence of solutions (y_ε) of (12.1) is uniformly bounded, and the obtained limit y for $\varepsilon \to 0$ solves (10.1); see Lemma 12.11.

4. For $\varepsilon > 0$, $g_0 \equiv 0$ or $g_0 \equiv -1$, the sequence of solutions (y_ε) of (12.1) converges to a solution y of (10.1), if all y_ε are uniformly (with respect to ε) bounded away from zero, i.e., there exists a constant C_0 (independent of $\varepsilon > 0$) such that $y_\varepsilon \geq C_0$; see Lemma 12.12.

13. Analysis of the optimization problem without regularization

In this chapter, we want to show solvability for the original optimization Problem 10.1 and derive necessary optimality conditions. This seems to be possible for a general potential g_0 which satisfies Hypothesis 12.3 and Hypothesis 12.9. The analysis in the following chapters relies on Lemmas 12.11 and 12.12, which is only shown for $g_0 \equiv 0$ and $g_0 \equiv -1$ (hence also for $g_0 \equiv c$ for $c < 0$). In order to keep arguments and calculations as easy as possible, we set $g_0 \equiv 0$ from now on, but the results are also valid (at least) for $g_0 \equiv -1$.

In order to use the Lagrange multiplier theorem, we have to ensure a certain regularity for the optimal solution of the optimization problem stated below. Since the control u in (10.1) is only in $L^2(\Omega_T)$ (due to the structure in the cost functional J), the desired regularity of the corresponding optimal state may not be reached. This regularity is a crucial property needed in the next chapter. To overcome this issue, we restrict proper controls in (10.1) to those of the form $\left(\Delta^{-1} u_x\right)_x$ instead of u_x; see (13.2). The choice of this particular term is not immediate, but crucial for the rest of this part: First, in order to exclude a trivial optimization problem, we want the mass to be conserved (otherwise an optimal control would lead to local evaporation or wetting effects), hence the modified term needs to be in divergence form. To determine the involved amount of derivatives in the modified term, a deep look into the proof of Theorem 14.4 is needed in order to uniformly bound all emerging terms there.

As we already discussed in the introduction of this part, an alternative is the use of an additional term $\|u_x\|^2$ in the functional, which also ensures the desired regularity. However, this would lead to second spatial derivatives of u in the optimality condition (13.13g). This equation is later used in order to show uniform bounds of all involved functions, but since we do not know if second derivatives of u are uniformly bounded, this alternative approach does not seem promising.

We now state the modified form of Problem 10.1.

Problem 13.1
Let $\alpha > 0$, $\tilde{y} \in L^2(\Omega_T)$. Minimize

$$J(y, u) := \frac{1}{2} \int_0^T \int_\Omega |y - \tilde{y}|^2 \,\mathrm{d}x \,\mathrm{d}t + \frac{\alpha}{2} \int_0^T \int_\Omega |u|^2 \,\mathrm{d}x \,\mathrm{d}t$$

subject to $y \geq C_0$ and

$$y_t = -\left(f(y) y_{xxx}\right)_x + q_x, \tag{13.1}$$

$$-q_{xx} = u_x, \tag{13.2}$$

together with initial condition $y(0,.) = y_0$, boundary conditions $y_x = y_{xxx} = 0$ and $q = 0$ in a, b.

Theorem 13.2
Problem 13.1 has at least one solution.

PROOF
Step 1. By Hypothesis 12.9, there exist at least $\underline{q} \in L^2(H^1)$ and $\underline{u} \in L^2(\Omega_T)$ such that all side constraints (i.e., the equation (13.1), (13.2), and $y \equiv y(\underline{q}(\underline{u})) \geq C_0$ in Ω_T) are satisfied. Therefore, we have

$$\inf J(y, u) =: J^* > -\infty,$$

where the infimum is taken over all feasible pairs (y, u).

Step 2. By the first step, there exists a sequence $\{(y_i, q_i, u_i)\}$ fulfilling (13.1), (13.2), and $y_i \geq C_0$ with $J(y_i, u_i) \searrow J^*$. Therefore, u_i is bounded in $L^2(\Omega_T)$ and there exists a $u \in L^2(\Omega_T)$ such that $u_i \rightharpoonup u$ weakly in $L^2(\Omega_T)$ (up to subsequences).

Since q_i solves the Poisson equation, we know that q_i is bounded in $L^2(H^1)$ and there exists a $q \in L^2(H^1)$ with $q_i \rightharpoonup q$ weakly in $L^2(H^1)$ (up to subsequences). It is clear that (q, u) solve (13.2).

By Lemma 12.12, there exists a $y \in L^2(H^4) \cap H^1(L^2)$ such that $y_i \rightharpoonup y$ weakly in $L^2(H^4) \cap H^1(L^2)$ and $y_i \to y$ uniformly in Ω_T, and y solves (13.1). Moreover, we have $y \geq C_0$.

Step 3. By the weak lower semicontinuity of the functional J, (y, u) is a minimum of Problem 13.1. □

By the nonlinearity of the leading equation (13.1), it is clear that a minimum need not be unique. In the remainder of this chapter, we will derive necessary optimality conditions for a minimum obtained by Theorem 13.2. A problem here is that a classical Slater type result (i.e., the feasible set must have nonempty interior) cannot be used, since the Slater type condition of a state constraint $y \geq C_0$ requires pointwise information (i.e., in $\mathcal{C}(\Omega_T)$; but this is not available since we need to get it on the set of solutions which has no interior; cf. [4]).

The key step to derive this is the following abstract result about optimal control problems with state constraints, which is obtained in [4].

Lemma 13.3
Let X, V, W be Banach spaces, U be a separable Banach space, let $J : X \times U \to \mathbb{R}$, $G : X \times U \to V$, $H : X \to W$ be mappings, and $C \subseteq W$ be a set.

Let $(\bar{x}, \bar{u}) \in X \times U$ be a minimum of the optimal control problem

$$J(\bar{x}, \bar{u}) = \min_{(x,u) \in S} J(x, u)$$

with

$$S := \{(x, u) \in X \times U : \; G(x, u) = 0, \; H(x) \in C\}$$

and let the following assumptions be true.

1. $G : X \times U \to V$ is Frechet differentiable at $(\bar{x}, \bar{u})$,
2. $H : X \to W$ is Frechet differentiable at $\bar{x}$,
3. $\varnothing \neq C \subseteq W$ is a convex subset with nonempty interior (measured in the topology of W),
4. $G'_x(\bar{x}, \bar{u}) : X \to V$ is surjective.

Then there exist $(p, \mu, \zeta) \in V^* \times W^* \times \mathbb{R}$ such that

$$\zeta \langle J'_x(\bar{x}, \bar{u}), x \rangle_{X,X^*} + \langle p, G'_x(\bar{x}, \bar{u})x \rangle_{V,V^*} + \langle \mu, H'(\bar{x})x \rangle_{W,W^*} = 0 \quad \forall x \in X, \tag{13.3a}$$

$$\zeta \langle J'_u(\bar{x}, \bar{u}), u \rangle_{U,U^*} + \langle p, G'_u(\bar{x}, \bar{u})u \rangle_{V,V^*} = 0 \quad \forall u \in U, \tag{13.3b}$$

$$\zeta \geq 0, \tag{13.3c}$$

$$\langle \mu, w - H(\bar{x}) \rangle_{W,W^*} \leq 0 \quad \forall w \in C \tag{13.3d}$$

and if $\zeta = 0$ then $\langle \mu, w \rangle_{W,W^*} \neq 0$ for some $w \in C$.

If we additionally assume that there exists $(\underline{x}, \underline{u}) \in X \times U$ such that

$$G'_x(\bar{x}, \bar{u})\underline{x} + G'_u(\bar{x}, \bar{u})(\underline{u} - \bar{u}) = 0, \tag{13.4a}$$

$$H(\bar{x}) + H'(\bar{x})\underline{x} \in \operatorname{int} C, \tag{13.4b}$$

then we can take $\zeta = 1$.

We now apply this general result to our setup in Problem 13.1. We define the spaces

$$X := X_y \times X_q, \qquad X_y := L^2(H^4) \cap H^1(L^2), \qquad X_q := L^2(H^1_0),$$

as well as the spaces $U := L^2(L^2)$, $V := L^2(L^2) \times L^2(H^{-1})$, $W := X_y$, and the set $C := \{v \in W : \ v \geq C_0 \text{ in } \Omega_T\}$. Since $W \subset \mathcal{C}(\overline{\Omega_T})$ by Sobolev embeddings, the set C is well-defined.

The function G is given by

$$G((y, q), u) := \begin{pmatrix} y_t + \left(f(y) y_{xxx}\right)_x - q_x \\ -q_{xx} - u_x \end{pmatrix},$$

while H is given by $H(y, q) := y$. We omit initial conditions and boundary conditions in G, which may be treated by standard methods; see, e.g., [40, Section 2.6].

Lemma 13.4

1. The function $G : X \times U \to V$ is well-defined.
2. The function $H : X \to W$ is well-defined.
3. The set C is convex with nonempty interior (measured in the topology of W).

Proof

1. This follows from Lemma 12.6.

2. Clear by definition.

3. Clearly, the set C is convex, since it is the intersection of two convex sets. We note that the set $\tilde{C} := \{v \in \mathcal{C}(\overline{\Omega_T}) : \; v \geq C_0 \text{ in } \Omega_T\}$ has nonempty interior (e.g., $\hat{v} \equiv 2C_0$ is an interior point), i.e., there exist a point $\hat{v} \in C$ and $r > 0$ such that $B_r(\hat{v}) \subset \tilde{C}$. Without loss of generality, we can assume that $\hat{v} \in W$ due to the density of $W \subset \mathcal{C}(\overline{\Omega_T})$. Since the embedding mapping id : $W \to \mathcal{C}(\overline{\Omega_T})$ is continuous by Sobolev embeddings, the preimage $\text{id}^{-1}(B_r(\hat{v})) \subset C$ is open, hence there exists an open neighborhood of $\text{id}^{-1}(\hat{v})$, which means that C has nonempty interior in the topology of W. □

Remark 13.5

As of this place, it seems non straight-forward to use

$$W = L^2(H^4) \cap H^1(L^2), \qquad C = \{v \in W : \; v \geq C_0 \text{ in } \Omega_T\},$$

instead of simply using $W = \mathcal{C}(\overline{\Omega_T})$ and C accordingly.

This particular choice will be evident in the proof of Theorem 14.4, where we need to bound the Lagrange multipliers μ_ε associated to the state constraint $y \geq C_0$ uniformly with respect to $\varepsilon > 0$, i.e., we need to bound some dual pairings $\langle \mu_\varepsilon, \varphi \rangle$ for all φ with $\|\varphi\|_W \leq 1$. If we choose $W = \mathcal{C}(\overline{\Omega_T})$, we would only know

$$\sup_{(t,x) \in \overline{\Omega_T}} |\varphi(t,x)| \leq 1,$$

which is not enough to bound all emerging terms. However, the choice $W = L^2(H^4) \cap H^1(L^2)$ allows to bound all those terms and thus to prove Theorem 14.4.

We now check that the remaining assumptions in Lemma 13.3 are valid. In order to write down (13.3), we have to show that $G'_x(\bar{x}, \bar{u}) : X \to V$ is surjective, which is done in the following.

Lemma 13.6

1. The function G_1, which is defined as

$$G_1\big((y,q),u\big) := y_t + (f(y) y_{xxx})_x - q_x,$$

has the following Frechet derivative

$$\begin{aligned}
\Big\langle G'_{1,y}\big((\bar{y},\bar{q}),\bar{u}\big), \delta y \Big\rangle &= (\delta y)_t + \big(\langle f'(\bar{y}), \delta y\rangle \bar{y}_{xxx}\big)_x + \big(f(\bar{y})(\delta y)_{xxx}\big)_x && \forall \delta y \in X_y,\\
\Big\langle G'_{1,q}\big((\bar{y},\bar{q}),\bar{u}\big), \delta q \Big\rangle &= -(\delta q)_x && \forall \delta q \in X_q,\\
\Big\langle G'_{1,u}\big((\bar{y},\bar{q}),\bar{u}\big), \delta u \Big\rangle &= 0 && \forall \delta q \in U.
\end{aligned}$$

2. The function G_2, which is defined as

$$G_2\big((y,q),u\big) := -q_{xx} - u_x,$$

has the following Frechet derivative

$$\begin{aligned}
\left\langle G'_{2,y}((\bar{y},\bar{q}),\bar{u}),\delta y\right\rangle &= 0 && \forall \delta y\in X_y,\\
\left\langle G'_{2,q}((\bar{y},\bar{q}),\bar{u}),\delta q\right\rangle &= -(\delta q)_{xx} && \forall \delta q\in X_q,\\
\left\langle G'_{2,u}((\bar{y},\bar{q}),\bar{u}),\delta u\right\rangle &= -(\delta u)_{x} && \forall \delta u\in U.
\end{aligned}$$

PROOF
The function G is smooth and the derivation of it is a straight forward calculation. □

Lemma 13.7

1. For every $\Phi\in L^2(L^2)$ and every $w\in L^2(H_0^1)$, there exists a $v\in L^2(H^4)\cap H^1(L^2)$ such that
$$\left\langle G'_{1,y}((\bar{y},\bar{q}),\bar{u}),v\right\rangle+\left\langle G'_{1,q}((\bar{y},\bar{q}),\bar{u}),w\right\rangle=\Phi \tag{13.5}$$
together with the initial conditions $v(0,.)=0$ and $w(0,.)=0$ as well as the boundary conditions $v_x=v_{xxx}=0$ and $w=0$ in a,b.
2. For every $\Psi\in L^2(H^{-1})$ and every $v\in L^2(H^4)\cap H^1(L^2)$, there exists a $w\in L^2(H_0^1)$ such that
$$\langle G_{2,y}((\bar{y},\bar{q}),\bar{u}),v\rangle+\left\langle G'_{2,q}((\bar{y},\bar{q}),\bar{u}),w\right\rangle=\Psi \tag{13.6}$$
together with the initial conditions $v(0,.)=0$ and $w(0,.)=0$ as well as the boundary conditions $v_x=v_{xxx}=0$ and $w=0$ in a,b.

PROOF

1. Inserting the derivative of G_1 with respect to y by Lemma 13.6, equation (13.5) reads as
$$v_t+\left(f(\bar{y})v_{xxx}\right)_x+\text{ lower order terms }=\Phi. \tag{13.7}$$
For a test function $\varphi\in X_y$, we write
$$\left\langle\left(f(\bar{y})v_{xxx}\right)_x,\varphi\right\rangle=-\langle f(\bar{y})v_{xxx},\varphi_x\rangle=\langle f(\bar{y})v_{xx},\varphi_{xx}\rangle+\langle f'(\bar{y})\bar{y}_xv_{xx},\varphi_x\rangle. \tag{13.8}$$
Since $f'(\bar{y})=3\bar{y}^{-1}f(\bar{y})$, on nothing that $\bar{y}\geq C_0>0$, we can estimate the last term in (13.8) as follows
$$\langle f'(\bar{y})\bar{y}_xv_{xx},\varphi_x\rangle\leq\sigma\|f(\bar{y})v_{xx}\|^2+C(\sigma)\|\bar{y}\bar{y}_x\varphi_x\|^2$$
with $\sigma>0$. The remaining term in (13.8) is either uniformly H^2-coercive (since $\bar{y}\geq C_0$) or is of lower order. Therefore, there exists a solution $v\in L^2(H^2)\cap H^1(H^{-1})$ of (13.7).

   As in the proof of Lemma 12.6, we can write
$$\left(f(\bar{y})v_{xxx}\right)_x=f(\bar{y})v_{xxxx}+f'(\bar{y})\bar{y}_xv_{xxx}, \tag{13.9}$$

i.e., the leading part of the equation (13.7) is uniformly elliptic since $\bar{y} \geq C_0$. Similar to the proof in Lemma 12.6, it is possible to multiply the equation with v_{xxxx} and to absorb the lower order terms into the leading term in (13.9). Therefore, it is possible to show that the solution v is as regular as claimed.

The choice of w is arbitrary, since it is of lower order and the related can be put into the right hand side Φ in the first place.

2. By Lemma 13.6, equation (13.6) reads as the Poisson equation, which is surjective on the corresponding spaces. □

We will now show that the regular point conditions (13.4a) and (13.4b) from Lemma 13.3 are fulfilled. For this goal, it is important to make use of the surjectivity of the derivative of G.

Lemma 13.8
There exists $(\underline{x}, \underline{u}) \in X \times U$ such that (13.4a) and (13.4b) are fulfilled.

Proof

Step 1. First, we note that $\operatorname{int} C - H(\bar{x}) = \{f \in \mathcal{C}(\Omega_T) : \; f > C_0 - \bar{y}\}$. Since $H'(\bar{x})\underline{x} = \underline{y}$, we have to choose $\underline{y} \in X_y$ such that $\underline{y} > C_0 - \bar{y}$ in Ω_T to meet (13.4b), which is always possible (e.g., we can choose $\underline{y} = 2C_0$).

Step 2. Now, we take a look at the first component of the equation (13.4a)

$$G'_x(\bar{x}, \bar{u})\underline{x} + G'_u(\bar{x}, \bar{u})(\underline{u} - \bar{u}) = 0,$$

which can be written as

$$\left\langle G'_{1,y}((\bar{y}, \bar{q}), \bar{u}), \underline{y} \right\rangle = \underline{q}_x \tag{13.10}$$

due to Lemma 13.6. By Lemma 13.7, the left-hand side of (13.10) is surjective, i.e., there exists a $\underline{q} \in X_q$ such that (13.10) holds.

Step 3. Finally, we take a look at the second component of (13.4a), which reads as

$$-\underline{q}_{xx} = \underline{u}_x - \bar{u}_x =: \tilde{u}_x. \tag{13.11}$$

This equation is the Poisson equation, which is known to be surjective on the corresponding spaces, i.e., there exists $\tilde{u}_x \in L^2(H^{-1})$ such that (13.11) holds. Since $\bar{u}_x$ is known and we do not have additional constraints on u, there exists a $\underline{u}_x \in L^2(H^{-1})$ such that (13.11) holds. But this implies the existence of $\underline{u} \in L^2(L^2) = U$ such that (13.11) holds.

To summarize, we have constructed $\left((\underline{y}, \underline{q}), \underline{u}\right) \in X \times U$ such that both conditions (13.4a) and (13.4b) hold. □

Remark 13.9
For a leading equation of second order (instead of the fourth order equation, which we have here), the proof of Lemma 13.8 would work in a much more general setting: In (13.4b), we

have to show that there exists a $\underline{y} \in X_y$ such that $\bar{y} + \underline{y} > C_0$ and $\underline{y}$ is a solution of the linearized equation (13.4a). Since u can be chosen arbitrarily, (13.4a) reads as

$$G'_x(\bar{x}, \bar{u})\underline{x} = \Phi, \tag{13.12}$$

where Φ can have an arbitrary sign (There are no additional constraints on u). If G contains an parabolic equation of second order, equation (13.12) would read as an linear parabolic equation of second order. There holds a maximum principle for such equations, i.e., if Φ has a certain sign, we can guarantee that $\underline{x}$ has also a sign making it easier to show (13.4b), where this information is useful.

Theorem 13.10
Let (y, q, u) be a solution of Problem 13.1. Then, there exist $z \in L^2(L^2)$, $\eta \in L^2(H^1_0)$, and $\mu \in (L^2(H^4) \cap H^1(L^2))^*$ such that the following optimality conditions are fulfilled.

$$\begin{aligned}
y_t =& \; - \left([f(y)]y_{xxx}\right)_x + q_x, && (13.13a)\\
-q_{xx} =& \; u_x, && (13.13b)\\
y \geq& \; C_0 && (13.13c)\\
0 \geq& \; \langle w - y, \mu\rangle \quad \forall X_y \ni w \geq C_0, && (13.13d)\\
0 =& \; \langle y - \tilde{y}, \varphi\rangle + \langle z, \varphi_t + (f'(y)y_{xxx}\varphi)_x\rangle && (13.13e)\\
& + \langle z, \left([f(y)]\varphi_{xxx}\right)_x\rangle + \langle \varphi, \mu\rangle \quad \forall \varphi \in X_y, &&\\
0 =& \; z_x - \eta_{xx}, && (13.13f)\\
0 =& \; \alpha u + \eta_x && (13.13g)
\end{aligned}$$

together with initial conditions $y(0, .) = y_0$, $z(T, .) = 0$; boundary conditions $y_x = y_{xxx} = z_x = z_{xxx} = 0$ in a, b; as well as $q = \eta = 0$ in a, b.

Proof
We use Lemma 13.3; the conditions there are fulfilled by Lemma 13.4, Lemma 13.6, Lemma 13.7, and Lemma 13.8. □

14. Optimization with regularization in the equation

In this chapter, we consider a modification of Problem 13.1, where the state equation is regularized; the functional remains the same. After having shown solvability and having derived corresponding optimality conditions in Theorem 14.2 and Theorem 14.3, respectively, we will show that solutions of this problem converge to objects which solve (13.13), i.e., we show that solutions of the modified problem convergence to those of the original problem in a certain sense.

Problem 14.1
Let $\varepsilon > 0$. Suppose $\tilde{y} \in L^2(\Omega_T)$. Minimize

$$J(y,u) := \frac{1}{2}\int_0^T \int_\Omega |y-\tilde{y}|^2 \,\mathrm{d}x\,\mathrm{d}t + \frac{\alpha}{2}\int_0^T \int_\Omega |u|^2 \,\mathrm{d}x\,\mathrm{d}t$$

subject to $y \geq C_0$, (12.1), and (13.2), together with the initial condition $y(0) = y_0$.

Theorem 14.2
Problem 14.1 has at least one solution.

PROOF
The proof uses the the same argument as the proof of Theorem 13.2. □

Theorem 14.3
Let (y,q,u) be a minimum of Problem 14.1. Then, there exist Lagrange multipliers $z \in L^2(L^2)$, $\eta \in L^2(H_0^1)$, and $\mu \in (L^2(H^4) \cap H^1(L^2))^*$ such that the following equations are fulfilled.

$$\begin{aligned}
y_t =& -\left([f(y)+\varepsilon]y_{xxx}\right)_x + q_x, && (14.1a)\\
-q_{xx} =& \, u_x, && (14.1b)\\
y \geq& \, C_0 && (14.1c)\\
0 \geq& \, \langle w-y,\mu\rangle \quad \forall X_y \ni w \geq C_0, && (14.1d)\\
0 =& \, \langle y-\tilde{y},\varphi\rangle + \langle z, \varphi_t + (f'(y)y_{xxx}\varphi)_x\rangle && (14.1e)\\
&+ \langle z, \left([f(y)+\varepsilon]\varphi_{xxx}\right)_x\rangle + \langle \varphi,\mu\rangle \quad \forall \varphi \in X_y, &&\\
0 =& \, z_x - \eta_{xx}, && (14.1f)\\
0 =& \, \alpha u + \eta_x && (14.1g)
\end{aligned}$$

together with initial conditions $y(0,.) = y_0$, $z(T,.) = 0$; boundary conditions $y_x = y_{xxx} = z_x = z_{xxx} = 0$ in a,b; as well as $q = \eta = 0$ in a,b.

Proof
This is a direct consequence of Lemma 13.3. The details are similar to the proof of Theorem 13.10. □

Theorem 14.4
Let $\{(y_\varepsilon, q_\varepsilon, u_\varepsilon)\}$ be a sequence of solutions of Problem 14.1. Then, there exists $(y^*, q^*, u^*) \in (H^1(L^2) \cap L^2(H^4)) \times L^2(H_0^1) \times L^2(L^2)$ such that $(y_\varepsilon, q_\varepsilon, u_\varepsilon) \rightharpoonup (y^*, q^*, u^*)$ weakly in $(H^1(L^2) \cap L^2(H^4)) \times L^2(H_0^1) \times L^2(L^2)$ (up to a subsequence). The limit functions (y^*, q^*, u^*) are a solution of (13.13).

Proof
Step 1. First we prove that (u_ε) is uniformly bounded in $L^2(L^2)$: To do so, we want to find a function $\bar{u}$ and a corresponding solution $(\bar{y}_\varepsilon, \bar{q}_\varepsilon)$, which is feasible for every $\varepsilon > 0$ small enough, i.e., which is solving (12.1) and (13.2), together with $\bar{y}_\varepsilon \geq C_0$. For $u \equiv 0$ (hence $q \equiv 0$) and $\varepsilon = 0$, there exists a solution $\bar{y}$ of (10.1) (See Hypothesis 12.9), which satisfies $\bar{y} \geq 2C_0$. Let $y_\varepsilon^{(0)}$ be the solution of (12.1) for $q \equiv 0$ (hence $u \equiv 0$). Then there exists $y : \Omega_T \to \mathbb{R}$ such that $\bar{y}_\varepsilon^{(0)} \to y$ uniformly for $\varepsilon \to 0$, cf. Lemma 12.11. Hence there exists an $\varepsilon_0 > 0$ such that $\bar{y}_\varepsilon^{(0)} \geq C_0$ for every $0 < \varepsilon \leq \varepsilon_0$.

Since $\bar{y}_\varepsilon$ is uniformly bounded (with respect to $\varepsilon > 0$) in $L^2(H^4) \cap H^1(L^2)$ by a constant depending on the fixed norm of $\bar{u} \equiv 0$, we may deduce that the solution $(y_\varepsilon, q_\varepsilon, u_\varepsilon)$ of Problem 14.1 satisfies $J(y_\varepsilon, u_\varepsilon) \leq J(y_\varepsilon^{(0)}, 0) < \infty$, i.e., by construction of the functional J, the sequence (u_ε) is bounded uniformly in $L^2(L^2)$. Hence there exists a $u^* \in L^2(L^2)$ such that $u_\varepsilon \rightharpoonup u^*$ weakly in $L^2(L^2)$.

Step 2. By standard estimates for the Poisson equation, there exists a constant $C > 0$ such that for the solution q_ε of (13.2) holds

$$\|q_\varepsilon\|_{L^2(H_0^1)} \leq C \|u_\varepsilon\|_{L^2(L^2)},$$

i.e., there exists an $q^* \in L^2(H_0^1)$ such that $q \rightharpoonup q^*$ weakly in $L^2(H_0^1)$ thanks to the boundedness of $\{u_\varepsilon\}$ from the last step.

Step 3. By Lemma 12.6, the solution y_ε of (12.1) is uniformly bounded (with respect to $\varepsilon > 0$) in $L^2(H^4) \cap H^1(L^2)$, i.e., there exists $y^* \in L^2(H^4) \cap H^1(L^2)$ such that $y_\varepsilon \rightharpoonup y^*$ weakly in $L^2(H^4) \cap H^1(L^2)$. Since all $y_\varepsilon \geq C_0$, we have $y^* \geq C_0$ and y^* solves (13.1) by Lemma 12.12.

Step 4. We have shown so far that there exists (y^*, q^*, u^*) in the given spaces and they solve (13.13a), (13.13b) and (13.13c). It remains to show that the Lagrange multipliers $(\mu_\varepsilon, z_\varepsilon, \eta_\varepsilon)$ are uniformly bounded (with respect to ε) and that their limits solve the remaining equations in (13.13).

Step 5. Since (u_ε) is bounded in $L^2(L^2)$, we have (η_ε) bounded in $L^2(H_0^1)$ by (14.1g). By (14.1f), $(z_{\varepsilon,x})$ is bounded uniformly in $L^2(H^{-1})$. We will now consider (14.1e) and may show that μ_ε is uniformly bounded in $(L^2(H^4) \cap H^1(L^2))^*$, i.e., we have to show that

$$\|\mu_\varepsilon\|_{(L^2(H^4)\cap H^1(L^2))^*} = \sup_{\substack{\psi \in L^2(H^4)\cap H^1(L^2) \\ \|\psi\|_{L^2(H^4)\cap H^1(L^2)} \leq 1}} |\langle \mu_\varepsilon, \psi \rangle|$$

is bounded independently from $\varepsilon > 0$. Since μ_ε is on the right-hand side of (14.1e), we can represent μ_ε by means of y_ε and z_ε, i.e., we have

$$\begin{aligned}|\langle \mu_\varepsilon, \psi\rangle| &\leq |\langle y_\varepsilon - \tilde{y}, \psi\rangle| + |\langle z_\varepsilon, \psi_t\rangle| + |\langle z_{\varepsilon,x}, f'(y_\varepsilon) y_{\varepsilon,xxx}\psi\rangle| \\ &\quad + |\langle z_{\varepsilon,x}, [f(y_\varepsilon) + \varepsilon]\psi_{xxx}\rangle| =: I_1 + I_2 + I_3 + I_4.\end{aligned}$$

We estimate those terms as follows and use the bounds from the first steps (and Sobolev embeddings),

$$\begin{aligned}I_1 &\leq \|y_\varepsilon - \tilde{y}\|\|\psi\| \leq C, \\ I_2 &= |\langle z_{\varepsilon,x}, \psi_t\rangle| \leq \|z_{\varepsilon,x}\|\|\psi_t\| \leq C, \\ I_3 &\leq \|z_{\varepsilon,x}\|f'(y_\varepsilon) y_{\varepsilon,xxx}\psi\| \leq C, \\ I_4 &\leq \|z_{\varepsilon,x}\|\|[f(y_\varepsilon) + \varepsilon]\psi_{xxx}\| \leq C,\end{aligned}$$

where we used that $\|\psi\|_{L^2(H^4)\cap H^1(L^2)} \leq 1$.

Adding up, we arrive at $\sup_{\varepsilon>0} \|\mu_\varepsilon\|_{(L^2(H^4)\cap H^1(L^2))^*} \leq C$, i.e., $\{\mu_\varepsilon\}$ is uniformly bounded with respect to $\varepsilon > 0$.

Step 6. By the bounds from the previous step, there exist $\eta^* \in L^2(H_0^1)$, $z_x^* \in L^2(H^{-1})$, and $\mu^* \in (L^2(H^4) \cap H^1(L^2))^*$ such that $\eta_\varepsilon \rightharpoonup \eta^*$ weakly in $L^2(H_0^1)$, $z_{\varepsilon,x} \rightharpoonup z_x^*$ weakly in $L^2(H^{-1})$, and $\mu_\varepsilon \rightharpoonup \mu^*$ weakly in $(L^2(H^4) \cap H^1(L^2))^*$.

Step 7. With the bounds and the convergence from the last step, it is possible to show that by taking the limit in (14.1d), (14.1e), (14.1f), and (14.1g), respectively, that (η^*, z^*, μ^*) solve (13.13d), (13.13e), (13.13f) and (13.13g), respectively. This concludes the proof. □

15. Penalty approximation

In this chapter, we investigate a penalty approximation of Problem 14.1. The main idea is to add an additional non-negative term to the functional, which increases in value in cases where the state constraint $y \geq C_0$ does not apply. This additional term allows us to get rid of the state constraint, hence we can get rid of the non-regular Lagrange multiplier μ in the optimality system (14.1). On the opposite, a drawback is that in general this generates non-feasible solutions (with respect to the constraint $y \geq C_0$). When considering a well-posed equation, this is not that crucial, but in our case the original equation (13.1) may degenerate, and non-feasible solutions might even not exist. That is the reason for introducing the intermediate Problem 14.1 with the regularized equation (12.1).

We now introduce a penalty approximation of Problem 14.1, and prove the existence of a corresponding minimum, as well as convergence of minimizers to a minimum of Problem 14.1 for a fixed $\varepsilon > 0$. Then we derive optimality conditions, which are the starting point for numerical studies in chapter 16. For more details to the penalty approximation we refer the reader to [24, Section 1.10] and [58, Section 3].

Problem 15.1
Let $\varepsilon, \gamma > 0$. We define the functional

$$J_\gamma(y,u) := J(y,u) + \frac{1}{2\gamma}\int_0^T \int_\Omega \left|(C_0 - y)^+\right|^2 \,\mathrm{d}x\,\mathrm{d}t. \tag{15.1}$$

Find $(y_\gamma, q_\gamma, u_\gamma)$ as the minimum of J_γ subject to (12.1) and (13.2).

Remark 15.2
As mentioned in the introduction, the penalty method is not the only method for the regularization of the state constraint in Problem 14.1. We add a few words about two other prominent methods.

1. In the Moreau-Yosida approximation [44], we consider the function

$$J_\gamma(y,u) := J(y,u) + \frac{1}{2\gamma}\int_0^T \int_\Omega \left|(\bar\chi + \gamma(C_0 - y))^+\right|^2 \,\mathrm{d}x\,\mathrm{d}t,$$

   where $\bar\chi \in L^2(L^2)$ is given and $\gamma > 0$ should tend to zero. Note that the scaling is different from that in (15.1). For the Moreau-Yosida based approximation, a class of effective solvers are available, cf. [43]. However, it is not clear if the convergence result in [44] also holds for the nonlinear equation in (10.1).

2. In the Lavrentiev approximation [64], the state constraint $y \geq C_0$ is replaced by a mixed control-state constraint $\gamma u + y \geq C_0$ for some $\gamma > 0$, which should tend to zero. It is standard to show existence of an optimum subject to the mixed constraints

instead of pure state constraints. The necessary optimality conditions now involve a Lagrange multiplier $\mu_\gamma \in L^2(\Omega_T)$ with $\mu_\gamma \geq 0$ and the complementary condition [64, Theorem 3.3]

$$(\mu_\gamma, C_0 - \gamma\bar{u} - \bar{y}) = 0, \tag{15.2}$$

which is then solved together with the remaining part of the optimality condition with a semi-smooth Newton method. As in the case of the Moreau-Yosida approximation, the we are not aware if the Lavrentiev converges in the case of the governing equation (10.1), and thus is left open at this place.

Theorem 15.3
There exists at least a solution $(y_\gamma, q_\gamma, u_\gamma)$ of Problem 15.1.

Proof
Similar to the proof of Theorem 13.2 and Theorem 14.2. □

Theorem 15.4
Let $\varepsilon > 0$, and $\{(y_\gamma, q_\gamma, u_\gamma)\}$ be a sequence of solutions of Problem 15.1. Then, there exist $y^* \in L^2(H^4) \cap H^1(L^2)$, $q \in L^2(H_0^1)$, and $u^* \in L^2(L^2)$ such that $y_\gamma \rightharpoonup y^*$ weakly in $L^2(H^4) \cap H^1(L^2)$, $q_\gamma \rightharpoonup q^*$ weakly in $L^2(H_0^1)$, and $u_\gamma \rightharpoonup u^*$ weakly in $L^2(L^2)$ for $\gamma \to 0$. Moreover, (y^*, q^*, u^*) is a solution of Problem 14.1.

Proof
Let $(y_\gamma, q_\gamma, u_\gamma)$ be the solution of Problem 15.1.

Step 1. We first show that the functional is uniformly bounded (with respect to $\gamma > 0$): Let $(\bar{y}, \bar{q}, \bar{u})$ be the solution of Problem 14.1, i.e., $(\bar{y}, \bar{q}, \bar{u})$ solve (12.1) and (13.2), $\bar{y} \geq C_0$ and $J(\bar{y}, \bar{u})$ is minimal for all such (y, q, u). Since $\bar{y} \geq C_0$, we have $J_\gamma(\bar{y}, \bar{u}) = J(\bar{y}, \bar{u})$ independent of $\gamma > 0$.

By the minimizing property of $(y_\gamma, q_\gamma, u_\gamma)$, there holds

$$J_\gamma(y_\gamma, u_\gamma) \leq J_\gamma(\bar{y}, \bar{u}) = J(\bar{y}, \bar{u}) < \infty.$$

Hence, $J_\gamma(y_\gamma, u_\gamma)$ is uniformly bounded with respect to $\gamma > 0$.

Step 2. We want to get weak limit functions: From the definition of J_γ, we derive a uniform (with respect to $\gamma > 0$) bound for u_γ in the $L^2(L^2)$-norm. By a-priori estimates for the Poisson equation (13.2), q_γ is uniformly bounded in $L^2(H_0^1)$. By the a-priori estimates from Lemma 12.6, y_γ is uniformly (with respect to $\gamma > 0$) bounded in the $L^2(H^4) \cap H^1(L^2)$-norm. Therefore, there exists $(y^*, q^*, u^*) \in \big(L^2(H^4) \cap H^1(L^2)\big) \times L^2(H_0^1) \times L^2(L^2)$ such that $(y_\gamma, q_\gamma, u_\gamma) \rightharpoonup (y^*, q^*, u^*)$ weakly in the corresponding spaces.

Step 3. We want to show that the limit functions (y^*, q^*, u^*) are feasible for Problem 14.1: It is easy to verify that (y^*, q^*, u^*) solves (12.1) and (13.2) like it was done, e.g., in the proof of Theorem 13.2. It remains to show that $y^* \geq C_0$. Since $J_\gamma(y_\gamma, u_\gamma) \leq C$ uniformly in $\gamma > 0$, we know that for $\gamma \to 0$,

$$\int_0^T \int_\Omega \left|(C_0 - y_\gamma)^+\right|^2 \,\mathrm{d}t\,\mathrm{d}x \to 0,$$

i.e., we have $(C_0 - y_\gamma)^+ \to 0$ a.e. in Ω_T, which means $y^* \geq C_0$.

Step 4. Finally, we show that (y^*, q^*, u^*) is a solution of Problem 14.1: We have to show that $J(y^*, u^*) \leq J(y, u)$ for every (y, q, u) solving (12.1) and $y \geq C_0$.

Let $(\bar{y}, \bar{q}, \bar{u})$ be a solution of Problem 14.1. By the first parts of the proof, we know that (y^*, q^*, u^*) is feasible for Problem 14.1, i.e., we have $J_\gamma(y^*, u^*) = J(y^*, u^*)$. Since $(y_\gamma, q_\gamma, u_\gamma) \rightharpoonup (y^*, q^*, u^*)$ weakly in the corresponding spaces by the second part of the proof, and J is weakly lower semi-continuous, we have

$$J(y^*, u^*) \leq \liminf_{\gamma \to 0} J_\gamma(y_\gamma, u_\gamma) \leq J(\bar{y}, \bar{u}), \tag{15.3}$$

where we used the first part which relies on $(y_\gamma, q_\gamma, u_\gamma)$ being a solution of Problem 15.1.

Since $(\bar{y}, \bar{u})$ is a minimum of J, all quantities in (15.3) must be equal, i.e., (y^*, q^*, u^*) is a solution of Problem 14.1. □

As in the last chapter, we can now derive an analogon to (13.13) and (14.1), respectively.

Theorem 15.5

Let (y, q, u) be a minimum of Problem 15.1. Then, there exist Lagrange multiplier $z \in L^2(L^2)$ and $\eta \in L^2(H_0^1)$ such that the following equations are fulfilled.

$$y_t = -\left([f(y) + \varepsilon] y_{xxx}\right)_x + q_x, \tag{15.4a}$$

$$-q_{xx} = u_x, \tag{15.4b}$$

$$\begin{aligned} 0 = {} & \langle y - \tilde{y}, \varphi \rangle + \langle z, \varphi_t + (f'(y) y_{xxx} \varphi)_x \rangle \\ & + \langle z, \left([f(y) + \varepsilon] \varphi_{xxx}\right)_x \rangle + \frac{1}{\gamma} \langle \varphi, (C_0 - y)^+ \mu \rangle \quad \forall \varphi \in X_y, \end{aligned} \tag{15.4c}$$

$$0 = z_x - \eta_{xx}, \tag{15.4d}$$

$$0 = \alpha u + \eta_x, \tag{15.4e}$$

together with initial conditions $y(0, .) = y_0$, $z(T, .) = 0$; boundary conditions $y_x = y_{xxx} = z_x = z_{xxx} = 0$ in a, b; as well as $q = \eta = 0$ in a, b.

16. Computational studies

In order to study numerical experiments for the optimal control Problem 15.1, we first have to discretize the optimization problem to obtain a finite dimensional problem: We use the "first discretize, then optimize" ansatz, which has several advantages such as that the system of necessary optimality conditions is well-posed, and the adjoint equation inherits a discretization from the discretization of the state equation. As in the last chapters, we also consider here $g_0 \equiv 0$.

16.1. Discretization of the equation

We use the following space-time discretization scheme for (12.1), which was originally suggested for (10.1) in [15].

Let $hN_{\text{space}} = b - a$ and $x_i := a + ih$ for $i = 0, \ldots, N_{\text{space}}$ denote the set of spatial nodes. Define the standard finite element space V_h, containing piecewise linear functions, via

$$V_h := \left\{ v_h \in \mathcal{C}([a,b]) : \left. v_h \right|_{[x_i, x_{i+1}]} \in P_1 \right\},$$

cf. [23]. The function $P_h : L^2 \to V_h$ denotes the projection onto V_h with respect to the L^2 scalar product.

Let $kN_{\text{time}} = T$, and let $t_n := nk$ for $n = 0, \ldots, N_{\text{time}}$ denote the nodal points of a time grid which covers $[0, T]$.

We will use the following notation for discrete functions: The notation $\{V^n\} \subseteq X_h$ describes a family of finite element functions evaluated at subsequent times t_n, while $V : \Omega_T \to \mathbb{R}$ stands for the piecewise affine, globally continuous time interpolant of $\{V^n\}$. Sometimes, we also write $V(t = t_n)$ instead of V^n.

The discrete version of (12.1) (for $g_0 \equiv 0$) reads as follows.

Problem 16.1
Let $Y_0 := P_h y_0 \in V_h$. Set $Y^0 := Y_0$, find $P^0 \in V_h$ such that

$$(Y_x^0, \Phi_x) - (P^0, \Phi) = 0 \quad \forall \Phi \in V_h.$$

Then for $n = 1, \ldots, N_{\text{time}} - 1$ find $Y^{n+1} \in V_h$, $P^{n+1} \in V_h$ and $P^{n+1} \in V_h$, such that

$$\frac{1}{k}(Y^{n+1} - Y^n, \Phi) + (f_\varepsilon(Y^{n+1}) P_x^{n+1}, \Phi_x) = (U_x(t_{n+1}), \Phi) \quad \forall \Phi \in V_h, \tag{16.1a}$$

$$(Y_x^{n+1}, \Phi_x) - (P^{n+1}, \Phi) = 0 \quad \forall \Phi \in V_h. \tag{16.1b}$$

Remark 16.2
Note that in this whole chapter, we will not include the term regularization $\left(\Delta^{-1}u_x\right)_x$, which we introduced in chapter 13 in order to cope with spatial regularity and convergence issues, but use instead the term u_x. In all our experiments, we do not study the effects in the discretization parameter h dealing with spatial resolution, i.e., $h > 0$ is kept fixed. On a finite dimensional level, the problem with the easier term u_x (from Problem 10.1) is related to the properly scaled Problem 13.1.

In [50], the author has provided simulations for the same problems, where the regularization term $\left(\Delta^{-1}u_x\right)_x$ is also dropped, but the functional being minimized consists of an $L^2(H^1)$-norm of the control u instead of the $L^2(L^2)$-norm of it.

The coupled system (16.1) is solved by Newton's method with exact derivatives, and all terms (which are polynomials of higher order) are assembled exactly using an accurate quadrature rule.

Lemma 12.7 motivates solvability of (16.1) for $\varepsilon > 0$. However, for small $\varepsilon > 0$, the system matrix has a high condition number in the presence of related large values of the approximation of $U_x(t_n)$ and small values of $\{Y^n\}$ due to the algebraic form of f_ε. We encountered this problem in the form of a singular system matrix on the level of numerical linear algebra. Smaller values of k, bigger values of ε and—in the context of optimal control—state constraints help to overcome this issue.

For all experiments in this chapter, we choose $\lambda = 1.0$ and Newton's method as nonlinear algebraic solver stops if the difference of two consecutive iterations is less than 10^{-10}, or if the maximum number of iterations exceeds 1 000. However, except for those experiments with singular system matrices, the observed number of iterates was well below (We needed in average 2–5 iterations and as maximum 30 iterations, which highly depends on the specific experiment).

16.2. Simulations of the equation

For the first experiment, we take $[a,b] = [0,5]$, $T = 1.0$, $N_{\text{space}} = 8$, $N_{\text{time}} = 5\,000$, and $\varepsilon = 0$; we take a fixed right-hand side U and solve (16.1). The output is displayed in Figure 16.1. In this experiment we see that the solution takes negative values for a general function u, which motivates that the state constraint in the optimization Problem 10.1 is really needed. For comparison, we included corresponding simulations for $U \equiv 0$, where we know from Hypothesis 12.9 that the solution stays positive.

In order to exclude effects concerning spatial discretization, we change data to $N_{\text{space}} = 30$, $\varepsilon = 0.03$ and repeat the experiment for a given right-hand side (actually the same like above, but the magnitude is decreased by a factor of 0.7); see Figure 16.2. The change in the right-hand side and the small, but positive value of ε have to be done since the system matrix was singular otherwise. For comparison, we included corresponding simulations for $U \equiv 0$ and $\varepsilon = 0.03$ as well as for $\varepsilon = 0$, which both stay positive for the whole time. We see that small values of ε do have only a small effect on the solution being negative if at all.

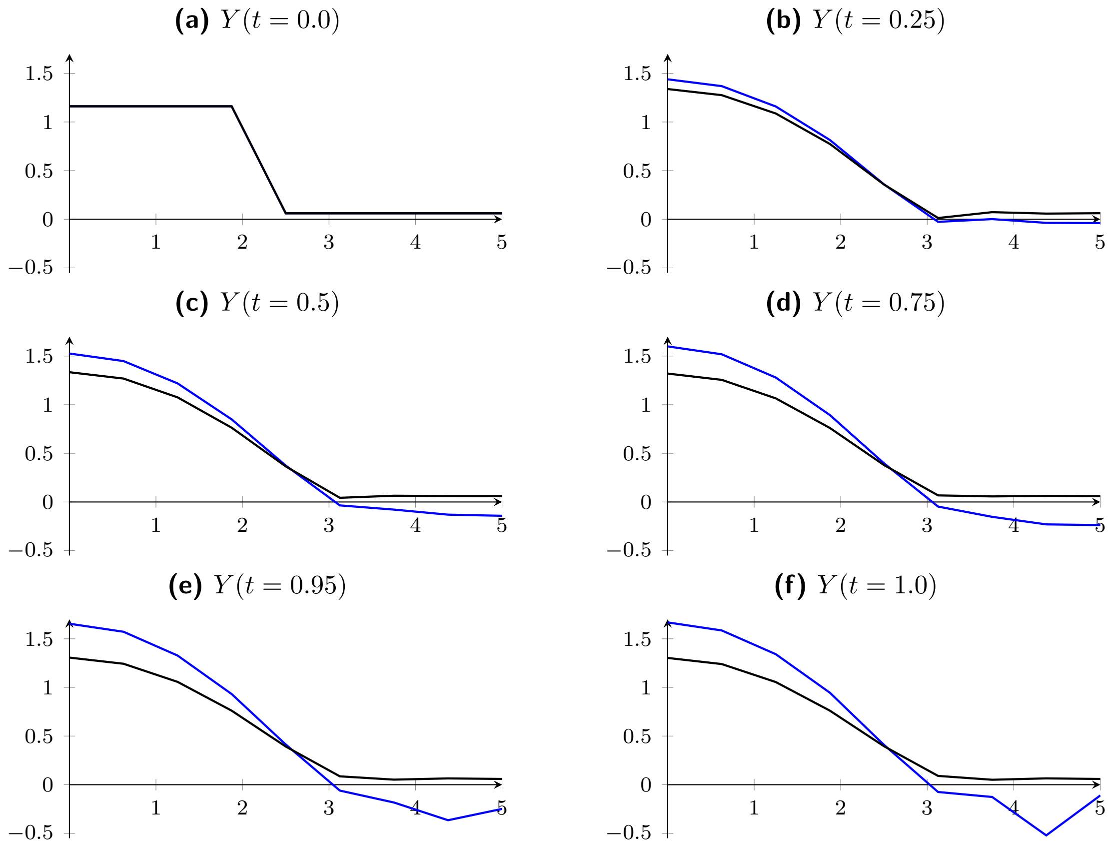


Figure 16.1. Solution Y at different times for a given right-hand side $U \neq 0$ (blue line) and $U \equiv 0$ (black line).

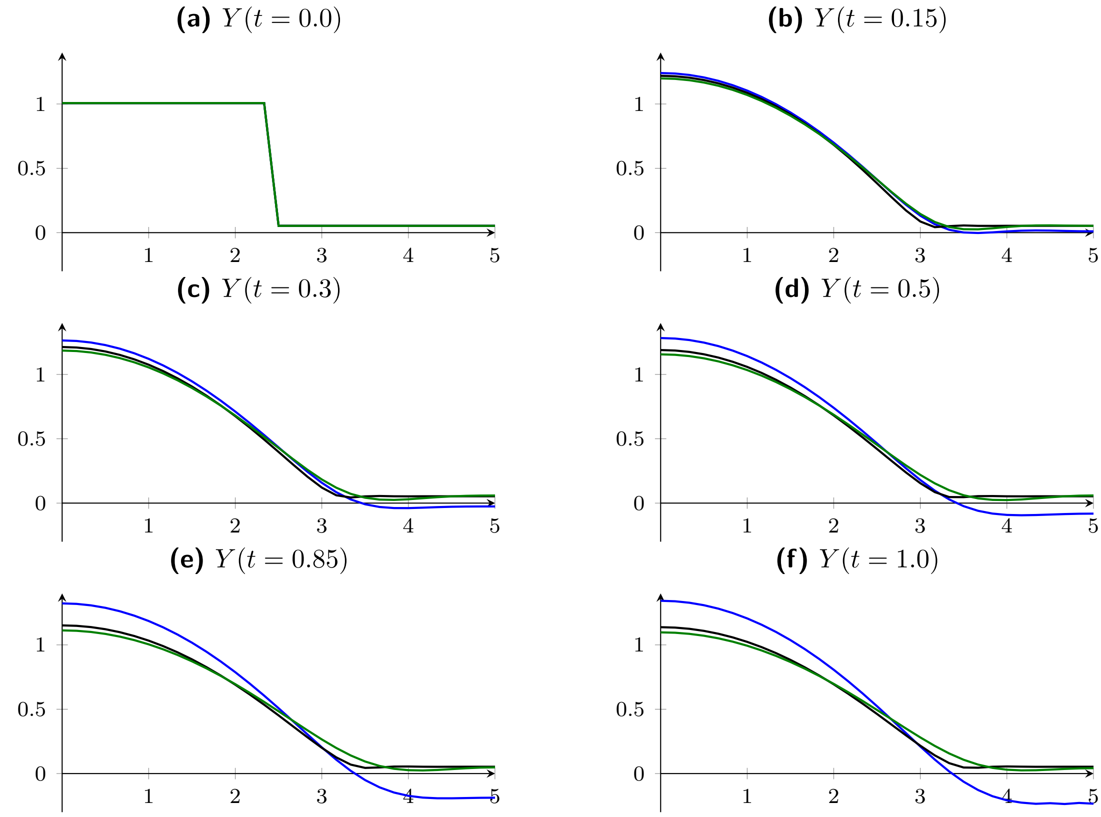


Figure 16.2. Solution Y at different times for a given right-hand side $U \neq 0$ and $\varepsilon = 0.03$ (——), for $U \equiv 0$ and $\varepsilon = 0.03$ (——), and for $U \equiv 0$ and $\varepsilon = 0$ (——).

16.3. Discretization of the optimal control problem

We use a "first discretize, then optimize" (cf. [46]) approach to state the following discrete version of Problem 15.1.

Problem 16.3
Let $\varepsilon > 0$, $\gamma \geq 0$, and let t_k like above. Define $J_{\gamma,\text{disc}} : V_h^{N_{\text{time}}+1} \times V_h^{N_{\text{time}}+1} \to \mathbb{R}$ via

$$J_{\gamma,\text{disc}}(Y, U) := \frac{k}{2} \sum_{n=0}^{N_{\text{time}}} \|Y^n - \tilde{Y}^n\|^2 + \frac{\alpha k}{2} \sum_{n=0}^{N_{\text{time}}} \|U^n\|^2 + \frac{k}{2\gamma} \sum_{n=0}^{N_{\text{time}}} \|(C_0 - Y^n)^+\|^2,$$

where the last term is ignored if we set $\gamma = 0$. If $\tilde{Y}^n \notin V_h$, we instead insert the interpolation of it into $J_{\gamma,\text{disc}}$.

Find (Y, U) as the minimum of $J_{\gamma,\text{disc}}$ subject to (16.1).

Theorem 16.4
Let $\varepsilon > 0$ and $\gamma \geq 0$. Then there exists a solution of Problem 16.3.

Theorem 16.5
Let $(Y, U) \in V_h^{N_{\text{time}}+1} \times V_h^{N_{\text{time}}+1}$ be a minimum of Problem 16.3. Then, there exist Lagrange multipliers $Z \in V_h^{N_{\text{time}}+1}$ and $S \in V_h^{N_{\text{time}}+1}$, such that for all $n = 1, \ldots, N_{\text{time}} - 1$ the following equations are fulfilled:

$$\frac{1}{k}(Y^{n+1} - Y^n, \Phi) + (f_\varepsilon(Y^{n+1})P_x^{n+1}, \Phi_x) = (U_x(t_{n+1}), \Phi) \quad \forall \Phi \in V_h, \tag{16.2a}$$

$$(Y_x^{n+1}, \Phi_x) - (P^{n+1}, \Phi) = 0 \quad \forall \Phi \in V_h, \tag{16.2b}$$

$$\frac{1}{k}(\Phi, Z^n) + (f'(Y^{n+1})\Phi P_x^{n+1}, Z_x^n) + (\Phi_x, S_x) = \frac{1}{k}(\Phi, Z^{n+1}) + (\Phi, \tilde{Y}^{n+1} - Y^{n+1}) + \frac{1}{\gamma}\Big(\Phi, (C_0 - Y^{n+1})^+\Big) \quad \forall \Phi \in V_h, \tag{16.2c}$$

$$(f_\varepsilon(Y^{n+1})\Phi_x, Z^n) - (\Phi, S^n) = 0 \quad \forall \Phi \in V_h, \tag{16.2d}$$

$$\alpha(U, \Phi) + (Z_x, \Phi) = 0 \quad \forall \Phi \in V_h, \tag{16.2e}$$

together with initial conditions $Y^0 = Y_0$, $Z^{N_{\text{time}}} = 0$. Conditions (16.2b), (16.2d), and (16.2e) are also valid for $n = 0$.

By the uniqueness of solutions for the continuous equation (12.1) (which is valid at least for $g_0 \equiv 0$) as well for the discrete version of it, (16.1) (which can be shown for $k > 0$ is small enough), the operator $U \mapsto Y(U)$ is well-defined. Therefore, we can use a steepest descent algorithm in order to solve Problem 16.3 numerically instead of addressing directly (16.2) with, e.g., a SQP-algorithm, which suffers from a huge system matrix for the present evolutionary problem.

We write $Y(U)$ for the solution of (16.1) for a given U and can restate Problem 16.3 by minimizing the functional

$$\tilde{J}(U) := J_{\gamma,\text{disc}}(Y(U), U)$$

without any constraints. From (16.2e) we know that the gradient of $\tilde{J}$ is given by the finite element projection of $\alpha U + Z_x$, which we use as search direction for the steepest descent method, in combination with an Armijo step size rule, which is very flexible and ensures by its selection of valid step sizes a monotone decrease of $\tilde{J}(U_r)$ as $r \to \infty$. For general details regarding our used method and a recent overview about the theoretical background we refer the reader to [41, 46].

The corresponding algorithm reads as follows.

Algorithm 16.6
Set $U_0 \equiv 0$ and fix $\sigma_* > 0$, $0 < \beta < 1$, $\delta_{\text{tol}} > 0$. Compute (Y_1, P_1) from solving (16.1), then compute (Z_1, S_1) from solving (16.2c) and (16.2d). Repeat for $r \geq 0$:

1. Evaluate $\nabla\tilde{J}(U_r) = \alpha U_r + (Z_r)_x$ and evaluate $\tilde{J}(U_r)$.
2. Repeat for $s \geq 0$:
   a) Define $U_{r+1}^{(s)} := U_r - \beta^s \nabla\tilde{J}(U_r)$.
   b) Compute $(Y_{r+1}^{(s)}, P_{r+1}^{(s)})$ from solving (16.1) for $U_{r+1}^{(s)}$ as right-hand side.
   c) STOP, if
   $$\tilde{J}(U_{r+1}^{(s)}) - \tilde{J}(U_r) \leq -\sigma_* \beta^s \|\nabla\tilde{J}(U_r)\|^2, \tag{16.3}$$
   and set $U_{r+1} := U_{r+1}^{(s)}$.
3. Compute (Z_{r+1}, S_{r+1}) from solving (16.2c) and (16.2d).
4. STOP, if $\|\nabla\tilde{J}(U_{r+1})\|^2 \leq \delta_{\text{tol}}$ and set $U_{\text{opt}} = U_{r+1}$, $Y_{\text{opt}} = Y_{r+1}$.

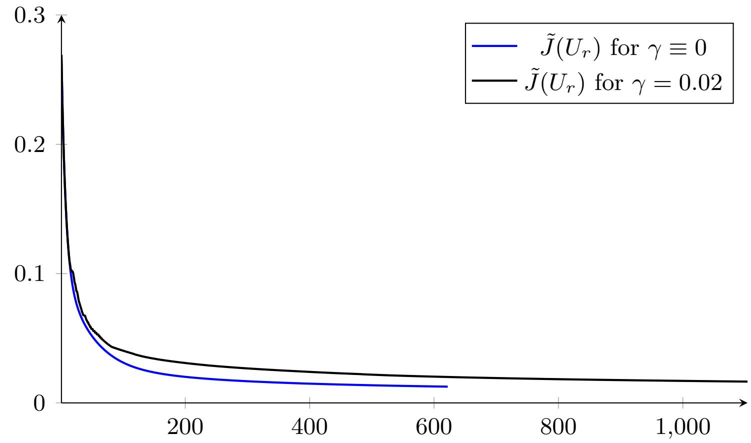
In all the studies below, we set $\sigma_* := 10^{-5}$ and $\beta := 0.15$. The stopping condition is set to be $\delta_{\text{tol}} := 5 \cdot 10^{-5}$, which is obtained after 700 up to 50 000 iterations. The number of iterations highly depends on the given data (i.e., on Y_0, $\tilde{Y}$, and on $\alpha, \varepsilon, \gamma > 0$). A typical evaluation with respect to the number of iterations of the functional $\tilde{J}$ and the gradient $\nabla\tilde{J}$ with respect to the number of iterations is shown in Figure 16.3. For the majority of the steps in this example, the biggest step size was considered as being suitable, i.e., (16.3) was fulfilled for $s = 0$. In cases where more nested iterations in Algorithm 16.6 are needed, the values of $\tilde{J}$ and $\|\nabla\tilde{J}\|^2$ decrease more slowly with respect to the number of iterations.

In Figure 16.3, we plotted two different scenarios depending on $\gamma > 0$: In both scenarios, $r \mapsto \tilde{J}(U_r)$ is monotonously decreasing, thanks to the definition of step size. This is also the case for $r \mapsto \|\nabla\tilde{J}(U_r)\|^2$ if $\gamma \equiv 0$. For $\gamma > 0$, the function $\tilde{J}$ is more complicated, and so is the norm of its gradient; see Figure 16.3b. The time dynamics of the optimal solutions of this particular experiment is displayed in Figure 16.8 and the experiment is explained in section 16.6.

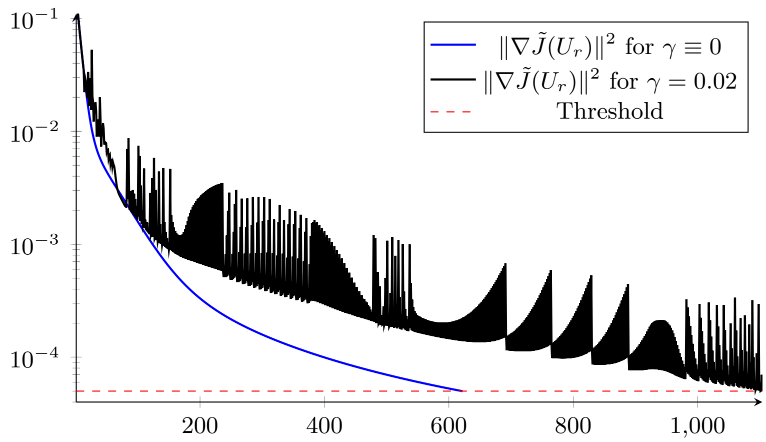
It is clear that the choice of β is one of the most crucial parameters for the performance of the algorithm. Smaller values of β rule out bigger step sizes, leading to more iterations. Bigger values of β allow for potentially bigger step sizes, which could also lead to a longer runtime since it takes more nested iterations to obtain a valid step size in the spirit of

(16.3). In the context of an optimal control problem with a degenerate equation such as in the present case, it is important not to choose β too big: Then, the new potential control $U_{\text{new}} = U_{\text{old}} - \beta^s \nabla \tilde{J}(U_{\text{old}})$ can be too "destructive", in a sense that the corresponding system matrix for the potential new state $Y(U_{\text{new}})$ is (close to) singular and hence would affect the linear algebra solver. With our choice of β, we observed a reasonable behavior of Algorithm 16.6.

We note that the performance of Algorithm 16.6 can be improved in the following way: First solve Problem 16.3 with coarse discretization parameters $h, k > 0$. Then, transfer the solution U_{opt} to U_0, and solve Problem 16.3 with the finer discretization with the different start value for U_0. Clearly, the number of iterations can be rapidly decreased in this way.



(a) Evolution of $\tilde{J}(U_r)$ with respect to the number of iterations of Algorithm 16.6 for $\gamma \equiv 0$ (——) and $\gamma = 0.02$ (——).



(b) Evolution of $\|\nabla \tilde{J}(U_r)\|^2$ with respect to the number of iterations of Algorithm 16.6 for $\gamma \equiv 0$ (——) and $\gamma = 0.02$ (——).

Figure 16.3. Typical behavior of the gradient algorithm for $\gamma \equiv 0$ (——) and $\gamma = 0.02$ (——); corresponding optimal states are displayed in Figure 16.8.

16.4. Comparison of the parameter ε

In the next experiment we take $[a, b] = [0, 5]$, $T = 1.0$, $N_{\text{space}} = 30$, $N_{\text{time}} = 5\,000$, and $\alpha = 10^{-6}$; and we solve (16.1) for $U \equiv 0$ to study the dependencies on $\varepsilon > 0$; see Figure 16.4. The bigger the value of ε, the more dissipative is the evolution, and the solution becomes almost flat after a short time. In contrast to this, for a small value of ε, the solution needs longer to approach a flat profile.

For a large value of ε, the solution is slightly negative in some regions; see Figures 16.4c, 16.4e, and 16.4f. This is due to the fact that there is no maximum principle for the bi-harmonic problem, which would force the solution to stay positive. This effect vanishes for decreasing values of ε.

We repeat the above experiment with the same parameters in the context of optimal control Problem 16.3 for $\gamma \equiv 0$; see Figure 16.5. In contrast to the previous experiment from Figure 16.4, there is not such a big difference between the computed evolution of the optimal states, depending on the value of ε. This is due to the fact that the optimal state $Y = Y(\varepsilon)$ belongs to different optimal controls $U = U(\varepsilon)$ which force the solution to obtain the given target profile $\tilde{Y}$. The experiment which is shown in Figure 16.5 demonstrates that relevant controls are active since the dynamics of the solutions completely differs from the case without control which was shown in Figure 16.4.

16.5. Comparison of the parameter α

In this experiment, we take $\varepsilon = 0.05$, $\gamma \equiv 0$, $N_{\text{space}} = 54$, $N_{\text{time}} = 5\,000$, and compare different values of $\alpha > 0$; see Figure 16.6 (state) and Figure 16.7 (control). Here, $\tilde{Y}$ is constant in time. We can see that a small value of α allows for bigger controls; see Figure 16.7. The optimal state Y (with small α) almost agrees with the target state $\tilde{Y}$ after a very short time, while the optimal state Y (with bigger α) needs more time for that. The snapshot in Figure 16.6e shows the first time when the optimal state Y coincides with the target state $\tilde{Y}$ for all values of α.

We note that the optimal controls displayed in Figure 16.7 are typical for many experiments: The control acts near the spatial boundary, i.e., it could be worth to consider Problem 10.1 with boundary control instead of a distributed control. Also, the amplitude of the controls decreases in time, which is typical for parabolic optimal control problems with a constant target profile $\tilde{y}$: A large control in the beginning of the experiment enforces the solution to be near the target profile $\tilde{y}$, which decreases immediately the tracking term $\|y - \tilde{y}\|$ in the functional, while a large control near $t = T$ has almost no impact on the optimal state y, but increases the cost term $\|u\|^2$ in the functional.

16.6. Comparison of the parameter γ

In this experiment, we take $C_0 = 0.01$, $\alpha = 10^{-6}$, $\varepsilon = 0.1$, $N_{\text{space}} = 42$, $N_{\text{time}} = 5\,000$ and simulate different values of $\gamma > 0$; see Figure 16.8. Here, $\tilde{Y}$ is constant in time and the

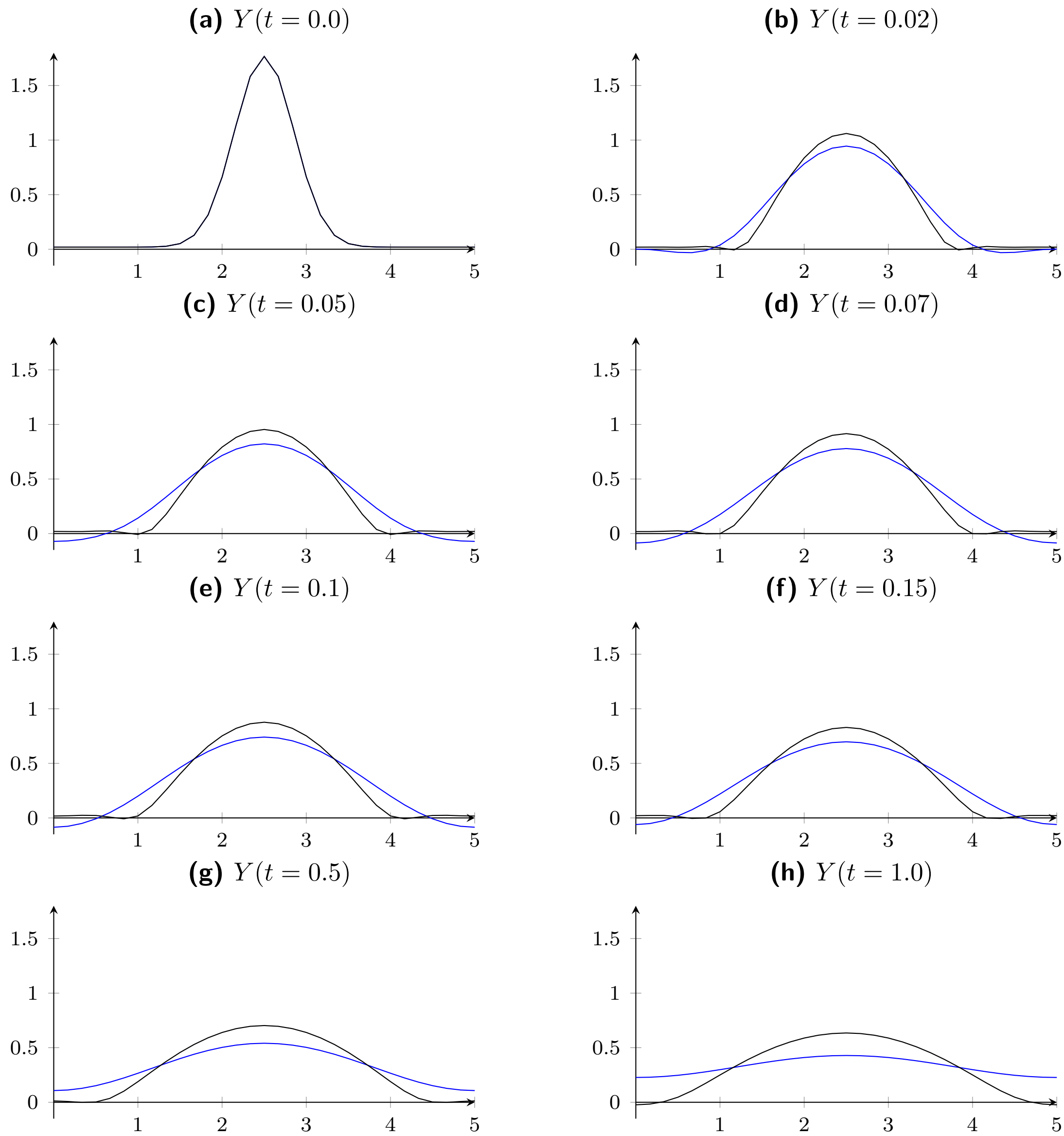


Figure 16.4. Solution Y of (16.1) for $U \equiv 0$, and for $\varepsilon = 0.5$ (——) and $\varepsilon = 0.005$ (——) at different times.

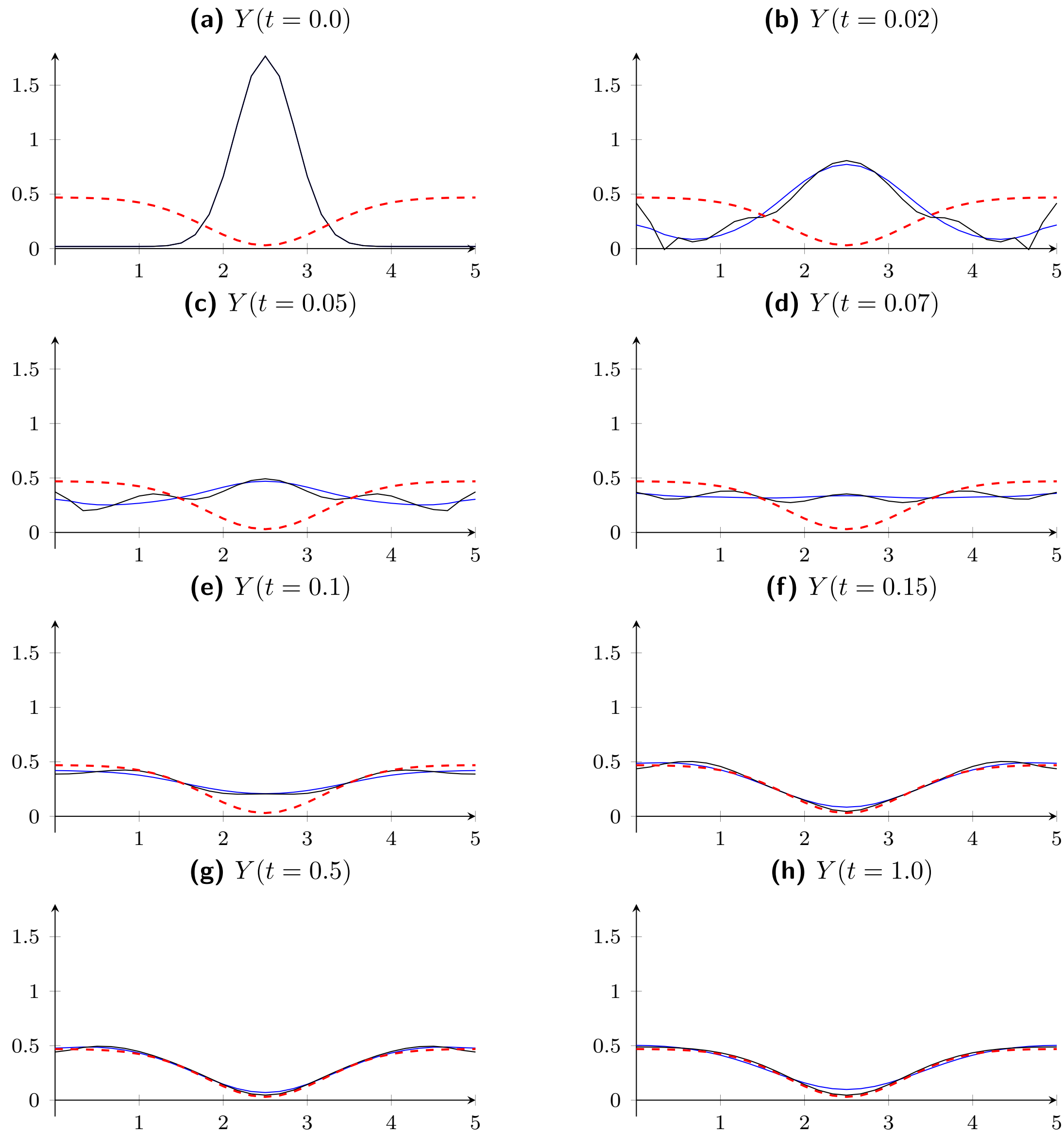


Figure 16.5. Target $\tilde{Y}$ (- - -), and optimal state Y for $\varepsilon = 0.5$ (——) and $\varepsilon = 0.005$ (——) at different times.

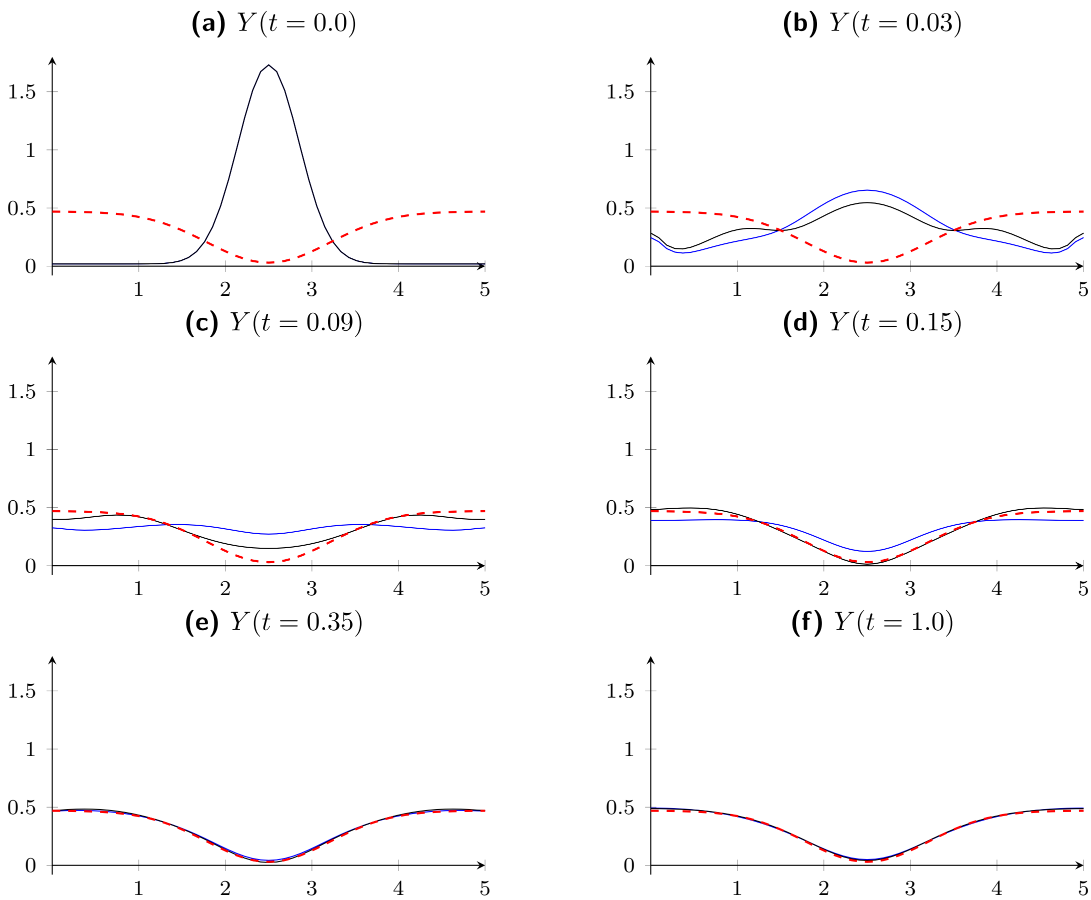


Figure 16.6. Target $\tilde{Y}$ (- - -) and optimal states Y for $\alpha = 10^{-2}$ (——) and $\alpha = 10^{-10}$ (——) at different times; corresponding controls are displayed in Figure 16.7.

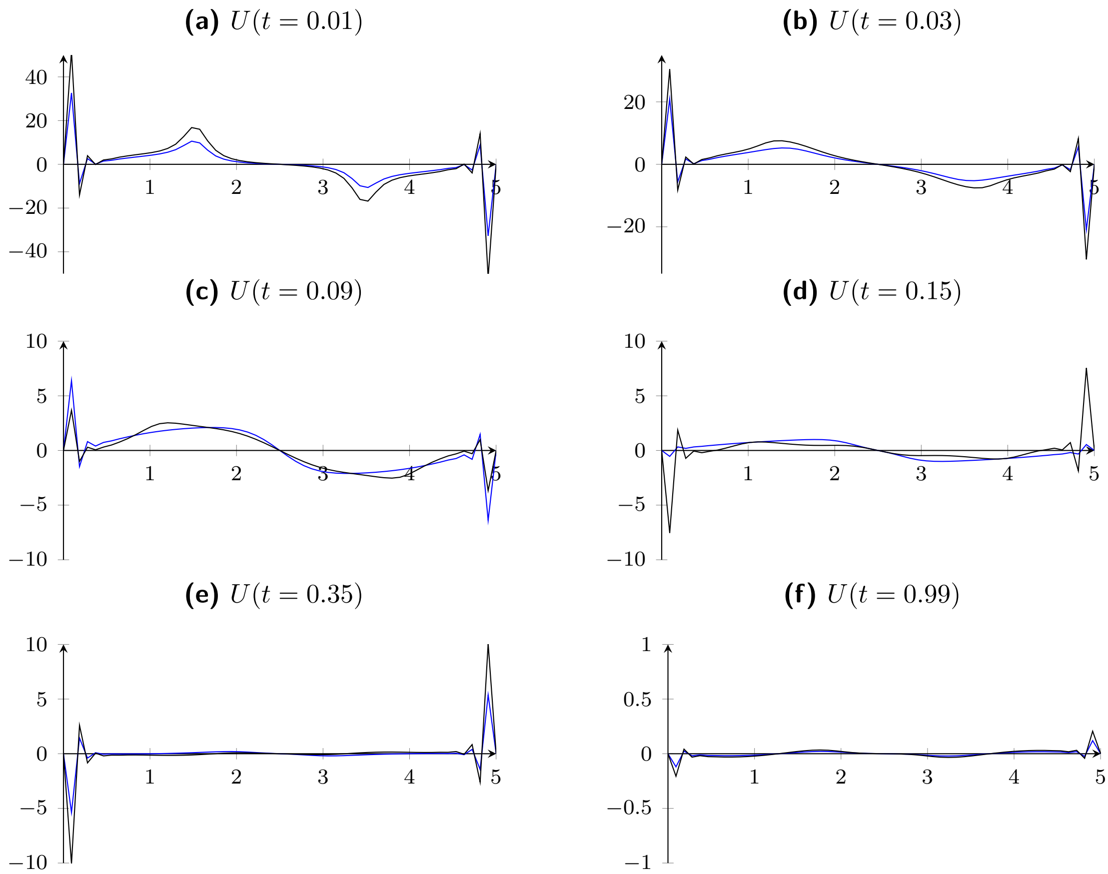


Figure 16.7. Control U for $\alpha = 10^{-2}$ (——) and $\alpha = 10^{-10}$ (——) at different times; corresponding optimal states are displayed in Figure 16.6. Note that the different plots are scaled by different factors.

profile is given in the figure. We can see that even for a moderate choice of $\gamma > 0$, this parameter has a significant effect on the simulation: If this penalization term is missing, the solution ceases to be positive, while the solution is positive (except for some single points) over the whole simulation if the penalization is active. As we have noted before, the condition number is increasing for decreasing values of γ. This leads to a longer runtime for smaller γ; sometimes it also happens that matrices are identified as singular by the linear algebra solver due to this increasing condition number.

Vice versa, for bigger values of γ, the state condition is not resolved properly, i.e., system matrices can also become singular in this case. This leads to the conclusion that – as long as $\varepsilon > 0$ is kept small, and as long as no sophisticated linear algebra solvers are used – there is only a small range for $\gamma > 0$ where simulations are likely to terminate in a reasonable amount of time.

Also, the more complex structure of the functional $\tilde{J}$ for $\gamma > 0$ leads to an increase of the needed amount of iterations in the steepest descent algorithm. The corresponding evolution of $r \mapsto \tilde{J}(U_r)$ and $r \mapsto \|\nabla \tilde{J}(U_r)\|^2$ are displayed for both values of γ in Figure 16.3.

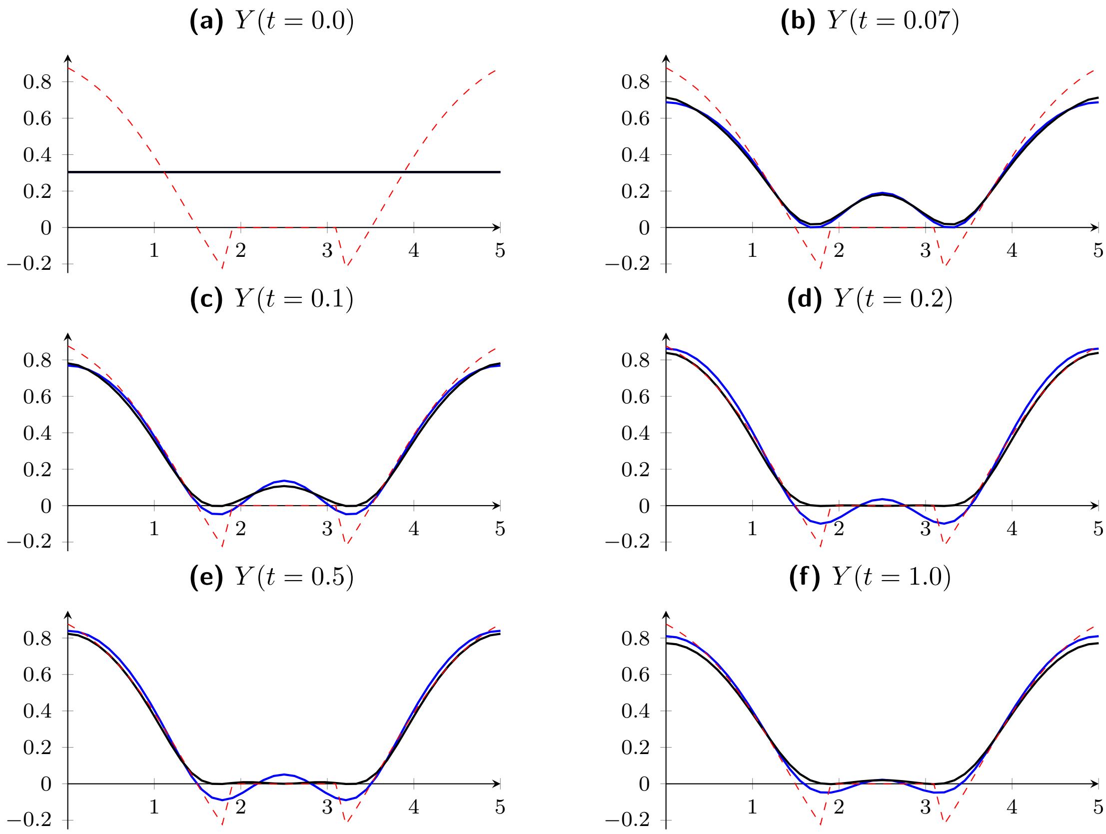


Figure 16.8. Target $\tilde{Y}$ (- - -) and optimal states Y for $\gamma \equiv 0$ (——) and $\gamma = 0.02$ (——) at different times.

16.7. Dewetting application

In the last experiment, we simulate the solution of a simplified version of the problem arising in [17]: Given a constant initial value Y_0, a profile $\tilde{Y}$ should be accomplished, where there is a bigger region (nearly) without any fluid; this evolution is refered to as dewetting procedure. We set $[a, b] = [0, 5]$, $T = 1.0$, $N_{\text{space}} = 54$, $N_{\text{time}} = 25\,000$, $\alpha = 10^{-6}$, $C_0 = 0.01$, and $\gamma = 0.01$; see Figure 16.9.

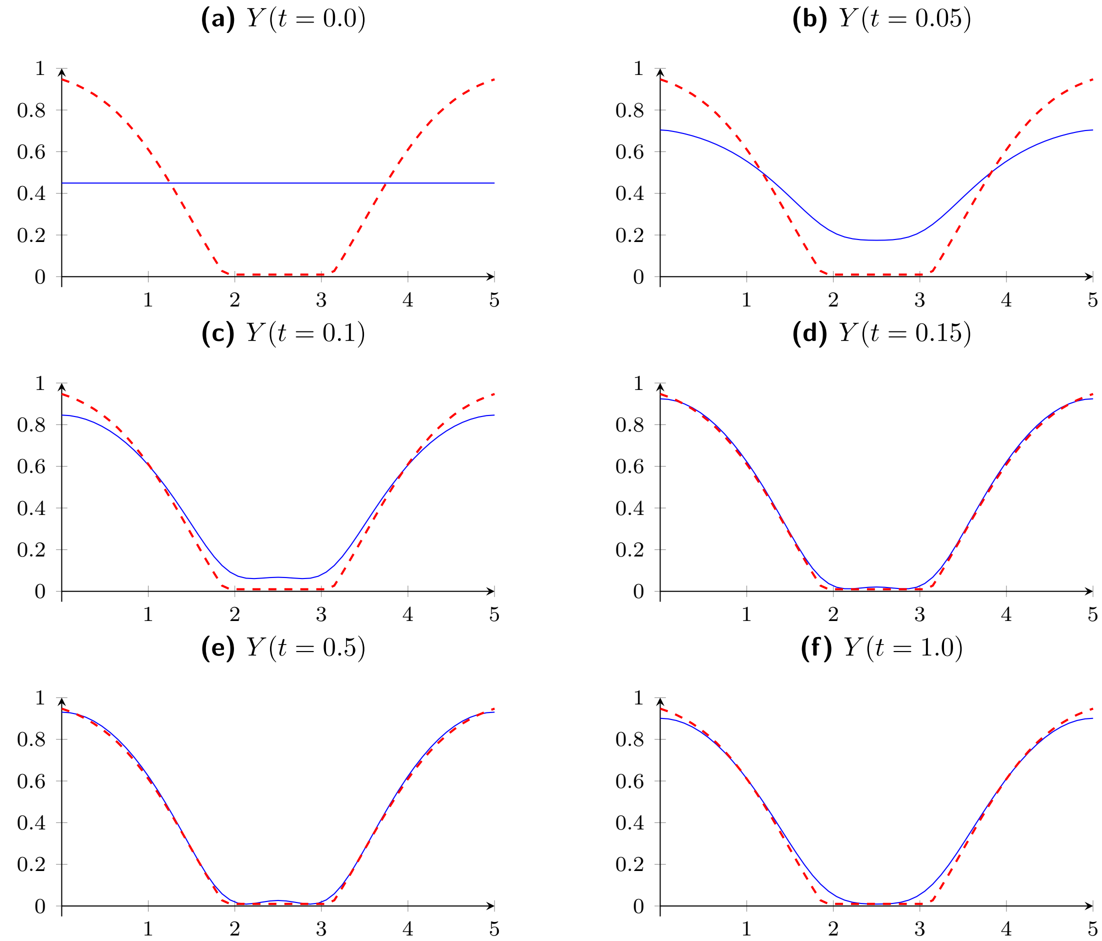


Figure 16.9. Target $\tilde{Y}$ (- - -) and optimal states Y (——) at different times.

A. Gamma-convergence

The next definition is standard, cf., e.g., [22, Chapter 3].

Definition A.1
Let X be an normed space and $f, f_i : X \to [-\infty, \infty]$. We say that $f_i \xrightarrow{\Gamma} f$ if for all $u \in X$

1. for every sequence $u_j \to u$, there holds

$$f(u) \leq \liminf_{j\to\infty} f_j(u_j), \tag{A.1}$$

2. there exists a sequence $\bar{u}_j \to u$ such that

$$f(u) \geq \limsup_{j\to\infty} f_j(\bar{u}_j).$$

It is clear that uniform convergence implies Γ-convergence, provided all involved functions are continuous.

The next lemma shows that Γ-convergences is not generally inherited by functions when dealing with constraints.

Lemma A.2
There exist functions $f_i, f : \mathbb{R} \to \mathbb{R}$ such that

1. $f_i \xrightarrow{\Gamma} f$ for $i \to \infty$,
2. The functions $\tilde{f}_i, \tilde{f} : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, defined by

$$\tilde{f}_i(x) := \begin{cases} f_i(x), & g_i(x) = 0, \\ \infty, & \text{else} \end{cases},$$

$$\tilde{f}(x) := \begin{cases} f(x), & g(x) = 0, \\ \infty, & \text{else} \end{cases},$$

   where $g_i, g : \mathbb{R} \to \mathbb{R}$ are given functions, do not hold $\tilde{f}_i \xrightarrow{\Gamma} f$ for $i \to \infty$.

Proof
Let $f, f_q : \mathbb{R} \to \mathbb{R}$, defined by $f(x) \equiv 1$ and $f_q \equiv 1$ for $q \in \mathbb{Q}$. It is clear that $f_q \to f$ uniformly, hence we have $f_q \xrightarrow{\Gamma} f$ for $q \to \infty$ (which means that $\text{Count}(q) \to \infty$, where $\text{Count} : \mathbb{Q} \to \mathbb{N}$ is the mapping counting the rational numbers).

We now introduce corresponding constraints $g, g_q : \mathbb{R} \to \mathbb{R}$, defined by $g(x) := x$ and $g_q(x) := x - q$, hence we have

$$\tilde{f}(x) = \begin{cases} 1, & x = 0, \\ \infty, & \text{else} \end{cases}, \qquad \tilde{f}_q(x) = \begin{cases} 1, & x = q, \\ \infty, & \text{else} \end{cases}.$$

Let $0 < \varepsilon \notin \mathbb{Q}$ and $\{x_i\} \subset \mathbb{Q}$ with $x_i \to \varepsilon$. It is clear that $\tilde{f}(\varepsilon) = \infty$ and for the subsequence $\tilde{f}_{x_i}$ there holds by definition $\tilde{f}_{x_i}(x_i) = 1$ for all $i \in \mathbb{N}$, hence we have

$$\liminf_{q \to \infty} \tilde{f}_q(x_q) \leq 1,$$

i.e., (A.1) does not hold for ε and therefore, the sequence $(\tilde{f}_q)$ does not Gamma-converge to $\tilde{f}$.
$\square$

B. Derivation of the thin-film equation

In this chapter, we want to repeat the formal derivation of the thin film equation (TF) from the Navier–Stokes equation for the sake of completeness. This derivation was done several times in the literature [16, 57, 59]. In principle we follow the presentation given in [16].

For a fixed time in $[0, t]$, we assume the domain $\Omega \subset \mathbb{R}^2$ being the intersection of the subgraph of a nonnegative continuous function $u : [a, b] \to [0, \infty)$ (with $u(a) = u(b) = 0$; if this is not true, the domain (a, b) has be increased such that the whole dynamics takes place inside it) with the positive half space, i.e.,

$$\Omega = \{(x, y) \in (a, b) \times \mathbb{R} : \ 0 < y < u(x)\}.$$

In Ω (which changes over the time), we assume the the classical one-fluid Navier–Stokes equation with a conservative force as right-hand side (where g_0 is a scalar potential) to be fulfilled,

$$\partial_t \boldsymbol{v} + [\boldsymbol{v} \cdot \nabla]\boldsymbol{v} - \Delta \boldsymbol{v} + \nabla q = -\nabla g_0, \tag{B.1a}$$

$$\operatorname{div} \boldsymbol{v} = 0. \tag{B.1b}$$

The boundary of each such Ω contains of a lower (planar) part $\underline{\partial}\Omega = [a, b] \times \{0\}$ and an upper boundary $\overline{\partial}\Omega = \operatorname{graph} u$. We assume zero Dirichlet boundary conditions $\boldsymbol{v} = 0$ on $\underline{\partial}\Omega$ (so-called no slip boundary conditions). On the upper boundary, we assume surface tension, i.e., we for $\boldsymbol{n}$ being a normal vector on $\overline{\partial}\Omega$ and $\boldsymbol{r}$ being a tangent vector on $\overline{\partial}\Omega$, we assume

$$\mathbb{S}\boldsymbol{n} \cdot \boldsymbol{n} = -\sigma h, \qquad \mathbb{S}\boldsymbol{n} \cdot \boldsymbol{r} = 0,$$

where $\sigma > 0$ is the (constant) surface tension, h is the mean curvature on the point of evaluation, and $\mathbb{S}$ is the stress tensor,

$$\mathbb{S} = \Big[(\nabla \boldsymbol{v})^T + \nabla \boldsymbol{v}\Big] - q \operatorname{id}.$$

The aim is to derive a differential equation for the "height function" u; precisely we want to derive the thin-film equation (B.14).

We decompose all quantifies into two components: A tangential direction and a normal direction with respect to the "ground" (a, b). We decompose all points $\boldsymbol{x} \in \Omega$ into $\boldsymbol{x} = (x, y)$ and the velocity $\boldsymbol{v} = (\boldsymbol{v}_x, \boldsymbol{v}_y)$. It is crucial to later transform all quantifies into a new coordinate system, where each component is scaled differently.

First, we note that a change of the height function u comes with a velocity in y-direction, i.e., we have

$$v_y = \frac{d}{dt} u(t, x(t)).$$

Taking the derivative of this quantify, we arrive at

$$v_y = u_t + v_y \partial_x u. \tag{B.2}$$

We define $\tau = \frac{\text{height}}{\text{width}} = \frac{\max u}{b-a} \ll 1$ and consider the following scaled variables (all denoted by capital letters),

$$\begin{aligned} X &= \tau x, & Y &= y, & \boldsymbol{X} &= (X, Y), \\ T &= \tau t, & U &= u, \\ \Sigma &= \tau \sigma, & H &= h, \\ Q &= \tau q, & G_0 &= \tau g_0. \end{aligned}$$

With these definitions, the transformed velocities $V = \frac{d}{dt}(X, Y)$ and the derivatives for the transformed variables scale as follows.

$$\begin{aligned} V_X &= v_x, & V_Y &= \frac{1}{\tau} v_y, & \boldsymbol{V} &= (V_X, V_Y), \\ \partial_X &= \frac{1}{\tau}\partial_x, & \partial_Y &= \partial_y, & \partial_T &= \frac{1}{\tau}\partial_t, & \nabla_{\boldsymbol{X}} &= (\partial_X, \partial_Y). \end{aligned}$$

We now insert these variables into (B.1). The component in normal (y-)direction of (B.1a) reads in the capital variables as

$$0 = \tau^2 \partial_T V_Y + \tau^2 [\boldsymbol{V} \cdot \nabla_{\boldsymbol{X}}] V_Y - \tau \partial_Y^2 V_X - \tau^3 \partial_X^2 V_Y + \frac{1}{\tau} \partial_Y (Q + G_0). \tag{B.3}$$

A similar calculation in tangential (x-)direction of (B.1a) reads as

$$0 = \tau \partial_T V_X + \tau [\boldsymbol{V} \cdot \nabla_{\boldsymbol{X}}] V_X - \partial_Y^2 V_X - \tau^2 \partial_X^2 V_X + \partial_X (Q + G_0). \tag{B.4}$$

Finally, we transform (B.1b) and arrive at

$$0 = \varepsilon \operatorname{div}_{\boldsymbol{X}} \boldsymbol{V}. \tag{B.5}$$

We restrict ourselves with the lowest order terms in (B.3), (B.4), and (B.5), and we get

$$\partial_Y (Q + G_0) = 0, \tag{B.6a}$$
$$\partial_X (Q + G_0) = \partial_Y^2 V_X, \tag{B.6b}$$
$$\operatorname{div}_{\boldsymbol{X}} \boldsymbol{V} = 0. \tag{B.6c}$$

We now take a look at conditions on the boundary: On $\underline{\partial}\Omega$, we get directly

$$\boldsymbol{V} = 0 \quad \text{on } \underline{\partial}\Omega. \tag{B.7}$$

In order to have a corresponding condition on $\overline{\partial}\Omega$, we first calculate the transformation of the stress tensor,

$$\mathbb{S} = \begin{pmatrix} 2\tau \partial_X V_X - \tau^{-1} Q & \partial_Y V_X + \tau^2 \partial_X V_Y \\ \\ \partial_Y V_X + \tau^2 \partial_X V_Y & 2\tau \partial_Y V_Y - \tau^{-1} Q \end{pmatrix}.$$

We scale the normal and tangential vectors as follows,

$$n = \frac{(\tau N_X, N_Y)}{\|(\tau N_X, N_Y)\|}, \qquad r = \frac{(R_X, \tau R_Y)}{\|(R_X, \tau R_Y)\|}.$$

The condition $-\sigma h = \mathbb{S}\boldsymbol{n} \cdot \boldsymbol{n}$ corresponds to the transformed version,

$$\begin{aligned} -\|(\tau N_X, N_Y)\|^2 \tau^{-1} \Sigma H = {} & 2\tau^3 \partial_X V_X (N_X)^2 - Q(N_X)^2 \\ & + 2\tau \partial_Y V_X (N_X N_Y) + 2\tau^3 \partial_X V_Y (N_X N_Y) \\ & + 2\tau \partial_Y V_Y (N_Y)^2 - \tau^{-1} Q(N_Y)^2. \end{aligned}$$

Neglecting all higher order terms, we get

$$Q = \Sigma H \quad \text{on } \overline{\partial}\Omega. \tag{B.8}$$

We perform a similar argumentation with the condition $\mathbb{S}\boldsymbol{n} \cdot \boldsymbol{r}$, use the orthogonality relation between the normal and tangential vector, and we arrive at

$$\partial_Y V_X = 0 \quad \text{on } \overline{\partial}\Omega. \tag{B.9}$$

The condition (B.2) is also valid on $\overline{\partial}\Omega$, hence there holds

$$V_Y = \partial_T U + V_X \partial_X U. \tag{B.10}$$

Since (B.5) holds pointwise, we can integrate this term in Y direction,

$$0 = \int_0^U \operatorname{div}_{\boldsymbol{X}} \boldsymbol{V} = \int_0^U \partial_Y V_Y + \int_0^U \partial_X V_X.$$

We integrate the latter term and get

$$0 = V_Y(U) - V_Y(0) + \int_0^U \partial_X V_X.$$

With the help of (B.7) and (B.10), we arrive at

$$0 = \partial_T U + V_X \partial_X U + \int_0^U \partial_X V_X = \partial_T U + \partial_X \left(\int_0^U V_X \right). \tag{B.11}$$

For the generalized pressure, $P := Q + G_0$, we see from (B.6a) that $Y \mapsto P(.,Y)$ is constant. We do now want to derive conditions for V_X. From (B.6b), (B.9), and (B.7), we may deduce

$$\partial_Y^2 V_X(X,Y) = \partial_X P(X) = \text{const}(X), \tag{B.12a}$$

$$\partial_X V_X(X,U) = 0, \tag{B.12b}$$

$$V_X(X,0) = 0. \tag{B.12c}$$

This system (B.12) is a boundary value problem for V_X in the variable Y, but since the right-hand side does not depend on Y, the expression $y \mapsto V_X(.,Y)$ is a polynomial of degree 2. The coefficients are given by solving (B.12), and we end up with

$$V_X(X,Y) = \partial_X P(X,Y) \left(\frac{1}{2} Y^2 - UY \right).$$

Inserting this particular form of V_X into (B.11) and using the concrete formula for the mean curvature of a graph (in the transformed variable), we arrive at

$$\partial_T U = \partial_X \left(\frac{U^3}{3} \partial_X P \right), \tag{B.13a}$$

$$P = -\tau^2 \Sigma \partial_X \left(\frac{\partial_X U}{\sqrt{1 + \tau^2 |\partial_X U|^2}} \right) + G_0. \tag{B.13b}$$

Finally, we can transform (B.13) back into original (lower case) variables, linearize the mean curvature to the second derivative, and end up with the desired form of the thin film equation,

$$\partial_t u = \partial_x \left(\frac{u^3}{3} \partial_x p \right), \tag{B.14a}$$

$$p = -\sigma \partial_x^2 u + g_0. \tag{B.14b}$$

This equation coincides with equation (TF) from the introduction of this thesis, when we set $\lambda = \sigma$ and scale the every quantity with the factor 3.

Bibliography

[1] H. Abels, H. Garcke, and G. Grün. *Thermodynamically consistent, frame indifferent diffuse interface models for incompressible two-phase flows with different densities. Math. Models Methods Appl. Sci.* 22.3 (2012), pp. 1150013, 40. doi: 10.1142/S0218202511500138.

[2] F. Abergel and R. Temam. *On some Control Problems in Fluid Mechanics. Theoretical and Computational Fluid Dynamics* 1.6 (1990), pp. 303–325. doi: 10.1007/BF00271794.

[3] R. A. Adams and J. J. F. Fournier. *Sobolev spaces.* Second. Vol. 140. Pure and Applied Mathematics. Elsevier/Academic Press, Amsterdam, 2003. isbn: 0-12-044143-8.

[4] J.-J. Alibert and J.-P. Raymond. *A Lagrange Multiplier Theorem for Control Problems with State Constraints. Numer. Funct. Anal. Optim.* 19.7-8 (1998), pp. 697–704. doi: 10.1080/01630569808816854.

[5] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems.* Oxford University Press, Oxford, 2000. isbn: 0-19-850245-1.

[6] D. M. Anderson, G. B. McFadden, and A. A. Wheeler. "Diffuse-interface Methods in Fluid Mechanics". In: vol. 30. Annu. Rev. Fluid Mech. Annual Reviews, Palo Alto, CA, 1998, pp. 139–165. doi: 10.1146/annurev.fluid.30.1.139.

[7] F. Andreu-Vaillo, V. Caselles, and J. M. Mazón. *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals.* Vol. 223. Progress in Mathematics. Birkhäuser, Basel, 2004. doi: 10.1007/978-3-0348-7928-6.

[8] J. Atecia and D. J. Beebe. *Controlled Microfluidic Interfaces. Nature* 437 (2004), pp. 648–655. doi: 10.1038/nature04163.

[9] Ľ. Baňas, M. Klein, and A. Prohl. *Control of Interface Evolution in Multiphase Fluid Flows. SIAM J. Control Optim.* 52.4 (2014), pp. 2284–2318. doi: 10.1137/120896530.

[10] Ľ. Baňas and R. Nürnberg. *A Posteriori Estimates for the Cahn–Hilliard Equation with Obstacle Free Energy. M2AN Math. Model. Numer. Anal.* 43.5 (2009), pp. 1003–1026. doi: 10.1051/m2an/2009015.

[11] Ľ. Baňas and R. Nürnberg. *Finite Element Approximation of a Three Dimensional Phase Field Model for Void Electromigration. J. Sci. Comp.* 37.2 (2008), pp. 202–232. doi: 10.1007/s10915-008-9203-y.

[12] Ľ. Baňas and A. Prohl. *Convergent Finite Element Discretization of the Multi-Fluid Nonstationary Incompressible Magnetohydrodynamics Equations. Math. Comp.* 79.272 (2010), pp. 1957–1999. doi: 10.1090/S0025-5718-10-02341-0.

[13] J. W. Barrett, R. Nürnberg, and V. Styles. *Finite Element Approximation of a Phase Field Model for Void Electromigration. SIAM J. Numer. Anal.* 42.2 (2004), pp. 738–772. doi: 10.1137/S0036142902413421.

[14] J. Barzilai and J. M. Borwein. *Two Point Step Size Gradient Methods. IMA J. Numer. Anal.* 8 (1988), pp. 141–148. doi: 10.1093/imanum/8.1.141.

[15] J. Becker and G. Grün. *The Thin-Film Equation: Recent Advances and some new Perspectives. Journal of Physics: Condensed Matter* 17.9 (2005), S291–S307. doi: 10.1088/0953-8984/17/9/002.

[16] J. Becker, G. Grün, M. Lenz, and M. Rumpf. *Numerical Methods for Fourth Order Nonlinear Degenerate Diffusion Problems. Appl. Math.* 47.6 (2002). Mathematical Theory in Fluid Mechanics (Paseky, 2001), pp. 517–543. doi: 10.1023/B:APOM.0000034537.55985.44.

[17] J. Becker, G. Grün, R. Seemann, H. Mantz, K. Jacobs, K. R. Mecke, and R. Blossey. *Complex Dewetting Scenarios Captured by Thin-Film Models. Nature Materials* 2.1 (2002), pp. 59–63. doi: 10.1038/nmat788.

[18] R. Becker, X. Feng, and A. Prohl. *Finite Element Approximations of the Ericksen–Leslie Model for Nematic Liquid Crystal Flow. SIAM J. Numer. Anal.* 46.4 (2008), pp. 1704–1731. doi: 10.1137/07068254X.

[19] F. Bernis and A. Friedman. *Higher Order Nonlinear Degenerate Parabolic Equations. J. Differential Equations* 83.1 (1990), pp. 179–206. doi: 10.1016/0022-0396(90)90074-Y.

[20] A. L. Bertozzi and M. Pugh. *The Lubrication Approximation for Thin Viscous Films: The Moving Contact Line with a "Porous Media" Cut-off of van der Waals Interactions. Nonlinearity* 7.6 (1994), pp. 1535–1564. doi: 10.1088/0951-7715/7/6/002.

[21] A. L. Bertozzi. *The Mathematics of Moving Contact Lines in Thin Liquid Films. Notices Amer. Math. Soc.* 45.6 (1998), pp. 689–697.

[22] A. Braides. *Approximation of Free Discontinuity Problems.* Lecture Notes in Mathematics 1694. Springer, Berlin, 1998. isbn: 3-540-64771-6.

[23] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods.* Vol. 15. Texts in Applied Mathematics. Springer, New York, 1994. isbn: 0-387-94193-2. doi: 10.1007/978-1-4757-4338-8.

[24] A. E. Bryson and Y. C. Ho. *Applied Optimal Control.* Optimization, Estimation, and Control. Hemisphere Publishing Corp. Washington, D. C., 1975. isbn: 0-89116-228-3.

[25] bwGRiD (http://www.bw-grid.de). *Member of the German D-Grid initiative, funded by the Ministry of Education and Research (Bundesministerium für Bildung und Forschung) and the Ministry for Science, Research and Arts Baden-Württemberg (Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg).* Tech. rep. Universities of Baden-Württemberg, 2007-2010.

[26] C. Clason and B. Kaltenbacher. *Avoiding Degeneracy in the Westervelt Equation by State Constrained Optimal Control. Evol. Equ. Control Theory* 2.2 (2013), pp. 281–300. doi: 10.3934/eect.2013.2.281.

[27] C. Clason and B. Kaltenbacher. *On the Use of State Constraints in Optimal Control of Singular PDEs. Systems Control Lett.* 62.1 (2013), pp. 48–54. doi: 10.1016/j.sysconle.2012.10.006.

[28] C. Clason and B. Kaltenbacher. *Optimal Control of a Singular PDE modeling Transient MEMS with Control or State Constraints. J. Math. Anal. Appl.* 410.1 (2014), pp. 455–468. doi: 10.1016/j.jmaa.2013.08.058.

[29] R. J. DiPerna and P.-L. Lions. *Ordinary Differential Equations, Transport Theory and Sobolev Spaces. Invent. Math.* 98.3 (1989), pp. 511–547. doi: 10.1007/BF01393835.

[30] T. Dunst, M. Klein, A. Prohl, and A. Schäfer. *Optimal Control in Evolutionary Micromagnetism. IMA J. Numer. Anal.* first published online (August 14, 2014). doi: 10.1093/imanum/dru034.

[31] C. L. Fefferman. "Existence and Smoothness of the Navier–Stokes Equation". In: *The Millennium Prize Problems.* Clay Math. Inst., Cambridge, 2006, pp. 57–67.

[32] J.-F. Gerbeau, C. Le Bris, and T. Lelièvre. *Mathematical Methods for the Magnetohydrodynamics of Liquid Metals.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2006. isbn: 0-19-856665-4. doi: 10.1093/acprof:oso/9780198566656.001.0001.

[33] V. Girault and P.-A. Raviart. *Finite Element Approximation of the Navier-Stokes Equations.* Vol. 749. Lecture Notes in Mathematics. Springer, Berlin, 1979. isbn: 3-540-09557-8.

[34] G. Grün. *Degenerate Parabolic Differential Equations of Fourth Order and a Plasticity Model with non-local Hardening. Z. Anal. Anwendungen* 14.3 (1995), pp. 541–574. doi: 10.4171/ZAA/639.

[35] G. Grün. *On the Convergence of Entropy Consistent Schemes for Lubrication Type Equations in Multiple Space Dimensions. Math. Comp.* 72.243 (2003), 1251–1279 (electronic). doi: 10.1090/S0025-5718-03-01492-3.

[36] G. Grün and M. Rumpf. *Nonnegativity Preserving Convergent Schemes for the Thin Film equation. Numer. Math.* 87.1 (2000), pp. 113–152. doi: 10.1007/s002110000197.

[37] G. Grün and M. Rumpf. *Simulation of Singularities and Instabilities Arising in Thin Film Flow. European J. Appl. Math.* 12.3 (2001). The dynamics of Thin Fluid Films, pp. 293–320. doi: 10.1017/S0956792501004429.

[38] F. Guillén-González and J. Gutiérrez-Santacreu. *Conditional Stability and Convergence of a Fully Discrete Scheme for Three-dimensional Navier–Stokes Equations with Mass Diffusion. SIAM J. Numer. Anal.* 46.5 (2008), pp. 2276–2308. doi: 10.1137/07067951X.

[39] M. D. Gunzburger, ed. *Flow Control.* Vol. 68. The IMA Volumes in Mathematics and its Applications. Springer, New York, 1995. isbn: 0-387-94445-1.

[40] M. D. Gunzburger. *Perspectives in Flow Control and Optimization.* Vol. 5. Advances in Design and Control. SIAM, Philadelphia, 2003. isbn: 0-89871-527-X.

[41] R. Herzog and K. Kunisch. *Algorithms for PDE-Constrained Optimization. GAMM Mitt.* 33.2 (2010), pp. 163–176. doi: 10.1002/gamm.201010013.

[42] J. G. Heywood and R. Rannacher. *Finite Element Approximation of the Nonstationary Navier–Stokes problem. I. Regularity of Solutions and Second-Order Error Estimates for Spatial Discretization. SIAM J. Numer. Anal.* 19.2 (1982), pp. 275–311. doi: 10.1137/0719018.

[43] M. Hintermüller and M. Hinze. *Moreau-Yosida Regularization in State Constrained Elliptic Control Problems: Error Estimates and parameter Adjustment. SIAM J. Numer. Anal.* 47.3 (2009), pp. 1666–1683. doi: 10.1137/080718735.

[44] M. Hintermüller and K. Kunisch. *Feasible and Noninterior Path-Following in Constrained Minimization with low Multiplier Regularity. SIAM J. Control Optim.* 45.4 (2006), pp. 1198–1221. doi: 10.1137/050637480.

[45] M. Hinze. *A Variational Discretization Concept in Control Constrained Optimization: The Linear-Quadratic Case. Comput. Optim. Appl.* 30 (1 2005), pp. 45–61. doi: 10.1007/s10589-005-4559-5.

[46] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints.* Vol. 23. Mathematical Modelling: Theory and Applications. Springer, New York, 2009. isbn: 978-1-4020-8838-4.

[47] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications.* Advances in Design and Control 15. SIAM, Philadelphia, 2008. isbn: 978-0-898716-49-8. doi: 10.1137/1.9780898718614.

[48] G. Kitavtsev, P. Laurençot, and B. Niethammer. *Weak Solutions to Lubrication Equations in the Presence of Strong Slippage. Methods Appl. Anal.* 18.2 (2011), pp. 183–202. doi: 10.4310/MAA.2011.v18.n2.a4.

[49] M. Klein and A. Prohl. *Optimal Control for the Thin-Film Equation: Convergence of a Multi-Parameter Approach to track State Constraints avoiding Degeneracies* (2014). arXiv: 1410.2077v1 [math.OC].

[50] M. Klein and A. Prohl. *Optimal Control for the Thin-Film Equation: Convergence of a Multi-Parameter Approach to track State Constraints avoiding Degeneracies* (2014). arXiv: 1410.2077v2 [math.OC].

[51] S. Korotov and M. Křížek. *Acute Type Refinements of Tetrahedral Partitions of Polyhedral Domains. SIAM J. Numer. Anal.* 39.2 (2001), pp. 724–733. doi: 10.1137/S003614290037040X.

[52] K. Kunisch and X. Lu. *Optimal Control for Multi-Phase Fluid Stokes Problems. Nonlinear Anal.* 74.2 (2011), pp. 585–599. doi: 10.1016/j.na.2010.09.014.

[53] G. M. Lieberman. *Second order parabolic differential equations.* World Scientific, Singapore, 1996. isbn: 981-02-2883-X. doi: 10.1142/3302.

[54] P.-L. Lions. *Mathematical Topics in Fluid Mechanics.* Vol. 1: Incompressible models. Oxford Lecture Series in Mathematics and its Applications. Clarendon Press, Oxford University Press, New York, 1996. isbn: 0-19-851487-5.

[55] C. Liu and N. J. Walkington. *Convergence of Numerical Approximations of the Incompressible Navier–Stokes Equations with Variable Density and Viscosity. SIAM J. Numer. Anal.* 45.3 (2007), pp. 1287–1304. doi: 10.1137/050629008.

[56] D. G. Luenberger. *Optimization by Vector Space Methods.* Series in Decision and Control: Wiley Professional Paperback Series. John Wiley, New York, 1969. isbn: 0-471-55359-X.

[57] A. Münch, B. Wagner, and T. P. Witelski. *Lubrication Models with small to large Slip Lengths. J. Engrg. Math.* 53.3-4 (2005), pp. 359–383. doi: 10.1007/s10665-005-9020-3.

[58] I. Neitzel and F. Tröltzsch. *On Convergence of Regularization Methods for Nonlinear Parabolic Optimal Control Problems with Control and State Constraints. Control Cybernet.* 37.4 (2008), pp. 1013–1043.

[59] A. Oron, S. H. Davis, and S. G. Bankoff. *Long-scale Evolution of Thin Liquid Films. Rev. Mod. Phys.* 69.3 (1997), pp. 931–980. doi: 10.1103/RevModPhys.69.931.

[60] C. S. Peskin. *Numerical Analysis of Blood Flow in the Heart. Journal of Computational Physics* 25.3 (1977), pp. 220 –252. doi: 10.1016/0021-9991(77)90100-0.

[61] S. Qian and H. H. Bau. *Magneto-hydrodynamics Based Microfluidics. Mechanics Research Communications* 36.1 (2009), pp. 10–21. doi: 10.1016/j.mechrescom.2008.06.013.

[62] J. Simon. *Sobolev, Besov and Nikol′skiĭ Fractional Spaces: Imbeddings and Comparisons for Vector Valued Spaces on an Interval. Ann. Mat. Pura Appl.* 157 (1990), pp. 117–148. doi: 10.1007/BF01765315.

[63] R. Temam. *Navier-Stokes Equations. Theory and Numerical Analysis.* Vol. 2. Studies in Mathematics and its Applications. North-Holland Publishing Co., Amsterdam, 1977. isbn: 0-7204-2840-8.

[64] F. Tröltzsch and I. Yousept. *A regularization Method for the Numerical Solution of Elliptic Boundary Control Problems with Pointwise State Constraints. Comput. Optim. Appl.* 42.1 (2009), pp. 43–66. doi: 10.1007/s10589-007-9114-0.

[65] A. I. Vol′pert. *Spaces BV and Quasilinear Equations. Mat. Sb. (N.S.)* 73 (115) (1967), pp. 255–302.

[66] N. J. Walkington. *Compactness Properties of the DG and CG Time Stepping Schemes for Parabolic Equations. SIAM J. Numer. Anal.* 47.6 (2010), pp. 4680–4710. doi: 10.1137/080728378.

[67] X. Zhao and C. Liu. *Optimal Control of a Fourth-order Parabolic Equation Modeling Epitaxial Thin Film Growth. Bull. Belg. Math. Soc. Simon Stevin* 20.3 (2013), pp. 547–557. url: http://projecteuclid.org/euclid.bbms/1378314515.