Numerische Simulation der Bildung fluider Strukturen auf inhomogenen Oberflächen

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Jürgen Becker

aus

Trier

Bonn 2004

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: HD Dr. Günther Grün
2. Referent: Prof. Dr. Jens Frehse

Tag der Promotion: 15.12.2004

Zusammenfassung

Die Dynamik eines auf einer Oberfläche evolvierenden dünnen Flüssigkeitsfilms kann durch die degeneriert parabolische Gleichung 4. Ordnung

$$\partial_t u - \operatorname{div}\left(u^n \nabla(-\Delta u + w_{,u}(u,x))\right) = q(u), \tag{1}$$

die sogenannte Dünne-Filme-Gleichung, beschrieben werden. Dabei bezeichnet u die Dicke des Films. Adhäsionskräfte zwischen Flüssigkeit und Substrat beeinflussen die Dynamik des Films. Sie werden durch das effektive Grenzflächenpotential w beschrieben. w ist häufig singulär bzgl. der Filmdicke u, und im Fall einer inhomogenen Substratoberfläche hängt w nichtstetig von der Ortsvariablen x ab.

In dieser Arbeit wird ein Finite-Elemente-Verfahren zu Gleichung (1) entwickelt, implementiert und auf seine Konvergenzeigenschaften untersucht. Es wird gezeigt, dass die Lösungen dieses Verfahrens unter geeigneten Voraussetzungen an w und q natürliche a priori Integralabschätzungen erfüllen. Mit ihrer Hilfe lässt sich zeigen, dass diskrete Lösungen nichtnegativ sind und im Grenzfall verschwindender Gitter- und Zeitschrittweiten gegen eine Funktion u konvergieren. Diese Funktion u löst Gleichung (1) in einer geeigneten schwachen Formulierung. Auf diese Weise ergibt sich zugleich ein erstes Existenzresultat für schwache Lösungen der Evolutionsgleichung dünner Filme auf heterogenen Grenzflächen.

Ein Vergleich der in numerischer Simulation und physikalischem Experiment beobachteten fluiden Strukturen zeigt sowohl qualitativ als auch quantitativ eine sehr gute Übereinstimmung.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Mit Flüssigkeit benetzte Oberflächen begegnen einem im Alltag an verschiedenen Stellen: eine frisch lackierte Oberfläche, Fett in einer Pfanne, eine beschlagene Glasscheibe, Tinte auf einer Folie, etc. Dabei lassen sich zwei gänzlich unterschiedliche Effekte beobachten: in manchen Fällen perlt die Flüssigkeit von der Oberfläche ab und bildet kleine Tröpfchen, in anderen Fällen verteilt sie sich gleichmäßig auf der Substratoberfläche. Welcher Effekt eintritt, wird dabei nicht nur von der Art der Flüssigkeit und der Dicke des Flüssigkeitsfilms bestimmt, sondern auch von der Beschaffenheit der Substratoberfläche: So "beschlägt" eine Brille, wenn sich kondensierendes Wasser in kleinsten Tropfen sammelt. Ist die Brille mit einem Anti-Beschlagmittel behandelt, so wird die Tröpfchenbildung verhindert, vielmehr bildet sich ein Film, durch den man hindurchsehen kann.

Über diese alltäglichen Situationen hinausgehend gibt es aber noch weitere Gründe, das Phänomen der Benetzung intensiver zu betrachten. Im Zuge der Miniaturisierung in Industrie und Technik gibt es Bestrebungen, chemische Reaktionen auf einem Chip ablaufen zu lassen. Da diese häufig in der flüssigen Phase stattfinden, ist dazu das Bewegen kleinster Flüssigkeitsmengen vonnöten. Dabei spielen Benetzungseigenschaften eine entscheidende Rolle. Deshalb wird versucht, das Fließverhalten der Flüssigkeiten durch Präparation geeigneter chemisch oder topologisch inhomogener Substratoberflächen zu steuern. Die Benetzungseigenschaften solcher Oberflächen sind daher in den letzten Jahren in den Fokus verschiedenster physikalischer Experimente gerückt: So untersuchen z. B. Gau et al. [16] die Benetzung eines Substrates aus parallelen hydrophilen und hydrophoben Streifen mit Wasser (siehe Abbildung 1.1), Schäfle et al. [38, 39] beschäftigen sich mit Kondensation und Evaporation auf chemisch heterogenen Substraten, Darhuber et al. [13] untersuchen die Ausbreitung von Flüssigkeiten entlang hydrophiler Streifen und Rehse et al. [35] studieren die Benetzung topologisch inhomogener Substrate.

Grundlagen der Theorie zur Benetzung von Oberflächen bilden die Arbeiten von Young [46] und Reynolds [36]. In der Youngschen Gleichung wird der Zusammenhang zwischen dem Kontaktwinkel eines Tropfens im stationären Zustand und der Grenzflächenspannung für homogene und glatte Oberflächen beschrieben. Die Dynamik dünner Filme auf homogenen Substraten wird durch die auf Reynolds zurückgehende Dünne-Filme-Gleichung, eine Differentialgleichung für die Höhe des Flüssigkeitsfilmes, welche sich aus den Navier-Stokes-Gleichungen herleiten lässt, beschrieben. Der von Reynolds benutzte Ansatz lässt sich auch zur Herleitung einer Dünne-Filme-Gleichung für chemisch heterogene Substrate verwenden,

was in Kapitel 2 gezeigt wird. Dabei wird vorausgesetzt, dass die Substratoberfläche glatt ist, topologische Unebenheiten können mit diesem Ansatz nicht beschrieben werden

Ziel dieser Arbeit ist es daher, ein zuverlässiges und stabiles numerisches Verfahren zu entwickeln, welches die Dünne-Filme-Gleichung für chemisch inhomogene Substrate löst. Das in Kapitel 3 definierte Verfahren baut auf einem von Grün und Rumpf [21] für den homogenen Fall entwickelten Entropiekonsistenten Finite-Elemente-Verfahren auf. Im Gegensatz zu den in [21] und den nachfolgenden Arbeiten [22, 19] beschriebenen Verfahren erlaubt das in Kapitel 3 definierte Finite-Elemente-Schema auch Kondensation und Evaporation von Masse.

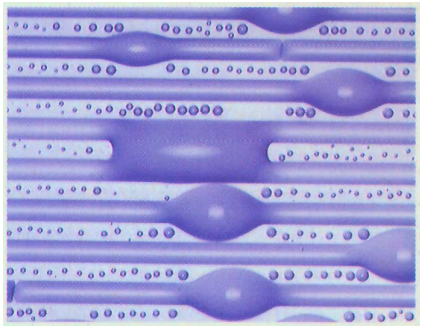


Abbildung 1.1: Benetzung eines abwechselnd hydrophoben und hydrophilen Substrates mit Wasser. Die Streifen sind etwa 60 μm breit (Experiment von Gau et al. [16])

Die diskreten Lösungen dieses Verfahrens zeichnen sich dadurch aus, dass sie verschiedene Integralabschätzungen erfüllen. Diese werden in Kapitel 4 bewiesen. Mit Hilfe dieser Abschätzungen kann in Kapitel 5 gezeigt werden, dass die diskreten Lösungen für gegen Null strebende Gitterweiten gegen eine nichtnegative, schwache Lösung des kontinuierlichen Problems konvergieren. Damit setzt sich das hier verwendete numerische Verfahren von anderen Verfahren (siehe z.B. [10, 27, 29, 43]) ab, für die ein solcher Konvergenzbeweis nicht möglich ist. Durch den Beweis der Konvergenz ist gleichzeitig auch die Existenz einer schwachen kontinuierlichen Lösung bewiesen.

Kapitel 6 zeigt die Ergebnisse numerischer Simulationen und vergleicht diese mit den Ergebnissen physikalischer Experimente. Näher untersucht werden hier vor allem die Entstehung von Satellitenlöchern in Polystyrolfilmen und die Kondensation auf chemisch strukturierten Substraten. Zur Simulation wurde das Programmpaket EConLub2D benutzt, welches eine Implementierung des in dieser Arbeit vorgestellten Verfahrens ist. Das Paket wird in Anhang A näher beschrieben.

Die in dieser Arbeit verwendeten Bezeichnungen sind in Anhang B zusammengefasst.

Kapitel 2

Mathematische Beschreibung dünner Filme

Dieses Kapitel gibt einen Überblick über die Herleitung der verwendeten Gleichungen und Potentiale. Abschnitt 2.1 zeigt, dem Übersichtsartikel von Oron, Davis und Bankoff [34] folgend, wie ausgehend von den Navier-Stokes-Gleichungen die Dünne-Filme-Gleichung hergeleitet wird. Dabei wird vorausgesetzt, dass Dichte und Viskosität der Flüssigkeit konstant sind, die Flüssigkeit also inkompressibel und newtonsch ist.

Der Einfluß der Struktur der Oberfläche auf die Gleichung wird in den Abschnitten 2.2 und 2.3 erläutert.

2.1 Die Dünne-Filme-Gleichung

Wir betrachten die folgende Situation: Die Menge $\{(x, y, z) \in \mathbb{R}^3 | z < 0\}$ beschreibe einen Festkörper mit glatter Oberfläche. Auf dieser Oberfläche befinde sich ein flüssiger Film. Der Fluss $v := (v_x, v_y, v_z)^T$ und der hydrostatische Druck $\mathfrak{p}$ innerhalb des Films werden durch die Navier-Stokes-Gleichung

$$\rho \left(\partial_t v + (v \cdot \nabla) v\right) = -\nabla \mathfrak{p} + \eta \Delta v - \nabla \phi \tag{2.1}$$

bestimmt. Dabei ist η die Viskosität und ϕ die Energiedichte eines äußeren Kraftfeldes. Da die Dichte ρ der Flüssigkeit als konstant angenommen wird, gilt die Kontinuitäts-Gleichung

$$\operatorname{div}(v) = 0. \tag{2.2}$$

An der Grenzfläche zwischen Flüssigkeit und Festkörper gelten die folgenden Randbedingungen: Neben der selbstverständlichen Bedingung $v_z = 0$ (kein Fluss in die Oberfläche hinein) gilt für den horizontalen Anteil der Geschwindigkeit $v_\parallel = (v_x, v_y)^T$:

$$v_\parallel|_{z=0} = \beta \partial_z v_\parallel. \tag{2.3}$$

Der Parameter $\beta \in \mathbb{R}_0^+$ wird als Schlupflänge bezeichnet, im Fall $\beta = 0$ (no-slip-Bedingung) haftet der Film an der Substratoberfläche. Unter der Annahme, dass die Oberflächenspannung ς konstant sei, gilt an der Flüssigkeits-Gas-Grenzfläche die Bedingung

$$\mathfrak{T} \vec{n} \cdot \vec{n} = -\kappa \varsigma. \tag{2.4}$$

Dabei ist der Spannungstensor $\mathfrak{T}$ definiert durch $\mathfrak{T} = \eta(\nabla v + (\nabla v)^T) + p\,\mathrm{Id}$, κ ist die mittlere Krümmung und $\vec{n}$ die Normale an der Flüssigkeits-Gas-Grenzfläche. Für Vektoren $\vec{t}$ tangential zur Flüssigkeit-Gas-Grenzfläche gilt $\mathfrak{T}\vec{n} \cdot \vec{t} = 0$. Außerdem soll keine Masse zwischen flüssiger und gasförmiger Phase ausgetauscht werden, d.h an der Oberfläche des Flüssigkeitsfilms gilt zusätzlich

$$\rho(v - v_i)\vec{n} = 0, \tag{2.5}$$

wobei v_i die Geschwindigkeit der Grenzfläche bezeichnet.

Ziel der auf Reynolds [36] zurückgehenden Lubrikationsapproximation ist es nun, unter der Annahme, dass erstens die Oberfläche des Films als eine Funktion $u(x, y)$ darstellbar und zweitens der Film dünn ist, also die horizontale Ausdehnung um einige Größenordnungen größer ist als die Dicke des Films, eine vereinfachte Differentialgleichung für u herzuleiten. Da dieser Ansatz in der Literatur ausführlich beschrieben ist (siehe z.B. [3, 34]), sollen hier nur kurz die wesentlichen Ideen beschrieben werden. Zuerst werden die Ortsvariablen durch dimensionslose Größen $\tilde{x}, \tilde{y}, \tilde{z}$ und die Zeit durch eine dimensionslose Größe $\tilde{t}$ ersetzt. Dazu sei h_0 die mittlere Filmdicke, l_0 sei eine charakteristische Längenskala und v_0 eine charakteristische Größe für die Geschwindigkeit. Nun setzen wir:

$$\tilde{x} = \frac{x}{l_0}, \quad \tilde{y} = \frac{y}{l_0}, \quad \tilde{z} = \frac{z}{h_0}, \quad \tilde{t} = \frac{v_0 t}{l_0}. \tag{2.6}$$

Unter der Annahme, dass $\varepsilon = \frac{h_0}{l_0} \ll 1$ ist, lassen sich nun die Gleichungen (2.1)-(2.5) deutlich vereinfachen, indem man nur Terme nullter Ordnung in ε betrachtet. Man erkennt dabei, dass $v_\parallel$ in z-Richtung ein parabolisches Profil aufweist (Poiseuille-Fluss). Integration über z führt schließlich zu einer Gleichung, welche nur noch von der Höhe $u(x, y)$ des Filmes abhängig ist, nämlich

$$\eta\partial_t u - \mathrm{div}_\parallel \left((\tfrac{1}{3}u^3 + \beta u^2)\nabla_\parallel(-\varsigma\Delta_\parallel u + \phi|_{z=u}) \right) = 0. \tag{2.7}$$

Dies ist die Dünne-Filme-Gleichung. Der Term $-\varsigma\Delta_\parallel u + \phi|_{z=u}$ wird auch reduzierter Druck genannt. Hier bezeichnen die Operatoren $\nabla_\parallel, \mathrm{div}_\parallel, \Delta_\parallel$ die auf die (x, y)-Koordinaten eingeschränkten Differentialoperatoren: $\nabla_\parallel = (\partial_x, \partial_y)^T, \mathrm{div}_\parallel = (\partial_x, \partial_y), \Delta_\parallel = \mathrm{div}_\parallel \nabla_\parallel$.

Im Falle von Evaporation und Kondensation gilt die Annahme der Massenerhaltung an der Flüssigkeits-Gas-Grenzfläche nicht mehr. Daher wird (2.5) ersetzt durch

$$\rho(v - v_i)\vec{n} = j. \tag{2.8}$$

Dieser Ansatz führt zu einer Dünne-Filme-Gleichung mit rechter Seite ungleich Null, nämlich

$$\eta\partial_t u - \mathrm{div}_\parallel \left((\tfrac{1}{3}u^3 + \beta u^2)\nabla_\parallel(-\varsigma\Delta_\parallel u + \phi|_{z=u}) \right) = \frac{\eta j}{\rho}. \tag{2.9}$$

Der Massefluß j normal zur Grenzfläche ist dabei nicht konstant, sondern abhängig von der Höhe u. Man setzt als Energiebilanz für $z = u$ an:

$$jL = -k_{th}\nabla\theta\vec{n}, \tag{2.10}$$

d.h. die gesamte zur Grenzfläche transportierte Wärme $-k_{th}\nabla\theta\vec{n}$ wird in latente Wärme der Evaporation umgewandelt. Hier bezeichnet θ die Temperatur, die Konstante k_{th} beschreibt die Wärmeleitfähigkeit der Flüssigkeit, L ist die latente Wärme der Evaporation

pro Masseneinheit. Die Temperatur θ wird innerhalb der Flüssigkeit bestimmt durch die Wärmeleitungsgleichung

$$\rho c \frac{d}{dt}\theta = k_{th}\Delta\theta \tag{2.11}$$

mit den Randwerten $\theta|_{z=0} = \theta_0$ und $\theta|_{z=h} - \theta_\infty = Kj = -\frac{K}{L}k_{th}\nabla\theta\vec{n}$, wobei θ_0 die Temperatur des Substrates und θ_∞ die Temperatur der Gasphase ist. Beide werden als konstant angenommen [34]. Die Konstante c beschreibt die spezifische Wärme der Flüssigkeit, die Konstante K setzt sich aus weiteren physikalischen Konstanten zusammen (Details siehe [34], Gleichung 2.82).

Dieser Ansatz liefert nach Übergang zu skalierten Größen und Grenzübergang $\varepsilon \to 0$ (Details siehe [34]) die folgende Formel für j:

$$j(u) = \frac{k_{th}(\theta_0 - \theta_\infty)}{Lu + Kk_{th}}. \tag{2.12}$$

Für $\theta_0 > \theta_\infty$ findet also Kondensation statt, für $\theta_0 < \theta_\infty$ Verdunstung. Für die Temperatur ergibt sich die Gleichung $\theta(u) = \theta_\infty + \frac{Kk_{th}(\theta_0-\theta_\infty)}{Lu+Kk_{th}}$. Man erhält also als Differentialgleichung für die Höhe u:

$$\eta\partial_t u - \mathrm{div}_\parallel\left(\left(\tfrac{1}{3}u^3 + \beta u^2\right)\nabla_\parallel\left(-\varsigma\Delta_\parallel u + \phi|_{z=u}\right)\right) = \frac{\eta}{\rho}\frac{k_{th}(\theta_0 - \theta_\infty)}{Lu + Kk_{th}}. \tag{2.13}$$

2.2 Molekulare Wechselwirkungen

Wechselwirkungen zwischen zwei Molekülen im Abstand r werden im allgemeinen durch ein Potential der Form $\frac{C}{r^n}$ beschrieben. Die Konstante C kann dabei je nach Art der Wechselwirkungen positiv oder negativ sein und hat die physikalische Einheit Jm^n. Van-der-Waals-Wechselwirkungen zum Beispiel werden mit $n = 6$ modelliert [24].

Unter der Annahme von Additivität kann die gesamte Wechselwirkungsenergie ψ zwischen einem Flüssigkeits-Molekül an der Stelle (x, y, z) und dem Substrat $\{z < 0\}$ berechnet werden. Für ein homogenes Substrat mit konstanter Molekül dichte ϱ_s ist

$$\psi(x, y, z) = \int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \int_{-\infty}^{0} dz' \frac{C\varrho_s}{((x' - x)^2 + (y' - y)^2 + (z' - z)^2)^{n/2}}. \tag{2.14}$$

Nach Transformation in Zylinderkoordinaten r, h, φ durch

$$z' - z = h, \quad y' - y = r\cos\varphi, \quad x' - x = r\sin\varphi$$

lässt sich dieses Integral berechnen und man erhält für $n > 3$:

$$\psi(x, y, z) = \int_0^{\infty} r dr \int_0^{2\pi} d\varphi \int_{-\infty}^{-z} dh \frac{C\varrho_s}{(r^2 + h^2)^{n/2}} = \frac{2\pi}{(n-2)(n-3)} C\varrho_s \frac{1}{z^{n-3}}. \tag{2.15}$$

Im Fall von van-der-Waals-Kräften erhält man also das Gesamtpotential

$$\psi(z) = \frac{\pi}{6} C\varrho_s \frac{1}{z^3}. \tag{2.16}$$

Ist das Substrat inhomogenen, so sind C und ϱ_s abhängig von den Ortsvariablen x', y', z'. Wir betrachten zunächst einmal den einfachen Fall eines aus zwei homogenen Materialien zusammengesetzten Substrats:

$$\begin{aligned}\Omega_1 &:= \{(x,y) \in \mathbb{R}^2 : x < 0\},\\ \Omega_2 &:= \{(x,y) \in \mathbb{R}^2 : x > 0\}.\end{aligned} \tag{2.17}$$

$$C(x',y',z')\varrho(x',y',z') = \begin{cases} C_1\varrho_{s_1} & \text{falls } (x',y') \in \Omega_1 \\ C_2\varrho_{s_2} & \text{falls } (x',y') \in \Omega_2. \end{cases} \tag{2.18}$$

Für die Berechnung des Integrals ψ ist es unerheblich, wie C und ϱ auf $\Gamma := \{(x,y) \in \mathbb{R}^2 : x = 0\}$ gewählt werden. Sei also z. B.

$$C(x',y',z')\varrho(x',y',z') = C_1\varrho_{s_1}, \quad \text{falls } (x',y') \in \Gamma.$$

Dann folgt, dass sich $\psi(x,y,z)$ aufspalten lässt in die beiden Integrale

$$\begin{aligned}\psi(x,y,z) &= \psi_-(x,y,z) + \psi_+(x,y,z)\\ &= \int_{-\infty}^{0} dx' \int_{-\infty}^{\infty} dy' \int_{-\infty}^{0} dz' \frac{C_1\varrho_{s_1}}{((x'-x)^2+(y'-y)^2+(z'-z)^2)^{n/2}}\\ &\quad + \int_{0}^{\infty} dx' \int_{-\infty}^{\infty} dy' \int_{-\infty}^{0} dz' \frac{C_2\varrho_{s_2}}{((x'-x)^2+(y'-y)^2+(z'-z)^2)^{n/2}}.\end{aligned} \tag{2.19}$$

Im Fall $x = 0$ lässt sich dies nach Transformation in Zylinderkoordinaten lösen und man erhält:

$$\begin{aligned}\psi_-(x,y,z) &= \frac{\pi}{(n-2)(n-3)} C_1\varrho_{s_1} \frac{1}{z^{n-3}},\\ \psi_+(x,y,z) &= \frac{\pi}{(n-2)(n-3)} C_2\varrho_{s_2} \frac{1}{z^{n-3}}.\end{aligned} \tag{2.20}$$

Im Fall $x \neq 0$ ergibt sich nach Integration über y

$$\psi_-(x,y,z) = C_1\varrho_{s_1}\sqrt{\pi}\frac{\Gamma(\frac{n}{2}-\frac{1}{2})}{\Gamma(\frac{n}{2})} \int_{-\infty}^{0} dx' \int_{-\infty}^{0} dz' \frac{1}{((x'-x)^2+(z'-z)^2)^{\frac{n-1}{2}}}, \tag{2.21}$$

wobei Γ die Eulersche Gammafunktion ist. Nun substituieren wir nach einer Idee von Dietrich und Rauscher [14] $\tilde{x} = -\frac{x'}{|x|}$ und $\tilde{z} = -\frac{z'}{z}$ (wir setzen $z > 0$ als gegeben voraus) und erhalten

$$\psi_-(x,y,z) = C_1\varrho_{s_1}\sqrt{\pi}\frac{\Gamma(\frac{n}{2}-\frac{1}{2})}{\Gamma(\frac{n}{2})}\frac{1}{z^{n-3}} \int_{0}^{\infty} \int_{0}^{\infty} \frac{|s| d\tilde{x}\, d\tilde{z}}{(s^2(\tilde{x}+\operatorname{sgn}(s))^2+(\tilde{z}+1)^2)^{\frac{n-1}{2}}}.$$

Analog erhalten wir für ψ_+ (hier substituieren wir $\tilde{x} = \frac{x'}{|x|}$ und $\tilde{z} = -\frac{z'}{z}$):

$$\psi_+(x,y,z) = C_2\varrho_{s_2}\sqrt{\pi}\frac{\Gamma(\frac{n}{2}-\frac{1}{2})}{\Gamma(\frac{n}{2})}\frac{1}{z^{n-3}} \int_{0}^{\infty} \int_{0}^{\infty} \frac{|s| d\tilde{x}\, d\tilde{z}}{(s^2(\tilde{x}-\operatorname{sgn}(s))^2+(\tilde{z}+1)^2)^{\frac{n-1}{2}}},$$

wobei s jeweils $\frac{x}{z}$ bezeichnet. Es gilt also

$$\psi(x,y,z) = \frac{2\pi}{(n-2)(n-3)}\frac{1}{z^{n-3}}\left(C_1\varrho_{s_1}\alpha_n^{(1)}(\tfrac{x}{z}) + C_2\varrho_{s_2}\alpha_n^{(2)}(\tfrac{x}{z})\right) \tag{2.22}$$

mit

$$\begin{aligned}\alpha_n^{(1)}(s) &:= \frac{(n-2)(n-3)\Gamma(\frac{n}{2}-\frac{1}{2})}{2\sqrt{\pi}\Gamma(\frac{n}{2})}\int_0^\infty\int_0^\infty \frac{|s|d\tilde{x}\,d\tilde{z}}{(s^2(\tilde{x}+\mathrm{sgn}(s))^2+(\tilde{z}+1)^2)^{\frac{n-1}{2}}},\\ \alpha_n^{(2)}(s) &:= \frac{(n-2)(n-3)\Gamma(\frac{n}{2}-\frac{1}{2})}{2\sqrt{\pi}\Gamma(\frac{n}{2})}\int_0^\infty\int_0^\infty \frac{|s|d\tilde{x}\,d\tilde{z}}{(s^2(\tilde{x}-\mathrm{sgn}(s))^2+(\tilde{z}+1)^2)^{\frac{n-1}{2}}}.\end{aligned} \tag{2.23}$$

Durch Vergleich mit der Lösung für das homogene Substrat erkennen wir, dass $\alpha_n^{(1)}(s) + \alpha_n^{(2)}(s) = 1$ gelten muss. Ebenso ergibt sich aus (2.20), dass $\lim_{s\to 0}\alpha_n^{(1)}(s) = \frac{1}{2}$ ist. Für van-der-Waals-Wechselwirkungen ergibt die Berechnung der obigen Integrale

$$\begin{aligned}\alpha_6^{(1)}(s) &:= \frac{1}{2}+\frac{1}{2s^3}-\frac{2+s^2+2s^4}{4s^3\sqrt{1+s^2}},\\ \alpha_6^{(2)}(s) &:= \frac{1}{2}-\frac{1}{2s^3}+\frac{2+s^2+2s^4}{4s^3\sqrt{1+s^2}}.\end{aligned} \tag{2.24}$$

Hier gilt außerdem noch $\lim_{s\to\infty}\alpha_6^{(1)}(s) = 0$ und $\lim_{s\to-\infty}\alpha_6^{(1)}(s) = 1$, was bedeutet, dass für $x \gg z$ die Inhomogenität für den Wert des Potentials ψ keine Rolle mehr spielt[1].

Auf diese Art und Weise kann das Potential ψ auch für ein aus parallelen Streifen unterschiedlicher Materialien bestehendes Substrat berechnet werden, da sich diese Situation auf die obige zurückführen lässt. Insbesondere lässt sich damit ψ für alle zweidimensionalen Probleme (als Substrat dient hier die x-Achse) berechnen.

Für kompliziertere Geometrien ist ψ häufig nicht explizit berechenbar. Dennoch lässt sich auf einfache Art und Weise eine gute Näherung bestimmen, wie im nächsten Abschnitt beschrieben wird.

2.3 Das effektive Grenzflächenpotential

Das im vorherigen Abschnitt berechnete Potential ψ beschreibt die Lageenergie eines Punktes. Die Energiedichte ϕ aus der Navier-Stokes-Gleichung (2.1) bestimmt sich nun durch

$$\phi = \varrho_f \psi, \tag{2.25}$$

wobei ϱ_f die Moleküldichte der Flüssigkeit ist. Das effektive Grenzflächenpotential[2] w bestimmt sich aus

$$\partial_z w(x,y,z)|_{z=u} = \phi(x,y,u(x,y)). \tag{2.26}$$

Also lautet die Dünne-Filme-Gleichung (2.7) auch

$$\eta\partial_t u - \mathrm{div}_\parallel\left((\tfrac{1}{3}u^3+\beta u^2)\nabla_\parallel(-\varsigma\Delta_\parallel u + \partial_z w(x,y,u))\right) = 0. \tag{2.27}$$

[1] Es lässt sich vermuten – und mit Hilfe von MAPLE für den Einzelfall auch nachrechnen – dass auch für allgemeines $n \in \mathbb{N}$, $n > 3$, $\lim_{s\to\infty}\alpha_n^{(1)}(s) = 0$ und $\lim_{s\to-\infty}\alpha_n^{(1)}(s) = 1$ gültig sind.

[2] Der Name 'Grenzflächenpotential' rührt daher, dass w im Fall eines homogenen Substrats für parallele Flüssigkeits-Gas- und Festkörper-Flüssigkeits-Grenzflächen die Energie pro Einheitsfläche beschreibt. w hat daher die physikalische Einheit $\mathrm{J/m^2}$.

Auf homogenen Substraten ist w unabhängig von x und y. Statt $\partial_z w(x,y,u)$ kann man daher auch $w'(u)$ schreiben. Van-der-Waals-Kräfte werden also durch das Grenzflächenpotential

$$w(u) = -\frac{A}{12\pi}u^{-2} \tag{2.28}$$

beschrieben, wobei $A = \pi^2 C \varrho_s \varrho_f$ die Hamakerkonstante ist. Für positive Hamakerkonstanten ist dieses Potential destabilisierend. Für das gleichzeitige Darstellen von destabilierenden van-der-Waals-Kräften und stabilisierenden kurzreichweitigen Wechselwirkungen gibt es verschiedene Ansätze. Oron, Davis und Bankoff [34] schlagen Potentiale der Form

$$w(u) = -a_2 u^{-2} + a_3 u^{-3}, \quad a_2, a_3 > 0 \tag{2.29}$$

vor. Der Lennard-Jones-Ansatz [24], welcher van-der-Waals-Kräfte mit $n = 6$ und kurzreichweitige Bornsche Abstoßung mit $n = 12$ modelliert, führt zu dem Grenzflächenpotential

$$w(u) = -\frac{A}{12\pi}u^{-2} + \varepsilon u^{-8}, \tag{2.30}$$

mit einer positiven Konstanten ε, welche die Stärke der kurzreichweitigen Kräfte beschreibt. Falls das Substrat selbst aus unterschiedlichen Schichten besteht, so spielen, wenn die oberste Schicht dünn genug ist, auch darunterliegende Materialien noch eine Rolle. So erhält man, von einem Lennard-Jones-Ansatz ausgehend, Potentiale der Form

$$w(u) = -\frac{A_1}{12\pi}u^{-2} + \frac{A_1 - A_2}{12\pi}(u+d)^{-2} + \varepsilon u^{-8}. \tag{2.31}$$

Hier beschreibt d die Dicke und A_1 die Hamakerkonstante der obersten Schicht. A_2 ist die Hamakerkonstante der darunterliegenden Schicht und $\varepsilon > 0$. (siehe Seemann, Herminghaus und Jacobs [40]).

Das Grenzflächenpotential w kann aber nicht nur molekulare Wechselwirkungen beschreiben, sondern auch noch weitere physikalische Kräfte. So lässt sich z. B. mit Hilfe von Potentialen der Form

$$w(u) = cu^2 \tag{2.32}$$

Gravitation beschreiben.

Auf inhomogenen Substraten ist w abhängig von den Ortsvariablen x, y. Falls das Substrat wie in (2.17) definiert ist, ergibt sich für van-der-Waals-Kräfte ($n = 6$) das Potential[3]

$$w(x,y,u) = -\frac{A_1}{12\pi}\frac{1}{u^2}\beta_6^{(1)}(\tfrac{x}{u}) - \frac{A_2}{12\pi}\frac{1}{u^2}\beta_6^{(2)}(\tfrac{x}{u}), \tag{2.33}$$

wobei die Hamakerkonstanten A_i, $i \in \{1,2\}$ gegeben sind durch $A_i = \pi^2 C_i \varrho_{s_i} \varrho_f$.
$\beta_6^{(i)}$ ist eine Partition der Eins wie folgt:

$$\begin{aligned} \beta_6^{(1)}(s) &= \frac{1}{2} - \frac{1}{s^3} + \frac{1}{s}\sqrt{1+s^2}\left(\frac{1}{s^2} - \frac{1}{2}\right), \\ \beta_6^{(2)}(s) &= \frac{1}{2} + \frac{1}{s^3} - \frac{1}{s}\sqrt{1+s^2}\left(\frac{1}{s^2} - \frac{1}{2}\right). \end{aligned} \tag{2.34}$$

[3]Das Potential (2.33) lässt sich auch mit Hilfe der Dichtefunktionaltheorie herleiten, siehe dazu [2, 28].

Gleichung (2.26) gilt, da $\beta_6^{(i)}(s) + \frac{1}{2}\partial_s\beta_6^{(i)}(s)s = \alpha_6^{(i)}(s)$ und damit

$$\partial_z w(x,y,u) = \frac{A_1}{6\pi}\frac{1}{u^3}\alpha_6^{(1)}(\tfrac{x}{u}) + \frac{A_2}{6\pi}\frac{1}{u^3}\alpha_6^{(2)}(\tfrac{x}{u}). \tag{2.35}$$

Ähnliche Formeln lassen sich auch für andere Werte von n berechnen, aus (2.23) lässt sich jedoch keine allgemeine n-abhängige Formel herleiten. Von einem 6-12 Lennard-Jones-Ansatz ausgehend, ergibt sich ein Potential der Form

$$w(x,y,z) = -\frac{A_1}{12\pi}\frac{1}{u^2}\beta_6^{(1)}(\tfrac{x}{u}) - \frac{A_2}{12\pi}\frac{1}{u^2}\beta_6^{(2)}(\tfrac{x}{u}) + \varepsilon_1 u^{-8}\beta_{12}^{(1)}(\tfrac{x}{u}) + \varepsilon_2 u^{-8}\beta_{12}^{(2)}(\tfrac{x}{u}) \tag{2.36}$$

mit $\beta_{12}^{(1)} + \beta_{12}^{(2)} \equiv 1$.

Für andere Geometrien der Oberfläche lässt sich $w(x,y,u)$ häufig nicht explizit berechnen. Daher wird Folgendes als Ansatz für ein Gebiet $\Omega \subset \mathbb{R}^2$, welches sich durch $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma$ disjunkt in zwei Gebiete Ω_1 und Ω_2 aus unterschiedlichen chemischen Materialien und eine Grenze $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ aufteilen lässt, gewählt[4]:

$$w(x,y,u) = \begin{cases} w_1(u) & \text{falls } (x,y) \in \Omega_1 \cup \Gamma, \\ w_2(u) & \text{falls } (x,y) \in \Omega_2. \end{cases} \tag{2.37}$$

Es lässt sich zeigen, dass dieser Ansatz zumindest für das in (2.17) definierte Substrat eine sehr gute Näherung darstellt. Der Grund dafür ist, dass Gleichung (2.33) eine exakte Darstellung des Grenzflächenpotentials ist, die Differentialgleichung (2.27) aber das Problem für den Grenzfall $\varepsilon = 0$ löst. Für die dimensionslosen Größen aus (2.6), d.h. $\tilde{x} = \frac{x}{l_0}$ und $\tilde{u} = \frac{u}{h_0}$, gilt $\frac{x}{u} = \frac{\tilde{x}}{\varepsilon\tilde{u}}$ und dann ist mit $\tilde{s} := \frac{\tilde{x}}{\tilde{u}}$

$$\beta_6^{(1)}(\tfrac{x}{u}) = \beta_6^{(1)}(\tfrac{\tilde{s}}{\varepsilon}) = \frac{1}{2} - \frac{\varepsilon^3}{\tilde{s}^3} + \frac{1}{\tilde{s}}\sqrt{\varepsilon^2+\tilde{s}^2}\left(\frac{\varepsilon^2}{\tilde{s}^2} - \frac{1}{2}\right). \tag{2.38}$$

Dies konvergiert für $\varepsilon \to 0$ gegen $\frac{1}{2} - \frac{1}{2}\operatorname{sgn}(\tilde{s})$, also:

$$\lim_{\varepsilon\to 0}\beta_6^{(1)}(\tfrac{x}{u}) = \begin{cases} 1 & \text{falls } x < 0, \\ 0 & \text{falls } x > 0. \end{cases} \tag{2.39}$$

Im Grenzfall $\varepsilon = 0$ der Dünne-Filme-Gleichung gilt (2.37) also für $x \neq 0$.

Mathematisch exakt zeigen dies Dietrich und Rauscher [14] durch *asymptotic matching*. Für die äußeren Gebiete $\{x < -\delta\}$ und $\{x > \delta\}$ leiten sie wie in Abschnitt 2.1 beschrieben die Dünne-Filme-Gleichung her. Im inneren Gebiet $\{-\delta < x < \delta\}$ wird eine andere Skalierung als (2.6) gewählt. Anschließend werden innere und äußere Lösung abgeglichen und der Grenzübergang $\delta \to 0$ betrachtet. So erhalten sie für den Fall $\theta_0 = \theta_\infty$ und konstanter Oberflächenspannung ς das folgende Resultat:

$$\begin{aligned} \eta\partial_t u - \operatorname{div}_\parallel\left((\tfrac{1}{3}u^3 + \beta u^2)\nabla_\parallel(-\varsigma\Delta_\parallel u + w_1'(u))\right) = 0 \quad & \text{falls } x < 0, \\ \eta\partial_t u - \operatorname{div}_\parallel\left((\tfrac{1}{3}u^3 + \beta u^2)\nabla_\parallel(-\varsigma\Delta_\parallel u + w_2'(u))\right) = 0 \quad & \text{falls } x > 0, \end{aligned} \tag{2.40}$$

mit den Anschlussbedingungen

$$\begin{aligned} u|_{x\nearrow 0} = u|_{x\searrow 0}, \qquad & p_1|_{x\nearrow 0} = p_2|_{x\searrow 0}, \\ \nabla_\parallel u|_{x\nearrow 0} = \nabla_\parallel u|_{x\searrow 0}, \quad & \nabla_\parallel p_1|_{x\nearrow 0} = \nabla_\parallel p_2|_{x\searrow 0}, \end{aligned} \tag{2.41}$$

[4] Für die analytischen Betrachtungen spielt der Wert von w auf Γ keine Rolle.

wobei p_i den reduzierten Druck $-\varsigma\Delta_\parallel u + w_i'(u)$ bezeichnet. Dies bedeutet, dass $\Delta_\parallel u$ nicht stetig in $x = 0$ ist, da p stetig ist und w nicht. Dies zeigt auch, welche Regularität wir für eine Lösung von (2.27) maximal erwarten können.

2.4 Zusammenfassung

Unter der Annahme einer konstanten Oberflächenspannung ς gilt die folgende Differentialgleichung für die Filmhöhe u:

$$\eta\partial_t u - \operatorname{div}_\parallel \left((\tfrac{1}{3}u^3 + \beta u^2)\nabla_\parallel p \right) = q(u). \tag{2.42}$$

Der reduzierte Druck p ist dabei gegeben durch

$$p = -\varsigma\Delta_\parallel u + \phi|_{z=u}. \tag{2.43}$$

Unter der Annahme von Massenerhaltung gilt $q(u) = 0$, und bei Kondensation oder Evaporation gilt:

$$q(u) = \frac{\eta}{\rho}\frac{k_{th}(\theta_0 - \theta_\infty)}{Lu + Kk_{th}}. \tag{2.44}$$

Mit Hilfe des effektiven Grenzflächenpotentials w lässt sich der reduzierte Druck auch schreiben als

$$p = -\varsigma\Delta_\parallel u + \partial_z w(x, y, u). \tag{2.45}$$

Das effektive Grenzflächenpotential selbst ist auf inhomogenen Substraten der Form $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma$, durch

$$w(x, y, u) = \begin{cases} w_1(u) & \text{falls } (x, y) \in \Omega_1 \cup \Gamma, \\ w_2(u) & \text{falls } (x, y) \in \Omega_2 \end{cases} \tag{2.46}$$

näherungsweise bestimmt. Dabei sind $w_1(u)$ und $w_2(u)$ die aus dem homogenen Fall bekannten Grenzflächenpotentiale.

Kapitel 3

Das Entropie-konsistente Finite-Elemente-Verfahren

In diesem Kapitel wird ein numerisches Verfahren vorgestellt, welches die Dünne-Filme-Gleichung auf inhomogenen Substraten löst. Die diskreten Lösungen dieses Verfahrens zeichnen sich dadurch aus, dass sie zwei Integralabschätzungen, welche Energie- und Entropieabschätzung genannt werden, erfüllen. Integralabschätzungen dieser Art sind in der Theorie der degeneriert parabolischen Differentialgleichungen häufig eingesetzte Hilfsmittel. Für die eindimensionale Dünne-Filme-Gleichung

$$u_t + \partial_x(u^n u_{xxx}) = 0 \tag{3.1}$$

ohne Potential- und Quellterm leiteten Bernis und Friedman [6] Energie- und Entropieabschätzungen her und benutzten diese um für $n > 1$ die Existenz nichtnegativer schwacher Lösungen zu zeigen.

Die Energieabschätzung zu Gleichung (3.1) ergibt sich – formal – durch Multiplikation von (3.1) mit u_{xx} und Integration über $(0,T) \times \Omega$. Nach partieller Integration erhält man eine Abschätzung für die Energie $\int_\Omega u_x^2(T)dx$ zur Zeit T,

$$\frac{1}{2}\int_\Omega u_x^2(T)dx + \int_0^T \int_\Omega u^n u_{xxx}^2 dxdt = \frac{1}{2}\int_\Omega u_x^2(0)dx, \tag{3.2}$$

womit eine Integralabschätzung für die erste Ableitung gegeben ist.

Die Entropieabschätzung ist einerseits eine Integralabschätzung für die zweiten Ableitungen, andererseits schätzt sie die Entropie $G(u)$ ab, welche durch

$$G(s) = \int_A^s g(r)dr, \quad g(s) = \int_A^s \frac{1}{r^n}dr \tag{3.3}$$

definiert ist. Formal erhält man die Entropieabschätzung zu (3.1), indem man Gleichung (3.1) mit $G'(u)$ multipliziert, über $(0,T) \times \Omega$ integriert, die Gleichheit

$$\partial_x G'(u)u^n = u_x \tag{3.4}$$

ausnutzt und das Resultat partiell integriert. Dann erhält man

$$\int_\Omega G(u(T))dx + \int_0^T \int_\Omega u_{xx}^2 dxdt = \int_\Omega G(u(0))dx. \tag{3.5}$$

Die Beschränktheit des Entropie-Terms $\int_\Omega G(u(T))dx$ ist nun der Ausgangspunkt des Nichtnegativitätsbeweises in [6].

Allgemeinere α-Entropieabschätzungen, in denen an Stelle von $\int_\Omega G(u(T))$ der Term $\int_\Omega u^{\alpha+1}(T)$ für $\frac{1}{2} < \alpha+n < 2$ abgeschätzt wird, und lokale Versionen der obigen Abschätzungen ermöglichen Existenz- und Nichtnegativitätsresultate auch im Fall $0 < n < 1$ und in höheren Raumdimensionen (siehe [5, 8, 12, 17]).

Diese Ideen können auch numerisch genutzt werden. Das von Grün und Rumpf [21] für die Gleichung (3.1) entwickelte Finite-Elemente-Verfahren ist so konstruiert, dass die diskrete Lösung diskrete Versionen der Abschätzungen (3.2) und (3.5) erfüllt. Dadurch wird Nichtnegativität der diskreten Lösung auf eine natürliche Art und Weise sichergestellt, wodurch Konvergenz gegen eine nichtnegative kontinuierliche Lösung gezeigt werden kann. Diese Strategie kann nicht nur auf Gleichung (3.1), sondern auch auf die mehrdimensionale Gleichung

$$\partial_t u - \operatorname{div}\left(u^n \nabla(-\Delta u + w'(u))\right) = 0 \tag{3.6}$$

angewandt werden (siehe [22] und [19]).

In diesem Kapitel wird nun, unter Ausnutzung dieser Ideen, ein numerisches Verfahren für Gleichungen der Form

$$\partial_t u - \operatorname{div}\left(u^n \nabla(-\Delta u + w_{,u}(u,x))\right) = q(u) \tag{3.7}$$

definiert. Im nächsten Kapitel wird gezeigt, dass auch die diskreten Lösungen dieses Verfahrens eine Energie- und eine Entropieabschätzung erfüllen.

Bemerkung zur Notation: In Abschätzungen auftretende Konstanten werden der Übersichtlichkeit halber häufig zu einer Konstanten C zusammengefasst. Dabei kann sich die Größe von C von Zeile zu Zeile einer Abschätzung ändern, ohne dass dies explizit erwähnt wird.

3.1 Diskretisierung mit Finiten Elementen

Das in Schema 3.2.2 definierte Verfahren benutzt als Diskretisierung für ein polygonal berandetes Gebiet $\Omega \subset \mathbb{R}^d$ eine zulässige und rechtwinklige Triangulierung. Diese ist definiert durch (siehe auch Ciarlet [11]):

Definition 3.1.1 *(zulässige und rechtwinklige Triangulierung)*

Sei $\Omega \subset \mathbb{R}^d$ polygonal berandet. Eine Triangulierung $\mathcal{T}_h$ von Ω heißt zulässig, wenn sie die folgenden Bedingungen erfüllt:

(T1) Alle $E \in \mathcal{T}_h$ sind nicht-degenerierte d-Simplizes[1] *und es gilt $\overline{\Omega} = \bigcup_{E\in\mathcal{T}_h} E$.*

(T2) Für zwei beliebige $E_i, E_j \in \mathcal{T}_h$ gilt: $E_i \cap E_j$ ist ein Untersimplex oder $\emptyset$.

(T3) Für alle E gilt: $\frac{h(E)}{\rho(E)} \le C < \infty$. Dabei ist $h(E) = \operatorname{diam}(E)$ und $\rho(E)$ der Durchmesser der Inkugel von E.

[1] d.h. es gibt Punkte $x_0(E), \dots, x_d(E) \in R^d$, so dass $x_1(E) - x_0(E), \dots, x_d(E) - x_0(E)$ linear unabhängig sind und $E = \{x \in \mathbb{R}^d : x = \sum_{i=0}^d \lambda_i x_i(E), 0 \le \lambda_i \le 1, \sum_{i=0}^d \lambda_i = 1\}$ gilt.

$\mathcal{T}_h$ heißt rechtwinklig, wenn zusätzlich gilt:

(T4) Für alle $E \in \mathcal{T}_h$ gibt es einen Knoten $x_0(E) \in E$, so dass die Kanten, welche x_0 mit den anderen Eckpunkten von E verbinden, rechtwinklig zueinander stehen.

Zulässige Triangulierungen sind also insbesondere Triangulierungen ohne hängende Knoten, deren Dreiecke nicht beliebig spitz werden können.

Mit der Triangulierung $\mathcal{T}_h$ ist gleichzeitig auch die Menge $\mathcal{N}_h$ der Knotenpunkte $\mathfrak{x}_i$ festgelegt. Die Anzahl dieser Punkte, wir bezeichnen sie mit D, bestimmt die Dimension des Finite-Element-Raums V^h der stetigen, stückweise linearen Funktionen. Funktionen aus V^h werden stets mit Großbuchstaben bezeichnet. Die Standardbasis $\{\varphi_i\}_{i=1,\dots,D}$ von V^h wird durch die "Hütchenfunktionen" gebildet, welche durch $\varphi_i(\mathfrak{x}_j) = \delta_{ij}$ definiert sind. Jede diskrete Funktion $U \in V^h$ lässt sich mit Hilfe der Standardbasis als Koeffizientenvektor $(U_1, \dots, U_D)^T \in \mathbb{R}^D$ darstellen:

$$U(x) = \sum_{i=1}^{D} U_i \varphi_i(x). \tag{3.8}$$

Da $U_i = U(\mathfrak{x}_i)$ gilt, wird dieser Koeffizientenvektor zur Vereinfachung der Schreibweise im folgenden ebenfalls mit U bezeichnet.

Der Operator $\mathcal{I}_h$ definiert eine Projektion $C^0(\overline{\Omega}) \to V^h$ durch:

$$\mathcal{I}_h u = \sum_{i=1}^{D} u(\mathfrak{x}_i)\varphi_i. \tag{3.9}$$

Mit Hilfe von $\mathcal{I}_h$ lässt sich ein auf V^h zum L^2-Skalarprodukt äquivalentes, aber einfacher zu berechnendes Skalarprodukt definieren:

Definition 3.1.2 *(verdichtete Massen Skalarprodukt)*

Für $\Phi, \Psi \in C^0(\overline{\Omega})$ sei das verdichtete Massen (engl.: lumped masses) Skalarprodukt $(\cdot,\cdot)_h$ definiert durch:

$$(\Phi, \Psi)_h := \int_\Omega \mathcal{I}_h(\Phi(x)\Psi(x))\, dx. \tag{3.10}$$

Die zugehörige Norm $\|\Psi\|_h := \sqrt{(\Psi,\Psi)_h}$ ist auf V^h äquivalent zur L^2-Norm. Ferner gibt es eine Konstante C_m, so dass für alle $U, V \in V^h$ die Abschätzung

$$|(U,V) - (U,V)_h| \le C_m h^{l+1} |U|_{l,2} |V|_{1,2} \tag{3.11}$$

gilt.

Ein Beweis der Ungleichung (3.11) findet sich z.B bei Thomée [44] (Lemma 15.1).

Das im folgenden betrachtete Zeitintervall $[0,T]$ sei durch $0 = \mathfrak{t}_0 < \mathfrak{t}_1 < \cdots < \mathfrak{t}_K = T$ unterteilt. Das Finite-Elemente-Schema 3.2.2 berechnet die numerische Lösung an den diskreten Zeitpunkten $\mathfrak{t}_k$, die Zeitschrittweiten werden mit $\tau_k = \mathfrak{t}_k - \mathfrak{t}_{k-1}$ bezeichnet. Die errechneten numerischen Lösungen sind also in der Menge

$$S^{-1,0}(V^h) := \left\{ U : [0,T] \to V^h \;\middle|\; U(t) = U(\mathfrak{t}_{k+1}) \text{ für } \mathfrak{t}_k < t \le \mathfrak{t}_{k+1}, k = 0, \dots, K-1 \right\} \tag{3.12}$$

enthalten, da zu jeder Folge $U^k \in V^h$, $k = 0, \ldots, K$ durch $U|_{(t_k, t_{k+1}]} = U^{k+1}$ eine Funktion $U \in S^{-1,0}(V^h)$ definiert ist, welche stückweise konstant in der Zeit ist. Man beachte, dass U auf jedem Zeitintervall den Wert des *rechten* Randes annimmt. Für $U(t)(x)$ schreiben wir auch $U(t, x)$.

3.2 Definition des Verfahrens

Sei Ω ein Lipschitz-Gebiet im R^d, $d \in \{1, 2\}$, $T \in \mathbb{R}$ und $\Omega_T = (0, T) \times \Omega$. Das Gebiet Ω unterteile sich in zwei[2] Teilgebiete:

$$\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma. \tag{3.13}$$

Dabei sind Ω_1, Ω_2 und Γ paarweise disjunkt, Ω_1 und Ω_2 seien Lipschitz-Gebiete und $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$. Gesucht ist nun eine numerische Lösung des folgenden dimensionslosen Anfangs-Randwertproblems[3] für u und p:

$$\partial_t u - \operatorname{div}(m(u)\nabla p) = q(u) \text{ in } \Omega_T, \tag{3.14}$$

$$p = -\Delta u + w_{,u}(u, x) \text{ in } \Omega_T, \tag{3.15}$$

$$\frac{\partial u}{\partial \nu} = \frac{\partial p}{\partial \nu} = 0 \text{ auf } \partial\Omega \times [0, T], \tag{3.16}$$

$$u(0) = u_0 \text{ auf } \Omega. \tag{3.17}$$

Dabei sei $u_0 \in H^1(\Omega) \cap C^0(\overline{\Omega})$ mit $u_0(x) \geq 0$ für alle $x \in \overline{\Omega}$. Die Randbedingungen (3.16) legen fest, dass kein Massefluss aus dem Gebiet heraus oder in das Gebiet hinein stattfindet. Die Mobilität $m(u)$ sei gegeben durch

$$m(u) = u^n, n \in \mathbb{R}^+. \tag{3.18}$$

Das effektive Grenzflächenpotential $w : \mathbb{R} \times \Omega \to \mathbb{R}$ erfülle eine der Bedingungen:

(w0) $w \in C^0(\mathbb{R}^+ \times \Omega)$ lasse sich in Funktionen $w = w^+ + w^-$ aufspalten, so dass $w^+(\cdot, x) \in C^1(\mathbb{R}^+)$ konvex ist für alle $x \in \Omega$ und $w^-(\cdot, x) \in C^1(\mathbb{R}^+)$ konkav ist für alle $x \in \Omega$. Außerdem gebe es eine Konstante $C \geq 0$, so dass $w(u, x) \geq -C$ für alle $u \in \mathbb{R}^+, x \in \Omega$ gilt.

(w1) w sei gegeben durch

$$w(u, x) = \begin{cases} w_1(u) & \text{falls } x \in \Omega_1 \cup \Gamma, \\ w_2(u) & \text{falls } x \in \Omega_2. \end{cases}$$

Dabei lassen sich die $w_i \in C^0(\mathbb{R}^+)$ in Funktionen $w_i = w_i^+ + w_i^-$ aufspalten, so dass $w_i^+ \in C^1(\mathbb{R}^+)$ konvex ist und $w_i^- \in C^1(\mathbb{R}^+)$ konkav ist. Außerdem gebe es eine Konstante $C \geq 0$, so dass $w_i(u) \geq -C$ für alle $u \in \mathbb{R}^+$ gilt.

[2] Besteht das Substrat aus drei oder mehr verschiedenen Materialien (=Teilgebieten), so entstehen dadurch keine zusätzlichen Schwierigkeiten. Um die Notation übersichtlich zu halten, wird hier aber von nur zwei Teilgebieten ausgegangen.

[3] Im Fall $\beta = 0, n = 3$ entsprechen die Gleichungen (3.14) und (3.15) nach Reskalierung den Gleichungen (2.42) und (2.45).

(w2) w sei gegeben durch

$$w(u,x) := \begin{cases} -a_{11}u^{-l_{11}} + a_{12}u^{-l_{12}} & \text{falls } x \in \Omega_1 \cup \Gamma, \\ -a_{21}u^{-l_{21}} + a_{22}u^{-l_{22}} & \text{falls } x \in \Omega_2. \end{cases}$$

Dabei sei für $i,j \in \{1,2\}$: $a_{i1} \geq 0$, $a_{i2} > 0$ und $l_{ij} \in \mathbb{N}$ mit $l_{i2} > \max\{l_{i1}, 2\}$.

Mit Bedingung (w0) wird also ein stetiges Potential beschrieben. Ein Potential der Art (w1) ist in Punkten $x \in \Gamma$ i.a. nicht stetig. Bedingung (w2) beschreibt einen Spezialfall von (w1) mit

$$w_1^+ = a_{12}u^{-l_{12}}, \quad w_1^- = -a_{11}u^{-l_{11}}, \quad w_2^+ = a_{22}u^{-l_{22}}, \quad w_2^- = -a_{21}u^{-l_{21}}.$$

Das in Kapitel 2.3 hergeleitete kontinuierliche Grenzflächenpotential (2.36) erfüllt die Bedingung (w0), eine durch (2.37) definierte Näherung die Bedingungen (w1) und (w2).

Der Quellterm q auf der rechten Seite von (3.14) erfülle eine der Bedingungen

(q0) $q \equiv 0$.

(q1) $q \in W^{1,\infty}(\mathbb{R})$ mit $q(s) \geq 0 \ \forall s \in \mathbb{R}$.

(q2) $q \in W^{1,\infty}(\mathbb{R})$ mit $q(s) \leq 0 \ \forall s \in \mathbb{R}$.

Ist q durch (2.44) definiert, so erfüllt q im Fall von Kondensation Bedingung (q1) und im Fall von Evaporation Bedingung (q2).

Um eine diskrete Lösung berechnen zu können ist es notwendig, Funktionen W und Q zu definieren, welche im Gegensatz zu den Funktionen w und q auch für negative Argumente definiert sind, da a priori Positivität der diskreten Lösung nicht gegeben ist. Die Näherung W von w muss daher in Analogie zu den Bedingungen (w0)–(w2) eine der folgenden Bedingungen erfüllen:

(W0) $W \in C^0(\mathbb{R}\times\Omega)$ lasse sich in Funktionen $W = W^+ + W^-$ aufspalten, so dass $W^+(\cdot,x) \in C^1(\mathbb{R})$ konvex ist für alle $x \in \Omega$ und $W^-(\cdot,x) \in C^1(\mathbb{R})$ konkav ist für alle $x \in \Omega$. Außerdem gebe es eine Konstante $C \geq 0$, so dass $W(u,x) \geq -C$ für alle $u \in \mathbb{R}, x \in \Omega$ gilt.

(W1) W sei gegeben durch

$$W(u,x) = \begin{cases} W_1(u) & \text{falls } x \in \Omega_1 \cup \Gamma, \\ W_2(u) & \text{falls } x \in \Omega_2. \end{cases}$$

Dabei lassen sich die $W_i \in C^0(\mathbb{R})$ in Funktionen $W_i = W_i^+ + W_i^-$ aufspalten, so dass $W_i^+ \in C^1(\mathbb{R})$ konvex ist und $W_i^- \in C^1(\mathbb{R})$ konkav ist. Außerdem gebe es eine Konstante $C \geq 0$, so dass $W_i(u) \geq -C$ für alle $u \in \mathbb{R}$ gilt.

(W2) w erfülle die Bedingung (w2), es sei $\varepsilon_w > 0$ und W sei gegeben durch

$$W(u,x) = \begin{cases} w(u,x) & \text{falls } u \geq \varepsilon_w, \\ w(\varepsilon_w,x) + (x-\varepsilon_w)w_{,u}(\varepsilon_w,x) & \text{falls } u < \varepsilon_w. \end{cases}$$

Auch hier gilt, dass im Fall (W0) das Potential stetig ist, und im Fall (W1) i.a. nicht stetig ist entlang der Grenze zwischen den Teilmengen Ω_1 und Ω_2. Wiederum ist (W2) ein Spezialfall von (W1). Da einige Aussagen noch stärkere Voraussetzungen an W benötigen, definieren wir noch zusätzlich die stärkeren Bedingungen

(W0') W erfülle (W0) und es sei $W_{,u}^+ \in L^\infty(\mathbb{R}\times\Omega)$ und $W_{,u}^- \in L^\infty(\mathbb{R}\times\Omega)$.

(W1') W erfülle (W1) und es sei $W_{,u}^+ \in L^\infty(\mathbb{R}\times\Omega)$ und $W_{,u}^- \in L^\infty(\mathbb{R}\times\Omega)$.

Erfüllt W die Bedingung (W2), so sind $W_{,u}^+ \in L^\infty(\mathbb{R}\times\Omega)$ und $W_{,u}^- \in L^\infty(\mathbb{R}\times\Omega)$ automatisch gegeben.

Q sei eine Näherung von q derart, dass in Analogie zu den Bedingungen (q0)-(q2) eine der folgenden Bedingungen erfüllt ist:

(Q0) Es sei $Q \equiv 0$.

(Q1) Es sei $Q \in W^{1,\infty}(\mathbb{R})$ mit $\|Q\|_{1,\infty} = C_q$ und $Q(s) \geq 0$ für alle $s \in \mathbb{R}$. Außerdem gebe es ein $\delta_q > 0$ und eine Konstante $c_q > 0$, so dass $Q(s) \geq c_q$ für alle $s \in (-\infty, \delta_q]$ gilt.

(Q2) Es sei $Q \in W^{1,\infty}(\mathbb{R})$ mit $\|Q\|_{1,\infty} = C_q$ und $Q(s) \leq 0$ für alle $s \in \mathbb{R}$. Außerdem gebe es ein $\delta_q > 0$, so dass $Q(s) = 0$ für alle $s \in (-\infty, \delta_q]$ gilt.

Ist q durch (2.44) definiert und erfüllt (q1), ist eine Fortsetzung Q von q, die (Q1) erfüllt, zum Beispiel durch

$$Q(s) = \begin{cases} q(s) & \text{falls } s > 0, \\ q(0) & \text{falls } s \leq 0 \end{cases} \tag{3.19}$$

gegeben. Erfüllt q (q2), dann ist eine Näherung Q, welche (Q2) erfüllt, zum Beispiel gegeben durch

$$Q(s) = q(s)\frac{2}{\pi}\arctan(\frac{s-\delta_q}{\delta_q})\chi_{[s\geq\delta_q]}. \tag{3.20}$$

Das numerische Verfahren soll so definiert sein, dass die diskrete Lösung eine Energie- und eine Entropieabschätzung erfüllt. Um eine diskrete Entropieabschätzung zu ermöglichen ist es notwendig, die Mobilität $m(u)$ so zu diskretisieren, dass zur diskreten Mobilität M eine Entropie G existiert, welche die Identität (3.4) im Diskreten widerspiegelt. Diese Überlegungen motivieren die folgende, von Grün und Rumpf [21] vorgenommene Definition:

Definition 3.2.1 *(zulässiges Entropie-Mobilitäts-Paar)*

Sei $\mathcal{T}_h$ eine zulässige und rechtwinklige Triangulierung, $A > 0$ und $m_\sigma : \mathbb{R} \to \mathbb{R}_0^+$ eine Approximation von m wie folgt:

$$m_\sigma(s) := \begin{cases} s^n & \textit{falls } s \geq \sigma \textit{ und } n \geq 1, \\ \sigma^n & \textit{falls } s < \sigma \textit{ und } n \geq 1, \\ s^n & \textit{falls } s \geq 0 \textit{ und } n < 1, \\ \sigma(-s)^n & \textit{falls } s < 0 \textit{ und } n < 1. \end{cases} \tag{3.21}$$

Dann heißt ein Paar von Funktionen $G_\sigma : \mathbb{R} \to \mathbb{R}_0^+$, $M_\sigma : V^h \to \bigotimes_{k=1}^{|\mathcal{T}_h|} \mathbb{R}^{d\times d}$ zulässiges Entropie-Mobilitäts-Paar *zur Triangulierung $\mathcal{T}_h$, falls die folgenden Axiome erfüllt sind:*

(M1) $M_\sigma : V^h \to \bigotimes_{k=1}^{|\mathcal{T}_h|} \mathbb{R}^{d\times d}$ *ist stetig.*

(M2) Für alle $E \in \mathcal{T}_h$ *ist* $M_\sigma(U)|_E = m_\sigma(u)\,\mathrm{Id}$*, falls* $U|_E$ *konstant ist.*

(M3) $M_\sigma(U)\nabla \mathcal{I}_h G'_\sigma(U) = \nabla U$ *mit* $G_\sigma(s) = \int_A^s g_\sigma(r)dr$, $g_\sigma(s) = \int_A^s m_\sigma(r)^{-1}dr$.

(M4) $M_\sigma(U)|_E$ *ist symmetrisch positiv semidefinit auf jedem Element* $E \in \mathcal{T}_h$.

Die reellwertige Funktion $m(u)$ wird also durch eine matrixwertige Funktion $M_\sigma(U)$ ersetzt. Dabei ist $M_\sigma(U)$ auf jedem Element der Triangulierung konstant.
Im Fall $d = 1$ erfüllt die Definition

$$M_\sigma(U)|_{[\mathfrak{x}_i,\mathfrak{x}_{i+1}]} = \rho_\sigma(U(\mathfrak{x}_i), U(\mathfrak{x}_{i+1})) \text{ für ein Element } [\mathfrak{x}_i, \mathfrak{x}_{i+1}] \in \mathcal{T}_h \tag{3.22}$$

die obigen Bedingungen. Dabei ist ρ_σ die Funktion

$$\rho_\sigma(a, b) := \left(\fint_a^b \frac{ds}{m_\sigma(s)} \right)^{-1}. \tag{3.23}$$

Im Fall $d = 2$ gilt für das Referenzdreieck $\hat{E} = (0, e_1, e_2)$, dass

$$\hat{M} := \begin{pmatrix} \rho_\sigma(\hat{U}(0), \hat{U}(e_1)) & 0 \\ 0 & \rho_\sigma(\hat{U}(0), \hat{U}(e_2)) \end{pmatrix} \tag{3.24}$$

die obigen Axiome erfüllt. e_1, e_2 ist hier die kanonische Basis des $\mathbb{R}^2$. Sei nun $E \in \mathcal{T}_h$ ein beliebiges, rechtwinkliges Dreieck. Dann gibt es eine affine Abbildung $f : \hat{E} \to E$, $f(\hat{x}) = A\hat{x} + x_0(E)$. Die Mobilitätsmatrix auf E ist nun gegeben durch

$$M = A^{-T}\hat{M}A^T. \tag{3.25}$$

Ein Entropie-konsistentes Finite-Elemente-Verfahren lautet nun wie folgt:

Schema 3.2.2 *(Finite-Elemente-Verfahren)*

Sei Ω *polygonal berandet und mit* $\mathcal{T}_h$ *eine zulässige und rechtwinklige Triangulierung von* Ω *gegeben. Durch* M_σ *und* G_σ *sei ein zulässiges Entropie-Mobilitäts-Paar zur Triangulierung* $\mathcal{T}_h$ *gegeben. Durch* $0 = \mathfrak{t}_0 < \mathfrak{t}_1 < \cdots < \mathfrak{t}_K = T$ *sei eine diskrete Folge von Zeitpunkten gegeben, dabei sei* $c_\tau\tau \le \tau_k \le \tau$ *mit einer Konstanten* $c_\tau > 0$.

Sei $U^0 = \mathcal{I}_h u_0$. *Zu einer bereits berechneten Lösung* $U^k \in V^h$ *zum Zeitpunkt* $\mathfrak{t}_k$ *bestimme eine Lösung* $U^{k+1}, P^{k+1} \in V^h$ *zum nächsten Zeitpunkt* $\mathfrak{t}_{k+1}$*, welche*

$$(U^{k+1} - U^k, \Theta)_h + \tau_{k+1}(M_\sigma(U^{k+1})\nabla P^{k+1}, \nabla\Theta) = \tau_{k+1}(Q(U^{k+1}), \Theta)_h, \tag{3.26}$$

$$(P^{k+1}, \Psi)_h = (\nabla U^{k+1}, \nabla\Psi) + (W^+_{,u}(U^{k+1}, \cdot), \Psi)_h + (W^-_{,u}(U^k, \cdot), \Psi)_h \tag{3.27}$$

für alle $\Theta, \Psi \in V^h$ *erfüllt.*

Da die diskrete Funktion $U^k \in V^h$ als Vektor im $\mathbb{R}^D$ dargestellt wird, lässt sich mit Hilfe der $D \times D$-Matrizen

$$\begin{aligned} M_h &:= ((\varphi_i, \varphi_j)_h)_{i,j=1,\ldots,D}\,, \\ L_h &:= \left(\int_\Omega \nabla\varphi_i \nabla\varphi_j\right)_{i,j=1,\ldots,D}, \\ L_h^M(U) &:= \left(\int_\Omega M_\sigma(U)\nabla\varphi_i\nabla\varphi_j\right)_{i,j=1,\ldots,D} \end{aligned}$$

das Verfahren auch als Problem im $\mathbb{R}^D$ schreiben:

Zu einer bereits berechneten Lösung $U^k \in \mathbb{R}^D$ zum Zeitpunkt $\mathfrak{t}_k$ bestimme eine Lösung $U^{k+1} \in \mathbb{R}^D$ zum nächsten Zeitpunkt $\mathfrak{t}_{k+1}$, so dass gilt:

$$\begin{aligned} U^{k+1} - U^k + \tau_{k+1} M_h^{-1} L_h^M(U^{k+1})\big[M_h^{-1} L_h U^{k+1} \\ + \mathcal{I}_h W_{,u}^+(U^{k+1}) + \mathcal{I}_h W_{,u}^-(U^k)\big] = \tau_{k+1}\mathcal{I}_h Q(U^{k+1}). \end{aligned} \quad (3.28)$$

Dabei bezeichnet $\mathcal{I}_h W_{,u}^+(U^{k+1})$ den die Funktion $\mathcal{I}_h W_{,u}^+(U^{k+1}(x), x) \in V^h$ darstellenden Koeffizientenvektor

$$\left(W_{,u}^+(U^{k+1}(\mathfrak{x}_1), \mathfrak{x}_1), \ldots, W_{,u}^+(U^{k+1}(\mathfrak{x}_D), \mathfrak{x}_D)\right).$$

Analog sind $\mathcal{I}_h W_{,u}^-(U^k)$ und $\mathcal{I}_h Q(U^{k+1})$ definiert. Die Matrizen M_h, L_h^M und L_h haben die folgenden Eigenschaften:

i) M_h ist positiv definit und hat Diagonalgestalt.

ii) $L_h^M(U)$ ist positiv semidefinit für alle $U \in \mathbb{R}^D$.

iii) L_h ist positiv semidefinit.

3.3 Existenz der diskreten Lösung

Ziel dieses Abschnittes ist es, die Existenz von diskreten Funktionen $U \in S^{-1,0}(V^h)$ und $P \in S^{-1,0}(V^h)$ zu zeigen, welche die in Schema 3.2.2 aufgestellten Gleichungen lösen. Die folgenden Sätze geben hinreichende Bedingungen für die Existenz von Lösungen an.

Satz 3.3.1 *(Existenz einer diskreten Lösung für $Q \equiv 0$)*

Es gelte (Q0) und das effektive Grenzflächenpotential erfülle eine der Bedingungen (W0) oder (W1). Dann gibt es diskrete Funktionen $U \in S^{-1,0}(V^h)$ und $P \in S^{-1,0}(V^h)$, so dass die Funktionen $U^k := U(\mathfrak{t}_k), P^k := P(\mathfrak{t}_k)$, $k = 0, \ldots, K$ Schema 3.2.2 lösen.

Beweis : Wir definieren den Vektorraum V der diskreten Funktionen mit Mittelwert 0 und die zugehörige Norm $\|.\|_V$ durch

$$\begin{aligned} V &:= \{Z \in \mathbb{R}^D | \mu(Z) = 0\}, \\ \|Z\|_V &:= \sqrt{\langle Z\,, L_h Z\rangle}, \end{aligned}$$

wobei μ durch

$$\mu(Z) := \fint_\Omega Z = \frac{\langle \mathbb{1}\,, M_h Z\rangle}{\langle \mathbb{1}\,, M_h \mathbb{1}\rangle}$$

definiert ist. $\mathbb{1}$ bezeichnet dabei den Vektor $(1,\dots,1)^T \in \mathbb{R}^D$. Wir setzen nun $Z^k = U^k - \mu^0, \mu^0 = \mu(U^0)$ und betrachten das modifizierte Problem:

$$\begin{aligned}(Z^{k+1} - Z^k, \Theta)_h + \tau_{k+1}(M_\sigma(Z^{k+1}+\mu^0)\nabla P^{k+1}, \nabla\Theta) = 0 \quad \forall \Theta \in V^h,\\ (P^{k+1},\Psi)_h = (\nabla Z^{k+1}, \nabla\Psi) + (W^+_{,u}(Z^{k+1}+\mu^0,\cdot),\Psi)_h + (W^-_{,u}(Z^k+\mu^0,\cdot),\Psi)_h \quad \forall \Psi \in V^h.\end{aligned}$$

Falls also mit $Z^k \in V$ eine Lösung für den k-ten Zeitschritt gegeben ist, so bestimmt man eine Lösung $Z^{k+1} \in V$ für den nächsten Zeitschritt wie folgt:

Suche zu gegebenen $Z^k \in V$ eine Lösung $Z^{k+1} \in V$ von:

$$\begin{aligned}F(Z^{k+1}) = Z^{k+1} - Z^k + \tau_{k+1} M_h^{-1} L_h^M (Z^{k+1}+\mu^0)\big[M_h^{-1} L_h Z^{k+1}\\ + \mathcal{I}_h W^+_{,u}(Z^{k+1}+\mu^0) + \mathcal{I}_h W^-_{,u}(Z^k+\mu^0)\big] = 0.\end{aligned}$$

Um die Existenz eines solchen Z^{k+1} zu zeigen, wenden wir den Brouwerschen Fixpunktsatz an. Wir betrachten die beschränkte und abgeschlossene Menge

$$V_R := \{Z \in V : \|Z\|_V \le R\}$$

und definieren eine Funktion $G : V_R \to V_R$ durch

$$G(Z) = \frac{-RF(Z)}{\|F(Z)\|_V}.$$

Wir nehmen nun an, dass $\|F(Z)\|_V \neq 0$ für alle $Z \in V_R$ gilt und führen diese Annahme zum Widerspruch. Falls diese Annahme gilt, ist G stetig und es gibt einen Fixpunkt $X \in V_R$ mit $X = G(X)$ und $\|X\|_V = R$. Nun wählen wir $Y \in V$ so, dass

$$\begin{aligned}M_h^{-1} L_h Y = \mathcal{I}_h W^+_{,u}(X+\mu^0) - \mu(\mathcal{I}_h W^+_{,u}(X+\mu^0))\mathbb{1}\\ + \mathcal{I}_h W^-_{,u}(Z^k+\mu^0) - \mu(\mathcal{I}_h W^-_{,u}(Z^k+\mu^0))\mathbb{1} \quad (3.29)\end{aligned}$$

gilt. Dies ist möglich, da $M_h^{-1} L_h : V \to V$ invertierbar ist.

Wenn R groß genug gewählt wird, können wir zeigen:

i) $\langle X\,, L_h(X+Y)\rangle > 0.$

ii) $\langle F(X)\,, L_h(X+Y)\rangle > 0.$

Nachdem diese beiden Aussagen bewiesen sind, ergibt sich der Widerspruch

$$0 < \langle X\,, L_h(X+Y)\rangle = \langle G(X)\,, L_h(X+Y)\rangle = -\frac{R}{\|F(X)\|_V}\langle F(X)\,, L_h(X+Y)\rangle < 0,$$

und der Satz ist bewiesen.

Bevor wir nun die beiden Aussagen beweisen, stellen wir zunächst einmal fest, dass für eine beliebige Funktion $\phi \in V$ gilt:

$$\langle \phi \, , L_h Y \rangle = \langle \phi \, , M_h \mathcal{I}_h W^+_{,u}(X + \mu^0) + M_h \mathcal{I}_h W^-_{,u}(Z^k + \mu^0) \rangle, \tag{3.30}$$

denn die restlichen Terme verschwinden wegen $\langle \phi \, , M_h \mathbb{1} \rangle = \langle M_h \phi \, , \mathbb{1} \rangle = 0$.

Beweis zu i): Mit Hilfe von (3.30) gilt:

$$\begin{aligned} \langle X \, , L_h(X+Y) \rangle &= \langle X \, , L_h X \rangle + \langle X \, , L_h Y \rangle \\ &= R^2 + \langle X \, , M_h \mathcal{I}_h W^+_{,u}(X + \mu^0) \rangle + \langle X \, , M_h \mathcal{I}_h W^-_{,u}(Z^k + \mu^0) \rangle. \end{aligned}$$

Wir können mit Hilfe der Youngschen Ungleichung $\langle a \, , b \rangle \geq -\varepsilon \langle a \, , a \rangle - \frac{1}{4\varepsilon} \langle b \, , b \rangle$ und der aus der in endlichdimensionalen Vektorräumen gültigen Äquivalenz von Normen ableitbaren Abschätzung $\|X\| \leq c \|X\|_V$ abschätzen:

$$\langle X \, , M_h \mathcal{I}_h W^-_{,u}(Z^k + \mu^0) \rangle \geq -\varepsilon c R^2 - C.$$

Der andere Term lässt sich wie folgt umformen:

$$\begin{aligned} &\langle X \, , M_h \mathcal{I}_h W^+_{,u}(X + \mu^0) \rangle \\ &\quad = \langle X + \mu^0 \mathbb{1} - \mu^0 \mathbb{1} \, , M_h \mathcal{I}_h W^+_{,u}(X + \mu^0) - M_h \mathcal{I}_h W^+_{,u}(\mu^0) \rangle + \langle X \, , M_h \mathcal{I}_h W^+_{,u}(\mu^0) \rangle \\ &\quad =: a_1 + a_2. \end{aligned}$$

Nun gilt wegen der Konvexität von W^+:

$$a_1 = \sum_{i=1}^{D} (X_i + \mu^0 - \mu^0)(M_h)_{ii} \left(W^+_{,u}(X_i + \mu^0, \mathfrak{x}_i) - W^+_{,u}(\mu^0, \mathfrak{x}_i) \right) \geq 0.$$

Eine Abschätzung von a_2 erhalten wir wiederum mit Hilfe der Youngschen Ungleichung und der Äquivalenz der Normen $\|.\|_V$ und $\|.\|$:

$$a_2 \geq -\varepsilon \|X\|^2 - \frac{1}{4\varepsilon} \|M_h \mathcal{I}_h W^+_{,u}(\mu^0)\|^2 \geq -\varepsilon c R^2 - C.$$

Damit folgt insgesamt:

$$\langle X \, , L_h(X+Y) \rangle \geq R^2(1 - 2\varepsilon c) - C.$$

Wenn ε klein genug und R groß genug gewählt wird, ist dies positiv und damit ist i) bewiesen.

Beweis zu ii): Sei

$$\Phi := M_h^{-1} L_h X + \mathcal{I}_h W^+_{,u}(X + \mu^0) + \mathcal{I}_h W^-_{,u}(Z^k + \mu^0). \tag{3.31}$$

Dann gilt:

$$\begin{aligned} &\langle F(X) \, , L_h(X+Y) \rangle \\ &\quad = \langle X - Z^k \, , L_h X \rangle + \langle X - Z^k \, , L_h Y \rangle + \tau_{k+1} \langle M_h^{-1} L_h^M (X + \mu^0) \Phi \, , L_h(X+Y) \rangle \\ &\quad =: (a) + (b) + (c). \end{aligned}$$

Term (a) lässt sich mit Hilfe der Youngschen Ungleichung abschätzen:

$$(a) = \|X\|_V^2 - \langle Z^k \,, L_h X\rangle \geq \|X\|_V^2 - \varepsilon\|X\|_V^2 - \frac{1}{4\varepsilon}\|Z^k\|_V^2 \geq (1-\varepsilon)R^2 - C.$$

Bei der Abschätzung von Term (b) nutzen wir zunächst die Beziehung (3.30) aus und anschließend die Konvexität/Konkavität der Funktionen W^+ bzw W^-:

$$\begin{aligned}
(b) &= \langle X - Z^k \,, M_h \mathcal{I}_h W^+_{,u}(X+\mu^0) + M_h \mathcal{I}_h W^-_{,u}(Z^k+\mu^0)\rangle \\
&= \sum_{i=1}^{D} (X_i - Z_i^k)(M_h)_{ii} \left(W^+_{,u}(X_i+\mu^0, \mathfrak{x}_i) + W^-_{,u}(Z_i^k+\mu^0, \mathfrak{x}_i)\right) \\
&= \sum_{i=1}^{D} (M_h)_{ii} \left((X_i+\mu^0) - (Z_i^k+\mu^0)\right) W^+_{,u}(X_i+\mu^0, \mathfrak{x}_i) \\
&\quad + \sum_{i=1}^{D} (M_h)_{ii} \left((X_i+\mu^0) - (Z_i^k+\mu^0)\right) W^-_{,u}(Z_i^k+\mu^0, \mathfrak{x}_i) \\
&\geq \sum_{i=1}^{D} (M_h)_{ii} \left(W^+(X_i+\mu^0, \mathfrak{x}_i) - W^+(Z_i^k+\mu^0, \mathfrak{x}_i)\right) \\
&\quad + \sum_{i=1}^{D} (M_h)_{ii} \left(W^-(X_i+\mu^0, \mathfrak{x}_i) - W^-(Z_i^k+\mu^0, \mathfrak{x}_i)\right) \\
&= \langle \mathbb{1} \,, M_h \mathcal{I}_h W(X+\mu^0) - M_h \mathcal{I}_h W(Z^k+\mu^0)\rangle \\
&= \int_\Omega \mathcal{I}_h W(X+\mu^0, \cdot) - \int_\Omega \mathcal{I}_h W(Z^k+\mu^0, \cdot) \\
&\geq -C,
\end{aligned}$$

wobei die letzte Ungleichung aus der Tatsache folgt, dass das effektive Grenzflächenpotential nach unten beschränkt ist.

Da $\langle M_h M_h^{-1} L_h^M(X+\mu^0)\Phi \,, \mathbb{1}\rangle = 0$ für jedes beliebige $\Phi \in V^h$ gilt, ist $M_h^{-1} L_h^M(X+\mu^0)\Phi \in V$. Also können wir bei der Abschätzung von (c) Gleichung (3.30) ausnutzen und erhalten:

$$\begin{aligned}
(c) &= \tau_{k+1}\langle M_h^{-1} L_h^M(X+\mu^0)\Phi \,, L_h X\rangle \\
&\quad + \tau_{k+1}\langle M_h^{-1} L_h^M(X+\mu^0)\Phi \,, M_h \mathcal{I}_h W^+_{,u}(X+\mu^0) + M_h \mathcal{I}_h W^-_{,u}(Z^k+\mu^0)\rangle \\
&= \tau_{k+1}\langle L_h^M(X+\mu^0)\Phi \,, \Phi\rangle.
\end{aligned}$$

Da L_h^M positiv semidefinit ist, ist dieser Term nichtnegativ. Damit gilt insgesamt, wenn ε klein genug und R groß genug gewählt wird,

$$\langle F(X) \,, L_h(X+Y)\rangle = R^2(1-\varepsilon) - C > 0,$$

und ii) ist bewiesen. □

Im Fall $Q \not\equiv 0$ reicht die Bedingung (W0) bzw. (W1) nicht aus, um die Existenz einer diskreten Lösung zu zeigen. Es werden vielmehr die stärkeren Voraussetzungen (W2) benötigt, da der Beweis der folgenden Sätze im Gegensatz zum Beweis von Satz 3.3.1 ausnutzt, dass Vorzeichen und Wachstumsverhalten der Terme W^+ und W^- bekannt sind. Im Fall von Evaporation existiert eine diskrete Lösung dabei nicht notwendigerweise bis zur Zeit T, sondern nur solange noch Masse vorhanden ist.

Satz 3.3.2 *(Existenz einer diskreten Lösung bei Kondensation)*

Es gelte (W2) und Q erfülle die Bedingung (Q1). Dann gibt es diskrete Funktionen $U \in S^{-1,0}(V^h)$ und $P \in S^{-1,0}(V^h)$, so dass die Funktionen $U^k := U(\mathfrak{t}_k), P^k := P(\mathfrak{t}_k)$, $k = 0, \ldots, K$ Schema 3.2.2 lösen.

Satz 3.3.3 *(Existenz einer diskreten Lösung bei Evaporation)*

Es gelte (W2) und Q erfülle die Bedingung (Q2). Dann existiert zu einer bereits berechneten Lösung U^k an einem Zeitpunkt $\mathfrak{t}_k$ eine Lösung U^{k+1} am Zeitpunkt $\mathfrak{t}_{k+1}$, welche die Gleichungen (3.26) und (3.27) erfüllt, falls für die Zeitschrittweite τ_{k+1} gilt:

$$\tau_{k+1} \leq \frac{1}{C_q} \fint_\Omega U^k. \tag{3.32}$$

Beweis (Satz 3.3.2 und Satz 3.3.3): In jedem Zeitschritt ist also zu einem gegebenen Vektor $U^k \in V^h$ eine Lösung $U^{k+1} \in V^h$ gesucht, welche (3.28) erfüllt. Nun definieren wir in Analogie zum vorherigen Beweis eine Funktion $F : V^h \to V^h$ durch

$$\begin{aligned} F(U) := U - U^k - \tau_{k+1}\mathcal{I}_h Q(U) \\ + \tau_{k+1} M_h^{-1} L_h^M(U) \left(M_h^{-1} L_h U + \mathcal{I}_h W_{,u}^+(U) + \mathcal{I}_h W_{,u}^-(U^k) \right) \end{aligned}$$

und eine Funktion $G : \{U \in V^h : \|U\|_\sim \leq R\} \to \{U \in V^h : \|U\|_\sim \leq R\}$ durch

$$G(U) := \frac{-RF(U)}{\|F(U)\|_\sim}.$$

Die Norm $\|.\|_\sim$ und das zugehörige Skalarprodukt $\langle .\,, .\rangle_\sim$ auf V^h sind gegeben durch

$$\begin{aligned} \|U\|_\sim &:= \sqrt{\langle U\,, U\rangle_\sim}, \\ \langle U\,, V\rangle_\sim &:= \langle U\,, L_h V\rangle + \langle \mathbb{1}\,, M_h U\rangle \langle \mathbb{1}\,, M_h V\rangle. \end{aligned}$$

Unter der Widerspruchs-Annahme, dass $F(U) \neq 0$ für alle $U \in V^h$ mit $\|U\|_\sim \leq R$ gilt, gibt es nach dem Brouwerschen Fixpunktsatz einen Fixpunkt X mit $X = G(X)$ und $\|X\|_\sim = R$. Sei nun $Y \in V^h$ so gewählt, dass $\langle \mathbb{1}\,, M_h Y\rangle = 0$ ist und

$$M_h^{-1} L_h Y = \mathcal{I}_h W_{,u}^+(X) + \mathcal{I}_h W_{,u}^-(U^k) - \mu(\mathcal{I}_h W_{,u}^+(X))\mathbb{1} - \mu(\mathcal{I}_h W_{,u}^-(U^k))\mathbb{1}. \tag{3.33}$$

Y ist dadurch eindeutig bestimmt. Wenn R groß genug gewählt wird, können wir zeigen:

i) $\langle X\,, X + Y\rangle_\sim > 0$,

ii) $\langle F(X)\,, X + Y\rangle_\sim > 0$,

und führen damit die Annahme zum Widerspruch.

Dazu stellen wir zunächst einmal fest, dass für den Mittelwert von $\mu(X) := \fint_\Omega X$ gilt:

$$\begin{aligned} \mu(X) = \mu(G(X)) = \mu\left(\frac{-RF(X)}{\|F(X)\|_\sim}\right) &= \frac{-R}{\|F(X)\|_\sim}\mu(F(X)) \\ &= \frac{-R}{\|F(X)\|_\sim}\left(\mu(X) - \mu(U^k) - \tau_{k+1}\mu(\mathcal{I}_h Q(X))\right). \end{aligned}$$

Also gilt

$$\mu(X) = \frac{1}{1 + \frac{\|F(X)\|_\sim}{R}} \left(\mu(U^k) + \tau_{k+1} \mu(\mathcal{I}_h Q(X)) \right),$$

woraus $|\mu(X)| \leq C$ folgt, da Q beschränkt ist. Weiterhin gilt $\mu(X) > 0$, da im Fall (Q1) $\mu(U^k)$ und $\mathcal{I}_h Q(X)$ positiv sind und im Fall (Q2) $\tau_{k+1}|\mu(\mathcal{I}_h Q(X))| \leq |\mu(U^k)|$ ist aufgrund der Bedingung (3.32).

Beweis zu i):

$$\begin{aligned}
\langle X\,, X+Y \rangle_\sim &= R^2 + \langle X\,, Y \rangle_\sim \\
&= R^2 + \langle X\,, L_h Y \rangle_\sim \\
&= R^2 + \langle M_h \mathcal{I}_h W^+_{,u}(X)\,, X - \mu(X)\mathbb{1} \rangle + \langle M_h \mathcal{I}_h W^-_{,u}(U^k)\,, X - \mu(X)\mathbb{1} \rangle
\end{aligned}$$

Mit Hilfe der Youngschen Ungleichung, der Beschränktheit von $\mu(X)$ und der Äquivalenz $\|X\| \leq c\|X\|_\sim$ der Normen $\|\cdot\|_\sim$ und $\|\cdot\|$ lässt sich abschätzen:

$$\langle M_h \mathcal{I}_h W^-_{,u}(U^k)\,, X - \mu(X)\mathbb{1} \rangle \geq -\tfrac{1}{4\varepsilon} C - \varepsilon c R^2 - C.$$

Unter Ausnutzung der Konvexität von W^+ und der Positivität von $\mu(X)$ und W^+ gilt:

$$\begin{aligned}
\langle M_h \mathcal{I}_h W^+_{,u}(X)\,, X - \mu(X)\mathbb{1} \rangle &\geq \int_\Omega \mathcal{I}_h W^+(X,\cdot) - \mathcal{I}_h W^+(\mu(X),\cdot) \\
&\geq \int_\Omega \mathcal{I}_h W^+(X,\cdot) - \mathcal{I}_h W^+(0,\cdot) \\
&\geq 0 - C.
\end{aligned}$$

Damit ergibt sich insgesamt, falls erst ε klein genug und dann R groß genug gewählt wird:

$$\langle X\,, X+Y \rangle_\sim \geq R^2(1 - \varepsilon c) - C > 0.$$

Beweis zu ii):

$$\begin{aligned}
\langle F(X)\,, X+Y \rangle_\sim &= \langle X - U^k - \tau_{k+1} \mathcal{I}_h Q(X)\,, X \rangle_\sim \\
&\quad + \langle X - U^k - \tau_{k+1} \mathcal{I}_h Q(X)\,, Y \rangle_\sim \\
&\quad + \tau_{k+1} \langle M_h^{-1} L_h^M(X)\Phi\,, X+Y \rangle_\sim \\
&=: (a) + (b) + (c)
\end{aligned}$$

Hierin ist Φ wie in (3.31) definiert. Da L_h^M positiv semidefinit ist, gilt

$$(c) = \tau_{k+1} \langle L_h^M(X)\Phi\,, \Phi \rangle \geq 0.$$

Term (a) lässt sich mit der Youngschen Ungleichung, der Beschränktheit von $Q(X)$ und der Aquivalenz der Normen $\|\cdot\|_\sim$ und $\|\cdot\|$ abschätzen. Wählt man ε in der Youngschen Ungleichung klein genug, so gilt:

$$\begin{aligned}
(a) &\geq R^2(1 - \varepsilon c) - C - \tau_{k+1} \langle \mathcal{I}_h Q(X)\,, X \rangle_\sim \\
&\geq R^2(1 - \varepsilon c) - C - \tau_{k+1} \varepsilon R^2 - \tau_{k+1} C \\
&\geq \frac{1}{2} R^2 - C.
\end{aligned}$$

Term (b) schließlich lässt sich umformen zu

$$\begin{aligned}(b) &= \langle M_h\mathcal{I}_hW^+_{,u}(X) + M_h\mathcal{I}_hW^-_{,u}(U^k)\,, X - U^k - \mu(F(X))\mathbb{1}\rangle \\ &\quad -\tau_{k+1}\langle M_h\mathcal{I}_hW^+_{,u}(X) + M_h\mathcal{I}_hW^-_{,u}(U^k)\,, \mathcal{I}_hQ(X)\rangle \\ &=: (b1) + (b2).\end{aligned}$$

Im Fall (Q1) ist $\mathcal{I}_hQ(X)$ positiv, $\mathcal{I}_hW^+_{,u}(X,\cdot)$ ist negativ. Daher gilt:

$$-\tau_{k+1}\langle M_h\mathcal{I}_hW^+_{,u}(X)\,, \mathcal{I}_hQ(X)\rangle \geq 0.$$

Ferner ist $\mathcal{I}_hW^-_{,u}(U^k)$ konstant und $\mathcal{I}_hQ(X)$ beschränkt. Also ist insgesamt

$$(b2) \geq 0 - C.$$

Im Fall (Q2) gilt für alle $1 \leq i \leq D$, dass entweder $Q(X_i) = 0$ oder $|W^+_{,u}(X_i, \mathfrak{x}_i)| \leq \max_{j\in\{1,2\}} |W^+_{j,u}(\delta_q)|$ ist. Also gilt auch hier:

$$(b2) \geq -C.$$

Der Term (b1) lässt sich, analog zum Vorgehen im Fall $Q \equiv 0$, durch

$$\begin{aligned}(b1) \geq \int_\Omega \mathcal{I}_hW^+(X,\cdot) - \int_\Omega \mathcal{I}_hW^+(U^k + \mu(F(X)),\cdot) \\ + \int_\Omega \mathcal{I}_hW^-(X - \mu(F(X)),\cdot) - \int_\Omega \mathcal{I}_hW^-(U^k,\cdot)\end{aligned}$$

abschätzen. Wir wissen, dass $\mu(F(X)) < 0$ ist, also gilt wegen der Monotonie von W^- punktweise:

$$\mathcal{I}_hW^-(X - \mu(F(X)),\cdot) \geq \mathcal{I}_hW^-(X,\cdot).$$

Ebenso gilt punktweise, da $\mu(X) > 0$ und W^+ monoton fallend ist:

$$\begin{aligned}&\mathcal{I}_hW^+(U^k + \mu(X) - \mu(U^k) - \tau_{k+1}\mu(\mathcal{I}_hQ(X)),\cdot) \\ &\qquad \leq \mathcal{I}_hW^+(U^k - \mu(U^k) - \tau_{k+1}\mu(\mathcal{I}_hQ(X)),\cdot) \leq \mathcal{I}_hW^+(U^k - \mu(U^k) - C,\cdot).\end{aligned}$$

Damit lässt sich (b1) abschätzen durch

$$(b1) \geq \int_\Omega \mathcal{I}_hW(X,\cdot) - C.$$

Da W nach unten beschränkt ist, gilt also insgesamt:

$$\langle F(X)\,, X + Y\rangle_\sim \geq \frac{1}{2}R^2 - C.$$

Wenn R groß genug gewählt wird, ist dies positiv und damit ist der Widerspruch gezeigt. □

Korollar 3.3.4 *(Existenzintervall bei Evaporation)*
Es gelte (Q2) und W erfülle (W2). Für

$$T < \frac{1}{C_q} \fint_\Omega U^0 \tag{3.34}$$

existiert eine Lösung von Schema 3.2.2 auf dem ganzen Intervall $[0, T]$.

Beweis : Wir setzen $\Theta \equiv 1$ in (3.26) und erhalten

$$\int_\Omega U^{k+1} - \int_\Omega U^k = \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}).$$

Nach Summation über k ergibt sich daraus für die Masse der diskreten Lösung die Gleichung:

$$\int_\Omega U^i = \int_\Omega U^0 + \sum_{k=1}^{i} \tau_k \int_\Omega \mathcal{I}_h Q(U^k).$$

Mit Hilfe der Bedingung (Q2) lässt sich abschätzen:

$$\left| \sum_{k=1}^{i} \tau_k \int_\Omega \mathcal{I}_h Q(U^k) \right| \leq T C_q |\Omega|.$$

Nun wählen wir Zeitschrittweiten τ_k, welche kleiner sind als die maximale erlaubte Zeitschrittweite $\tau := \frac{1}{C_q} \fint_\Omega U^0 - T$. Wegen (3.34) ist τ positiv. Daraus folgt:

$$\frac{1}{C_q} \fint_\Omega U^i \geq \frac{1}{C_q |\Omega|} \left(\int_\Omega U^0 - T C_q |\Omega| \right) = \tau,$$

und damit ist die Bedingung (3.32) für alle Zeitschritte $\mathfrak{t}_i$ erfüllt. □

Kapitel 4

A priori Abschätzungen

Im letzten Kapitel wurde gezeigt, dass eine diskrete Lösung $U \in S^{-1,0}(V^h)$ des in 3.2.2 definierten Verfahrens existiert. Im Falle von Evaporation existiert diese allerdings nicht für jedes beliebig große Zeitintervall, sondern nur für $T < \frac{1}{C_q} \fint_\Omega U^0$. In diesem Kapitel wird gezeigt, dass diese Lösung verschiedene a priori Abschätzungen erfüllt.

4.1 Die diskrete Energieabschätzung

Die Energieabschätzung schätzt die Energie $\frac{1}{2}\int_\Omega |\nabla U(T)|^2 + \mathcal{I}_h W(U(T), \cdot)$ zum Zeitpunkt T gegen die Energie zum Zeitpunkt 0 ab. Kombiniert mit einer Abschätzung für die Masse, kann dieses Resultat genutzt werden, um gleichmäßige Beschränktheit der $L^\infty(0,T;H^1(\Omega))$-Norm von U zu folgern. Daher beginnt dieser Abschnitt mit einer Abschätzung für die Masse von U.

Satz 4.1.1 *(Beschränkung der Masse)*
Die Lösung U von Schema 3.2.2 erfüllt die Ungleichungen

$$\begin{aligned}
\int_\Omega U^0 = \int_\Omega U(T) \qquad & \text{falls (Q0) gilt,} \\
\int_\Omega U^0 \le \int_\Omega U(T) \le \int_\Omega U^0 + C_q|\Omega_T| \qquad & \text{falls (Q1) gilt,} \\
0 \le \int_\Omega U(T) \le \int_\Omega U^0 \qquad & \text{falls (Q2) und } T < \frac{1}{C_q} \fint_\Omega U^0 \text{ gilt.}
\end{aligned}$$

Beweis : Wir setzen $\Theta \equiv 1$ in (3.26) und erhalten

$$\int_\Omega U^{k+1} - \int_\Omega U^k = \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}). \tag{4.1}$$

Im Fall $Q \equiv 0$ folgt daraus nach Summation über k sofort die Behauptung.

Im Fall (Q1) gilt nach Summation über k:

$$\int_\Omega U(T) - \int_\Omega U^0 = \sum_{k=0}^{K-1} \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}).$$

Da sich die rechte Seite dieser Gleichung durch

$$0 \leq \sum_{k=0}^{K-1} \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}) \leq |\Omega_T| C_q$$

abschätzen läßt, folgt die behauptete Ungleichung im Fall (Q1).

Im Fall (Q2) ist

$$\sum_{k=0}^{K-1} \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}) \leq 0,$$

woraus unmittelbar $\int_\Omega U(T) \leq \int_\Omega U^0$ folgt. Die Bedingung (3.32) garantiert die Positivität des Mittelwertes der Lösung, denn sie impliziert

$$\left| \tau_{k+1} \int_\Omega \mathcal{I}_h Q(U^{k+1}) \right| \leq \left| \int_\Omega U^k \right|.$$

Kombiniert mit (4.1) folgt daraus offensichtlich: $\int_\Omega U^{k+1} \geq 0$ für alle k. □

Korollar 4.1.2 *(L^2-Abschätzung)*

Die Lösung U des Verfahrens 3.2.2 erfüllt die Ungleichung

$$\|U(T)\|_{L^2(\Omega)} \leq c_p \|\nabla U(T)\|_{L^2(\Omega)} + \left| \fint_\Omega U(T) \right| |\Omega|^{\frac{1}{2}}. \tag{4.2}$$

Beweis : Wir betrachten die Funktion $U(T,\cdot) - \fint_\Omega U(T,x)dx \in H^1(\Omega)$. Da diese Funktion Mittelwert 0 hat, gilt die Poincaré-Ungleichung

$$\|U(T,\cdot) - \fint_\Omega U(T,x)dx\|_{L^2(\Omega)} \leq c_p \|\nabla U(T)\|_{L^2(\Omega)}.$$

Mit Hilfe der Dreiecksungleichung folgt daraus die Behauptung. □

Die L^2-Norm von $U(T)$ kann also abgeschätzt werden, wenn eine Abschätzung für die L^2-Norm von $\nabla U(T)$ bekannt ist. Diese liefern die folgenden Energieabschätzungen.

Satz 4.1.3 *(Diskrete Energieabschätzung für $Q \equiv 0$)*

Es gelte (Q0) und W erfülle eine der Bedingungen (W0) oder (W1). Dann erfüllt eine diskrete Lösung $U, P \in S^{-1,0}(V^h)$ von Schema 3.2.2 die Ungleichung:

$$\frac{1}{2}\int_\Omega |\nabla U^K|^2 + \int_\Omega \mathcal{I}_h W(U^K,\cdot) + \int_0^T (M_\sigma(U)\nabla P, \nabla P) + \frac{1}{2}\sum_{k=0}^{K-1} \int_\Omega |\nabla U^{k+1} - \nabla U^k|^2$$
$$\leq \frac{1}{2}\int_\Omega |\nabla U^0|^2 + \int_\Omega \mathcal{I}_h W(U^0,\cdot). \tag{4.3}$$

Beweis : Wir setzen $\Theta = P^{k+1}$ in Gleichung (3.26) und erhalten:

$$(U^{k+1} - U^k, P^{k+1})_h + \tau_{k+1}(M_\sigma(U^{k+1})\nabla P^{k+1}, \nabla P^{k+1}) = 0. \tag{4.4}$$

Der erste Term läßt sich unter Ausnutzung von (3.27) weiter umformen:

$$\begin{aligned}(U^{k+1} - U^k, P^{k+1})_h = (\nabla U^{k+1}, \nabla U^{k+1} - \nabla U^k) \\ + (W^+_{,u}(U^{k+1}, \cdot), U^{k+1} - U^k)_h + (W^-_{,u}(U^k, \cdot), U^{k+1} - U^k)_h.\end{aligned}$$

Mit Hilfe der Konvexität läßt sich nun abschätzen:

$$\begin{aligned}(W^+_{,u}(U^{k+1}, \cdot), U^{k+1} - U^k)_h &= \langle M_h \mathcal{I}_h W^+_{,u}(U^{k+1}), U^{k+1} - U^k \rangle \\ &= \sum_{i=1}^{D} (M_h)_{ii} W^+_{,u}(U^{k+1}(\mathfrak{x}_i), \mathfrak{x}_i) \left(U^{k+1}(\mathfrak{x}_i) - U^k(\mathfrak{x}_i) \right) \\ &\geq \sum_{i=1}^{D} (M_h)_{ii} \left(W^+(U^{k+1}(\mathfrak{x}_i), \mathfrak{x}_i) - W^+(U^k(\mathfrak{x}_i), \mathfrak{x}_i) \right) \\ &= \int_\Omega \mathcal{I}_h \left(W^+(U^{k+1}, \cdot) - W^+(U^k, \cdot) \right).\end{aligned}$$

Analog folgt für den konkaven Term:

$$(W^-_{,u}(U^k, \cdot), U^{k+1} - U^k)_h \geq \int_\Omega \mathcal{I}_h \left(W^-(U^{k+1}, \cdot) - W^-(U^k, \cdot) \right).$$

Nutzt man nun noch die Formel $a(a-b) = \frac{1}{2}(a^2 - b^2 - (a-b)^2)$ aus, so erhält man

$$\begin{aligned}(U^{k+1} - U^k, P^{k+1})_h \geq \frac{1}{2}\int_\Omega |\nabla U^{k+1}|^2 - \frac{1}{2}\int_\Omega |\nabla U^k|^2 + \frac{1}{2}\int_\Omega |\nabla U^{k+1} - \nabla U^k|^2 \\ + \int_\Omega \mathcal{I}_h W(U^{k+1}, \cdot) - \int_\Omega \mathcal{I}_h W(U^k, \cdot).\end{aligned}$$

Dies setzen wir in (4.4) ein und summieren die Ungleichung über k:

$$\begin{aligned}\sum_{k=0}^{K-1} \tau_{k+1}(M_\sigma(U^{k+1})\nabla P^{k+1}, \nabla P^{k+1}) + \sum_{k=0}^{K-1} \frac{1}{2}\int_\Omega |\nabla U^{k+1} - \nabla U^k|^2 \\ + \frac{1}{2}\int_\Omega |\nabla U^K|^2 - \frac{1}{2}\int_\Omega |\nabla U^0|^2 + \int_\Omega \mathcal{I}_h W(U^K, \cdot) - \int_\Omega \mathcal{I}_h W(U^0, \cdot) \leq 0.\end{aligned}$$

Dies aber ist (4.3). □

Satz 4.1.4 *(Diskrete Energieabschätzung für $Q \not\equiv 0$)*

Es gelte (W2) und Q erfülle eine der beiden Bedingungen (Q1) oder (Q2). Im Fall (Q2) gelte zusätzlich die Bedingung $T < \frac{1}{C_q} \fint_\Omega U^0$. Dann erfüllt eine diskrete Lösung $U, P \in S^{-1,0}(V^h)$ von Schema 3.2.2 die Ungleichung:

$$\begin{aligned}\frac{1}{2}\int_\Omega |\nabla U^K|^2 + \int_\Omega \mathcal{I}_h W(U^K, \cdot) + \int_0^T (M_\sigma(U)\nabla P, \nabla P) + \frac{1}{2}\sum_{k=0}^{K-1}\int_\Omega |\nabla U^{k+1} - \nabla U^k|^2 \\ \leq \frac{1}{2}\int_\Omega |\nabla U^0|^2 + \int_\Omega \mathcal{I}_h W(U^0, \cdot) + C_q \int_{\Omega_T} |\nabla U|^2 + R.\end{aligned} \tag{4.5}$$

Dabei ist die Konstante R gegeben durch

$$R = C_q|\Omega_T|\left(\max_{i\in\{1,2\}}|W^-_{i,u}(\delta_q)| + \max_{i\in\{1,2\}}|W^+_{i,u}(\delta_q)|\right) + \begin{cases} C_q\tau\int_\Omega \mathcal{I}_h W^-_{,u}(U^0,\cdot) + C_R & \textit{falls (Q1) gilt,} \\ 0 & \textit{falls (Q2) gilt,} \end{cases}$$

mit einer weiteren Konstanten $C_R > 0$, welche von $c_\tau, a_{ij}, l_{ij}, c_q, C_q$ abhängt.

Beweis : Analog zum Beweis im Fall $Q = 0$ testet man Gleichung (3.26) mit $\Theta = P^{k+1}$. Die Terme auf der linken Seite der Gleichung ändern sich dadurch nicht und können wie im Beweis von Satz 4.1.3 abgeschätzt werden. Zusätzlich erhält man nun auf der rechten Seite der Gleichung noch den Term

$$\sum_{k=0}^{K-1}\tau_{k+1}(Q(U^{k+1}), P^{k+1})_h. \tag{4.6}$$

Dieser läßt sich mit Hilfe von (3.27) umformen zu

$$\sum_{k=0}^{K-1}\tau_{k+1}(\nabla U^{k+1}, \nabla\mathcal{I}_h Q(U^{k+1})) + \sum_{k=0}^{K-1}\tau_{k+1}(Q(U^{k+1}), W^+_{,u}(U^{k+1},\cdot) + W^-_{,u}(U^k,\cdot))_h. \tag{4.7}$$

Sei nun $d = 2$. Zu jedem Dreieck $E_i \in \mathcal{T}_h$ gibt es ein Referenzdreieck $\hat{E}_i$ mit den Eckpunkten $0, \alpha^i_1 e_1, \alpha^i_2 e_2$ und eine orthogonale Matrix A_i, so dass die affine Abbildung $f_i(\hat{x}) = x_0(E_i) + A_i\hat{x}$ das Referenzdreieck $\hat{E}_i$ auf E_i abbildet. Sei ferner $\hat{\nabla}$ die Ableitung nach den $\hat{x}$-Koordinaten, und für $U \in V^h$ und $E_i \in \mathcal{T}_h$ bezeichne $\hat{U} : \hat{E}_i \to \mathbb{R}$ die Funktion $\hat{U}(\hat{x}) := U(f_i(\hat{x}))$. Dann gilt:

$$\begin{aligned}
(\nabla U^{k+1}, \nabla\mathcal{I}_h Q(U^{k+1})) &= \sum_{E_i\in\mathcal{T}_h}\int_{E_i}\langle\nabla U^{k+1}, \nabla\mathcal{I}_h Q(U^{k+1})\rangle \\
&= \sum_{E_i\in\mathcal{T}_h}\int_{\hat{E}_i}\langle A_i^{-T}\hat{\nabla}\hat{U}^{k+1}, A_i^{-T}\hat{\nabla}\mathcal{I}_h Q(\hat{U}^{k+1})\rangle|\det A| \\
&= \sum_{E_i\in\mathcal{T}_h}|E_i|\langle\hat{\nabla}\hat{U}^{k+1}, \hat{\nabla}\mathcal{I}_h Q(\hat{U}^{k+1})\rangle \\
&= \sum_{E_i\in\mathcal{T}_h}|E_i|\langle\hat{\nabla}\hat{U}^{k+1}, \begin{pmatrix}\partial_1 Q(\hat{U}^{k+1}) & 0 \\ 0 & \partial_2 Q(\hat{U}^{k+1})\end{pmatrix}\hat{\nabla}\hat{U}^{k+1}\rangle.
\end{aligned}$$

Dabei ist

$$\partial_j Q(\hat{U}^{k+1}) = \begin{cases} \frac{Q(\hat{U}^{k+1}(\alpha^i_j e_j)) - Q(\hat{U}^{k+1}(0))}{\hat{U}^{k+1}(\alpha^i_j e_j) - \hat{U}^{k+1}(0)} & \text{falls } \hat{U}^{k+1}(\alpha^i_j e_j) \neq \hat{U}^{k+1}(0), \\ 0 & \text{sonst.} \end{cases}$$

Da Q in $W^{1,\infty}(\mathbb{R})$ beschränkt ist, gilt insgesamt:

$$(\nabla U^{k+1}, \nabla\mathcal{I}_h Q(U^{k+1})) \leq C_q\sum_{E_i\in\mathcal{T}_h}|E_i|\langle\hat{\nabla}\hat{U}^{k+1}, \hat{\nabla}\hat{U}^{k+1}\rangle = C_q(\nabla U^{k+1}, \nabla U^{k+1}).$$

Dies gilt auch im einfacheren Fall $d = 1$, wie sich leicht überprüfen läßt. Damit gilt also für den ersten Term aus (4.7):

$$\sum_{k=0}^{K-1} \tau_{k+1}(\nabla U^{k+1}, \nabla \mathcal{I}_h Q(U^{k+1})) \leq C_q \int_0^T |\nabla U|^2.$$

Für den zweiten Term aus (4.7) gilt die Gleichheit:

$$\sum_{k=0}^{K-1} \tau_{k+1}(\mathcal{I}_h Q(U^{k+1}), W_{,u}^+(U^{k+1},\cdot) + W_{,u}^-(U^k,\cdot))_h$$
$$= \sum_{k=1}^{K} \tau_k (Q(U^k), W_{,u}^+(U^k,\cdot))_h + \sum_{k=0}^{K-1} \tau_{k+1}(Q(U^{k+1}), W_{,u}^-(U^k,\cdot))_h. \quad (4.8)$$

Wir betrachten nun getrennt die Bereiche

$$[U^k \geq \delta_q] := \{x \in \Omega : U^k(x) \geq \delta_q\} \quad \text{und} \quad [U^k < \delta_q] := \{x \in \Omega : U^k(x) < \delta_q\}.$$

Auf $[U^k \geq \delta_q]$ können wir $W_{,u}^+(U^k,\cdot)$ und $W_{,u}^-(U^k,\cdot)$ betragsmäßig gegen $\max_{i\in\{1,2\}} |W_{i,u}^+(\delta_q)|$ bzw. $\max_{i\in\{1,2\}} |W_{i,u}^-(\delta_q)|$ abschätzen, da (W2) die Monotonie der Funktionen $W_{,u}^+$ und $W_{,u}^-$ sicherstellt. Es gilt also:

$$\begin{aligned}
(Q(U^k), W_{,u}^+(U^k,\cdot))_h &\leq C_q|\Omega| \max_{i\in\{1,2\}} |W_{i,u}^+(\delta_q)| + \int_{[U^k<\delta_q]} \mathcal{I}_h \left(Q(U^k) W_{,u}^+(U^k,\cdot) \right), \\
(Q(U^{k+1}), W_{,u}^-(U^k,\cdot))_h &\leq C_q|\Omega| \max_{i\in\{1,2\}} |W_{i,u}^-(\delta_q)| + \int_{[U^k<\delta_q]} \mathcal{I}_h \left(Q(U^{k+1}) W_{,u}^-(U^k,\cdot) \right).
\end{aligned}$$

Dies setzen wir in (4.8) ein und erhalten:

$$\begin{aligned}
&\sum_{k=0}^{K-1} \tau_{k+1}(\mathcal{I}_h Q(U^{k+1}), W_{,u}^+(U^{k+1},\cdot) + W_{,u}^-(U^k,\cdot))_h \\
\leq\ & C_q|\Omega_T| \left(\max_{i\in\{1,2\}} |W_{i,u}^-(\delta_q)| + \max_{i\in\{1,2\}} |W_{i,u}^+(\delta_q)| \right) \\
&+ \underbrace{\sum_{k=1}^{K} \tau_k \int_{[U^k<\delta_q]} \mathcal{I}_h \left(Q(U^k) W_{,u}^+(U^k,\cdot) \right)}_{I} + \underbrace{\sum_{k=0}^{K-1} \tau_{k+1} \int_{[U^k<\delta_q]} \mathcal{I}_h \left(Q(U^{k+1}) W_{,u}^-(U^k,\cdot) \right)}_{II}.
\end{aligned}$$

Wir unterscheiden nun die Fälle (Q1) und (Q2). Im Fall (Q2) ist $Q(U^k(x)) = 0$ für $U^k(x) < \delta_q$ und damit ist $I = 0$. Da außerdem Q negativ und $W_{,u}^-$ positiv ist, ist $II \leq 0$. Somit ist der Term (4.6) abgeschätzt durch:

$$\sum_{k=0}^{K-1} \tau_{k+1}(Q(U^{k+1}), P^{k+1})_h \leq C_q|\Omega_T| \left(\max_{i\in\{1,2\}} |W_{i,u}^-(\delta_q)| + \max_{i\in\{1,2\}} |W_{i,u}^+(\delta_q)| \right).$$

Im Fall (Q1) ist Q positiv und für $U^k(x) < \delta_q$ gilt $Q(U^k(x)) \geq c_q$. Da $W_{,u}^+$ negativ ist, gilt

$$I \leq \sum_{k=1}^{K} c_q c_\tau \tau \int_{[U^k<\delta_q]} \mathcal{I}_h W_{,u}^+(U^k,\cdot).$$

Die Abschätzung für Term II nutzt aus, dass $W^+_{,u}$ positiv und Q durch C_q beschränkt ist:

$$II \leq \sum_{k=0}^{K-1} C_q \tau \int_{[U^k<\delta_q]} \mathcal{I}_h W^-_{,u}(U^k,\cdot).$$

Also gilt insgesamt:

$$\begin{aligned} I+II &\leq C_q\tau \int_{[U^k<\delta_q]} \mathcal{I}_h W^-_{,u}(U^0,\cdot) + c_q c_\tau \tau \int_{[U^k<\delta_q]} \mathcal{I}_h W^+_{,u}(U^N,\cdot) \\ &\quad + \sum_{k=1}^{K-1} \int_{[U^k<\delta_q]} \left(c_q c_\tau \tau \mathcal{I}_h W^+_{,u}(U^k,\cdot) + C_q \tau \mathcal{I}_h W^-_{,u}(U^k,\cdot) \right). \end{aligned}$$

Da der negative Term $W^+_{,u}(U^k,\cdot)$ von höherer Ordnung ist als der positive Term $W^-_{,u}(U^k,\cdot)$, konvergiert die Summe (falls ε_w klein genug gewählt wird) für $U^k(x) \to 0$ gegen $-\infty$. Insbesondere ist der Term nach oben beschränkt. Also gibt es eine Konstante C_R, so dass

$$\begin{aligned} &\sum_{k=0}^{K-1} \tau_{k+1} (Q(U^{k+1}), P^{k+1})_h \\ &\qquad \leq C_q |\Omega_T| \left(\max_{i\in\{1,2\}} |W^-_{i,u}(\delta_q)| + \max_{i\in\{1,2\}} |W^+_{i,u}(\delta_q)| \right) + C_q \tau \int_\Omega \mathcal{I}_h W^-_{,u}(U^0,\cdot) + C_R \end{aligned}$$

gilt, woraus die Behauptung im Fall (Q1) folgt. □

Mit (4.5) allein ist jedoch noch keine Abschätzung für die Energie gegeben, da auf der rechten Seite der Ungleichung noch der Term $\int_{\Omega_T} |\nabla U|^2$ vorkommt. Mit Hilfe des Lemmas von Gronwall läßt sich aber folgern:

Korollar 4.1.5 *Unter den Voraussetzungen von Satz 4.1.4 gibt es eine positive Konstante C_T abhängig von T, so dass*

$$\int_\Omega |\nabla U(t)|^2 \leq C_T \tag{4.9}$$

für alle $t \in [0,T]$ gilt.

Beweis : Da W nach unten durch eine Konstante beschränkt ist, können wir zu (4.5) eine Konstante C_w hinzuaddieren, so dass die Terme $\int_\Omega (\mathcal{I}_h W(U^K)+C_w)$ und $\int_\Omega (\mathcal{I}_h W(U^0)+C_w)$ positiv sind. Damit gilt also

$$\int_\Omega |\nabla U(T)|^2 \leq \int_\Omega |\nabla U(0)|^2 + \int_0^T \int_\Omega |\nabla U(t)|^2 dt + C,$$

dabei ist die Konstante C abhängig von den Parametern $c_q, C_q, \delta_q, a_{ih}, l_{ij}, |\Omega_t|, c_\tau, u_0$. Wir wenden nun das Lemma von Gronwall auf die Funktion

$$f(t) := \int_\Omega |\nabla U(t)|^2$$

an und erhalten mit $a_0 := \int_\Omega |\nabla U(0)|^2 + C$:

$$\int_\Omega |\nabla U(t)|^2 \leq a_0 + \int_0^t a_0 e^{t-s} ds \leq a_0 + T a_0 e^T,$$

woraus die Behauptung folgt. □

Damit kann der Term $\int_{\Omega_T} |\nabla U(t)|^2 dt$ auf der rechten Seite von (4.5) gegen $Ta_0 + T^2 a_0 e^T$ abgeschätzt werden, womit sich alle Terme auf der rechten Seite von (4.5) durch Konstanten abschätzen lassen.

4.2 Entropieabschätzung und Nichtnegativität

Die diskrete Entropieabschätzung ist eine a priori Abschätzung für die in Definition 3.2.1 definierte diskrete Entropie. Die Beschränktheit der Entropie kann – analog zum Nichtnegativitätsbeweis von Bernis [6] – dazu genutzt werden, die Nichtnegativität der diskreten Lösung zu zeigen. Darüberhinaus zeigt die Entropieabschätzung die Beschränktheit der zweiten Ableitung von U. Mit Hilfe der Funktion $U^- \in S^{-1,0}(V^h)$, welche durch

$$U^-(t) := U^k, \text{ falls } t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}], \tag{4.10}$$

auf dem Intervall $[0,T]$ definiert ist, lassen sich die folgenden diskreten Entropieabschätzungen zeigen:

Satz 4.2.1 *(Entropieabschätzung für $Q \equiv 0$)*
Es gelte (Q0) und eine der Bedingungen (W0) oder (W1). Dann erfüllt eine Lösung $U, P \in S^{-1,0}(V^h)$ von Schema 3.2.2 die Ungleichung:

$$\begin{aligned}&\int_\Omega \mathcal{I}_h G_\sigma(U(T)) + \frac{1}{4}\int_0^T \|\Delta_h U(t)\|_h^2 dt + \frac{1}{4}\int_0^T \|P(t)\|_h^2 dt \\ &\leq \int_\Omega \mathcal{I}_h G_\sigma(U^0) + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W_{,u}^+(U(t,x),x)|^2 + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W_{,u}^-(U(t,x),x)|^2. \end{aligned} \tag{4.11}$$

Dabei ist $\Delta_h U \in S^{-1,0}(V^h)$ definiert durch:

$$(\Delta_h U(t,\cdot), \phi)_h = (\nabla U(t,\cdot), \nabla\phi) \quad \forall \phi \in V^h, \forall t \in [0,T]. \tag{4.12}$$

Beweis : Wir testen Gleichung (3.26) mit $\Theta = \mathcal{I}_h G'_\sigma(U^{k+1})$ und erhalten für den parabolischen Term:

$$\begin{aligned}(U^{k+1} - U^k, \mathcal{I}_h G'_\sigma(U^{k+1}))_h &= \int_\Omega \mathcal{I}_h(U^{k+1} - U^k, G'_\sigma(U^{k+1}))_h \\ &\geq \int_\Omega \mathcal{I}_h G_\sigma(U^{k+1}) - \int_\Omega \mathcal{I}_h G_\sigma(U^k).\end{aligned}$$

Der elliptische Term, getestet mit $\Theta = \mathcal{I}_h G'_\sigma(U^{k+1})$, läßt sich mit Hilfe von (M3) und Gleichung (3.27) unformen:

$$\begin{aligned}\tau_{k+1}(M_\sigma(U^{k+1})\nabla P^{k+1}, \nabla \mathcal{I}_h G'_\sigma(U^{k+1})) &= \tau_{k+1}(\nabla P^{k+1}, \nabla U^{k+1}) \\ &= \tau_{k+1}(P^{k+1}, P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot))_h.\end{aligned}$$

Dies läßt sich nun auf zwei verschiedene Weisen weiter umformen. Die erste Möglichkeit nutzt die mit Hilfe der Youngschen Ungleichung $(a,b)_h \leq \frac{\varepsilon}{2}\|a\|_h^2 + \frac{1}{2\varepsilon}\|b\|_h^2$ gewonnene Abschätzung

$$\begin{aligned}&(P^{k+1}, P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot))_h \\ &\qquad = \|P^{k+1}\|_h^2 - (P^{k+1}, \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) + \mathcal{I}_h W^-_{,u}(U^k,\cdot))_h \\ &\qquad \geq \|P^{k+1}\|_h^2 - \frac{1}{2}\|P^{k+1}\|_h^2 - \|\mathcal{I}_h W^+_{,u}(U^{k+1},\cdot)\|_h^2 - \|\mathcal{I}_h W^-_{,u}(U^k,\cdot)\|_h^2\end{aligned}$$

und liefert nach Summation über $k = 0, \ldots, K-1$:

$$\begin{aligned}&\int_\Omega \mathcal{I}_h G_\sigma(U(T)) + \frac{1}{2}\int_0^T \|P(t)\|_h^2 dt \\ &\qquad \leq \int_\Omega \mathcal{I}_h G_\sigma(U^0) + \int_0^T \|\mathcal{I}_h W^+_{,u}(U(t),\cdot)\|_h^2 + \|\mathcal{I}_h W^-_{,u}(U^-(t),\cdot)\|_h^2 dt. \quad (4.13)\end{aligned}$$

Die zweite Möglichkeit nutzt die Abschätzung

$$\begin{aligned}&(P^{k+1}, P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot))_h \\ &\quad = \|P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot)\|_h^2 \\ &\qquad + (\mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) + \mathcal{I}_h W^-_{,u}(U^k,\cdot), P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot))_h \\ &\quad \geq (1 - \frac{1}{2})\|P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1},\cdot) - \mathcal{I}_h W^-_{,u}(U^k,\cdot)\|_h^2 \\ &\qquad - \|\mathcal{I}_h W^+_{,u}(U^{k+1},\cdot)\|_h^2 - \|\mathcal{I}_h W^-_{,u}(U^k,\cdot)\|_h^2\end{aligned} \quad (4.14)$$

und ergibt nach Summation über $k = 0, \ldots, K-1$:

$$\begin{aligned}&\int_\Omega \mathcal{I}_h G_\sigma(U(T)) + \frac{1}{2}\int_0^T \|P(t) - \mathcal{I}_h W^+_{,u}(U(t),\cdot) - \mathcal{I}_h W^-_{,u}(U^-(t),\cdot)\|_h^2 dt \\ &\qquad \leq \int_\Omega \mathcal{I}_h G_\sigma(U^0) + \int_0^T \|\mathcal{I}_h W^+_{,u}(U(t),\cdot)\|_h^2 + \|\mathcal{I}_h W^-_{,u}(U^-(t),\cdot)\|_h^2 dt. \quad (4.15)\end{aligned}$$

Addition von (4.13) und (4.15) ergibt Ungleichung (4.11), wenn man berücksichtigt, dass aufgrund von (3.27)

$$P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1}) - \mathcal{I}_h W^-_{,u}(U^k) = -\Delta_h U^{k+1}$$

ist. □

Bemerkung : Der Beweis der Entropieabschätzung für homogene Substrate (siehe [19]) schätzt Term (4.14) nicht mit Hilfe der Youngschen Ungleichung ab, sondern formt den

Term zunächst mit Hilfe von (3.27) weiter um:

$$\begin{aligned}(P^{k+1}, P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1}) - \mathcal{I}_h W^-_{,u}(U^k))_h \\ &= (\mathcal{I}_h W^+_{,u}(U^{k+1}) + \mathcal{I}_h W^-_{,u}(U^k), P^{k+1} - \mathcal{I}_h W^+_{,u}(U^{k+1}) - \mathcal{I}_h W^-_{,u}(U^k))_h \\ &= (\nabla U^{k+1}, \nabla \mathcal{I}_h W^+_{,u}(U^{k+1}) + \nabla \mathcal{I}_h W^-_{,u}(U^k)) \\ &= (\nabla U^{k+1}, \mathcal{I}_h W^+_{,uu}(U^{k+1}) \nabla U^{k+1}) + (\nabla U^{k+1}, \mathcal{I}_h W^-_{,uu}(U^k) \nabla U^k)\end{aligned}$$

Die so entstandenen Terme können mit Hilfe der Konvexität bzw. Konkavität von W^+ bzw. W^- weiter abgeschätzt werden. Würde man diesen Ansatz auch für inhomogene Substrate verfolgen, so müssten $W^+_{,u}$ und $W^-_{,u}$ nun zusätzlich partiell nach x abgeleitet werden. Im Fall (W1) ist dies nicht möglich, da W an der Grenze zwischen Ω_1 und Ω_2 nicht stetig sein muss. Im Fall eines stetigen Potentials (W0) müsste man, wenn man diesen Ansatz verfolgt, die Terme $(\nabla U^{k+1}, \mathcal{I}_h W^+_{,ux}(U^{k+1}, \cdot))$ und $(\nabla U^{k+1}, \mathcal{I}_h W^-_{,ux}(U^k, \cdot))$ abschätzen. Dies ist aber ohne weitere Bedingungen an W nicht möglich.

Satz 4.2.2 *(Entropieabschätzung für $Q \not\equiv 0$)*

Es gelte (W2) und eine der beiden Bedingungen (Q1) oder (Q2). Im Fall (Q1) gelte zusätzlich $n > 1$, im Fall (Q2) sei $T < \frac{1}{C_q} \fint_\Omega U^0$. Dann erfüllt eine Lösung $U, P \in S^{-1,0}(V^h)$ des Finite-Elemente-Verfahrens 3.2.2 die Ungleichung

$$\begin{aligned}&\int_\Omega \mathcal{I}_h G_\sigma(U(T)) + \frac{1}{4}\int_0^T \|\Delta_h U(t)\|_h^2 dt + \frac{1}{4}\int_0^T \|P(t)\|_h^2 dt \\ &\le \int_\Omega \mathcal{I}_h G_\sigma(U^0) + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W^+_{,u}(U(t,x),x)|^2 + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W^-_{,u}(U(t,x),x)|^2 + R.\end{aligned} \tag{4.16}$$

Dabei ist $R = 0$ im Fall (Q2) und $R = (C_q|\Omega| + 1)\frac{1}{n-1}A^{1-n}$ im Fall (Q1).

Beweis : Geht man wie im Beweis des Satzes 4.2.1 vor, so erhält man auf der rechten Seite zusätzlich den Term

$$\int_\Omega \mathcal{I}_h Q(U^{k+1}) \mathcal{I}_h G'_\sigma(U^{k+1}).$$

Es gilt:

$$G'_\sigma(U^{k+1}(x)) = g_\sigma(U^{k+1}(x)) = \int_A^{U^{k+1}(x)} \frac{dr}{m_\sigma(r)}.$$

Im Fall (Q1) ist $0 \le \mathcal{I}_h Q(U^{k+1}(x)) \le C_q$. Für $U^{k+1}(x) \ge A > \sigma$ gilt:

$$g_\sigma(U^{k+1}(x)) = \int_A^{U^{k+1}(x)} r^{-n} dr = -\frac{1}{n-1}(U^{k+1}(x))^{1-n} + \frac{1}{n-1}A^{1-n} \le \frac{1}{n-1}A^{1-n}.$$

Für $U^{k+1}(x) < A$ ist $g_\sigma(U^{k+1}(x)) \le 0$, und damit gilt:

$$\int_\Omega \mathcal{I}_h Q(U^{k+1}) \mathcal{I}_h G'_\sigma(U^{k+1}) \le C_q \frac{1}{n-1} A^{1-n}|\Omega|.$$

Im Fall (Q2) sei o.B.d.A. $\delta_q > A$. Nun gilt für alle Stützstellen $\mathfrak{x}_i$, dass entweder $Q(U^{k+1}(\mathfrak{x}_i)) = 0$ ist oder aber $g_\sigma(U^{k+1}(\mathfrak{x}_i)) > 0$ und $Q(U^{k+1}(\mathfrak{x}_i)) \leq 0$ ist, woraus unmittelbar

$$\int_\Omega \mathcal{I}_h Q(U^{k+1}) \mathcal{I}_h G'_\sigma(U^{k+1}) \leq 0$$

folgt. Dies ergibt, kombiniert mit dem Beweis von Satz 4.2.1, die Behauptung. □

Die Nichtnegativität der diskreten Lösung läßt sich nun wie in [21] zeigen, falls $\int_\Omega G_\sigma(U(T))$ gleichmäßig beschränkt ist. Dazu müssen die Terme auf der rechten Seite von (4.11) bzw. (4.16) unabhängig von h, τ, σ beschränkt sein. Dies gilt, falls W eine der stärkeren Bedingungen (W0'), (W1') oder (W2) erfüllt. Außerdem muss $\int_\Omega G_\sigma(U^0)$ unabhängig von σ beschränkt sein. Dies ist genau dann der Fall, wenn $u_0 \geq 0$ und $0 < n < 2$, oder aber $u_0 \geq c_0 > 0$ und $n > 2$. Es gilt der folgende Satz:

Satz 4.2.3 *(Nichtnegativität der diskreten Lösung)*
In jeder der vier Situationen

i) $Q \equiv 0$ *und das Potential* W *erfüllt die Bedingung (W0'),*

ii) $Q \equiv 0$ *und das Potential* W *erfüllt die Bedingung (W1'),*

iii) Q *erfüllt (Q1), es gilt* $n > 1$ *und das Potential* W *erfüllt die Bedingung (W2),*

iv) Q *erfüllt (Q2), es gilt* $T < \frac{1}{C_q} \fint_\Omega U^0$ *und das Potential* W *erfüllt die Bedingung (W2),*

gilt: Zu jedem $\varepsilon > 0$ *existiert ein* $\delta > 0$*, welches von* n, ε, h *und den Anfangswerten* $u_0 \in L^2(\Omega)$ *abhängig ist, so dass für alle* $0 < \sigma < \delta$ *die zu dem Entropie-Mobilitäts-Paar* M_σ, G_σ *gehörige diskrete Lösung* U_σ *des Finite-Elemente-Verfahrens 3.2.2 die folgende Eigenschaft hat:*

$$U_\sigma > \begin{cases} -\varepsilon & \text{falls } u_0 \geq 0 \text{ und } 0 < n < 2, \\ -\varepsilon & \text{falls } u_0 \geq \delta \text{ und } n = 2, \\ \frac{\sigma}{2} & \text{falls } u_0 \geq \delta \text{ und } n > 2. \end{cases} \tag{4.17}$$

Beweis : Wir definieren

$$C_e := \int_\Omega \mathcal{I}_h G_\sigma(U^0_\sigma) + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W^+_{,u}(U_\sigma(t,x),x)|^2 + |\Omega_T| \sup_{(t,x)\in\Omega_T} |W^-_{,u}(U_\sigma(t,x),x)|^2 + R.$$

In jeder der Situationen i) - iv) gilt: $C_e < \infty$.

Sei zunächst $1 \leq n$. Wir nehmen o.B.d.A. an, dass $A > \sigma$ gilt. Nun definieren wir für $s > 0$ Stammfunktionen $R_1(s)$ und $R_2(s)$ von $m^{-1}(s)$ durch:

$$\begin{aligned} R'_2(s) &= R_1(s), \quad & R_2(A) &= 0, \\ R'_1(s) &= \frac{1}{s^n}, & R_1(A) &= 0. \end{aligned}$$

Damit ist $R_2 \geq 0$ konvex und R_1 monoton steigend. Es gilt

$$G_\sigma(s) = \begin{cases} R_2(s) & \text{falls } s \geq \sigma, \\ R_2(\sigma) + (s-\sigma)R_1(\sigma) + \frac{1}{2}(s-\sigma)^2 \frac{1}{m(\sigma)} & \text{falls } s < \sigma. \end{cases}$$

Aufrund der Positivität von $R_2(\sigma)$ und $\frac{1}{2}(s-\sigma)^2 m(\sigma)^{-1}$ folgt, dass

$$G_\sigma(s) \geq (s-\sigma)R_1(\sigma)$$

für alle $s > 0$ gilt. Im Fall $s \geq \sigma$ folgt dies durch die Negativität von $R_1(\sigma)$. Also gilt für eine beliebige Indexmenge $I \subset \{1, \ldots, D\}$:

$$\sum_{i \in I} \int_\Omega \varphi_i(x)dx(U_\sigma(t, \mathfrak{x}_i) - \sigma)R_1(\sigma) \\ \leq \sum_{i \in I} \int_\Omega \varphi_i(x)dx G_\sigma(U_\sigma(t, \mathfrak{x}_i)) \leq \int_\Omega \mathcal{I}_h G_\sigma(U_\sigma(t)) \leq C_e. \quad (4.18)$$

Da sich die Masse der Basisfunktionen φ_i durch $\int_\Omega \varphi_i(x)dx \geq c\,h^d$, $c > 0$, abschätzen lässt, gilt insbesondere punktweise:

$$U_\sigma(t, \mathfrak{x}_i) \geq \frac{C_e}{c\,h^d R_1(\sigma)} + \sigma. \quad (4.19)$$

Für $\sigma \to 0$ konvergiert $R_1(\sigma)$ wie $\log \sigma$ (falls $n = 1$) bzw. wie $-\sigma^{1-n}$ (falls $n > 1$) gegen $-\infty$. Deshalb konvergiert die rechte Seite von Ungleichung (4.19) für $\sigma \to 0$ gegen 0. Also existiert zu jedem ε ein $\delta(\varepsilon, n, C_e, h)$ mit der geforderten Eigenschaft.

Im Fall $n > 2$ läßt sich dieses Resultat noch verbessern. Wir definieren die Menge

$$K_\beta(t) := \{\mathfrak{x}_i \in \mathcal{N}_h | U_\sigma(t, \mathfrak{x}_i) < \beta\}.$$

Also folgt aus (4.18):

$$-\frac{C_e}{R_1(\sigma)} \geq \sum_{\mathfrak{x}_i \in K_{\frac{\sigma}{2}}(t)} \int_\Omega \varphi_i(x)dx(\sigma - U_\sigma(t, \mathfrak{x}_i)) \geq \sum_{\mathfrak{x}_i \in K_{\frac{\sigma}{2}}(t)} c\,h^d(\sigma - \frac{1}{2}\sigma).$$

und es gilt für die Anzahl der Elemente von $K_{\frac{\sigma}{2}}(t)$:

$$|K_{\frac{\sigma}{2}}(t)| \leq \frac{2C_e}{c\,h^d} \frac{-1}{\sigma R_1(\sigma)}.$$

Für $\sigma \to 0$ konvergiert $\sigma R_1(\sigma)$ wie $-\sigma^{2-n}$ gegen $-\infty$. Deshalb gibt es ein genügend kleines δ, so dass $K_{\frac{\sigma}{2}}(t)$ für alle $\sigma < \delta$ eine leere Menge ist.

Im Fall $0 < n < 1$ definieren wir für $s \in \mathbb{R}_0^+$ Stammfunktionen $R_i(s)$ von $m^{-1}(s)$ durch:

$$R_2'(s) = R_1(s), \quad R_2(0) = 0, \\ R_1'(s) = \frac{1}{s^n}, \qquad R_1(0) = 0.$$

Nun gilt

$$G_\sigma(s) = \begin{cases} R_2(s) & \text{falls } s \geq 0, \\ \frac{1}{\sigma} R_2(-s) & \text{falls } s < 0. \end{cases}$$

Da $R_2(s)$ für $s > 0$ positiv und monoton wachsend ist, folgt daraus:

$$C_e \geq \sum_{\mathfrak{x}_i \in K_{-\varepsilon}(t)} \int_\Omega \varphi_i(x) dx \frac{1}{\sigma} R_2(-U_\sigma(t, \mathfrak{x}_i)) \geq c\, h^d \sum_{\mathfrak{x}_i \in K_{-\varepsilon}(t)} \frac{1}{\sigma} R_2(\varepsilon),$$

und damit ergibt sich:

$$|K_{-\varepsilon}(t)| \leq \frac{C_e \sigma}{c\, h^d R_2(\varepsilon)}.$$

Für $\sigma \to 0$ konvergiert die rechte Seite gegen 0, und damit ist der Satz bewiesen. □

Kapitel 5

Konvergenz des Verfahrens

In Kapitel 3 wurde gezeigt, dass zu jeder zulässigen und rechtwinkligen Triangulierung $\mathcal{T}_h$ und zu jeder Wahl von Zeitpunkten $\mathfrak{t}_0, \ldots, \mathfrak{t}_K$ eine diskrete Lösung von Problem 3.2.2 existiert. Mit $\{U_{\tau h}\}$ ist also eine Familie von Lösungen zu verschiedenen Gitterweiten und Zeitschrittweiten gegeben. Dabei bezeichnet h die Gitterweite

$$h = \max_{E \in \mathcal{T}_h} \{\mathrm{diam}(E)\}$$

der dazugehörigen Triangulierung und τ die Zeitschrittweite

$$\tau = \max_{k=1,\ldots,K} \tau_k$$

der dazugehörigen Wahl von Zeitschritten. Diese Schreibweise ist etwas vereinfachend, da $U_{\tau h}$ nicht allein von τ und h abhängig ist, sondern von $\mathfrak{t}_0, \ldots, \mathfrak{t}_K$ und $\mathcal{T}_h$. Außerdem ist die Lösung zunächst einmal abhängig von der Wahl des Parameters σ. Durch Satz 4.2.3 ist aber implizit eine Funktion $\sigma(h)$ mit $\sigma(h) \xrightarrow{h \to 0} 0$ gegeben, welche σ so klein wählt, dass die Nichtnegativität der diskreten Lösung im Sinne von Satz 4.2.3 gegeben ist. Wir nehmen daher in diesem Kapitel an, dass σ durch dieses $\sigma(h)$ gegeben sei.

In diesem Kapitel wird nun gezeigt, dass wir aus der Familie der $U_{\tau h}$ eine Folge auswählen können mit $h, \tau \to 0$, so dass diese Folge gegen eine schwache Lösung des kontinuierlichen Problems konvergiert. Gleichzeitig wird so die Existenz einer schwachen kontinuierlichen Lösung gezeigt.

5.1 Konvergenz in Raumdimension $d = 1$

Im Fall $d = 1$ werden wir zwei unterschiedliche Konvergenzresultate beweisen. Das erste der beiden nun folgenden Konvergenztheoreme geht von einem Potential w der Form (w0) oder (w1) und $q \equiv 0$ aus. In diesem Fall lässt sich Konvergenz gegen eine nichtnegative Lösung zeigen, wenn die Anfangsdaten u_0 die in Satz 4.2.3 benötigten Bedingungen erfüllen. Da das Potential $w(u, x)$ für $u = 0$ singulär sein kann, u aber nicht notwendigerweise positiv ist, ist die Differentialgleichung lediglich für die Näherung W erfüllt.

Theorem 5.1.1 *(Konvergenz gegen eine nichtnegative schwache Lösung)*

Sei $\Omega \subset \mathbb{R}$ ein offenes und beschränktes Intervall und $T \in \mathbb{R}$. Es gelte $q \equiv 0$ und das Grenzflächenpotential w erfülle eine der Bedingungen (w0) oder (w1). Die Funktion W sei eine Näherung von w derart, daß (W0') bzw. (W1') erfüllt ist. Für die Anfangsdaten u_0 gelte:

$$u_0 \in H^1(\Omega) \cap C^0(\bar{\Omega}) \text{ mit } \begin{cases} u_0 \geq 0 & \text{falls } 0 < n < 2, \\ u_0 \geq c_0 > 0 & \text{falls } n \geq 2. \end{cases} \tag{5.1}$$

Dann gibt es eine Funktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T) \cap L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;H^2(\Omega))$ mit einer schwachen Ableitung $\partial_t u \in L^2(0,T;H^1(\Omega)')$ und eine Funktion $p \in L^2(\Omega_T)$ mit $\partial_x p \in L^2([u>\delta])$ für alle $\delta > 0$, so dass

$$\int_0^T \langle \partial_t u\,,\psi\rangle_{H^1(\Omega)'\times H^1(\Omega)} + \int_{[u>0]} m(u)\partial_x p \partial_x \psi = 0 \tag{5.2}$$

für alle $\psi \in L^2(0,T;H^1(\Omega))$ gilt. Dabei ist

$$p = -\partial_{xx} u + W_{,u}(u,\cdot). \tag{5.3}$$

u ist nichtnegativ, es gilt $\lim_{t\to 0} u(t,\cdot) = u_0(\cdot)$, und u erfüllt für fast alle $t \in (0,T)$ die Ungleichung

$$\int_\Omega |\partial_x u(t)|^2 + \int_\Omega W(u(t),\cdot) + \int_\Omega G(u(t)) \leq C. \tag{5.4}$$

Ferner gibt es eine Folge diskreter Lösungen $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Schemas 3.2.2, so dass für $\tau, h \to 0$ gilt:

$$U_{\tau h} \to u \text{ gleichmäßig in } \Omega_T, \tag{5.5}$$

$$\Delta_h U_{\tau h} \rightharpoonup \partial_{xx} u \text{ schwach in } L^2(\Omega_T), \tag{5.6}$$

$$P_{\tau h} \rightharpoonup p \text{ schwach in } L^2(\Omega_T), \tag{5.7}$$

$$\partial_x P_{\tau h} \rightharpoonup \partial_x p \text{ schwach in } L^2([u>\delta]) \text{ für alle } \delta > 0. \tag{5.8}$$

Das zweite Konvergenztheorem betrachtet Potentiale der Art (w2) und verschiedene rechte Seiten q. Hier lässt sich unter Ausnutzung der Hölderstetigkeit von u und dem speziellen Wachstumsverhalten der $w_i(u)$ strikte Positivität der Lösung u zeigen. Dadurch können einige Aussagen von Theorem 5.1.1 verbessert werden, insbesondere kann W in (5.3) durch w ersetzt werden.

Theorem 5.1.2 *(Konvergenz gegen eine positive schwache Lösung)*

Sei $\Omega \subset \mathbb{R}$ ein offenes und beschränktes Intervall und $T \in \mathbb{R}$. Das effektive Grenzflächenpotential w erfülle die Bedingung (w2). Für die Anfangsdaten $u_0 \in H^1(\Omega) \cap C^0(\bar{\Omega})$ gelte $u_0 \geq c_0 > 0$. q erfülle eine der Bedingungen (q0), (q1) oder (q2), Q dementsprechend eine der Bedingungen (Q0), (Q1) oder (Q2). Im Fall (q1) gelte zusätzlich $n > 1$, im Fall (q2) sei $T < \frac{1}{C_q} ⨍_\Omega U^0$.

Dann gibt es eine Funktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T) \cap L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;H^2(\Omega))$ mit einer schwachen Ableitung $\partial_t u \in L^2(0,T;H^1(\Omega)')$ und eine Funktion $p \in L^2(0,T;H^1(\Omega))$, so dass

$$\int_0^T \langle \partial_t u\,,\psi\rangle_{H^1(\Omega)'\times H^1(\Omega)} + \int_{\Omega_T} m(u)\partial_x p \partial_x \psi = \int_{\Omega_T} Q(u)\psi \tag{5.9}$$

für alle $\psi \in L^2(0,T;H^1(\Omega))$ gilt. Dabei ist

$$p = -\partial_{xx}u + w_{,u}(u,\cdot). \tag{5.10}$$

Es gibt eine Konstante $\gamma > 0$, so dass $u(t,x) \geq \gamma$ für alle $(t,x) \in \Omega_T$ gilt. Es ist $\lim_{t\to 0} u(t,\cdot) = u_0(\cdot)$ und u erfüllt für fast alle $t \in (0,T)$ die Ungleichung

$$\int_\Omega |\partial_x u(t)|^2 + \int_\Omega w(u(t),\cdot) + \int_\Omega G(u(t)) \leq C. \tag{5.11}$$

Ferner gibt es eine Folge diskreter Lösungen $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Schemas 3.2.2, so dass für $\tau, h \to 0$ gilt:

$$U_{\tau h} \to u \text{ gleichmäßig in } \Omega_T, \tag{5.12}$$

$$\Delta_h U_{\tau h} \rightharpoonup \partial_{xx}u \text{ schwach in } L^2(\Omega_T), \tag{5.13}$$

$$P_{\tau h} \rightharpoonup p \text{ schwach in } L^2(0,T;H^1(\Omega)). \tag{5.14}$$

Bemerkung : Ist q durch $q(s) = \frac{C_1}{u+C_2}$ mit positiven Konstanten C_1 und C_2 gegeben, so kann Q in (5.9) durch q ersetzt werden.

In den folgenden zwei Abschnitten werden beide Theoreme gleichzeitig bewiesen. Im ersten Teil wird die Zeitkompaktheit der diskreten Lösungen und die daraus folgende gleichmäßige Konvergenz gegen eine hölderstetige Funktion u bewiesen. Dabei wird nur die Energieabschätzung benutzt, nicht jedoch die Entropieabschätzung. Im Fall (W2) lässt sich aus diesen Resultaten strikte Positivität der diskreten und der kontinuierlichen Lösung folgern.

Im zweiten Teil des Beweises werden die Energieabschätzung, die Entropieabschätzung und die im ersten Teil bewiesenen Resultate zur Zeitkompaktheit benutzt, um in den Gleichungen (3.26) und (3.27) den Grenzübergang $\tau, h \to 0$ durchzuführen.

5.1.1 Zeitkompaktheit und Hölderstetigkeit der Lösungen

Notwendige Voraussetzung zum Beweis der gleichmäßigen Konvergenz von $U_{\tau h}$ gegen eine hölderstetige Funktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$ ist die Gültigkeit einer Energieabschätzung. Im folgenden Lemma überzeugen wir uns davon, dass die Voraussetzungen der Theoreme garantieren, dass die diskreten Lösungen $U_{\tau h}$ existieren und eine Energieabschätzung mit konstanter rechter Seite erfüllen:

Lemma 5.1.3 *(Gültigkeit der Energieabschätzung)*

Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gilt: Zu jeder zulässigen Triangulierung $\mathcal{T}_h$ von Ω und jeder Wahl von Zeitschrittweiten τ_k existiert eine Lösung $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Verfahrens 3.2.2. Diese erfüllt für alle $0 < \tilde{t} \leq T$ die Ungleichung

$$\int_\Omega |\partial_x U_{\tau h}(\tilde{t})|^2 + \int_\Omega \mathcal{I}_h W(U_{\tau h}(\tilde{t}),\cdot) + \int_0^T (M_\sigma(U_{\tau h}(t))\partial_x P_{\tau h}(t), \partial_x P_{\tau h}(t))dt \leq C_g. \tag{5.15}$$

Dabei ist die Konstante C_g unabhängig von den Parametern τ, h, σ. In der Situation von Theorem 5.1.2 ist C_g insbesondere unabhängig von der Wahl des Parameters ε_w.

Beweis : Wenn wir die in den Theoremen formulierten Bedingungen an W, Q, T, U^0 und Ω mit den Voraussetzungen der Existenzsätze 3.3.1, 3.3.2, 3.3.3 und Korollar 3.3.4 vergleichen, stellen wir fest, dass eine diskrete Lösung des Finite-Elemente-Verfahrens existiert und diese die Energieabschätzung 4.1.3 bzw. 4.1.4 erfüllt. Daraus folgt sofort Gleichung (5.15), denn die auf der rechten Seite der Energieabschätzungen auftretenden Terme

$$\int_\Omega |\partial_x U_{\tau h}(0)|^2, \quad \int_\Omega \mathcal{I}_h W(U_{\tau h}(0), \cdot), \quad \int_\Omega \mathcal{I}_h W^-_{,u}(U^0)$$

sind aufgrund der Bedingungen an u_0 und W unabhängig von τ, h, σ beschränkt. Unter den Voraussetzungen von Theorem 5.1.2 ist diese Schranke sogar unabhängig von der speziellen Wahl der Näherung W von w, da die Anfangsdaten $u_0 \geq c_0 > 0$ positiv sind. Gleichung (5.15) gilt für alle $\tilde{t} < T$, da die entsprechenden Abschätzungen für jeden Zeitpunkt $\mathfrak{t}_i \leq T$ hergeleitet werden können. □

Ungleichung (5.15) zeigt, dass die Menge der $\{\partial_x U_{\tau h}\}$ in $L^\infty(0, T; L^2(\Omega))$ gleichmäßig beschränkt ist. Mit Hilfe von Korollar 4.1.2 folgt, dass $\{U_{\tau h}\}$ gleichmäßig beschränkt in $L^\infty(0, T; H^1(\Omega))$ ist. Da $H^1(\Omega)$ in Raumdimension $d = 1$ in $L^\infty(\Omega)$ eingebettet ist, folgt unmittelbar das folgende Lemma:

Lemma 5.1.4 *(Einige Abschätzungen)*

Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gibt es Konstanten C_1, $U_{\max}$ und M, so dass für alle diskreten Lösungen $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Schemas 3.2.2 gilt:

$$\|U_{\tau h}\|_{L^\infty(0,T;H^1(\Omega))} \leq C_1, \tag{5.16}$$

$$|U_{\tau h}(t, x)| \leq U_{\max} \quad \forall (t, x) \in \Omega_T, \tag{5.17}$$

$$\|M_\sigma(U_{\tau h})\|_{L^\infty(\Omega_T)} \leq U^n_{\max} = M, \tag{5.18}$$

$$\|M_\sigma(U_{\tau h})\partial_x P_{\tau h}\|_{L^2(\Omega_T)} \leq \sqrt{MC_g}. \tag{5.19}$$

Darüberhinaus ist $H^1(\Omega)$ für $d = 1$ in $C^{\frac{1}{2}}(\Omega)$ eingebettet. Daher ist die Familie der $\{U_{\tau h}\}$ gleichmäßig beschränkt in $L^\infty(0, T; C^{\frac{1}{2}}(\Omega))$, woraus unmittelbar folgt:

Lemma 5.1.5 *(Hölderstetigkeit bzgl. der Ortsvariablen)*

Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 ist die Menge der diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Schemas 3.2.2 gleichmäßig hölderstetig im Ort zum Exponenten $\frac{1}{2}$, d.h. es gibt eine Konstante C_2, so dass für alle $t \in (0, T)$, alle $x, y \in \Omega$ und jede Wahl von τ, h gilt:

$$|U_{\tau h}(t, x) - U_{\tau h}(t, y)| \leq C_2 |x - y|^{\frac{1}{2}}. \tag{5.20}$$

Etwas schwieriger ist es, die gewünschte Hölderstetigkeit zum Exponenten $\frac{1}{8}$ in der Zeitvariablen zu zeigen. Da die $U_{\tau h}$ stückweise konstant in der Zeit definiert sind, können sie natürlich weder stetig noch hölderstetig sein. Stattdessen lässt sich gleichmäßige diskrete Hölderstetigkeit zeigen. Dieser Beweis erfolgt in den nächsten beiden Sätzen.

Satz 5.1.6 *Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gibt es eine Konstante $C_3 > 0$, so dass für alle diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Schemas 3.2.2 für alle $i, j \in \mathbb{N}$ mit $0 \leq i < j \leq K$ gilt:*

$$\|U_{\tau h}(\mathfrak{t}_j) - U_{\tau h}(\mathfrak{t}_i)\|_h^2 \leq C_3 \sqrt{\mathfrak{t}_j - \mathfrak{t}_i}. \tag{5.21}$$

Beweis : Wir wählen $\Theta = U_{\tau h}^{k+1} - U_{\tau h}^i$ in Gleichung (3.26) und erhalten:

$$\begin{aligned}(U_{\tau h}^{k+1} - U_{\tau h}^k, U_{\tau h}^{k+1} - U_{\tau h}^i)_h + \tau_{k+1}(M_\sigma(U_{\tau h}^{k+1})\partial_x P_{\tau h}^{k+1}, \partial_x U_{\tau h}^{k+1} - \partial_x U_{\tau h}^i) \\ = \tau_{k+1}(Q(U_{\tau h}^{k+1}), U_{\tau h}^{k+1} - U_{\tau h}^i)_h.\end{aligned} \tag{5.22}$$

Der erste Term lässt sich umformen zu:

$$(U_{\tau h}^{k+1} - U_{\tau h}^k, U_{\tau h}^{k+1} - U_{\tau h}^i)_h = \frac{1}{2}\|U_{\tau h}^{k+1} - U_{\tau h}^i\|_h^2 + \frac{1}{2}\|U_{\tau h}^{k+1} - U_{\tau h}^k\|_h^2 - \frac{1}{2}\|U_{\tau h}^k - U_{\tau h}^i\|_h^2.$$

Summation dieser Gleichung über $k = i, \ldots, j-1$ ergibt:

$$\begin{aligned}\sum_{k=i}^{j-1} &(U_{\tau h}^{k+1} - U_{\tau h}^k, U_{\tau h}^{k+1} - U_{\tau h}^i)_h \\ &= \frac{1}{2}\|U_{\tau h}^j - U_{\tau h}^i\|_h^2 + \frac{1}{2}\sum_{k=i}^{j-1}\|U_{\tau h}^{k+1} - U_{\tau h}^k\|_h^2 \geq \frac{1}{2}\|U_{\tau h}^j - U_{\tau h}^i\|_h^2.\end{aligned} \tag{5.23}$$

Der elliptische Term aus (5.22), summiert über $k = i, \ldots, j-1$, lässt sich abschätzen:

$$\begin{aligned}\sum_{k=i}^{j-1} &-\tau_{k+1}(M_\sigma(U_{\tau h}^{k+1})\partial_x P_{\tau h}^{k+1}, \partial_x U_{\tau h}^{k+1} - \partial_x U_{\tau h}^i) \\ &\leq \left| \int_{\mathfrak{t}_i}^{\mathfrak{t}_j} \int_\Omega M_\sigma(U_{\tau h})\partial_x P_{\tau h}(\partial_x U_{\tau h} - \partial_x U_{\tau h}(\mathfrak{t}_i)) \right| \\ &\leq \left| \int_{\mathfrak{t}_i}^{\mathfrak{t}_j} \left(\int_\Omega M_\sigma(U_{\tau h})^2 |\partial_x P_{\tau h}|^2 \right)^{\frac{1}{2}} \left(\int_\Omega |\partial_x U_{\tau h} - \partial_x U_{\tau h}(\mathfrak{t}_i)|^2 \right)^{\frac{1}{2}} dt \right| \\ &\leq \left(\int_{\mathfrak{t}_i}^{\mathfrak{t}_j} \int_\Omega M_\sigma(U_{\tau h})^2 |\partial_x P_{\tau h}|^2 \right)^{\frac{1}{2}} \left(\int_{\mathfrak{t}_i}^{\mathfrak{t}_j} \int_\Omega |\partial_x U_{\tau h} - \partial_x U_{\tau h}(\mathfrak{t}_i)|^2 \right)^{\frac{1}{2}}.\end{aligned}$$

Der erste Term hiervon lässt sich mit Hilfe der Ungleichung (5.19) gegen $\sqrt{MC_g}$ abschätzen, der zweite Term mit Hilfe der Ungleichung (5.16) gegen $(\int_{\mathfrak{t}_i}^{\mathfrak{t}_j} 4C_1^2)^{\frac{1}{2}}$, so dass sich insgesamt ergibt:

$$\sum_{k=i}^{j-1} -\tau_{k+1}(M_\sigma(U_{\tau h}^{k+1})\partial_x P_{\tau h}^{k+1}, \partial_x U_{\tau h}^{k+1} - \partial_x U_{\tau h}^i) \leq 2C_1 \sqrt{MC_g}\sqrt{\mathfrak{t}_j - \mathfrak{t}_i}. \tag{5.24}$$

Die rechte Seite von (5.22), summiert über $k = i, \ldots, j-1$, lässt sich mit Hilfe von (5.17) und der aus der Bedingung (Q1) bzw. (Q2) folgenden L^∞-Beschränktheit von Q abschätzen

durch

$$\sum_{k=i}^{j-1} \tau_{k+1}(Q(U_{\tau h}^{k+1}), U_{\tau h}^{k+1} - U_{\tau h}^{i})_h \leq \left| \int_{\mathfrak{t}_i}^{\mathfrak{t}_j} (Q(U_{\tau h}), U_{\tau h} - U_{\tau h}(t_i))_h \right|$$
$$\leq C_q \int_{\mathfrak{t}_i}^{\mathfrak{t}_j} |\mathcal{I}_h(U_{\tau h} - U_{\tau h}(\mathfrak{t}_i))| \leq C_q 2U_{\max} T^{\frac{1}{2}} \sqrt{\mathfrak{t}_j - \mathfrak{t}_i}. \quad (5.25)$$

Insgesamt ergibt sich also aus (5.23), (5.24) und (5.25) die behauptete Ungleichung:

$$\frac{1}{2}\|U_{\tau h}^j - U_{\tau h}^i\|_h^2 \leq 2C_1\sqrt{MC_g}\sqrt{\mathfrak{t}_j - \mathfrak{t}_i} + C_q 2U_{\max} T^{\frac{1}{2}} \sqrt{\mathfrak{t}_j - \mathfrak{t}_i}.$$

□

Satz 5.1.7 *(Diskrete gleichmäßige Hölderstetigkeit)*

Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gilt: Falls τ und h die Bedingung

$$h^4 < \min_{1 \leq k \leq K} \tau_k \quad (5.26)$$

erfüllen, so ist die Menge der diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Schemas 3.2.2 diskret gleichmäßig hölderstetig zum Exponent $\frac{1}{8}$ in der Zeit. Dies bedeutet, dass es eine Konstante $C_4 > 0$ unabhängig von τ, h gibt, so dass für alle $i, j \in \mathbb{N}$, $0 \leq i < j \leq K$ und alle $x \in \Omega$ gilt:

$$|U_{\tau h}(\mathfrak{t}_j, x) - U_{\tau h}(\mathfrak{t}_i, x)| \leq C_4(\mathfrak{t}_j - \mathfrak{t}_i)^{\frac{1}{8}}. \quad (5.27)$$

Beweis : Wir wählen o.B.d.A. ein $\delta > 0$ so, dass $[x, x+\delta) \subset \Omega$ gilt (ansonsten betrachten wir $(x-\delta, x]$). Dann ist

$$\begin{aligned} |U_{\tau h}^j(x) - U_{\tau h}^i(x)| &= \Big| \fint_x^{x+\delta} U_{\tau h}^j(x) - U_{\tau h}^j(y)\, dy \\ &\quad + \fint_x^{x+\delta} U_{\tau h}^j(y) - U_{\tau h}^i(y)\, dy \\ &\quad + \fint_x^{x+\delta} U_{\tau h}^i(y) - U_{\tau h}^i(x)\, dy \Big| \\ &=: |(a) + (b) + (c)|. \end{aligned}$$

Die Hölderstetigkeit in der Ortsvariablen liefert uns eine Abschätzung für den ersten und letzten Summanden:

$$|(a) + (c)| \leq 2C_2\delta^{\frac{1}{2}}.$$

Für den zweiten Summanden gilt (hier nutzen wir Satz 5.1.6 und Ungleichung (3.11) aus):

$$\begin{aligned} |(b)| &\leq \left(\int_x^{x+\delta} \frac{1}{\delta^2}\right)^{\frac{1}{2}} \left(\int_\Omega |U_{\tau h}^j(y) - U_{\tau h}^i(y)|^2\, dy\right)^{\frac{1}{2}} \\ &= \delta^{-\frac{1}{2}} \left(\|U_{\tau h}^j - U_{\tau h}^i\|_{L^2(\Omega)}^2 - \|U_{\tau h}^j - U_{\tau h}^i\|_h^2 + \|U_{\tau h}^j - U_{\tau h}^i\|_h^2\right)^{\frac{1}{2}} \\ &\leq \delta^{-\frac{1}{2}} \left(C_m h^2 \|\partial_x U_{\tau h}^j - \partial_x U_{\tau h}^i\|_{L^2(\Omega)}^2 + C_3\sqrt{\mathfrak{t}_j - \mathfrak{t}_i}\right)^{\frac{1}{2}} \\ &\leq \delta^{-\frac{1}{2}} \left(4C_m C_1^2 h^2 + C_3\sqrt{\mathfrak{t}_j - \mathfrak{t}_i}\right)^{\frac{1}{2}}. \end{aligned}$$

Wir wählen nun $\delta = (\mathfrak{t}_j - \mathfrak{t}_i)^{\frac{1}{4}}$ und nutzen Bedingung (5.26) aus. Dann erhalten wir:

$$|(b)| \leq (\mathfrak{t}_j - \mathfrak{t}_i)^{-\frac{1}{8}} \left(4C_m C_1^2 (\mathfrak{t}_j - \mathfrak{t}_i)^{\frac{1}{2}} + C_3 (\mathfrak{t}_j - \mathfrak{t}_i)^{\frac{1}{2}}\right)^{\frac{1}{2}} = \left(4C_m C_1^2 + C_3\right)^{\frac{1}{2}} (\mathfrak{t}_j - \mathfrak{t}_i)^{\frac{1}{8}},$$

und die Behauptung des Satzes ist bewiesen. □

Aus den diskret gleichmäßig hölderstetigen Funktionen $U_{\tau h} \in S^{-1,0}(V^h)$ lassen sich auf einfache Art und Weise gleichmäßig hölderstetige Funktionen bilden, indem man eine in t stückweise lineare Funktion $\tilde{U}_{\tau h}$ definiert durch

$$\tilde{U}_{\tau h}(t) := U_{\tau h}^k + \frac{t - \mathfrak{t}_k}{\mathfrak{t}_{k+1} - \mathfrak{t}_k}(U_{\tau h}^{k+1} - U_{\tau h}^k) \quad \text{für } t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}]. \tag{5.28}$$

Für diese gilt der folgende Satz:

Satz 5.1.8 *(Gleichmäßige Hölderstetigkeit der $\tilde{U}_{\tau h}$)*

Es gelten die Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 und die Bedingung (5.26). Dann sind die Funktionen $\tilde{U}_{\tau h}$ gleichmäßig beschränkt in $C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$.

Beweis : Seien $s \in (\mathfrak{t}_l, \mathfrak{t}_{l+1}]$, $t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}]$ und $s < t$. Dann gilt, falls $l \neq k$:

$$\begin{aligned}
|\tilde{U}_{\tau h}(t) - \tilde{U}_{\tau h}(s)| &\leq |U_{\tau h}^k - U_{\tau h}^{l+1}| + \frac{t - \mathfrak{t}_k}{\tau_{k+1}}|U_{\tau h}^{k+1} - U_{\tau h}^k| + \frac{\mathfrak{t}_{l+1} - s}{\tau_{l+1}}|U_{\tau h}^{l+1} - U_{\tau h}^l| \\
&\leq C_4(\mathfrak{t}_k - \mathfrak{t}_{l+1})^{\frac{1}{8}} + (t - \mathfrak{t}_k)C_4\tau_{k+1}^{-\frac{7}{8}} + (\mathfrak{t}_{l+1} - s)C_4\tau_{l+1}^{-\frac{7}{8}} \\
&\leq C_4(t-s)^{\frac{1}{8}} + (t-s)^{\frac{1}{8}}C_4\left(\frac{t - \mathfrak{t}_k}{\tau_{k+1}}\right)^{\frac{7}{8}} + (t-s)^{\frac{1}{8}}C_4\left(\frac{\mathfrak{t}_{l+1} - s}{\tau_{l+1}}\right)^{\frac{7}{8}} \\
&\leq 3C_4(t-s)^{\frac{1}{8}}.
\end{aligned}$$

Im Fall $l = k$ gilt:

$$|\tilde{U}_{\tau h}(t) - \tilde{U}_{\tau h}(s)| = \left|\frac{t-s}{\tau_{k+1}}(U_{\tau h}^{k+1} - U_{\tau h}^k)\right| \leq C_4(t-s)^{\frac{1}{8}}\left(\frac{t-s}{\tau_{k+1}}\right)^{\frac{7}{8}} \leq C_4(t-s)^{\frac{1}{8}}.$$

Damit ist $\tilde{U}_{\tau h}$ gleichmäßig hölderstetig in der Zeit zum Exponenten $\frac{1}{8}$ mit einer Hölderkonstanten $3C_4$. Die gleichmäßige Hölderstetigkeit der $\tilde{U}_{\tau h}$ im Ort folgt sofort aus der gleichmäßigen Hölderstetigkeit der $U_{\tau h}^k$ im Ort. □

Dank diesem Resultat sind wir nun in der Lage, die gleichmäßige Konvergenz von $U_{\tau h}$ gegen eine Funktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$ zu zeigen und damit eine Aussage der Theoreme zu beweisen.

Lemma 5.1.9 *(Gleichmäßige Konvergenz der $U_{\tau h}$)*

Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gilt: Es gibt eine Funktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$ und eine Folge von diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Schemas 3.2.2, so dass für $\tau, h \to 0$ gilt:

$$U_{\tau h} \to u \text{ gleichmäßig in } \Omega_T, \tag{5.29}$$

$$U_{\tau h}^- \to u \text{ gleichmäßig in } \Omega_T, \tag{5.30}$$

$$\tilde{U}_{\tau h} \to u \text{ in } C^{\alpha,\beta}(\Omega_T) \text{ für alle } \alpha < \tfrac{1}{8}, \beta < \tfrac{1}{2}. \tag{5.31}$$

Beweis : Sei also $U_{\tau h}$ eine Folge von diskreten Lösungen zu 3.2.2. Dann garantiert die in Satz 5.1.8 gezeigte gleichmäßige Beschränktheit für eine Teilfolge $\tau, h \to 0$ die Existenz eines schwachen Limes $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$. Man beachte, dass diese Teilfolge so gewählt sein muss, dass sie Bedingung (5.26) erfüllt. Der Satz von Arzela-Ascoli zeigt nun, dass, nach Übergang zu einer weiteren Teilfolge, $\tilde{U}_{\tau h}$ auch gleichmäßig in ganz Ω_T gegen u konvergiert. Nun können wir den folgenden Konvergenztrick anwenden (siehe Zeidler [47], Prop 21.57):

Seien $X \subset Y \subset Z$ Banachräume und (u_n) eine in X beschränkte und in Z konvergente Folge. Falls es eine Konstante $0 < \theta < 1$ gibt, so dass $\|u\|_Y \leq C\|u\|_X^{1-\theta}\|u\|_Z^{\theta}$ für alle $u \in X$ gilt, so folgt auch $u_n \to u$ in Y.

Dies wenden wir nun an auf $X = C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$, $Y = C^{\alpha,\beta}(\Omega_T)$, $\alpha < \frac{1}{8}, \beta < \frac{1}{2}$ und $Z = C^0(\Omega_T)$ und erhalten, dass $\tilde{U}_{\tau h} \to u$ in $C^{\alpha,\beta}(\Omega_T)$ konvergiert. Um die Aussagen (5.29) und (5.30) zu zeigen, benutzen wir, dass für $t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}]$ gilt:

$$|U_{\tau h}(t,x) - \tilde{U}_{\tau h}(t,x)| = \left(1 - \frac{t - \mathfrak{t}_k}{\tau_{k+1}}\right) |U_{\tau h}^{k+1} - U_{\tau h}^{k}| \leq C_4 \tau_{k+1}^{\frac{1}{8}}.$$

Damit folgt nun, dass

$$|U_{\tau h}(t,x) - u(t,x)| \leq |U_{\tau h}(t,x) - \tilde{U}_{\tau h}(t,x)| + |\tilde{U}_{\tau h}(t,x) - u(t,x)| \overset{\tau,h\to 0}{\longrightarrow} 0$$

gleichmäßig in ganz Ω_T gegen 0 konvergiert und dadurch auch

$$\begin{aligned} |U_{\tau h}^{-}(t,x) - u(t,x)| &= |U_{\tau h}^{k}(x) - u(t,x)| \\ &\leq |U_{\tau h}^{k}(x) - U_{\tau h}^{k+1}(x)| + |U_{\tau h}^{k+1}(x) - u(t,x)| \leq C_4 \tau_{k+1}^{\frac{1}{8}} + |U_{\tau h}(t,x) - u(t,x)| \overset{\tau,h\to 0}{\longrightarrow} 0 \end{aligned}$$

gleichmäßig für alle t, x konvergiert. □

Die bisher hergeleiteten Resultate gelten sowohl unter den Voraussetzungen von Theorem 5.1.1 als auch unter den Voraussetzungen von 5.1.2. Für letzteren Fall lässt sich strikte Positivität der diskreten Lösungen zeigen, womit auch im Fall (w2) die Gültigkeit der Abschätzung (5.33) sichergestellt ist.

Satz 5.1.10 *(Strikt positive Lösungen)*

In der Situation von Theorem 5.1.2 gibt es eine Konstante $\gamma > 0$, so dass für alle $\varepsilon_w < \gamma$, alle τ, h und alle dazugehörigen diskreten Lösungen $U_{\tau h}$ von Schema 3.2.2 gilt:

$$U_{\tau h}(t,x) \geq \gamma \quad \forall (t,x) \in \Omega_T. \tag{5.32}$$

Insbesondere gilt damit $W(U_{\tau h}(t,x),x) = w(U_{\tau h}(t,x),x)$ für alle $(t,x) \in \Omega_T$, und es gibt eine Konstante C, so dass $\sup_{(t,x)\in\Omega_T} W_{,u}^{\pm}(U_{\tau h}(t,x),x) \leq C$.

Beweis : Für $x \in \Omega_i$ gilt, falls $u \leq \min\{1, \frac{a_{i2}}{2a_{i1}}\}$:

$$w(u,x) = a_{i2}\left(1 - \tfrac{a_{i1}}{a_{i2}} u^{l_{i2}-l_{i1}}\right) u^{-l_{i2}} \geq \tfrac{a_{i2}}{2} u^{-l_{i2}}.$$

Also gilt für beliebige x, falls $u \leq \min\{\frac{a_{12}}{2a_{12}}, \frac{a_{22}}{2a_{21}}, 1\}$:

$$w(u,x) \geq \min\{\tfrac{a_{12}}{2}, \tfrac{a_{22}}{2}\} u^{-\min\{l_{12},l_{22}\}}.$$

Wir nehmen nun an, dass $U_{\tau h}(x_0,t_0) < \varepsilon_w$ sei. O.B.d.A. gelte $2\varepsilon_w \leq \min\{\frac{a_{12}}{2a_{12}}, \frac{a_{22}}{2a_{21}}, 1\}$. Dann gilt aufgrund der Hölderstetigkeit von $U_{\tau h}$:

$$U_{\tau h}(t_0,y) \leq 2\varepsilon_w \quad \forall y \in B_{(\varepsilon_w/C_2)^2}(x_0).$$

Mit Hilfe der Energieabschätzung können wir abschätzen:

$$\begin{aligned} C_g &\geq \int_\Omega \mathcal{I}_h W(U_{\tau h}(t_0,x),x)dx \\ &= \int_\Omega \left(\mathcal{I}_h W(U_{\tau h}(t_0,x),x) + C_w\right) dx - \int_\Omega C_w dx. \end{aligned}$$

Da W nach unten beschränkt ist, können wir eine Konstante C_w so groß wählen, dass der erste Integrand positiv ist. Also gilt

$$\begin{aligned} C_g + C_w|\Omega| &\geq \int_{B_{(\varepsilon_w/C_2)^2}(x_0)} \mathcal{I}_h W(U_{\tau h}(t_0,x),x)dx \\ &\geq \int_{B_{(\varepsilon_w/C_2)^2}(x_0)} \mathcal{I}_h W(2\varepsilon_w,x)dx \\ &\geq \left|B_{(\varepsilon_w/C_2)^2}(x_0)\right| \min\left\{\tfrac{a_{12}}{2}, \tfrac{a_{22}}{2}\right\} (2\varepsilon_w)^{-\min\{l_{12},l_{22}\}} \\ &\geq C\varepsilon_w^{2d-\min\{l_{12},l_{22}\}}. \end{aligned}$$

Somit gilt: $\varepsilon_w^{\min\{l_{12},l_{22}\}-2} \geq C$. Dies ist aber ein Widerspruch, da ε_w frei gewählt werden durfte. Also gibt es eine Konstante $\gamma > 0$, so dass für alle $\varepsilon_w \leq \gamma$ gilt: $U_{\tau h}(t,x) \geq \varepsilon_w$. Damit ist aber $W(U_{\tau h},\cdot) = w(U_{\tau h},\cdot)$, und deshalb ist die Lösung unabhängig von der Wahl von ε_w, woraus (5.32) folgt. □

Bemerkung : Der Beweis zeigt, dass die Größe der Konstanten γ abhängig ist von der Größe der Konstanten C_g und damit auch von δ_q. Ist q durch (2.44) gegeben und positiv, so ist dies unwichtig, da die durch (3.19) gegebene Funktion Q unabhängig von δ_q ist und δ_q damit beliebig gewählt werden kann. Beschreibt q Evaporation, so ist Q so zu wählen, dass $Q(u) = 0$ für $u \leq \delta_q$ gilt. Wir erkennen also, dass sich in diesem Fall Positivität nur zeigen lässt, weil auf dem Intervall $[0,\delta_q]$ keine Evaporation mehr stattfindet.

5.1.2 Konvergenz gegen eine schwache Lösung

Ziel dieses Abschnittes ist es, in den Gleichungen (3.26) und (3.27) mit $\tau, h \to 0$ zur Grenze überzugehen, und so Konvergenz gegen eine schwache, kontinuierliche Lösung zu zeigen. Dazu benötigen wir zusätzlich zur Energieabschätzung nun auch die Entropieabschätzung. Das folgende Lemma zeigt, dass diese sowohl unter den Voraussetzungen von Theorem 5.1.1 als auch unter den Voraussetzungen von Theorem 5.1.2 gültig ist.

Lemma 5.1.11 *(Gültigkeit der Entropieabschätzung)*
Unter den Voraussetzungen von Theorem 5.1.1 oder Theorem 5.1.2 gilt: Die Lösung $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Verfahrens 3.2.2 erfüllt für alle $0 < \tilde{t} \leq T$ die Ungleichung

$$\int_\Omega \mathcal{I}_h G_\sigma(U_{\tau h}(\tilde{t})) + \int_0^T \|\Delta_h U_{\tau h}(t)\|_h^2 dt + \int_0^T \|P_{\tau h}(t)\|_h^2 dt \leq C_e \tag{5.33}$$

mit einer von der Größe der Parameter τ, h, σ unabhängigen Konstante C_e. Unter den Voraussetzungen von Theorem 5.1.2 ist C_e außerdem unabhängig von der Wahl von ε_w.

Beweis : Die in den Theoremen vorausgesetzten Eigenschaften von W, Q, u_0, T und Ω stellen sicher, dass die Entropieabschätzung 4.2.1 bzw. 4.2.2 gilt. Daraus folgt die Gleichung (5.33), da erstens der Term

$$\int_\Omega \mathcal{I}_h G_\sigma(U^0)$$

aufgrund der Voraussetzungen an u_0 beschränkt ist, und zweitens die Terme

$$\sup_{(t,x)\in\Omega_T} W_{,u}^+(U_{\tau h}(t,x),x), \qquad \sup_{(t,x)\in\Omega_T} W_{,u}^-(U_{\tau h}(t,x),x)$$

unter den in Theorem 5.1.1 verlangten Bedingungen (W0') bzw. (W1') beschränkt sind. In der Situation von Theorem 5.1.2 gilt die in 5.1.10 bewiesene strikte Positivität der diskreten Lösung. Also sind die Suprema in diesem Fall sogar unabhängig von der Wahl der Näherung W. □

Die Lemmata des vorherigen Abschnittes und eine Anwendung des Prinzips der schwachen Kompaktheit beschränkter Mengen auf die Ungleichungen (5.15) und (5.33) zeigen uns, dass wir aus der Menge der diskreten Lösungen eine Teilfolge mit $\tau, h \to 0$ auswählen können, welche die folgenden Konvergenzresultate erfüllt:

$$U_{\tau h} \to u \text{ gleichmäßig in } \Omega_T, \tag{5.34}$$
$$U_{\tau h}^- \to u \text{ gleichmäßig in } \Omega_T, \tag{5.35}$$
$$\tilde{U}_{\tau h} \to u \text{ in } C^{\alpha,\beta} \text{ für } \alpha < \tfrac{1}{8}, \beta < \tfrac{1}{2}, \tag{5.36}$$
$$U_{\tau h} \rightharpoonup u \text{ schwach in } L^\infty(0,T;H^1(\Omega)), \tag{5.37}$$
$$P_{\tau h} \rightharpoonup p \text{ schwach in } L^2(\Omega_T), \tag{5.38}$$
$$-\Delta_h U_{\tau h} \rightharpoonup f \text{ schwach in } L^2(\Omega_T), \tag{5.39}$$
$$M_\sigma(U_{\tau h})\partial_x P_{\tau h} \rightharpoonup j \text{ schwach in } L^2(\Omega_T). \tag{5.40}$$

Dabei sind die Grenzwerte $u \in L^\infty(0,T;H^1(\Omega)) \cap C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$ und $j, p, f \in L^2(\Omega_T)$. Außerdem können wir mit Hilfe des Satzes von Lebesgue folgern:

$$\mathcal{I}_h Q(U_{\tau h}) \to Q(u) \text{ in } L^p(\Omega_T) \text{ für } 1 \leq p < \infty, \tag{5.41}$$
$$\mathcal{I}_h W_{,u}^+(U_{\tau h},\cdot) \to W_{,u}^+(u,\cdot) \text{ in } L^p(\Omega_T) \text{ für } 1 \leq p < \infty, \tag{5.42}$$
$$\mathcal{I}_h W_{,u}^-(U_{\tau h}^-,\cdot) \to W_{,u}^-(u,\cdot) \text{ in } L^p(\Omega_T) \text{ für } 1 \leq p < \infty. \tag{5.43}$$

Die gleichmäßige Konvergenz von $U_{\tau h}$ und $U^-_{\tau h}$ nämlich führt fast überall zu punktweiser Konvergenz von $\mathcal{I}_h W^+_{,u}(U_{\tau h}(t,x),x)$ und $\mathcal{I}_h W^-_{,u}(U^-_{\tau h}(t,x),x)$ gegen $W^+_{,u}(u(t,x),x)$ bzw. $W^-_{,u}(u(t,x),x)$. Außerdem sind die Funktionen $\mathcal{I}_h W^+_{,u}(U_{\tau h}(t,x),x)$ und $\mathcal{I}_h W^-_{,u}(U^-_{\tau h}(t,x),x)$ unter den Voraussetzungen (W0') oder (W1') global beschränkt, weshalb der Konvergenzsatz von Lebesgue anwendbar ist. Im Fall (W2) sind die Funktionen aufgrund der Positivität der Lösung beschränkt. Ebenso konvergiert $\mathcal{I}_h Q(U_{\tau h})$ punktweise gegen $Q(u)$ und Q ist global beschränkt. Der Satz von Lebesgue zeigt dann (5.41).

Aus (5.38), (5.39), (5.42) und (5.43) folgt unmittelbar, dass

$$f = p - W^+_{,u}(u,\cdot) - W^-_{,u}(u,\cdot). \tag{5.44}$$

Diese Konvergenzresultate wollen wir nun ausnutzen, um in (3.26) und (3.27) zur Grenze überzugehen.

Lemma 5.1.12 *(Grenzübergang in (3.26))*

Sei $U_{\tau h}, P_{\tau h}$ eine Folge von diskreten Lösungen des Finite-Elemente-Verfahrens 3.2.2, welche die Konvergenzaussagen (5.34)-(5.43) erfüllt. Dann gibt es eine schwache Ableitung $\partial_t u \in L^2(0,T;H^1(\Omega)')$, so dass

$$\int_0^T \langle \partial_t u\,,\phi\rangle_{H^1(\Omega)'\times H^1(\Omega)} + \int_{\Omega_T} j\partial_x\phi = \int_{\Omega_T} Q(u)\phi \tag{5.45}$$

für alle $\phi \in L^2(0,T;H^1(\Omega))$ gilt.

Beweis : Wir wählen eine Testfunktion $\theta \in C^1([0,T];C_0^\infty(\mathbb{R}))$ mit $\theta(T) \equiv 0$ und konstruieren dazu eine diskrete Testfunktion $\Theta_{\tau h} \in S^{-1,0}(V^h)$ durch

$$\Theta^k_{\tau h} = \mathcal{I}_h\theta(\mathfrak{t}_k), \quad k = 0,\dots,K.$$

Mit Hilfe der Interpolationsabschätzung $\|\partial_x\mathcal{I}_h\theta - \partial_x\theta\|_{0,p,\Omega} \le Ch|\partial_x\theta|_{1,p,\Omega}$ (siehe Ciarlet [11]) erkennt man, dass

$$\partial_x\Theta_{\tau h} \to \partial_x\theta \text{ in } L^2(\Omega_T) \tag{5.46}$$

konvergiert. Da die Ableitung von $\partial_t\theta$ beschränkt ist, ist auch der Differenzenquotient $\partial^-_\tau\Theta_{\tau h}$ beschränkt, denn eine Anwendung des Mittelwertsatzes zeigt für $t \in (\mathfrak{t}_k,\mathfrak{t}_{k+1}]$, dass $\partial^-_\tau\Theta_{\tau h}(t,x) = \mathcal{I}_h\partial_t\theta(\eta,x)$ für ein $\eta \in (\mathfrak{t}_k,\mathfrak{t}_{k+1})$ gilt. Dies zeigt auch, dass $\partial^-_\tau\Theta_{\tau h}$ punktweise gegen $\partial_t\theta$ konvergiert, und damit konvergiert wegen der Beschränktheit auch

$$\partial^-_\tau\Theta_{\tau h} \to \partial_t\theta \text{ in } L^2(\Omega_T). \tag{5.47}$$

Wir wählen nun $\Theta = \Theta^k_{\tau h}$ in (3.26) und summieren die Gleichung über $k = 0,\dots,K-1$:

$$\sum_{k=0}^{K-1}(U^{k+1}_{\tau h} - U^k_{\tau h}, \Theta^{k+1}_{\tau h})_h + \sum_{k=0}^{K-1}\tau_{k+1}(M_\sigma(U^{k+1}_{\tau h})\partial_x P^{k+1}_{\tau h}, \partial_x\Theta^{k+1}_{\tau h}) = \sum_{k=0}^{K-1}\tau_{k+1}(Q(U^{k+1}_{\tau h}),\Theta^{k+1}_{\tau h})_h.$$

Dies ist nach Umsortieren der ersten Summe äquivalent zu

$$-\sum_{k=0}^{K-1}(U_{\tau h}^k - U_{\tau h}^0, \Theta_{\tau h}^{k+1} - \Theta_{\tau h}^k)_h + (U_{\tau h}^K - U_{\tau h}^0, \Theta_{\tau h}^K)_h \\ + \sum_{k=0}^{K-1}\tau_{k+1}(M_\sigma(U_{\tau h}^{k+1})\partial_x P_{\tau h}^{k+1}, \partial_x \Theta_{\tau h}^{k+1}) = \sum_{k=0}^{K-1}\tau_{k+1}(Q(U_{\tau h}^{k+1}), \Theta_{\tau h}^{k+1})_h.$$

Da $\Theta_{\tau h}^K \equiv 0$ ist, entfällt der zweite Term in dieser Gleichung und wir erhalten

$$-\int_0^T (U_{\tau h}^- - U_{\tau h}^0, \partial_\tau^- \Theta_{\tau h})_h + \int_0^T (M_\sigma(U_{\tau h})\partial_x P_{\tau h}, \partial_x \Theta_{\tau h}) = \int_0^T (Q(U_{\tau h}), \Theta_{\tau h})_h.$$

Wir betrachten nun den Grenzübergang $\tau, h \to 0$. Dann folgt aus (5.46) und (5.40):

$$\int_0^T (M_\sigma(U_{\tau h})\partial_x P_{\tau h}, \partial_x \Theta_{\tau h}) \to \int_0^T (j, \partial_x \theta). \tag{5.48}$$

Aus (3.11) folgern wir:

$$\int_0^T |(U_{\tau h}^- - U_{\tau h}^0, \partial_\tau^- \Theta_{\tau h})_h - (U_{\tau h}^- - U_{\tau h}^0, \partial_\tau^- \Theta_{\tau h})| \\ \le Ch \int_0^T \|\partial_x U_{\tau h}^- - \partial_x U_{\tau h}^0\|_{L^2(\Omega)} \|\partial_\tau^- \Theta_{\tau h}\|_{L^2(\Omega)} \overset{\tau,h\to 0}{\longrightarrow} 0.$$

Mit Hilfe der Konvergenz von $U_{\tau h}^- \to u$ in $L^2(\Omega_T)$ folgt daraus:

$$\int_0^T (U_{\tau h}^- - U_{\tau h}^0, \partial_\tau^- \Theta_{\tau h})_h \to \int_0^T (u - u_0, \partial_t \theta). \tag{5.49}$$

Ebenso folgt aus (3.11), (5.41) und der starken L^2-Konvergenz von $\Theta_{\tau h}$ gegen θ, dass

$$\int_0^T (Q(U_{\tau h}), \Theta_{\tau h})_h \to \int_0^T (Q(u), \theta). \tag{5.50}$$

Kombiniert man (5.48), (5.49) und (5.50), so erhält man

$$-\int_{\Omega_T} (u - u_0)\partial_t \theta + \int_{\Omega_T} j \partial_x \theta = \int_{\Omega_T} Q(u)\theta. \tag{5.51}$$

Dies gilt für alle Testfunktionen $\theta \in C^1([0,T]; C_0^\infty(\mathbb{R}))$ mit $\theta(T) = 0$. Gleichung (5.2) folgt aus dieser Aussage wie folgt (ein analoger Beweis für den Fall $Q = 0$ findet sich in [18]): Da die Menge

$$\{\psi \in C^1([0,T]; C_0^\infty(\mathbb{R})) : \psi(T) = 0\}$$

dicht in

$$\mathcal{A} := \{\psi \in L^2(0,T; H^1(\Omega)) \cap W^{1,1}(0,T; L^2(\Omega)) \cap C(0,T; L^2(\Omega)) : \psi(T) = 0\}$$

liegt, gilt (5.51) auch für alle $\theta \in \mathcal{A}$. Nun definieren wir die beschränkten, also auch stetigen, linearen Funktionale $K_1 : L^2(0,T; H^1(\Omega)) \to \mathbb{R}$ durch

$$K_1(\phi) := \int_{\Omega_T} j \partial_x \phi - Q(u)\phi$$

und $K_2 : \mathcal{A} \to \mathbb{R}$ durch

$$K_2(\phi) := \int_{\Omega_T} (u - u_0)\partial_t \phi.$$

Für $\phi \in \mathcal{A}$ gilt nun $K_1(\phi) = K_2(\phi)$. Da $\mathcal{A}$ dichte Teilmenge von $L^2(0,T;H^1(\Omega))$ ist, kann $K_2(\phi)$ zu einem Funktional $\bar{K_2} \in L^2(0,T;H^1(\Omega)')$ erweitert werden. Wegen der Stetigkeit von K_1 und K_2 gilt:

$$K_1 = \bar{K_2} \text{ in } L^2(0,T;H^1(\Omega)').$$

Wir nennen nun $\bar{K_2} = -\partial_t u$. Also gilt

$$\int_0^T \langle \partial_t u\,, \phi \rangle_{H^1(\Omega)' \times H^1(\Omega)} + \int_{\Omega_T} j \partial_x \phi = \int_{\Omega_T} Q(u)\phi.$$

□

Lemma 5.1.13 *(Grenzübergang in (3.27))*

Sei $U_{\tau h}, P_{\tau h}$ eine Folge von diskreten Lösungen des Finite-Elemente-Verfahrens 3.2.2, welche die Konvergenzaussagen (5.34)-(5.43) erfüllt. Dann gilt $u \in L^2(0,T;H^2(\Omega))$ und

$$p = -\partial_{xx} u + W^+_{,u}(u,\cdot) + W^-_{,u}(u,\cdot). \tag{5.52}$$

Beweis : Wir wählen eine Testfunktion $\psi \in L^2(0,T;H^1(\Omega))$ und konstruieren eine diskrete Testfunktion $\Psi_{\tau h} \in S^{-1,0}(V^h)$ durch

$$\Psi^k_{\tau h} = \mathcal{R}_h \psi(\mathfrak{t}_k), \quad k = 0, \dots, K,$$

wobei $\mathcal{R}_h$ die Ritzprojektion ist, welche durch

$$(\partial_x \mathcal{R}_h v, \partial_x \varphi_h) = (\partial_x v, \partial_x \varphi_h) \quad \forall \varphi_h \in V^h$$

definiert ist. Die Ritzprojektion erfüllt die Abschätzungen:

$$\|\mathcal{R}_h v - v\|_{0,2} \le c h^s |v|_{s,2},$$
$$\|\mathcal{R}_h v - v\|_{1,2} \le c h^{s-1} |v|_{s,2}.$$

Also konvergiert

$$\Psi_{\tau h} \to \psi \text{ in } L^2(0,T;H^1(\Omega)). \tag{5.53}$$

Nun wählen wir $\Psi = \Psi^k_{\tau h}$ in (3.27), multiplizieren mit τ_{k+1}, und summieren über $k = 0, \dots, K-1$:

$$\sum_{k=0}^{K-1} \tau_{k+1} (\Psi^{k+1}_{\tau h}, P^{k+1}_{\tau h})_h = \sum_{k=0}^{K-1} \tau_{k+1} (\partial_x U^{k+1}_{\tau h}, \partial_x \Psi^{k+1}_{\tau h}) + \sum_{k=0}^{K-1} \tau_{k+1} (\mathcal{I}_h W^+_{,u}(U^{k+1}_{\tau h}, \cdot), \Psi^{k+1}_{\tau h})_h + \sum_{k=0}^{K-1} \tau_{k+1} (\mathcal{I}_h W^-_{,u}(U^{k}_{\tau h}, \cdot), \Psi^{k+1}_{\tau h})_h.$$

Dies ist äquivalent zu

$$\int_0^T (\partial_x U_{\tau h}, \partial_x \Psi_{\tau h}) = \int_0^T (P_{\tau h} + \mathcal{I}_h W^+_{,u}(U_{\tau h}, \cdot) + \mathcal{I}_h W^-_{,u}(U^-_{\tau h}, \cdot), \Psi_{\tau h})_h.$$

Nun konvergiert die linke Seite dieser Gleichung wegen (5.37) und (5.53), und die rechte Seite der Gleichung konvergiert wegen (5.39) und (5.53). Wir erhalten damit die Gleichung

$$\int_0^T (\partial_x u, \partial_x \psi) = \int_0^T (f, \psi) \quad \forall \psi \in L^2(0,T; H^1(\Omega)).$$

Also ist gemäß Definition der schwachen Ableitung $f = -\partial_{xx} u \in L^2(\Omega_T)$ und damit ist $u \in L^2(0,T; H^2(\Omega))$. Nach (5.44) ist dann also

$$-\partial_{xx} u = p - W^+_{,u}(u, \cdot) - W^-_{,u}(u, \cdot),$$

und das Lemma ist damit bewiesen. □

Lemma 5.1.14 *(Identifikation von j mit $m(u)\partial_x p$)*

Sei $U_{\tau h}, P_{\tau h}$ eine Folge von diskreten Lösungen des Finite-Elemente-Verfahrens 3.2.2, welche die Konvergenzaussagen (5.34)-(5.43) erfüllt. Dann konvergiert $\partial_x P_{\tau h}$ für alle $\delta > 0$ schwach gegen $\partial_x p$ in $L^2([u > \delta])$. Außerdem gilt:

$$\int_{\Omega_T} j \partial_x \phi = \int_{[u>0]} m(u) \partial_x p \partial_x \phi \tag{5.54}$$

für alle $\phi \in L^2(0,T; H^1(\Omega))$.

Beweis : Wir zeigen zunächst einmal, dass $M_\sigma(U_{\tau h}(t,x))$ gleichmäßig auf ganz Ω_T gegen $m(u(t,x))$ konvergiert.

Seien dazu $t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}]$ und $x \in [\mathfrak{x}_l, \mathfrak{x}_{l+1}]$. Dann zeigt eine Anwendung des Mittelwertsatzes, dass es einen Wert $\xi \in [U^{k+1}_{\tau h}(\mathfrak{x}_l), U^{k+1}_{\tau h}(\mathfrak{x}_{l+1})]$ gibt mit

$$M_\sigma(U_{\tau h}(t,x)) = M_\sigma(U^{k+1}_{\tau h}(x)) = m_\sigma(\xi).$$

Also gilt

$$\begin{aligned} &|m(u(t,x)) - M_\sigma(U_{\tau h}(t,x))| \\ &\qquad \le |m(u(t,x)) - m(\xi)| + |m(\xi) - m_\sigma(\xi)| \\ &\qquad \le \sup_{|s| \le U_{\max}} |m'(s)||u(t,x) - \xi| + \left\{ \begin{array}{ll} \sigma^n & \text{falls } n \ge 1 \\ \sigma\varepsilon^n & \text{falls } 0 < n < 1 \end{array} \right\} \overset{\sigma,\tau,h \to 0}{\longrightarrow} 0, \end{aligned}$$

wobei ε das in Satz 4.2.3 gewählte ε ist, welches für $\sigma, \tau, h \to 0$ beliebig klein wird.

Um nun j mit $m(u)\partial_x p$ zu identifizieren, unterteilen wir Ω_T in $[u > \delta]$ und $[u \le \delta]$. Wegen der gleichmäßigen Konvergenz der $U_{\tau h}$ können wir folgern, dass es τ_0, h_0 klein genug gibt, so dass für alle $\tau < \tau_0, h < h_0$ gilt: $U_{\tau h} > \frac{\delta}{2}$ auf $[u > \delta]$ und $U_{\tau h} < 2\delta$ auf $[u \le \delta]$.

Nun gilt auf $[u \leq \delta]$:

$$\begin{aligned}\int_{[u\leq\delta]} & M_\sigma(U_{\tau h})\partial_x P_{\tau h}\partial_x\Theta_{\tau h} \\ &\leq \|\sqrt{M_\sigma(U_{\tau h})}\|_{L^\infty([u\leq\delta])}\left(\int_{[u\leq\delta]} M_\sigma(U_{\tau h})|\partial_x P_{\tau h}|^2\right)^{\frac{1}{2}}\|\Theta_{\tau h}\|_{L^2(0,T;H^1(\Omega))} \\ &\leq (2\delta)^{\frac{n}{2}}\sqrt{C_g}\|\Theta_{\tau h}\|_{L^2(0,T;H^1(\Omega))}.\end{aligned} \tag{5.55}$$

Auf $[u > \delta]$ gilt die folgende Abschätzung:

$$\int_{[u>\delta]}\left(\frac{\delta}{2}\right)^n |\partial_x P_{\tau h}|^2 \leq \int_{[u>\delta]} M_\sigma(U_{\tau h})|\partial_x P_{\tau h}|^2 \leq C_g,$$

woraus wir folgern, dass zumindest eine Teilfolge von $\partial_x P_{\tau h}$ schwach in $L^2([u > \delta])$ konvergiert. Durch (5.38) können wir diesen schwachen Limes mit $\partial_x p$ identifizieren, und wir erkennen, dass sogar für die ganze Folge gilt:

$$\partial_x P_{\tau h} \rightharpoonup \partial_x p \text{ schwach in } L^2([u > \delta]).$$

Daraus wiederum folgt, dass

$$\int_{[u>\delta]} M_\sigma(U_{\tau h})\partial_x P_{\tau h}\partial_x\Theta_{\tau h} \to \int_{[u>\delta]} m(u)\partial_x p\partial_x\theta \tag{5.56}$$

für alle $\delta > 0$ konvergiert. (5.55) und (5.56) zusammen ergeben für $\delta \to 0$, dass

$$\int_{\Omega_T} M_\sigma(U_{\tau h})\partial_x P_{\tau h}\partial_x\Theta_{\tau h} \to \int_{[u>0]} m(u)\partial_x p\partial_x\theta,$$

woraus im Zusammenspiel mit (5.48) die Behauptung folgt. □

Lemma 5.1.15 *(Grenzübergang in den Abschätzungen)*

Sei $U_{\tau h}, P_{\tau h}$ eine Folge von diskreten Lösungen des Finite-Elemente-Verfahrens 3.2.2, welche die Konvergenzaussagen (5.34)-(5.43) erfüllt. Dann erfüllt die Lösung u für fast alle $t \in (0, T)$ die Ungleichung

$$\int_\Omega |\partial_x u(t)|^2 + \int_\Omega W(u(t), \cdot) + \int_\Omega G(u(t)) \leq C_e + C_g. \tag{5.57}$$

Beweis : Ungleichung (5.15) zeigt, dass $\|\partial_x U_{\tau h}(t)\|_{L^2(\Omega)}$ gleichmäßig beschränkt ist. Daher gibt es für alle t eine Funktion $q_t \in L^2(\Omega)$ und eine Teilfolge, so dass $\partial_x U_{\tau h}(t) \rightharpoonup q_t$ konvergiert. Da aber $U_{\tau h}(t, \cdot)$ für alle t gleichmäßig gegen $u(t, \cdot)$ konvergiert, können wir q_t durch

$$\int_\Omega \partial_x U_{\tau h}(t, \cdot)\phi = -\int_\Omega U_{\tau h}(t, \cdot)\partial_x\phi \to -\int_\Omega u(t, \cdot)\partial_x\phi \quad \forall\phi \in C_0^\infty(\Omega)$$

und

$$\int_\Omega \partial_x U_{\tau h}(t,\cdot)\phi \to \int_\Omega q_t \phi \quad \forall \phi \in C_0^\infty(\Omega)$$

mit $\partial_x u(t)$ identifizieren und daher gilt sogar für die ganze Folge:

$$\partial_x U_{\tau h}(t,\cdot) \rightharpoonup \partial_x u(t,\cdot) \text{ schwach in } L^2(\Omega). \tag{5.58}$$

Aufgrund der Unterhalbstetigkeit der $L^2(\Omega)$-Norm und der Energieabschätzung folgt

$$\int_\Omega |\partial_x u(t,\cdot)|^2 \le \liminf_{\tau,h\to 0} \int_\Omega |\partial_x U_{\tau h}(t)|^2. \tag{5.59}$$

Da das Potential $W(U_{\tau h}(t),\cdot)$ nach unten beschränkt ist und wegen der gleichmäßigen Konvergenz von $U_{\tau h}(t)$ punktweise gegen $W(u(t),\cdot)$ konvergiert, gilt mit Hilfe der Energieabschätzung und dem Lemma von Fatou:

$$\int_\Omega W(u(t),\cdot) = \int_\Omega \liminf_{\tau,h\to 0} W(U_{\tau h}(t),\cdot) \le \liminf_{\tau,h\to 0} \int_\Omega W(U_{\tau h}(t),\cdot) \le C_g. \tag{5.60}$$

Da $G_\sigma(U_{\tau h}(t))$ positiv und durch die Entropieabschätzung beschränkt ist, ist $G(u(t)) = \liminf_{\sigma,\tau,h\to 0} G_\sigma(U_{\tau h}(t))$ nach dem Lemma von Fatou integrabel und es gilt:

$$\int_\Omega G(u(t)) = \int_\Omega \liminf_{\tau,h\to 0} G_\sigma(U_{\tau h}(t)) \le \liminf_{\tau,h\to 0} \int_\Omega G_\sigma(U_{\tau h}(t)) \le C_e. \tag{5.61}$$

Kombination von (5.59), (5.60) und (5.61) ergibt die Behauptung. □

Durch diese Lemmata ist Theorem 5.1.1 nun vollständig bewiesen: Lemma 5.1.9 zeigt die Existenz einer Grenzfunktion $u \in C^{\frac{1}{8},\frac{1}{2}}(\Omega_T)$ und die gleichmäßige Konvergenz (5.5). Lemma 5.1.12 zeigt die Existenz der schwachen Zeitableitung $\partial_t u$ und zusammen mit der Identifikation $j = m(u)\partial_x p$ aus Lemma 5.1.14 die schwache Differentialgleichung (5.2). Lemma 5.1.13 zeigt die $L^2(0,T;H^2(\Omega))$-Regularität von u und die Gleichung (5.3), woraus zusammen mit (5.39) und (5.44) die Konvergenz der zweiten Ableitung (5.6) folgt. (5.7) ist durch (5.38) bewiesen und (5.8) wird in Lemma 5.1.14 bewiesen. Die stetige Annahme der Anfangsdaten ist eine direkte Folgerung aus Hölderstetigkeit und gleichmäßiger Konvergenz von u. Die Nichtnegativität von u folgt aus Satz 4.2.3 und der gleichmäßigen Konvergenz der diskreten Lösungen. Schließlich zeigt Lemma 5.1.15 noch die Gültigkeit der Ungleichung (5.4).

Theorem 5.1.2 folgt aus Theorem 5.1.1 und der strikten Positivität der $U_{\tau h}$. Da wegen der gleichmäßigen Konvergenz damit auch u strikt positiv ist, gilt $\Omega_T = [u \ge \gamma]$. Daher vereinfachen sich die Konvergenzaussagen (5.7) und (5.8) zu (5.14). Außerdem kann W durch w ersetzt werden, da beide Funktionen auf $[u \ge \gamma]$ identisch sind, falls $\varepsilon_w < \gamma$ gewählt wird.

5.2 Konvergenz in Raumdimension $d = 2$

In Raumdimension $d = 2$ kann selbst für den Fall eines Potentials vom Typ (w2) keine Positivität der Lösung mehr gezeigt werden. Daher ähneln die Resultate des folgenden Theorems denen von Theorem 5.1.1, es gibt aber einen gravierenden Unterschied: u ist nicht mehr hölderstetig in Bezug auf die Zeitvariable, die Konvergenz in $C^{\alpha,\beta}(\Omega_T)$ wird durch Konvergenz in $L^2(0,T;C^\beta(\Omega))$ ersetzt:

Theorem 5.2.1 *(Konvergenz gegen eine schwache Lösung)*

Das Gebiet $\Omega \subset \mathbb{R}^2$ sei konvex und lasse sich durch eine rechtwinklige Triangulierung unterteilen, es sei $T \in \mathbb{R}$ und die Anfangswerte u_0 erfüllen die Bedingungen

$$u_0 \in H^1(\Omega) \cap C^0(\overline{\Omega}) \text{ mit } \begin{cases} u_0 \geq 0 & \text{falls } 0 < n < 2, \\ u_0 \geq c_0 > 0 & \text{falls } n \geq 2. \end{cases} \tag{5.62}$$

Die rechte Seite Q erfülle eine der drei Bedingungen (Q0), (Q1) oder (Q2). Im Fall (Q0) gelte eine der Bedingungen (W0'), (W1') oder (W2). Im Fall (Q1) gelte (W2) und $n > 1$, im Fall (Q2) gelte (W2) und $T < \frac{1}{C_q} \fint_\Omega U^0$. Wir nehmen außerdem an, dass die Zeitschrittweite $\tau_k = \tau$ konstant sei.

Seien $2 < r < \infty$ und $\beta < 1$. Dann gibt es eine nichtnegative Funktion $u \in L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;W^{1,r}(\Omega)) \cap L^2(0,T;C^\beta(\Omega)) \cap C(0,T;L^2(\Omega))$ mit den schwachen Ableitungen $\partial_t u \in L^2(0,T;W^{1,r}(\Omega)')$ und $\Delta u \in L^2(\Omega_T)$ und eine Funktion $p \in L^2(\Omega_T)$ mit $\nabla p(t) \in L^2([u(t) > \delta])$ für fast alle $t \in (0,T)$ und alle $\delta > 0$, welche die Gleichung

$$\int_0^T \langle \partial_t u\,, \psi \rangle_{W^{1,r}(\Omega)' \times W^{1,r}(\Omega)} + \int_{[u>0]} m(u) \nabla p \nabla \psi = \int_{\Omega_T} Q(u) \psi \tag{5.63}$$

für alle $\psi \in L^2(0,T;W^{1,r}(\Omega))$ erfüllen. Ferner gilt

$$p = -\Delta u + W_{,u}(u,\cdot), \tag{5.64}$$

$$\lim_{t \to 0} u(t,\cdot) = u_0(\cdot) \text{ in } L^2(\Omega), \tag{5.65}$$

und die Lösung u erfüllt für fast alle $t \in (0,T)$ die Ungleichung

$$\int_\Omega |\nabla u(t)|^2 + \int_\Omega W(u(t),\cdot) + \int_\Omega G(u(t)) \leq C. \tag{5.66}$$

Außerdem gibt es eine Folge diskreter Lösungen $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Schemas 3.2.2, so dass für $\tau, h \to 0$ gilt:

$$U_{\tau h} \to u \text{ stark in } L^2(0,T;C^\beta(\Omega)), \tag{5.67}$$

$$\Delta_h U_{\tau h} \rightharpoonup \Delta u \text{ schwach in } L^2(\Omega_T), \tag{5.68}$$

$$P_{\tau h} \rightharpoonup p \text{ schwach in } L^2(\Omega_T), \tag{5.69}$$

$$\nabla P_{\tau h}(t) \rightharpoonup \nabla p(t) \text{ schwach in } L^2([u(t) > \delta]) \text{ für fast alle } t \in (0,T) \text{ und alle } \delta > 0. \tag{5.70}$$

Bevor wir den Beweis des Theorems 5.2.1 beginnen, stellen wir fest, dass die Voraussetzungen des Theorems die Existenz von diskreten Lösungen und die Gültigkeit von Energie- und Entropieabschätzung sicherstellen:

Lemma 5.2.2 *(Gültigkeit von Energie- und Entropieabschätzung)*

Unter den Voraussetzungen von Theorem 5.2.1 gilt: Zu jeder zulässigen und rechtwinkligen Triangulierung Ω und jeder Wahl einer Zeitschrittweite τ existiert eine Lösung $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Verfahrens 3.2.2. Diese erfüllt für alle $0 < \tilde{t} \leq T$ die Ungleichungen

$$\int_\Omega |\nabla U_{\tau h}(\tilde{t})|^2 + \int_\Omega \mathcal{I}_h W(U_{\tau h}(\tilde{t}),\cdot) + \int_0^T (M_\sigma(U_{\tau h}(t)) \nabla P_{\tau h}(t), \nabla P_{\tau h}(t)) dt \leq C_g, \tag{5.71}$$

$$\int_\Omega \mathcal{I}_h G_\sigma(U_{\tau h}(\tilde{t})) + \int_0^T \|\Delta_h U_{\tau h}(t)\|_h^2 dt + \int_0^T \|P_{\tau h}(t)\|_h^2 dt \leq C_e \tag{5.72}$$

mit Konstanten C_g und C_e unabhängig von der Wahl von τ, h, σ. Außerdem ist die diskrete Lösung nichtnegativ im Sinne von Satz 4.2.3.

Beweis : Ein Abgleich der Voraussetzungen an Q, W, Ω, u_0 und T in Theorem 5.2.1 mit den Voraussetzungen der Existenzsätze 3.3.1, 3.3.2, 3.3.3 und Korollar 3.3.4 zeigt, dass in allen Fällen eine diskrete Lösung existiert. Diese erfüllt die in Kapitel 4 bewiesenen Energie- und Entropieabschätzungen. Da die Bedingung (5.62) an die Anfangsdaten u_0 sicherstellt, dass die Terme

$$\int_\Omega |\nabla U_{\tau h}(0)|^2, \quad \int_\Omega \mathcal{I}_h W(U_{\tau h}(0), \cdot), \quad \int_\Omega \mathcal{I}_h G_\sigma(U_{\tau h}(0))$$

beschränkt sind, und zudem (W0') bzw. (W1') bzw. (W2) garantieren, dass

$$\sup_{(t,x)\in\Omega_T} W^+_{,u}(U_{\tau h}(t,x), x), \qquad \sup_{(t,x)\in\Omega_T} W^-_{,u}(U_{\tau h}(t,x), x)$$

gleichmäßig beschränkt sind, gelten die Ungleichungen (5.71) und (5.72). Außerdem gilt Satz 4.2.3 unter den Voraussetzungen des Theorems. □

Der Beweis des Theorems gliedert sich in zwei Teile. Im ersten Teil des Beweises werden Energie- und Entropieabschätzung genutzt, um ein Resultat zur Zeitkompaktheit zu zeigen. Im zweiten Teil kann mit Hilfe dieses Resultates in den Gleichungen (3.26) und (3.27) mit $\tau, h \to 0$ zur Grenze übergegangen werden.

5.2.1 Nikol'skii-Abschätzung und Zeitkompaktheit

In zwei Raumdimensionen können wir nicht mehr darauf hoffen, gleichmäßige Konvergenz von $U_{\tau h}$ gegen u zeigen zu können. Zwar zeigt uns die Energie-Abschätzung in Kombination mit Korollar 4.1.2 gleichmäßige Beschränktheit der diskreten Lösungen in $L^\infty(0,T;H^1(\Omega))$, eine Einbettung von $H^1(\Omega)$ nach $C^{\frac{1}{2}}(\Omega)$ existiert aber nicht mehr. Ersetzt wird die gleichmäßige Konvergenz durch starke Konvergenz in $L^2(0,T;C^\beta(\Omega))$. Um diese zu beweisen, zeigen wir zunächst einmal mit Hilfe der Entropieabschätzung, dass die Menge der $U_{\tau h}$ gleichmäßig beschränkt in $L^2(0,T;C^\beta(\Omega))$ ist. Dies geschieht in folgendem Lemma:

Lemma 5.2.3 *(Beschränktheit in $L^2(0,T;C^\beta(\Omega))$)*

Unter den Voraussetzungen von Theorem 5.2.1 ist die Menge der diskreten Lösungen $U_{\tau h}$ von Schema 3.2.2 für $1 < p < \infty$ und $0 < \beta < 1$ gleichmäßig beschränkt in $L^2(0,T;W^{1,p}(\Omega))$ und $L^2(0,T;C^\beta(\Omega))$.

Beweis : Um gleichmäßige Beschränktheit in $L^2(0,T;W^{1,p}(\Omega))$ zu zeigen, genügt es zu zeigen, dass es eine Konstante C_p gibt, so dass für alle $U_{\tau h}$ gilt:

$$\int_0^T \|\nabla U_{\tau h}\|^2_{0,p,\Omega} \le C_p, \tag{5.73}$$

denn die Energieabschätzung und die Einbettung $H^1(\Omega) \to L^p(\Omega)$ zeigen bereits die Beschränktheit von $\|U_{\tau h}\|_{L^2(0,T;L^p(\Omega))}$. Daraus folgt dann mit Hilfe der Einbettung $W^{1,p}(\Omega) \to C^\beta(\Omega)$, welche für $\beta < 1 - \frac{2}{p}$ gültig ist, auch die Beschränktheit in $L^2(0,T;C^\beta(\Omega))$.

Um (5.73) zu zeigen, nutzen wir das folgende auf konvexen Gebieten Ω gültige Regularitätsresultat für diskrete Funktionen (siehe Grün [19], Theorem 6.1):

Falls $F, U \in V^h$, $\int_\Omega U = 0$ und $(\nabla U, \nabla \Psi) = (F, \Psi)_h$ für alle $\Psi \in V^h$ gilt, so ist $\|\nabla U\|_{L^p(\Omega)} \le C\|F\|_{L^2(\Omega)}$ für alle $p < \infty$.

Nun definieren wir $\alpha(t) = \fint_\Omega U_{\tau h}(t)dx$, $Z_{\tau h}(t) = U_{\tau h}(t) - \alpha(t)$ und sehen, dass $\int_\Omega Z_{\tau h}(t) = 0$ und damit auch

$$(\nabla Z_{\tau h}(t), \nabla \Psi) = (\nabla U_{\tau h}(t), \nabla \Psi) = (P_{\tau h}(t) - \mathcal{I}_h W^+_{,u}(U_{\tau h}(t), \cdot) - \mathcal{I}_h W^-_{,u}(U^-_{\tau h}(t), \cdot), \Psi)_h.$$

Also gilt

$$\|\nabla Z_{\tau h}(t)\|_{L^p(\Omega)} \le C\|F_{\tau h}(t)\|_{L^2(\Omega)}$$

mit $F_{\tau h}(t) = P_{\tau h}(t) - \mathcal{I}_h W^+_{,u}(U_{\tau h}(t), \cdot) - \mathcal{I}_h W^-_{,u}(U^-_{\tau h}(t), \cdot)$ und damit

$$\int_0^T \|\nabla U_{\tau h}\|^2_{L^p(\Omega)} \le C \int_0^T \|F_{\tau h}\|^2_{L^2(\Omega)} \le C \int_0^T \|F_{\tau h}\|^2_h.$$

Wegen der Entropieabschätzung (5.72) lässt sich die rechte Seite nun durch C_e abschätzen, und daraus folgt die Aussage des Satzes. □

Da $H^1(\Omega)$ für Raumdimension $d = 2$ nur noch in $L^p(\Omega)$ mit $p < \infty$ und nicht mehr in $L^\infty(\Omega)$ einbettet ist, existiert a priori keine obere Schranke mehr für die $U_{\tau h}$. Daher lassen sich die Aussagen von Lemma 5.1.4 nicht übertragen und damit existiert auch keine Abschätzung mehr für $\|M_\sigma(U_{\tau h})\nabla P_{\tau h}\|_{L^2(\Omega_T)}$. Statt dessen gilt das etwas schwächere Resultat:

Lemma 5.2.4 *Unter den Voraussetzungen von Theorem 5.2.1 gibt es für $p < 2$ eine von τ, h unabhängige Konstante C, so dass für alle diskreten Lösungen $U_{\tau h}, P_{\tau h}$ des Finite-Elemente-Verfahrens 3.2.2 gilt:*

$$\|M_\sigma(U_{\tau h})\nabla P_{\tau h}\|_{L^2(0,T;L^p(\Omega))} \le C. \tag{5.74}$$

Beweis : Wir bezeichnen mit $|.|$ die euklidische Norm im $\mathbb{R}^2$ bzw. die dazugehörige Matrixnorm im $\mathbb{R}^{2\times 2}$. Da $M_\sigma(U_{\tau h})$ auf jedem Dreieck $E \in \mathcal{T}_h$ eine positiv definite 2×2 Matrix ist, gibt es eine Zerlegung $M_\sigma(U_{\tau h}) = M_\sigma(U_{\tau h})^{\frac{1}{2}} \cdot M_\sigma(U_{\tau h})^{\frac{1}{2}}$. Nun gilt

$$\begin{aligned}
&\|M_\sigma(U_{\tau h})\nabla P_{\tau h}\|^2_{L^2(0,T;L^p(\Omega))} \\
&\qquad = \int_0^T \left(\int_\Omega |M_\sigma(U_{\tau h})\nabla P_{\tau h}|^p \right)^{\frac{2}{p}} \\
&\qquad \le \int_0^T \left(\int_\Omega \left(|M_\sigma(U_{\tau h})^{\frac{1}{2}}||M_\sigma(U_{\tau h})^{\frac{1}{2}}\nabla P_{\tau h}| \right)^p \right)^{\frac{2}{p}} \\
&\qquad \le \int_0^T \left(\left(\int_\Omega |M_\sigma(U_{\tau h})^{\frac{1}{2}}\nabla P_{\tau h}|^2 \right)^{\frac{p}{2}} \left(\int_\Omega |M_\sigma(U_{\tau h})^{\frac{1}{2}}|^{\frac{2p}{2-p}} \right)^{\frac{2-p}{2}} \right)^{\frac{2}{p}}.
\end{aligned}$$

Um den zweiten Term abzuschätzen, benutzen wir, dass $|M_\sigma(U_{\tau h})^{\frac{1}{2}}| = \sqrt{\rho(M_\sigma(U_{\tau h}))}$ ist, wobei ρ den Spektralradius bezeichnet. Da $M_\sigma(U_{\tau h})$ auf jedem Dreieck gegeben ist durch $A^{-T}\hat{M}A^T$, genügt es also, die Eigenwerte von $\hat{M}$ abzuschätzen. Diese sind gegeben durch

$$\left(\fint_{\hat{U}_0}^{\hat{U}_i} \frac{ds}{m_\sigma(s)}\right)^{-1}, i = 1, 2.$$

Der Mittelwertsatz zeigt, dass es $\xi_i \in [\hat{U}_0, \hat{U}_i]$ gibt, so dass

$$m_\sigma(\xi_i) = \left(\fint_{\hat{U}_0}^{\hat{U}_i} \frac{ds}{m_\sigma(s)}\right)^{-1}, i = 1, 2.$$

Also gilt für den maximalen Eigenwert von $M_\sigma(U_{\tau h})$ auf einem beliebigen Dreieck $E \in \mathcal{T}_h$ mit den Eckpunkten x_0, x_1, x_2:

$$\varrho(M_\sigma(U_{\tau h})) \leq \max\{|U_{\tau h}(x_0)|, |U_{\tau h}(x_1)|, |U_{\tau h}(x_2)|, \sigma\}^n$$

und deshalb gilt für jede Zahl $\alpha \geq 1$

$$|M_\sigma(U_{\tau h})^{\frac{1}{2}}|^\alpha \leq \max_{i=0,1,2} |U_i|^{\frac{\alpha n}{2}} + \sigma^{\frac{\alpha n}{2}}.$$

Da $U_{\tau h}$ linear auf E ist, lässt sich das Maximum gegen den Mittelwert auf E abschätzen; es gilt also

$$|M_\sigma(U_{\tau h})^{\frac{1}{2}}|^\alpha \leq C \fint_E |U_{\tau h}|^{\frac{\alpha n}{2}} + \sigma^{\frac{\alpha n}{2}}.$$

Dabei ist die Konstante C nur abhängig von der Zahl $\frac{\alpha n}{2}$. Aufsummiert über alle Dreiecke E ergibt sich

$$\begin{aligned}\int_\Omega |M_\sigma(U_{\tau h})^{\frac{1}{2}}|^{\frac{2p}{2-p}} &\leq \sum_{E \in \mathcal{T}_h} |E| \left(C \fint_E |U_{\tau h}|^{\frac{pn}{2-p}} + \sigma^{\frac{pn}{2-p}}\right) \\ &\leq C \int_\Omega |U_{\tau h}|^{\frac{pn}{2-p}} + \sigma^{\frac{pn}{2-p}} |\Omega|.\end{aligned}$$

Damit gilt also insgesamt:

$$\begin{aligned}&\|M_\sigma(U_{\tau h})\nabla P_{\tau h}\|^2_{L^2(0,T;L^p(\Omega))} \\ &\qquad \leq C \int_0^T \left((M_\sigma(U_{\tau h})\nabla P_{\tau h}, \nabla P_{\tau h}) \left(\int_\Omega |U_{\tau h}|^{\frac{np}{2-p}} + 1\right)^{\frac{2-p}{p}}\right) \\ &\qquad \leq C \sup_{t \in [0,T]} \left(\int_\Omega |U_{\tau h}|^{\frac{np}{2-p}} + 1\right)^{\frac{2-p}{p}} \int_0^T (M_\sigma(U_{\tau h})\nabla P_{\tau h}, \nabla P_{\tau h}).\end{aligned}$$

Das Supremum existiert, da $U_{\tau h}$ in $L^\infty(0, T; L^{\frac{np}{2-p}}(\Omega))$ gleichmäßig beschränkt ist. Mit (5.71) folgt nun die Behauptung. □

Da Satz 5.1.6 die Beschränktheit von $M_\sigma(U_{\tau h})\nabla P_{\tau h}$ in $L^2(\Omega_T)$ benötigte, um diskrete gleichmäßige Hölderstetigkeit zu zeigen, ist dieses Resultat für $d = 2$ nicht mehr zu erzielen.

Kompaktheit in der Zeit wird nun mit Hilfe einer diskreten Nikol'skii-Abschätzung gezeigt. Der Nikol'skii-Raum (siehe [33]) $\mathcal{N}_2^r(0,T;X)$ zu einem Banachraum X ist für $0 < r < 1$ definiert als die Menge aller Funktionen aus $u \in L^2(0,T;X)$, für die für alle $0 < s < T$ gilt:

$$\left(\int_0^{T-s} \left\| \frac{u(t+s) - u(t)}{s^r} \right\|_X^2 \right)^{\frac{1}{2}} \leq C. \tag{5.75}$$

Der folgende Satz zeigt ein diskretes Analogon zu dieser Abschätzung.

Satz 5.2.5 *(Nikol'skii-Abschätzung)*

Unter den Voraussetzungen von Theorem 5.2.1 gibt es eine Konstante C_n, so dass für alle diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Verfahrens 3.2.2 für $0 < s < T$ gilt:

$$\int_0^{T-s} \|U_{\tau h}(t+s,\cdot) - U_{\tau h}(t,\cdot)\|_h^2 dt \leq C_n s. \tag{5.76}$$

Beweis : Seien zunächst $s = l\tau$, $l \in \mathbb{N}$, $l < K$. Wir testen (3.26) mit $\Theta = U_{\tau h}^{j+l} - U_{\tau h}^j$, summieren über $\sum_{k=j}^{j+l-1}$ und erhalten:

$$\sum_{k=j}^{j+l-1} (U_{\tau h}^{k+1} - U_{\tau h}^k, U_{\tau h}^{j+l} - U_{\tau h}^j)_h + \sum_{k=j}^{j+l-1} \tau (M_\sigma(U_{\tau h}^{k+1}) \nabla P_{\tau h}^{k+1}, \nabla (U_{\tau h}^{j+l} - U_{\tau h}^j))$$
$$= \sum_{k=j}^{j+l-1} \tau (Q(U_{\tau h}^{k+1}), U_{\tau h}^{j+l} - U_{\tau h}^j)_h.$$

Dies ist äquivalent zu:

$$(U_{\tau h}^{j+l} - U_{\tau h}^j, U_{\tau h}^{j+l} - U_{\tau h}^j)_h + \tau \sum_{k=1}^{l} (M_\sigma(U_{\tau h}^{j+k}) \nabla P_{\tau h}^{j+k}, \nabla U_{\tau h}^{j+l} - \nabla U_{\tau h}^j)$$
$$= \sum_{k=1}^{l} \tau (Q(U_{\tau h}^{j+k}), U_{\tau h}^{j+l} - U_{\tau h}^j)_h.$$

Nun multiplizieren wir diese Gleichung mit τ und summieren über $j = 1, \ldots, K-l$. Dann ergibt sich für den ersten Term:

$$\sum_{j=1}^{K-l} \tau (U_{\tau h}^{j+l} - U_{\tau h}^j, U_{\tau h}^{j+l} - U_{\tau h}^j)_h = \int_0^{T-l\tau} \|U_{\tau h}(t + l\tau) - U_{\tau h}(t)\|_h^2 dt.$$

Den zweiten Term können wir betragsmäßig mit Hilfe der Hölder-Ungleichung für $\frac{1}{p}+\frac{1}{p'}=1$, $p<2$, abschätzen:

$$\begin{aligned}
&\left|\tau\sum_{j=1}^{K-l}\tau\sum_{k=1}^{l}(M_\sigma(U_{\tau h}^{j+k})\nabla P_{\tau h}^{j+k},\nabla U_{\tau h}^{j+l}-U_{\tau h}^{j})\right| \\
&\quad\le \tau\sum_{k=1}^{l}\tau\sum_{j=1}^{K-l}\left(\int_\Omega |M_\sigma(U_{\tau h}^{j+k})\nabla P_{\tau h}^{j+k}|^p\right)^{\frac{1}{p}}\left(\int_\Omega |\nabla U_{\tau h}^{j+l}-\nabla U_{\tau h}^{j}|^{p'}\right)^{\frac{1}{p'}} \\
&\quad\le \tau\sum_{k=1}^{l}\left(\tau\sum_{j=1}^{K-l}\left(\int_\Omega |M_\sigma(U_{\tau h}^{j+k})\nabla P_{\tau h}^{j+k}|^p\right)^{\frac{2}{p}}\right)^{\frac{1}{2}}\left(\tau\sum_{j=1}^{K-l}\left(\int_\Omega |\nabla U_{\tau h}^{j+l}-\nabla U_{\tau h}^{j}|^{p'}\right)^{\frac{2}{p'}}\right)^{\frac{1}{2}} \\
&\quad\le \tau\sum_{k=1}^{l}\|M_\sigma(U_{\tau h})\nabla P_{\tau h}\|_{L^2(0,T;L^p(\Omega))}2\|\nabla U_{\tau h}\|_{L^2(0,T;L^{p'}(\Omega))} \\
&\quad\le C\tau l.
\end{aligned}$$

Da $\|Q\|_{L^\infty(\mathbb{R})}=C_q$ gilt und die Masse von $U_{\tau h}$ nach Satz 4.1.1 beschränkt ist, lässt sich die rechte Seite abschätzen durch:

$$\begin{aligned}
&\left|\tau\sum_{j=1}^{K-l}\sum_{k=1}^{l}\tau(Q(U_{\tau h}^{j+k}),U_{\tau h}^{j+l}-U_{\tau h}^{j})_h\right| \\
&\qquad\qquad\le \tau\sum_{j=1}^{K-l}\tau\sum_{k=1}^{l}C_q\int_\Omega\left|\mathcal{I}_h(U_{\tau h}^{j+l}-U_{\tau h}^{j})\right|\le\tau\sum_{j=1}^{K-l}\tau l C_q C\le CTC_q\tau l.
\end{aligned}$$

Insgesamt gilt also:

$$\int_0^{T-l\tau}\|U_{\tau h}(t+l\tau)-U_{\tau h}(t)\|_h^2dt\le C\tau l.$$

Damit ist die Behauptung für $s=l\tau$ bewiesen. Sei nun $s\in(l\tau,(l+1)\tau)$. Dann gibt es ein $r\in(l,l+1)$ und ein $\alpha\in(0,1)$, so dass $s=r\tau$ und $r=l+\alpha$ gilt. Dann ist, da $U_{\tau h}$ stückweise konstant in der Zeit ist,

$$U_{\tau h}(t+r\tau)=\begin{cases}U_{\tau h}(t+l\tau)=U_{\tau h}^{j+l+1} & \text{falls } t\in(j\tau,j\tau+(1-\alpha)\tau],\\ U_{\tau h}(t+(l+1)\tau)=U_{\tau h}^{j+l+2} & \text{falls } t\in(j\tau+(1-\alpha)\tau,(j+1)\tau].\end{cases}$$

Nun gilt

$$\begin{aligned}
&\int_0^{T-r\tau}\|U_{\tau h}(t+r\tau)-U_{\tau h}(t)\|_h^2 \\
&\qquad=\sum_{j=0}^{K-l-1}(1-\alpha)\tau\|U_{\tau h}^{j+l-1}-U_{\tau h}^{j+1}\|_h^2+\sum_{j=0}^{K-l-2}\alpha\tau\|U_{\tau h}^{j+l+2}-U_{\tau h}^{j+1}\|_h^2 \\
&\qquad\le(1-\alpha)l\tau C+\alpha(l+1)\tau C=(l+\alpha)\tau C=sC,
\end{aligned}$$

und damit ist die Behauptung auch für den Fall $s\neq l\tau$ bewiesen. □

Die Nikol'skii-Abschätzung und die gleichmäßige Beschränktheit können nun genutzt werden, um Kompaktheit von $\{U_{\tau h}\}$ in $L^2(0,T;C^\beta(\Omega))$ zu zeigen. Dazu benötigen wir das folgende Resultat von Simon [42]:

Satz 5.2.6 *Seien $X \subset B \subset Y$ Banachräume mit kompakter Einbettung $X \hookrightarrow B$ und $1 \leq p \leq \infty$. Falls*

i) F beschränkte Teilmenge von $L^p(0,T;X)$ ist,

ii) $\|f(t+\tau,x)-f(t,x)\|_{L^p(0,T-\tau;Y)} \to 0$ für $\tau \to 0$ gleichmäßig für alle $f \in F$ konvergiert,

so ist F relativ kompakt in $L^p(0,T;B)$ für $1 \leq p < \infty$ und relativ kompakt in $C(0,T;B)$ für $p = \infty$.

Lemma 5.2.7 *(Kompaktheit in $L^2(0,T;C^\beta(\Omega))$)*

Unter den Voraussetzungen von Theorem 5.2.1 gilt: Es gibt eine Funktion $u \in L^2(0,T;C^\beta(\Omega))$, $\beta < 1$ und eine Folge von diskreten Lösungen $U_{\tau h}$ von Schema 3.2.2, so dass für $\tau, h \to 0$ gilt:

$$U_{\tau h} \to u \text{ stark in } L^2(0,T;C^\beta(\Omega)), \tag{5.77}$$

$$U^-_{\tau h} \to u \text{ stark in } L^2(0,T;C^\beta(\Omega)). \tag{5.78}$$

Beweis : Zum Beweis der Aussage (5.77) wenden wir Satz 5.2.6 für den Fall $p = 2$ auf das Tripel $C^\alpha(\Omega) \hookrightarrow C^\beta(\Omega) \to L^2(\Omega)$ mit $0 < \beta < \alpha < 1$ an. Dies ist möglich, da Bedingung i) durch Satz 5.2.3 und Bedingung ii) durch die Nikol'skii-Abschätzung 5.2.5 gegeben sind.

(5.78) folgt nun mit Hilfe des Satzes von Fréchet-Kolmogorov (siehe z.B. [1]): Die Konvergenz (5.77) zeigt nämlich, dass $\{U_{\tau h}\}_h$ präkompakt in $L^2(0,T;C^\beta(\Omega))$ ist. Dann muss nach Fréchet-Kolmogorov auch gelten, dass

$$\sup_h \|U_{\tau h}(t+s) - U_{\tau h}(t)\|_{L^2(0,T;C^\beta(\Omega))} \overset{|s|\to 0}{\longrightarrow} 0.$$

Dies aber zeigt für $s = -\tau$ unmittelbar (5.78). □

5.2.2 Konvergenz gegen eine schwache Lösung

Wir nutzen nun die drei Bausteine Energieabschätzung, Entropieabschätzung und starke $L^2(0,T;C^\beta(\Omega))$-Konvergenz aus, um in der Differentialgleichung zur Grenze überzugehen. Dabei nutzen wir aus, dass wir aufgrund der Lemmata 5.2.2 und 5.2.7 und der schwachen Kompaktheit beschränkter Mengen aus der Menge der diskreten Lösungen eine Teilfolge auswählen können, welche für $\tau, h \to 0$ zu $\beta < 1$, $r > 2$ und $\frac{1}{r} + \frac{1}{r'} = 1$ die folgenden Konvergenzresultate erfüllt:

$$U_{\tau h} \to u \text{ stark in } L^2(0,T;C^\beta(\Omega)), \tag{5.79}$$

$$U^-_{\tau h} \to u \text{ stark in } L^2(0,T;C^\beta(\Omega)), \tag{5.80}$$

$$U_{\tau h} \rightharpoonup u \text{ schwach in } L^\infty(0,T;H^1(\Omega)), \tag{5.81}$$

$$U_{\tau h} \rightharpoonup u \text{ schwach in } L^2(0,T;W^{1,r}(\Omega)), \tag{5.82}$$

$$-\Delta_h U_{\tau h} \rightharpoonup f \text{ schwach in } L^2(\Omega_T), \tag{5.83}$$

$$M_\sigma(U_{\tau h})\nabla P_{\tau h} \rightharpoonup j \text{ schwach in } L^2(0,T;L^{r'}(\Omega)), \tag{5.84}$$

$$P_{\tau h} \rightharpoonup p \text{ schwach in } L^2(\Omega_T). \tag{5.85}$$

Darin sind die Limites $u \in L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;W^{1,r}(\Omega)) \cap L^2(0,T;C^\beta(\Omega))$, $f \in L^2(\Omega_T)$, $p \in L^2(\Omega_T)$ und $j \in L^2(0,T;L^{r'}(\Omega))$.

Für die Mobilität M_σ gilt das folgende Konvergenzresultat (siehe Grün [19], Lemma 8.5):

Lemma 5.2.8 *(Konvergenz von M_σ)*

Sei $\mathcal{T}_h$ eine zulässige und rechtwinklige Triangulierung und sei $U_{\tau h}$ eine stark konvergente Folge diskreter Lösungen, welche gegen $u \in L^2(0,T;C^\beta(\Omega))$ konvergieren. Außerdem sei $U_{\tau h}$ gleichmäßig beschränkt in $L^\infty(0,T;H^1(\Omega))$. Dann konvergiert für alle $p < \infty$:

$$M_\sigma(U_{\tau h}) \to m(u)Id \text{ stark in } L^p(\Omega_T). \tag{5.86}$$

Da $U_{\tau h}(t,x)$ und $U^-_{\tau h}(t,x)$ aufgrund der starken Konvergenz (5.79) bzw. (5.80) punktweise fast überall gegen $u(t,x)$ konvergieren und $\mathcal{I}_h W^+_{,u}$ und $\mathcal{I}_h W^-_{,u}$ unter den Voraussetzungen des Theorems beschränkt sind, gilt mit Hilfe des Satzes von Lebesgue:

$$\mathcal{I}_h W^+_{,u}(U_{\tau h}(t,x),x) \to W^+_{,u}(u(t,x),x) \text{ stark in } L^2(\Omega_T), \tag{5.87}$$

$$\mathcal{I}_h W^-_{,u}(U_{\tau h}(t-\tau,x),x) \to W^-_{,u}(u(t,x),x) \text{ stark in } L^2(\Omega_T). \tag{5.88}$$

Ebenso folgt für die rechte Seite:

$$\mathcal{I}_h Q(U_{\tau h}) \to Q(u) \text{ stark in } L^2(\Omega_T). \tag{5.89}$$

Die nun folgenden Lemmata zeigen, dass eine Teilfolge, welche die Konvergenzresultate (5.79)-(5.89) erfüllt, gegen eine schwache Lösung des kontinuierlichen Problems konvergiert.

Lemma 5.2.9 *(Grenzübergang in (3.26))*

Sei mit $U_{\tau h}, P_{\tau h}$ eine Folge diskreter Lösungen des Finite-Elemente-Verfahrens 3.2.2 gegeben, welche die Konvergenzaussagen (5.79)-(5.89) erfüllt. Dann ist $u \in C(0,T;L^2(\Omega))$ und die Funktionen u und j erfüllen die Gleichung

$$-\int_{\Omega_T} (u-u_0)\partial_t\psi + \int_{\Omega_T} j\nabla\psi = \int_{\Omega_T} Q(u)\psi \tag{5.90}$$

für alle $\psi \in C^1([0,T];W^{1,r}(\Omega))$ mit $\psi(T) = 0$. Außerdem gibt es eine schwache Ableitung $\partial_t u \in L^2(0,T;W^{1,r}(\Omega)')$, so dass

$$\int_0^T \langle \partial_t u\,,\psi\rangle_{W^{1,r}(\Omega)'\times W^{1,r}(\Omega)} + \int_{\Omega_T} j\nabla\psi = \int_{\Omega_T} Q(u)\psi \tag{5.91}$$

für alle $\psi \in L^2(0,T;W^{1,r}(\Omega))$ gilt.

Beweis : Wie im eindimensionalen Fall (Lemma 5.1.12) wählen wir eine Testfunktion $\theta \in C^1([0,T];C_0^\infty(\mathbb{R}^2))$ mit $\theta(T) = 0$ und konstruieren dazu eine diskrete Testfunktion $\Theta_{\tau h} \in S^{-1,0}(V^h)$ durch

$$\Theta^k_{\tau h} = \mathcal{I}_h\theta(\mathfrak{t}_k), \quad k = 0,\dots,K.$$

Für $\Theta_{\tau h}$ gelten die Konvergenzresultate:

$$\nabla\Theta_{\tau h} \rightarrow \nabla\theta \text{ in } L^r(\Omega_T), \tag{5.92}$$
$$\partial_\tau^- \Theta_{\tau h} \rightarrow \partial_t\theta \text{ in } L^2(\Omega_T). \tag{5.93}$$

Wir wählen nun wiederum Testfunktionen $\Theta = \Theta_{\tau h}^k$ in (3.26), summieren die Gleichung über $k = 0, \ldots, K-1$ und erhalten analog zum eindimensionalen Fall:

$$-\int_0^T (U_{\tau h}^- - U_{\tau h}^0, \partial_\tau^- \Theta_{\tau h})_h + \int_0^T (M_\sigma(U_{\tau h})\nabla P_{\tau h}, \nabla\Theta_{\tau h}) = \int_0^T (Q(U_{\tau h}), \Theta_{\tau h})_h.$$

Wir betrachten nun den Grenzübergang $h, \tau \rightarrow 0$. Es folgt aus (5.92) und (5.84), dass

$$\int_0^T (M_\sigma(U_{\tau h})\nabla P_{\tau h}, \nabla\Theta_{\tau h}) \rightarrow \int_0^T (j, \nabla\theta).$$

Die Konvergenz des parabolischen Terms folgt wie im eindimensionalen Fall. Die rechte Seite konvergiert aufgrund der Beschränktheit von Q und der punktweisen Konvergenz von $U_{\tau h}$ nach dem Satz von Lebesgue gegen u, so dass insgesamt gilt:

$$-\int_{\Omega_T} (u - u_0)\partial_t\theta + \int_{\Omega_T} j\nabla\theta = \int_{\Omega_T} Q(u)\theta.$$

Dies gilt für alle Testfunktionen $\theta \in C^1([0,T]; C_0^\infty(\mathbb{R}^2))$ mit $\theta(T) = 0$. In Analogie zum Beweis im eindimensionalen Fall lässt sich außerdem ein $\partial_t u \in L^2(0,T; W^{1,r}(\Omega)')$ finden, so dass

$$\int_0^T \langle \partial_t u\,, \psi\rangle_{W^{1,r}(\Omega)' \times W^{1,r}(\Omega)} + \int_{\Omega_T} j\nabla\psi = \int_{\Omega_T} Q(u)\psi$$

für alle Testfunktionen $\psi \in L^2(0,T; W^{1,r}(\Omega))$ gilt. Nun folgt $u \in C(0,T; L^2(\Omega))$, indem man den folgenden Satz (Zeidler [47], Seite 422) auf das Evolutionstripel[1] $W^{1,r}(\Omega) \subset L^2(\Omega) \subset W^{1,r}(\Omega)'$ und $p = 2$ anwendet:

Seien $V \subset H \subset V'$ ein Evolutionstripel, $1 < p < \infty$ und $\frac{1}{p} + \frac{1}{p'} = 1$. Dann ist die Menge

$$\{u : u \in L^p(0,T;V), \partial_t u \in L^{p'}(0,T;V')\}$$

stetig eingebettet in $C(0,T;H)$. □

Lemma 5.2.10 *(Grenzübergang in (3.27))*

Sei mit $U_{\tau h}, P_{\tau h}$ eine Folge diskreter Lösungen des Finite-Elemente-Verfahrens 3.2.2 gegeben, welche die Konvergenzaussagen (5.79)-(5.89) erfüllt. Dann ist $\Delta u \in L^2(\Omega_T)$ und es gilt

$$p = -\Delta u + W_{,u}^+(u,\cdot) + W_{,u}^-(u,\cdot). \tag{5.94}$$

[1] Ein Evolutionstripel V, H, V' ist gegeben, falls V ein reeller, separabler und reflexiver Banachraum und H ein reeller separabler Hilbertraum ist, ferner V dicht in H liegt und die Einbettung $V \rightarrow H$ stetig ist.

Beweis : Für den Grenzübergang $\tau, h \to 0$ in Gleichung (3.27) wählen wir wie im Beweis von Lemma 5.1.13 eine Testfunktion $\psi \in L^2(0,T;H^1(\Omega))$ und konstruieren eine diskrete Testfunktion $\Psi_{\tau h} \in S^{-1,0}(V^h)$ durch

$$\Psi_{\tau h}^k = \mathcal{R}_h \psi(\mathfrak{t}_k), \quad k = 0, \dots, K.$$

Wir erinnern uns, dass die diskrete Funktion $\Psi_{\tau h}$ in $L^2(0,T;H^1(\Omega))$ gegen ψ konvergiert. Nun setzen wir $\Psi = \Psi_{\tau h}^k$ in (3.27), multiplizieren mit τ_{k+1}, summieren über $k = 0, \dots, K-1$ und erhalten:

$$\int_0^T (\nabla U_{\tau h}, \nabla \Psi_{\tau h}) = \int_0^T (P_{\tau h} - \mathcal{I}_h W_{,u}^+(U_{\tau h}, \cdot) - \mathcal{I}_h W_{,u}^-(U_{\tau h}^-, \cdot), \Psi_{\tau h})_h.$$

In dieser Gleichung können wir nun zur Grenze übergehen, indem wir (5.81), (5.83) und die Konvergenz von $(\cdot,\cdot)_h$ gegen das L^2-Skalarprodukt ausnutzen. Wir erhalten so

$$\int_0^T (\nabla u, \nabla \psi) = \int_0^T (f, \psi).$$

Damit ist nach Definition der schwachen Ableitung $f = -\Delta u \in L^2(\Omega_T)$. Mit Hilfe der Konvergenzresultate (5.85), (5.87) und (5.88) folgt

$$-\Delta u = p - W_{,u}^+(u,\cdot) - W_{,u}^-(u,\cdot) \text{ in } L^2(\Omega_T),$$

womit das Lemma bewiesen ist. □

Lemma 5.2.11 *(Identifikation von j mit $m(u)\nabla p$)*

Sei mit $U_{\tau h}, P_{\tau h}$ eine Folge diskreter Lösungen des Finite-Elemente-Verfahrens 3.2.2 gegeben, welche die Konvergenzaussagen (5.79)-(5.89) erfüllt. Dann gibt es eine Menge $\mathcal{S} \subset (0,T)$ mit Lebesgue-Maß $\mu((0,T) \setminus \mathcal{S}) = 0$, so dass $\nabla p(t,\cdot) \in L^2([u(t) > \delta])$ für alle $t \in \mathcal{S}$ und $\delta > 0$ ist. Es gilt

$$\nabla P_{\tau h}(t,\cdot) \rightharpoonup \nabla p(t,\cdot) \text{ schwach in } L^2([u(t) > \delta]) \text{ für alle } t \in \mathcal{S}, \delta > 0, \tag{5.95}$$

und

$$j(t,\cdot) = \begin{cases} m(u(t,\cdot))\nabla p(t,\cdot) & \text{auf } [u(t) > 0] \\ 0 & \text{auf } [u(t) = 0] \end{cases} \quad \forall t \in \mathcal{S}. \tag{5.96}$$

Beweis : Wir zeigen zunächst einmal, dass eine Menge $\mathcal{S} \subset (0,T)$ mit $\mu((0,T) \setminus \mathcal{S}) = 0$ und eine Teilfolge $\tau, h \to 0$ existiert, so dass

$$M_\sigma(U_{\tau h}(t)) \to m(u(t)) \text{ stark in } L^2(\Omega) \text{ für alle } t \in \mathcal{S}, \tag{5.97}$$
$$\mathcal{I}_h W_{,u}^+(U_{\tau h}(t),\cdot) \to W_{,u}^+(u(t),\cdot) \text{ stark in } L^2(\Omega) \text{ für alle } t \in \mathcal{S}, \tag{5.98}$$
$$\mathcal{I}_h W_{,u}^-(U_{\tau h}^-(t),\cdot) \to W_{,u}^-(u(t),\cdot) \text{ stark in } L^2(\Omega) \text{ für alle } t \in \mathcal{S}, \tag{5.99}$$
$$U_{\tau h}(t) \to u(t) \text{ stark in } C^\beta(\Omega) \text{ für alle } t \in \mathcal{S}, \tag{5.100}$$
$$U_{\tau h}^-(t) \to u(t) \text{ stark in } C^\beta(\Omega) \text{ für alle } t \in \mathcal{S} \tag{5.101}$$

konvergiert, und es von t abhängige Konstanten C_t gibt, so dass

$$\int_\Omega |\Delta_h U_{\tau h}(t,\cdot)|^2 \leq C_t \text{ für alle } t \in \mathcal{S}, \tag{5.102}$$

$$(M_\sigma(U_{\tau h}(t,\cdot))\nabla P_{\tau h}(t,\cdot), \nabla P_{\tau h}(t,\cdot)) \leq C_t \text{ für alle } t \in \mathcal{S}. \tag{5.103}$$

Die Aussagen (5.97)-(5.101) folgen sofort aus den Aussagen (5.79),(5.80) und (5.86)-(5.88), da starke Konvergenz in L^p die Existenz einer punktweise fast überall konvergenten Teilfolge impliziert.

Um (5.102) zu beweisen, definieren wir die Menge

$$\mathcal{E} := \{t \in [0,T] : \liminf_{h\to 0} \int_\Omega |\Delta_h U_{\tau h}(t,\cdot)|^2 = +\infty\}$$

und das Symbol $[.]_L$ durch

$$[x]_L := \begin{cases} x & \text{falls } x < L, \\ L & \text{sonst.} \end{cases}$$

Mit Hilfe der Entropieabschätzung gilt nun

$$C_e \geq \int_{\mathcal{E}} |\Delta_h U_{\tau h}(t,\cdot)|^2 \geq \int_{\mathcal{E}} \left[|\Delta_h U_{\tau h}(t,\cdot)|^2\right]_L \overset{h\to 0}{\longrightarrow} L|\mathcal{E}|.$$

Da L beliebig groß gewählt werden kann, folgt daraus $|\mathcal{E}| = 0$, und (5.102) gilt damit für fast alle t. Ungleichung (5.103) zeigt man analog unter Ausnutzung der Energieabschätzung.

Aus (5.102) wird ersichtlich, dass für alle $t \in \mathcal{S}$ eine in $L^2(\Omega)$ schwach konvergente Teilfolge von $\Delta_h U_{\tau h}(t,\cdot)$ existieren muss. Den Grenzwert dieser Folge, zunächst eimal l_t genannt, können wir mit $\Delta u(t,.)$ identifizieren. Dazu bemerken wir zunächst einmal, dass unter Ausnutzung von (5.100) wie im eindimensionalen Fall (siehe Gleichung (5.58)) gilt:

$$\nabla U_{\tau h}(t,\cdot) \rightharpoonup \nabla u(t,\cdot) \text{ schwach in } L^2(\Omega).$$

Also gilt für alle Testfunktionen $\Psi \in V^h$:

$$(\Delta_h U_{\tau h}(t,\cdot), \Psi)_h = -(\nabla U_{\tau h}(t,\cdot), \nabla\Psi) \to -(\nabla u(t,\cdot), \nabla\Psi)$$

und

$$(\Delta_h U_{\tau h}(t,\cdot), \Psi)_h \to (l_t, \Psi),$$

und damit gilt $\Delta u(t,\cdot) = l_t(\cdot)$. Da dies für jede gewählte Teilfolge gilt, konvergiert sogar die ganze Folge gegen $\Delta u(t,\cdot)$. Also gilt

$$\Delta_h U_{\tau h}(t,\cdot) \rightharpoonup \Delta u(t,\cdot) \text{ in } L^2(\Omega) \text{ für alle } t \in \mathcal{S}.$$

Dann gilt für den Druck $P_{\tau h}$:

$$\begin{aligned} P_{\tau h}(t,\cdot) &= -\Delta_h U_{\tau h}(t,\cdot) + \mathcal{I}_h W^+_{,u}(U_{\tau h}(t,\cdot),\cdot) + \mathcal{I}_h W^-_{,u}(U^-_{\tau h}(t,\cdot),\cdot) \\ &\rightharpoonup -\Delta u(t,\cdot) + W^+_{,u}(u(t,\cdot),\cdot) + W^-_{,u}(u(t,\cdot),\cdot) = p(t,\cdot) \text{ in } L^2(\Omega) \quad \forall t \in \mathcal{S}. \end{aligned}$$

Da $U_{\tau h}(t,\cdot)$ stark in $C^\beta(\Omega)$ konvergiert, gilt auf $[u(t)>\delta]$ für τ, h klein genug:

$$(M_\sigma(U_{\tau h}(t,\cdot))\nabla P_{\tau h}(t,\cdot), \nabla P_{\tau h}(t,\cdot)) \geq C\left(\tfrac{\delta}{2}\right)^n (\nabla P_{\tau h}(t,\cdot), \nabla P_{\tau h}(t,\cdot)).$$

Also folgt aus (5.103), dass $\nabla P_{\tau h}(t,\cdot)$ in $L^2([u(t)>\delta])$, $\delta>0$ eine schwach konvergente Teilfolge besitzt. Da sich mit Hilfe von (5.85) zeigen lässt, dass alle solchen Teilfolgen gegen $\nabla p(t,\cdot)$ konvergieren, konvergiert die ganze Folge und damit gilt für alle $t\in\mathcal{S}$ und $\delta>0$:

$$\nabla P_{\tau h}(t,\cdot) \rightharpoonup \nabla p(t,\cdot) \text{ in } L^2([u(t)>\delta]).$$

Somit folgt für

$$J_{\tau h}(t,x) := M_\sigma(U_{\tau h}(t,x))\nabla P_{\tau h}(t,x)$$

und alle Testfunktionen $\phi\in W^{1,r}([u(t)>\delta])$:

$$\begin{aligned}\int_{[u(t)>\delta]} J_{\tau h}(t)\nabla\phi &= \int_{[u(t)>\delta]} M_\sigma(U_{\tau h}(t))\nabla P_{\tau h}(t)\nabla\phi \\ &\to \int_{[u(t)>\delta]} m(u(t))\nabla p(t)\nabla\phi,\end{aligned}$$

denn M_σ konvergiert wegen (5.97) stark in $L^{\frac{2r}{r-2}}([u(t)>\delta])$. Also gilt für alle $\delta>0$:

$$J_{\tau h}(t,\cdot) \rightharpoonup m(u(t,\cdot))\nabla p(t,\cdot) \text{ schwach in } L^{r'}([u(t)>\delta]).$$

Es bleibt nun noch zu zeigen, dass dieser Grenzwert für alle $t\in\mathcal{S}$ mit dem in (5.84) bestimmten Grenzwert $J_{\tau h}\rightharpoonup j\in L^2(0,T;L^{r'}(\Omega))$ übereinstimmt. Dies zeigen wir mit Hilfe des Konvergenzsatzes von Vitali (siehe z.B. [1]). Seien dazu $\phi\in C_0^\infty([u>\delta])$ und

$$\alpha_{\tau h}(t) := \int_{[u(t)>\delta]} J_{\tau h}(t,\cdot)\nabla\phi(t,\cdot).$$

Dann ist $\alpha_{\tau h}\in L^1(0,T;\mathbb{R})$ und konvergiert für alle $t\in\mathcal{S}$ gegen

$$\alpha_{\tau h}(t) \overset{\tau,h\to 0}{\longrightarrow} \alpha(t) := \int_{[u(t)>\delta]} m(u(t,\cdot))\nabla p(t,\cdot)\nabla\phi(t,\cdot).$$

Da

$$\begin{aligned}\sup_{\tau,h}\int_{t_1}^{t_2}\int_{[u(t)>\delta]} J_{\tau h}\nabla\phi &\leq \sup_{\tau,h}\left(\int_{t_1}^{t_2}\left(\int_\Omega |J_{\tau h}|^{r'}\right)^{\frac{2}{r'}}\right)^{\frac12}\left(\int_{t_1}^{t_2}\left(\int_\Omega|\nabla\phi|^r\right)^{\frac2r}\right)^{\frac12} \\ &\leq \sup_{\tau,h}\|J_{\tau h}\|_{L^2(0,T;L^{r'}(\Omega))}\|\nabla\phi\|_{L^\infty(0,T;L^r(\Omega))}(t_2-t_1)^{\frac12} \\ &\leq C(t_2-t_1)^{\frac12}\end{aligned}$$

für $t_2\to t_1$ gegen 0 konvergiert, gilt der Satz von Vitali und wir erhalten

$$\int_0^T \alpha_{\tau h}(t)dt \to \int_0^T \alpha(t)dt.$$

Da aber bereits aus (5.84) bekannt ist, dass $\int_0^T \alpha_{\tau h}(t)dt$ gegen $\int_0^T\int_{[u(t)>\delta]} j\nabla\phi$ konvergiert, ist auf diese Weise $j(t,\cdot)=m(u(t,\cdot))\nabla p(t,\cdot)$ auf $[u(t)>\delta]$ für alle $\delta>0$ gezeigt.

Auf $[u(t) = 0]$ gilt für eine Testfunktion $\phi \in W^{1,r}(\Omega)$

$$\int_{[u(t)=0]} J_{\tau h}(t,\cdot)\nabla\phi$$
$$\leq \left(\int_\Omega M_\sigma(U_{\tau h}(t))\nabla P_{\tau h}(t)\nabla P_{\tau h}(t) \right)^{\frac{1}{2}} \left(\int_\Omega |\nabla\phi|^2 \right)^{\frac{1}{2}} \sup_{x\in[u(t)=0]} M_\sigma(U_{\tau h}(t,x)) \overset{h\to 0}{\longrightarrow} 0,$$

da $U_{\tau h}(t,x)$ für h klein genug auf $[u(t) = 0]$ dank (5.100) gleichmäßig gegen 0 konvergiert. Damit folgt die Behauptung des Lemmas. □

Lemma 5.2.12 *(Annahme der Anfangswerte)*

Sei mit $U_{\tau h}$ eine Folge diskreter Lösungen des Finite-Elemente-Verfahrens 3.2.2 gegeben, welche die Konvergenzaussagen (5.79)-(5.89) erfüllt. Dann gilt

$$\lim_{t\to 0} u(t,\cdot) = u_0(\cdot) \text{ in } L^2(\Omega). \tag{5.104}$$

Zum Beweis benötigen wir den folgenden Satz von Simon [42]:

Satz 5.2.13 *Seien $X \subset B \subset Y$ Banachräume mit kompakter Einbettung $X \hookrightarrow B$. Falls*

i) F beschränkte Teilmenge von $L^\infty(0,T;X)$ ist,

ii) $\frac{\partial F}{\partial t} := \{v : \exists u \in F \text{ mit } v = \partial_t u\}$ beschränkt ist in $L^r(0,T;Y)$ für ein $r > 1$,

so ist F relativ kompakt in $C(0,T;B)$.

Beweis von Lemma 5.2.12: Wir betrachten die in (5.28) definierte Funktion $\tilde{U}_{\tau h}$. Für die Zeitableitung dieser Funktion gilt:

$$\partial_t \tilde{U}_{\tau h}(t) = \partial_\tau^- U_{\tau h}(t) = \frac{U_{\tau h}^{k+1} - U_{\tau h}^k}{\tau_{k+1}} \text{ für } t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}].$$

Wir möchten nun Satz 5.2.13 auf die Funktionenmenge $F = \{\tilde{U}_{\tau h} : h^2 \leq C\tau\}$ und die Banachräume $X = H^1(\Omega)$, $B = L^2(\Omega)$, $Y = W^{1,r}(\Omega)'$ anwenden. Zu i) müssen wir zeigen, dass $\tilde{U}_{\tau h}$ gleichmäßig beschränkt in $L^\infty(0,T;H^1(\Omega))$ ist. Dies folgt direkt aus der Definition der $\tilde{U}_{\tau h}$ und der durch die Energieabschätzung gegebenen Beschränktheit der $U_{\tau h}$ in $L^\infty(0,T;H^1(\Omega))$.

Zu ii) müssen wir zeigen, dass $\partial_t \tilde{U}_{\tau h}$ gleichmäßig beschränkt in $L^2(0,T;W^{1,r}(\Omega)')$ ist. Sei dazu $\phi \in W^{1,r}(\Omega)$. Dann gilt:

$$\begin{aligned}
(\partial_\tau^- U_{\tau h}(t),\phi) &\leq (\partial_\tau^- U_{\tau h}(t), \phi - \mathcal{I}_h\phi) + |(\partial_\tau^- U_{\tau h}(t), \mathcal{I}_h\phi) - (\partial_\tau^- U_{\tau h}(t), \mathcal{I}_h\phi)_h| \\
&\quad + (\partial_\tau^- U_{\tau h}(t), \mathcal{I}_h\phi)_h \\
&\leq \|\partial_\tau^- U_{\tau h}(t)\|_{0,2}\|\phi - \mathcal{I}_h\phi\|_{0,2} + Ch|\mathcal{I}_h\phi|_{1,2}\|\partial_\tau^- U_{\tau h}(t)\|_{0,2} \\
&\quad + |(M_\sigma(U_{\tau h}(t))\nabla P_{\tau h}(t), \nabla\mathcal{I}_h\phi)| + |(Q(U_{\tau h}(t)), \mathcal{I}_h\phi)_h| \\
&\leq Ch|\phi|_{1,2}\|\partial_\tau^- U_{\tau h}(t)\|_{0,2} \\
&\quad + C|\mathcal{I}_h\phi|_{1,r}\|M_\sigma(U_{\tau h}(t))\nabla P_{\tau h}(t)\|_{0,r'} + CC_q|\mathcal{I}_h\phi|_{0,2} \\
&\leq Ch|\phi|_{1,2}\|\partial_\tau^- U_{\tau h}(t)\|_h \\
&\quad + C|\phi|_{1,r}\|M_\sigma(U_{\tau h}(t))\nabla P_{\tau h}(t)\|_{0,r'} + C|\phi|_{0,2}.
\end{aligned}$$

Damit lässt sich abschätzen:

$$\|\partial_t \tilde{U}_{\tau h}\|_{L^2(0,T;W^{1,r}(\Omega)')} = \int_0^T \left(\sup_{\phi \in W^{1,r}(\Omega)} \frac{(\partial_\tau^- U_{\tau h}(t), \phi)}{\|\phi\|_{1,r}} \right)^2$$
$$\leq \int_0^T Ch^2 \|\partial_\tau^- U_{\tau h}\|_h^2 + \int_0^T \|M_\sigma(U_{\tau h}(t)) \nabla P_{\tau h}(t)\|_{0,r'}^2 + CT.$$

Der zweite Term ist nach Lemma 5.2.4 beschränkt. Der Erste lässt sich mit Hilfe der Nikol'skii-Abschätzung 5.2.5 abschätzen:

$$\int_0^T Ch^2 \|\partial_\tau^- U_{\tau h}\|_h^2 = Ch^2 \int_0^{T-\tau} \frac{1}{\tau^2} \|U_{\tau h}(t+\tau) - U_{\tau h}(t)\|_h^2 \leq Ch^2 \frac{1}{\tau^2} C_n \tau \leq C \frac{h^2}{\tau}.$$

Satz 5.2.13 zeigt nun, dass eine Funktion $\tilde{u} \in C(0,T;L^2(\Omega))$ existiert, so dass eine Teilfolge

$$\tilde{U}_{\tau h} \to \tilde{u} \text{ stark in } C(0,T;L^2(\Omega))$$

konvergiert. Es verbleibt zu zeigen, dass $\tilde{u} = u$ ist.

Wir wissen bereits aus den vorherigen Lemmata, dass $U_{\tau h}$ stark in $L^2(\Omega_T)$ gegen eine Funktion $u \in C(0,T;L^2(\Omega))$ konvergiert. Nun zeigen wir, dass auch $\tilde{U}_{\tau h}$ stark in $L^2(\Omega_T)$ gegen u konvergiert:

$$\begin{aligned}
\|\tilde{U}_{\tau h} - u\|_{L^2(\Omega_T)}^2 &\leq C \int_0^T \|\tilde{U}_{\tau h}(t) - U_{\tau h}(t)\|_h^2 + C \int_0^T \|U_{\tau h}(t) - u(t)\|_{0,2}^2 \\
&= \sum_{k=0}^{K-1} \int_{\mathfrak{t}_k}^{\mathfrak{t}_{k+1}} \left(1 - \frac{t - \mathfrak{t}_k}{\tau_{k+1}}\right)^2 \|U_{\tau h}^{k+1} - U_{\tau h}^k\|_h^2 + C \int_0^T \|U_{\tau h}(t) - u(t)\|_{0,2}^2 \\
&\leq \sum_{k=0}^{K-1} \int_{\mathfrak{t}_k}^{\mathfrak{t}_{k+1}} \|U_{\tau h}^{k+1} - U_{\tau h}^k\|_h^2 + C \int_0^T \|U_{\tau h}(t) - u(t)\|_{0,2}^2 \\
&\leq \int_0^{T-\tau} \|U_{\tau h}(t+\tau) - U_{\tau h}(t)\|_h^2 + C\tau + C \int_0^T \|U_{\tau h}(t) - u(t)\|_{0,2}^2 \\
&\leq C\tau + C \int_0^T \|U_{\tau h}(t) - u(t)\|_{0,2}^2 \overset{h,\tau \to 0}{\longrightarrow} 0.
\end{aligned}$$

Also gilt $\tilde{u} = u$ in $L^2(\Omega_T)$. Wir wissen aber bereits, dass $u \in C(0,T;L^2(\Omega))$ ist und ebenso $\tilde{u} \in C(0,T;L^2(\Omega))$. Somit gilt auch $u = \tilde{u}$ in $C(0,T;L^2(\Omega))$.

Damit haben wir:

$$\tilde{U}_{\tau h}(0,\cdot) \to u(0,\cdot) \text{ in } L^2(\Omega)$$

und

$$\tilde{U}_{\tau h}(0,\cdot) = \mathcal{I}_h u_0(\cdot) \to u_0(\cdot) \text{ in } L^2(\Omega).$$

Deshalb gilt aufgrund der Stetigkeit von u:

$$\lim_{t \to 0} u(t,\cdot) = u(0,\cdot) = u_0(\cdot).$$

□

Lemma 5.2.14 *(Grenzübergang in den Abschätzungen)*

Sei mit $U_{\tau h}, P_{\tau h}$ eine Folge von diskreten Lösungen des Finite-Elemente-Verfahrens 3.2.2 gegeben, welche die Konvergenzaussagen (5.79)-(5.89) erfüllt. Dann gilt für fast alle $t \in (0,T)$ die Ungleichung

$$\int_\Omega |\nabla u(t)|^2 + \int_\Omega W(u(t),\cdot) + \int_\Omega G(u(t)) \leq C_e + C_g. \tag{5.105}$$

Beweis : Der Beweis ist identisch zum Beweis im eindimensionalen Fall (Lemma 5.1.15). □

Durch diese Lemmata ist Theorem 5.2.1 vollständig bewiesen: Die starke Konvergenz (5.67) folgt aus Lemma 5.2.7, die schwache Konvergenz des Laplace-Operators (5.68) folgt aus der Aussage (5.83) und der in Lemma 5.2.10 erfolgten Identifizierung $f = -\Delta u$. Die schwache Konvergenz des Drucks (5.69) ist in (5.85) gezeigt, die schwache Konvergenz des Druckgradienten (5.70) in Lemma 5.2.11 bewiesen. Lemma 5.2.9 zeigt die Gültigkeit der schwachen Differentialgleichung (5.63), Lemma 5.2.10 beweist Gleichung (5.64). Die stetige Annahme der Anfangsdaten (5.65) wird in Lemma 5.2.12 bewiesen und die Abschätzung (5.66) in Lemma 5.2.14. u ist nichtnegativ, da die diskreten Lösungen $U_{\tau h}$ Satz 4.2.3 erfüllen und $U_{\tau h}$ stark konvergent ist.

Kapitel 6

Simulationen und physikalische Experimente - ein Vergleich

In diesem Kapitel werden numerische Ergebnisse des in 3.2.2 definierten Verfahrens präsentiert und mit den Resultaten physikalischer Experimente verglichen. Zunächst einmal werden jedoch noch einige Details des numerischen Verfahrens diskutiert.

6.1 Verwendete Methoden und Programme

Zur Berechnung der numerischen Simulationen wurde das Programmpaket EConLub2D erstellt und benutzt. Damit lässt sich das Anfangs-Randwert-Problem

$$\begin{aligned} \eta\partial_t u - \operatorname{div}(m(u)\nabla p) &= q(u) && \text{in } \Omega\times(0,T),\\ p &= -\varsigma\Delta u + w_{,u}(u,x) && \text{in } \Omega\times(0,T),\\ \frac{\partial}{\partial\nu}u = \frac{\partial}{\partial\nu}p &= 0 && \text{auf } \partial\Omega\times(0,T),\\ u(0,x) &= u_0(x) && \text{in } \Omega \end{aligned} \tag{6.1}$$

auf einem quadratischen Gebiet $\Omega = (0,l)^2$ und einem Zeitintervall $(0,T)$ lösen. Wie schon in den vorherigen Kapiteln beschreibt u hier die Höhe des Flüssigkeitsfilms, η die Viskosität und ς die Stärke der Oberflächenspannung. Die Mobilität m ist durch $m(s) = cs^n$ gegeben, und das Grenzflächenpotential w und die rechte Seite q erfüllen eine der Bedingungen aus Kapitel 3.2.

Das Programmpaket wird im Anhang A im Detail vorgestellt. Im folgenden werden kurz die wichtigsten Ideen zusammengefasst.

6.1.1 Berechnung der diskreten Lösung

Nach (3.28) berechnet sich die diskrete Lösung $U^{k+1} \in V^h$ des Finite-Elemente-Schemas 3.2.2 zum Zeitpunkt t_{k+1} aus der Lösung U^k zum vorherigen Zeitpunkt t_k mit Hilfe der

Gleichung

$$U^{k+1} - U^k + \tau_{k+1} M_h^{-1} L_h^M(U^{k+1}) \Big(M_h^{-1} L_h U^{k+1} + \mathcal{I}_h W_{,u}^+(U^{k+1}) + \mathcal{I}_h W_{,u}^-(U^k) \Big) = \tau_{k+1} \mathcal{I}_h Q(U^{k+1}). \quad (6.2)$$

Um die Lösung U^{k+1} dieser Gleichung zu finden, wird die folgende Fixpunktiteration durchgeführt. Man setzt $U_0^{k+1} = U^k$ und berechnet U_{i+1}^{k+1} aus U_i^{k+1}, indem eine Lösung für

$$\frac{1}{\tau_{k+1}}(U_{i+1}^{k+1} - U^k) + M_h^{-1} L_h^M(U_i^{k+1}) \Big(M_h^{-1} L_h U_{i+1}^{k+1} + \mathcal{I}_h W_{,u}^+(U_{i+1}^{k+1}) + \mathcal{I}_h W_{,u}^-(U^k) \Big) = \mathcal{I}_h Q(U_i^{k+1}). \quad (6.3)$$

gefunden wird. Die Lösung dieser Gleichung kann nun mit Hilfe eines Newton-Verfahrens berechnet werden. Zur Verbesserung der Konvergenz benutzt EConLub2D hierfür ein Newton-Verfahren mit Armijo-Schrittweitensteuerung. In jedem Iterationsschritt des Newton-Verfahrens ist ein lineares Gleichungssystem mit einer dünn besetzten, nichtsymmetrischen Matrix zu lösen. Dies geschieht durch ein BiCGstab-Verfahren [45]. Die Anzahl der nötigen Iterationsschritte des BiCGstab-Verfahrens wird durch einen BPX-ähnlichen Vorkonditionierer (siehe A.3.3.4) reduziert. Die Verwendung des Vorkonditionierers bringt im Fall homogener Substrate einen deutlichen Geschwindigkeitsvorteil (siehe auch [20]). Im Fall inhomogener Substrate wird die Zahl der nötigen Iterationen im allgemeinen weniger stark reduziert.

Da das Newton-Verfahren die Ableitung $\frac{d}{dX}(\mathcal{I}_h W_{,u}^+(X))$ berechnen muss, benutzt EConLub2D anstelle der linearen Näherung (W2) die Näherung

$$W(u,x) = \begin{cases} w(u,x) & \text{falls } u > \varepsilon_w, \\ w(\varepsilon_w,x) + (x-\varepsilon_w) w_{,u}(\varepsilon_w,x) + \frac{1}{2}(x-\varepsilon_w)^2 w_{,uu}(\varepsilon_w,x) & \text{falls } u \le \varepsilon_w. \end{cases} \quad (6.4)$$

Wenn ε_w klein genug gewählt wird, ändert dies die Lösung U nicht, was im folgenden Satz bewiesen wird:

Satz 6.1.1 *(Positivität der diskreten Lösung)*

Es gelte $d = 2$, (W2) und die Anfangsdaten seien strikt positiv: $u_0 \ge c_0 > 0$. Dann gibt es ein $\varepsilon_w > 0$, so dass für eine Lösung $U \in S^{-1,0}(V^h)$ von Schema 3.2.2 gilt:

$$U(t,x) \ge \varepsilon_w \quad \forall t \in (0,T), x \in \Omega. \quad (6.5)$$

Beweis : Mit Hilfe der Energieabschätzung folgt:

$$\begin{aligned} C_g &\ge \int_\Omega \mathcal{I}_h W(U(t,.),.) \\ &= \sum_{i=1}^{D} \left(\int_\Omega \left(W(U(t,\mathfrak{x}_i),\mathfrak{x}_i) + C_w \right) \varphi_i(x) dx - \int_\Omega C_w \varphi_i(x) dx \right). \end{aligned}$$

Dabei sei die Konstante C_w so groß gewählt, dass die obigen Integranden positiv sind. Also gilt für alle $U(t, \mathfrak{x}_i)$:

$$\int_\Omega W(U(t,\mathfrak{x}_i),\mathfrak{x}_i)\varphi_i(x)dx \le C_g + C_w|\Omega|.$$

Wir nehmen nun an, dass $U(t,\mathfrak{x}_i) \le \varepsilon_w$ sei, dabei sei o.B.d.A. $\varepsilon_w \le \min\{\frac{a_{12}}{2a_{11}}, \frac{a_{22}}{2a_{21}}, 1\}$: Dann gilt:

$$\int_\Omega W(U(t,\mathfrak{x}_i),\mathfrak{x}_i)\varphi_i(x)dx \ge \min\left\{\tfrac{a_{12}}{2}, \tfrac{a_{22}}{2}\right\} \varepsilon_w^{-\min\{l_{12},l_{22}\}} Ch^2.$$

Daraus folgt:

$$\varepsilon_w^{\min\{l_{12},l_{22}\}} \ge Ch^2,$$

und dies ist ein Widerspruch zur Annahme, falls ε_w klein genug ist. □

Damit ist die Lösung im Fall (W2) strikt positiv und größer als ε_w, falls die Schranke ε_w klein genug gewählt wird. U löst damit die Gleichung für $W = w$, ist also von der Art der Näherung unabhängig[1]. Aus Gründen der numerischen Stabilität ist es aber trotzdem erforderlich, W auf ganz $\mathbb{R}$ zu definieren, da dieses Positivitätsresultat nicht für jedes Zwischenergebnis des Iterationsverfahrens gelten muss. Somit kann die Art der Näherung durchaus das Konvergenzverhalten des Iterationsverfahrens beeinflußen, nicht aber die gesuchte Lösung.

6.1.2 Adaptivität

Das Verfahren verwendet ein adaptives Finite-Elemente-Gitter. Die Triangulierung $\mathcal{T}_h$ wird nach jedem Zeitschritt angepasst. In Ermangelung eines geeigneten a posteriori Fehlerschätzers wird als Entscheidungskriterium zum Verfeinern beziehungsweise Vergröbern die Differenz von ∇U auf zwei benachbarten Finiten Elementen betrachtet. Ist sie zu groß, so wird verfeinert. Ist die Differenz klein, so können beide Dreiecke zusammengefasst werden (siehe A.3.3.2).

Auch die Zeitschrittweite ist nicht fest gewählt. Die Formel zu ihrer Berechnung ist wie folgt motiviert (siehe auch [21]). Für einen sich ausbreitenden Tropfen mit kompaktem Träger soll die Zeitschrittweite τ_{k+1} so gewählt werden, dass sich der freie Rand $\partial[u > 0]$ in einem Zeitschritt nicht um mehr als einen Gitterpunkt weiterbewegt. Also sollte $\tau_{k+1} \le \frac{h}{v_{\max}(t_k)}$ sein, wenn $v_{\max}(t_k)$ die maximale Normalengeschwindigkeit des freien Randes zur Zeit t_k bezeichnet. Die Normalengeschwindigkeit des freien Randes an einer Stelle $x_0 \in \partial[u(t) > 0]$ berechnet sich formal durch

$$v_n(x_0) = \lim_{x\to x_0, x\in \mathrm{supp}(u(t,.))} \frac{m(u(t,x))}{u(t,x)} \frac{\partial}{\partial \nu} p(t,x). \tag{6.6}$$

Diese Formel ist exakt für selbstähnliche Quelllösungen des Anfangswertproblems

$$\begin{aligned} u_t + \mathrm{div}(u^n \nabla \Delta u) &= 0, \\ u(t,.) &\overset{t\to 0}{\longrightarrow} \delta_0. \end{aligned} \tag{6.7}$$

[1] Man beachte, dass dieses Resultat nicht genutzt werden kann um strikte Positivität der kontinuierlichen Lösung u aus Theorem 5.2.1 zu zeigen, da für $h \to 0$ auch $\varepsilon_w \to 0$ gewählt werden müßte.

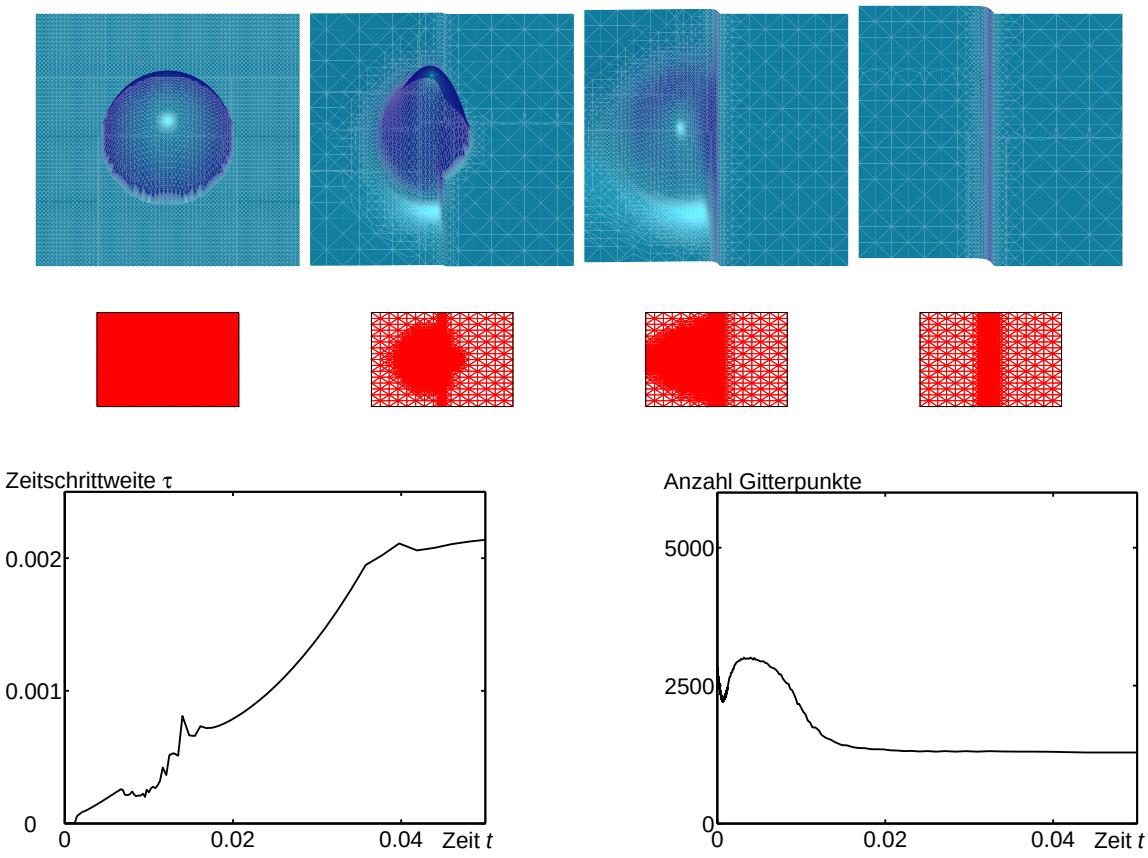


Abbildung 6.1: Beispiel zur Adaptivität: Die Abbildung zeigt die numerische Lösung der Gleichung $\partial_t u - \operatorname{div}(\frac{1}{3}u^3\nabla(-\Delta u + w_{,u}(u,x))) = 0$ auf dem Gebiet $[0,1]^2$. Das Potential ist gegeben durch $w(u,x) = 0.02u^{-3}$ für $x_1 < 0.5$ und $w(u,x) = -u^{-2} + 0.02u^{-3}$ für $x_1 \geq 0.5$. Das erste Bild zeigt die Anfangsdaten $u_0(x) = \max\{2(\frac{1}{16} - (x-M)^2)^{1/2}, 0.03\}$, dabei ist $M = (\frac{1}{2}, \frac{1}{2})$ der Mittelpunkt von Ω. Die weiteren Bilder zeigen die numerische Lösung zu den Zeitpunkten $0.00064, 0.0057, 0.05$. Die roten Abbildungen darunter zeigen das Gitter an dem jeweiligen Zeitpunkt, die Rechnung beginnt mit dem feinstmöglichen Gitter. Die beiden Graphen zeigen die verwendeten Zeitschrittweiten und Gittergrößen.

Diese Quelllösungen wurden von Bernis, Peletier und Williams [7] für $d = 1$ und Bernis und Ferreira [15] für $d > 1$ bestimmt; ein Beweis für Formel (6.6) befindet sich in [21]. Der Algorithmus berechnet nun auf jedem Dreieck E der Triangulierung den Wert $M_\sigma(\overline{U^k_{\tau h}}(E))|\nabla P^k| \,/\, \overline{U^k_{\tau h}}(E)$, wobei $\overline{U^k_{\tau h}}(E)$ den Mittelwert von $U^k_{\tau h}$ auf E bezeichnet, und berechnet die Zeitschrittweite aus dem Quotienten von h und dem Maximum dieser Werte (siehe Anhang A.3.3.2, Gleichungen (A.22) und (A.23)).

In Abbildung 6.1 wird die Funktionsweise von Zeitschrittweitensteuerung und adaptiver Gitterverfeinerung anhand eines Beispiels demonstriert.

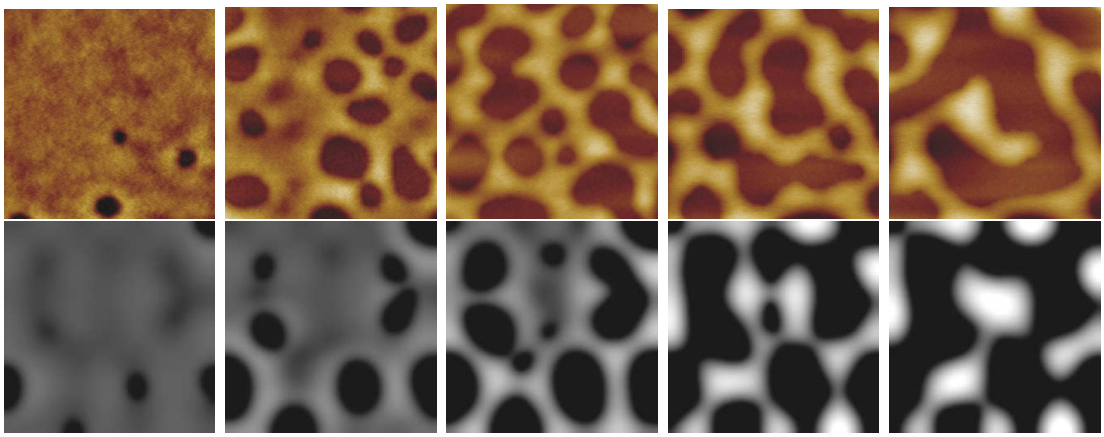


Abbildung 6.2: Spinodales Entnetzen eines 3.9 nm dicken PS(2k)-Films auf einem Siliziumsubstrat mit 191 nm dicker Siliziumoxidschicht. Die obere Bildreihe zeigt experimentelle Resultate von R. Seemann und K. Jacobs [41, 4], die Bilder zeigen einen Ausschnitt mit Seitenlänge 1.5μm zu den Zeiten $t = 704$ s, 2988 s, 4863 s, 6289 s, 7015 s. Die untere Bildreihe zeigt Ergebnisse der Simulation zu einem Potential vom Typ (6.10) und den Parametern: $A_{SiO} = 2.2 \cdot 10^{-20}$ J, $\varepsilon = 6.25 \cdot 10^{-76}$ Jm^6, $\eta = 12000$ Pa·s und $\varsigma = 0.0308$ N/m. Als Anfangsdaten wurde $u_0(x) = 3.9$ nm $(1 + S(x))$ mit einer Störung $|S(x)| \leq 0.02$ gewählt. Die Bilder zeigen den Film zu den Zeitpunkten $t = 5670$ s, 5892 s, 6092 s, 6605 s, 6917 s. Beachte: Im Experiment wurde zum Auftreten des ersten Loches die Zeit $t = 0$ gesetzt, in der Simulation bedeutet $t = 0$ den Beginn der Rechnung.

6.2 Entnetzung von Polymerfilmen

Das Modellsystem Polystyrol auf Silizium ist für physikalische Experimente in besonderer Weise geeignet, da sich bis zu 4 nm dünne Filme präparieren lassen (siehe [25]) und der Entnetzungsprozess in gut zu beobachtenden Zeitspannen abläuft (mehrere Stunden). Das Verhalten eines solchen Films wird durch die Gleichung

$$\eta \partial_t u - \operatorname{div}\left(\tfrac{1}{3} u^3 \nabla(-\varsigma \Delta u + w_{,u}(u))\right) = 0 \tag{6.8}$$

beschrieben, wobei η die Viskosität und ς die Oberflächenspannung des Films ist. Da der Silizium-Wafer von einer Siliziumoxid-Schicht bedeckt sein kann, ist das Grenzflächenpotential durch

$$w(u) = -\frac{A_{SiO}}{12\pi u^2} + \frac{A_{SiO} - A_{Si}}{12\pi (u+d)^2} + \varepsilon u^{-8} \tag{6.9}$$

gegeben.

Dabei ist A_{SiO} die Hamakerkonstante des Systems Luft/PS/SiO und A_{Si} die des Systems Luft/PS/Si, d ist die Dicke der Siliziumoxid-Schicht. Da A_{SiO} positiv und A_{Si} negativ ist (siehe [40]), ist die Art des Grenzflächenpotentials abhängig von der Dicke der Oxidschicht. Für $d = 0$ ist das Potential von Typ (1), es findet also keine Entnetzung statt. Für eine dünne Oxidschicht ist das Potential vom Typ (2), der Film ist daher metastabil. Bei einer dicken Oxidschicht kann der Term $\frac{A_{SiO} - A_{Si}}{12\pi(u+d)^2}$ vernachlässigt werden, das Potential ist deshalb vom Typ (3), näherungsweise durch

$$w(u) = -\frac{A_{SiO}}{12\pi u^2} + \varepsilon u^{-8} \tag{6.10}$$

gegeben, und der Polystyrolfilm ist instabil.

Die Werte von A_{SiO}, A_{Si} und ε lassen sich experimentell bestimmen. Darüberhinaus sind auch Oberflächenspannung und, wenn auch mit relativ großen Unsicherheiten (siehe [23]), die Viskosität des Polymerfilms bekannt. Dadurch wird es möglich, die Ergebnisse der numerischen Simulationen auch quantitativ mit den physikalischen Experimenten zu vergleichen.

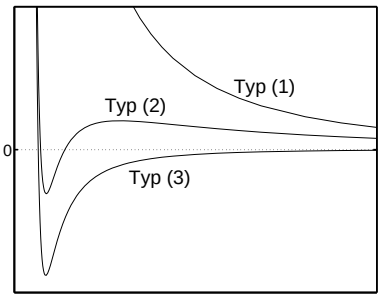


Abbildung 6.3: Skizze des effektiven Grenzflächenpotentials für stabile (1), metastabile (2) und instabile (3) Filme.

In Experimenten mit flüssigem Polystyrol (Molekulargewicht 2 kg/mol) werden, in Abhängigkeit von Filmdicke und Typ des Grenzflächenpotentials, verschiedene Entnetzungsmuster beobachtet.

Beim *spinodalen Entnetzen* reißt der Film nahezu gleichzeitig an mehreren Stellen auf und produziert dabei ein gleichmäßiges Loch- und Tropfenmuster. Spinodales Entnetzen kann bei Grenzflächenpotentialen vom Typ (2) oder (3) auftreten und wird einzig und allein durch die vom effektivem Grenzflächenpotential hervorgerufene Instabilität des Films ausgelöst. Eine gleichmäßig auf der Oberfläche verteilte Flüssigkeit ist nämlich energetisch ungünstiger als eine in mehreren Tropfen gesammelte Flüssigkeit.

Die obere Bildreihe in Abbildung 6.2 zeigt ein solches experimentelles Resultat. Der Silizium-Wafer ist hier mit einer 191 nm dicken Siliziumoxidschicht bedeckt, es liegt also ein Potential vom Typ (3) vor. Die untere Bildreihe zeigt die Ergebnisse der numerischen Rechnung zu einem durch (6.10) gegebenen Grenzflächenpotential, wobei die Parameter η, ς, A_{SiO} und ε entsprechend den experimentell bestimmten Daten gewählt wurden. Die Resultate stimmen sowohl qualitativ als auch quantitativ gut überein, wie eine Analyse der entstehenden Muster zeigt (siehe [4]).

Die Unterschiede in der Zeitskala zwischen Experiment und Simulation erklären sich wie folgt: Da es wegen zu großer Unsicherheiten in den Messungen nicht möglich ist herauszufinden, wie glatt der im Experiment präparierte Film wirklich ist, kann die Größe und die Art der Störung $S(x)$ der numerischen Anfangsdaten nicht übereinstimmend mit dem physikalischen Experiment gewählt werden. Dies beeinflusst die zeitliche Entwicklung zu Beginn der numerischen Rechnung, da ein nur gering gestörter Film wesentlich später aufreißt als ein stark gestörter. Des weiteren ist lediglich die Größenordnung der Viskosität bekannt, nicht aber der exakte Wert.

Bei der zweiten, insbesondere bei dickeren Filmen oder bei Grenzflächenpotentialen vom Typ (2) beobachteten Art der Entnetzung, tauchen vereinzelte Löcher in einem ansonsten noch nahezu homogenen Film auf (siehe Abbildung 6.4). Diese Löcher sind nicht allein durch die Instabilität verursacht, vielmehr geht man davon aus, dass sie durch Verunreinigungen des Substrats oder des Films entstehen. Daher nennt man diesen Entnetzungsprozess auch Entnetzung durch heterogene Nukleation. Das Zusammenspiel dieser beiden Effekte kann interessante Muster entstehen lassen, wie z.B. die Satellitenlöcher in Abb. 6.5. Obwohl in diesem Experiment ein Potential vom Typ (3) vorliegt, entstehen dort zunächst einmal durch heterogene Nukleation erzeugte Löcher, um welche sich dann weitere Satellitenlöcher bilden.

Dieser Effekt ist in der numerischen Simulation schwieriger zu realisieren: Simulationen zu homogenen Substraten und einem Grenzflächenpotential vom Typ (3), welche mit einem

flachen, nur leicht gestörten Film beginnen, zeigen stets spinodales Entnetzen. Simulationen zu einem Grenzflächenpotential vom Typ (2) zeigen, je nach Dicke des Ausgangsfilms, entweder spinodales Entnetzen oder einen stabilen Film.

Zur Simulation der Satellitenlochbildung sind zwei verschiedene Ansätze möglich. Die erste Möglichkeit ist, die Rechnung mit Anfangswerten zu beginnen, welche bereits eine Vertiefung oder ein Loch enthalten. Wie Abbildung 6.6 zeigt, entstehen dann um das durch die Anfangsdaten vorgegebene Loch herum weitere Satellitenlöcher. Der Mechanismus dabei ist der folgende: Hinter dem sich um das Loch herum bildenden Wulst entsteht ein ringförmiger Graben. Längs dieses Grabens findet nun ein aufgrund der geringeren Dicke des Filmes beschleunigter spinodaler Entnetzungsprozess statt, so dass dort eine Kette von weiteren Löchern entsteht.

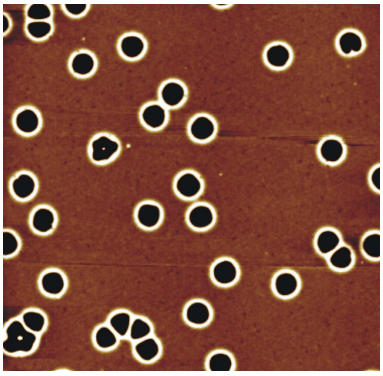


Abbildung 6.4: *Heterogene Nukleation:* Entnetzungsszenario eines 6.6 nm dicken PS(2k)-Films auf einem Si-Substrat mit einer ca. 2.4 nm dicken SiO-Schicht (Experiment von R. Seemann u. K. Jacobs).

Dass solche Vertiefungen im Ausgangsfilm auch im physikalischen Experiment Satellitenlöcher hervorrufen, ist durch Präparation von PS-Filmen, in die eine Rille gekratzt wurde, bestätigt worden [31, 32]. Bei diesem Ansatz zur Simulation bleibt allerdings unklar, wie die Löcher im ersten Bild der Abbildung 6.5 enstanden sind.

Eine mögliche Erklärung liefert hier der zweite Ansatz. Dabei startet die Rechnung mit einem flachen, nur leicht gestörten Film. Im Gegensatz zum vorherigen Ansatz ist jetzt aber die Substratoberfläche inhomogen, was hier durch unterschiedliche Hamakerkonstanten A_1, A_2 dargestellt wird:

$$w(u,x) = \begin{cases} -\frac{A_1}{12\pi}u^{-2} + \varepsilon u^{-8} & \text{falls } x \in B_r(x_o), \\ -\frac{A_2}{12\pi}u^{-2} + \varepsilon u^{-8} & \text{sonst .} \end{cases} \tag{6.11}$$

Wie Abbildung 6.7 zeigt, führt schon eine kleine Differenz in der Hamakerkonstanten zu einem Aufreißen des Films über dem hydrophoberen Bereich im Zentrum. Damit entsteht zunächst also eine Situation ähnlich der Ausgangssituation in Abb. 6.6. Im weiteren Verlauf der Rechnung entstehen dann auf die gleiche Art und Weise wie zuvor die Satellitenlöcher.

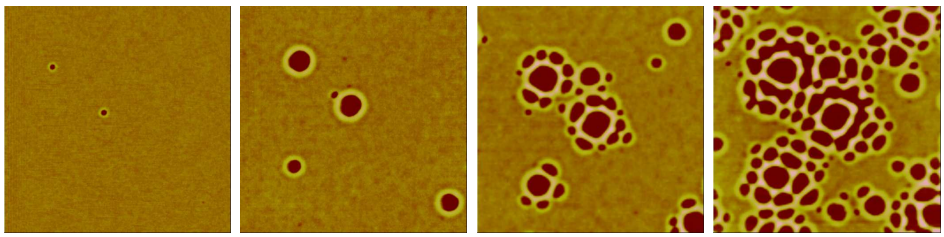


Abbildung 6.5: *Satellitenlöcher*: Experiment von R. Seemann und K. Jacobs [4]: Entnetzungsmuster eines 4.9 nm dicken PS(2k)-Films auf einem Siliziumsubstrat mit 191 nm dicker Siliziumoxidschicht. Die Bilder zeigen einen 10 μm×10 μm großen Ausschnitt eines AFM-Scans zu den Zeiten $t = 0$ s, 870 s, 1300 s, 3060 s.

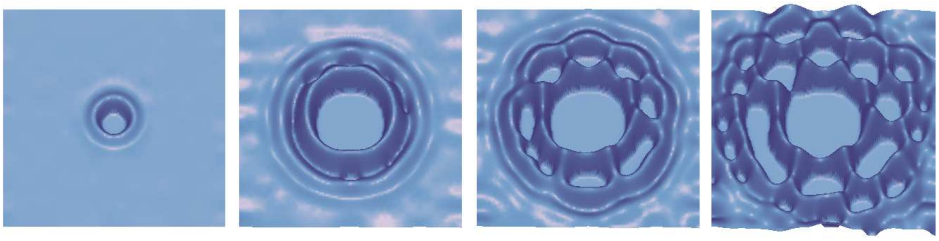


Abbildung 6.6: *Satellitenlochbildung um ein vorgegebenes Loch*: Numerische Simulation zu einem Grenzflächenpotential vom Typ (6.10) und den Parametern $A_{SiO} = 2.2{\cdot}10^{-20}$ J, $\varepsilon = 6.25{\cdot}10^{-76}$ Jm^6, $\eta = 3600$ Pa s, $\varsigma = 0.0308$ N/m. Als Anfangswert wurde gewählt: $u_0(x) = 1.25$ nm für $x \in B_{0.2\mu\mathrm{m}}(x_0)$ und $u_0(x) = 4.9$ nm $(1 + S(x))$ sonst. Dabei ist $x_0 = (2\mu\mathrm{m}, 2\mu\mathrm{m})$ der Mittelpunkt des Gebietes $\Omega = [0, 4\mu\mathrm{m}]^2$ und S eine Störung mit $|S(x)| \leq 0.02$. Die Bilder zeigen den Film zu den Zeiten $t =$ 64 s, 819 s, 1102 s, 1439 s.

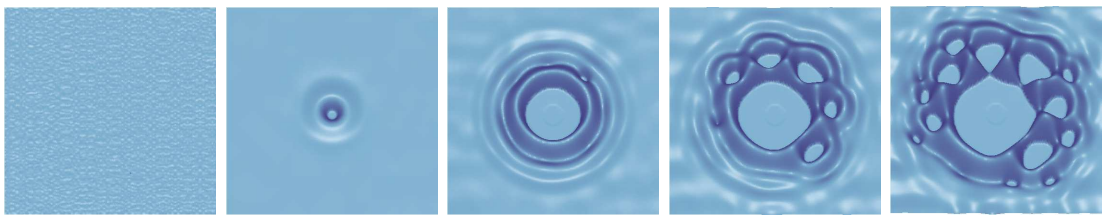


Abbildung 6.7: *Satellitenlochbildung bei inhomogenem Substrat*: Numerische Simulation zu den Parametern $\eta = 3600$ Pa s, $\varsigma = 0.0308$ N/m und Anfangswerten $u_0 = 4.9\mathrm{nm}(1 + S(x))$ mit einer Störung $|S(x)| \leq 0.01$. Das Grenzflächenpotential w ist gegeben durch $w(u, x) = -\frac{\tilde{A}}{12\pi}u^{-2} + \varepsilon u^{-8}$ für $x \in B_{0.2\mu\mathrm{m}}(x_0)$ und $w(u, x) = -\frac{A_{SiO}}{12\pi}u^{-2} + \varepsilon u^{-8}$ sonst. Dabei ist $\tilde{A} = 2.5 \cdot 10^{-20}$ J, $A_{SiO} = 2.2 \cdot 10^{-20}$ J, $\varepsilon = 6.25 \cdot 10^{-76}$ Jm^6, und $x_0 = (2\mu\mathrm{m}, 2\mu\mathrm{m})$ ist der Mittelpunkt des betrachteten Gebietes $\Omega = [0, 4\mu\mathrm{m}]^2$. Die Bilder zeigen den Film zu den Zeitpunkten $t = 0$ s, 700 s, 1342 s, 1671 s, 1931 s.

6.3 Kondensation und Evaporation auf chemisch strukturierten Substraten

Die in diesem Abschnitt präsentierten numerischen Ergebnisse sind Lösungen zu den dimensionslosen[2] Gleichungen

$$\partial_t u - \operatorname{div}(m(u)\nabla p) = q(u), \tag{6.12}$$

$$p = -\Delta u + w_{,u}(u, x). \tag{6.13}$$

Wir gehen von no-slip-Bedingungen an der Flüssigkeits-Festkörper-Grenzfläche aus, d.h. wir setzen $\beta = 0$ in Gleichung (2.42). Die Mobilität $m(u)$ ist also gegeben durch

$$m(u) = \frac{1}{3}u^3. \tag{6.14}$$

Kondensation wird modelliert durch

$$q(u) = \frac{C_1}{u + C_2}, \tag{6.15}$$

Evaporation durch

$$q(u) = \frac{-C_1}{u + C_2}\chi_{[u>0]}, \tag{6.16}$$

mit positiven Konstanten C_1 und C_2. Als Näherung Q wird im Fall von Kondensation

$$Q(s) = \begin{cases} q(s) & \text{falls } s \geq 0, \\ q(0) & \text{falls } s < 0 \end{cases} \tag{6.17}$$

verwendet. Ebensogut kann man jedoch trotz der Singularität bei $s = -C_2$ mit $Q(s) = q(s)$ rechnen – dies ist in der Praxis numerisch stabil, da die Lösungen positiv bleiben. Im Fall von Evaporation wählt man

$$Q(s) = q(s)\frac{2}{\pi}\arctan(\frac{s-\varepsilon}{\varepsilon})\chi_{[s\geq\varepsilon]}. \tag{6.18}$$

In den folgenden Beispielen ist dabei stets $\varepsilon = C_2/10$ verwendet worden.

Als erstes Beispiel betrachten wir Kondensation auf einem Substrat $\Omega = [0, 3]^2$. Das Substrat enthält einen kleinen hydrophilen[3] Kreis in einer vergleichsweise hydrophoben[3] Umgebung. Das effektive Grenzflächenpotential ist von der Art (w2), genauer

$$w(u, x) = \begin{cases} 10^{-6}u^{-8} & \text{falls } x \in B_{\frac{1}{5}}\left((\frac{3}{2}, \frac{3}{2})\right), \\ -u^{-2} + 10^{-6} - u^{-8} & \text{sonst .} \end{cases} \tag{6.19}$$

Die Näherung W von w ist durch (6.4) definiert, dabei wurde $\varepsilon_w = 0.031$ gewählt[4].

[2] Obwohl t, x und u hier also den in (2.6) eingeführten dimensionslosen Größen $\tilde{t}, \tilde{x}$ und $\tilde{u}$ entsprechen, wird in diesem Abschnitt der Einfachheit halber auf die Akzente $\sim$ verzichtet.

[3] Anstelle von hydrophil/hydrophob könnten auch die Begriffe lipophil/lipophob oder lyophil/lyophob verwendet werden, da die dimensionslose Gleichung verschiedene Arten von Flüssigkeiten beschreiben kann.

[4] Der Parameter ε_w wird von EConLub2D automatisch berechnet. Er ist $\frac{1}{4}\operatorname{argmin} w_{,u}(u)$.

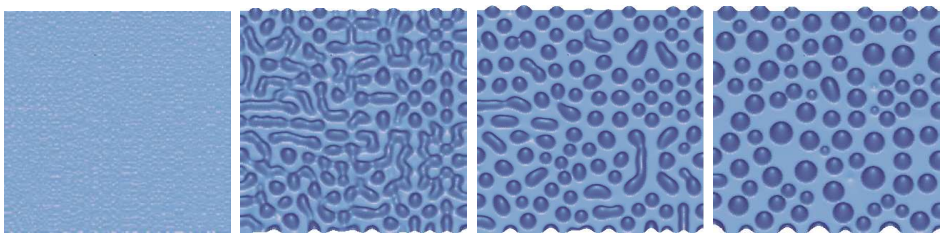


Abbildung 6.8: *Kondensation auf einem homogenem Substrat*: Numerische Simulation zu $\Omega = [0,3]^2$, $q(u) = \frac{0.25}{u+0.01}$, $W(u) = -u^{-2} + 10^{-6}u^{-8}$. Das erste Bild zeigt die Anfangsdaten: $u_0(x) = 0.2(1 + S(x))$, wobei die Störung $|S(x)| \leq 0.02$ ist. Die weiteren Bilder zeigen den Film zu den Zeitschritten $t = 0.0045, t = 0.0066$ und $t = 0.020$. Die Tropfen im letzten Bild erreichen die Höhe 0.8.

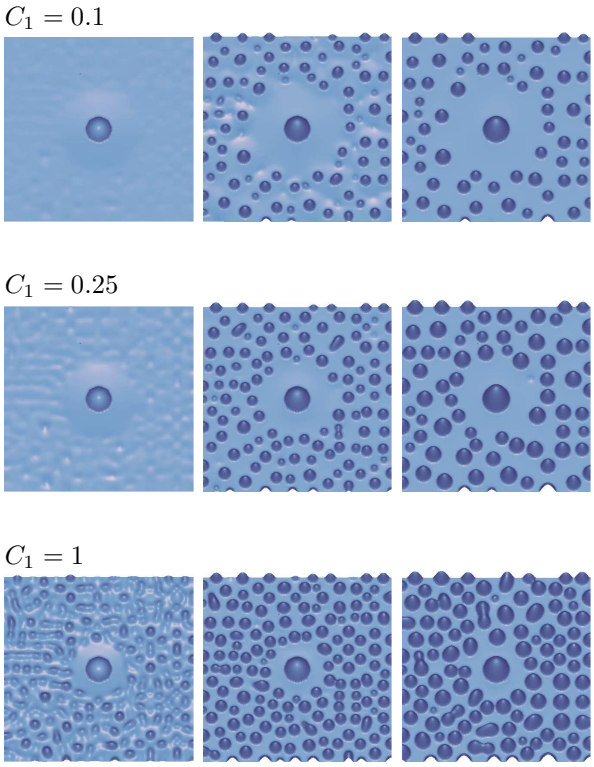


Abbildung 6.9: *Verarmungszonen*: Numerische Simulationen zu $\Omega = [0,3]^2$, $q(u) = \frac{C_1}{u+0.01}$ und einem Potential (6.19), jeweils zu den Anfangsdaten $u_0(x) = 0.15(1 + S(x))$ mit einer Störung $|S(x)| \leq 0.05$. Die drei Simulationen unterscheiden sich lediglich in der Wahl der Konstante C_1: Die obere Reihe zeigt die zeitliche Entwicklung des Films bei $C_1 = 0.1$, die mittlere Reihe bei $C_1 = 0.25$ und die untere Reihe bei $C_1 = 1.0$.

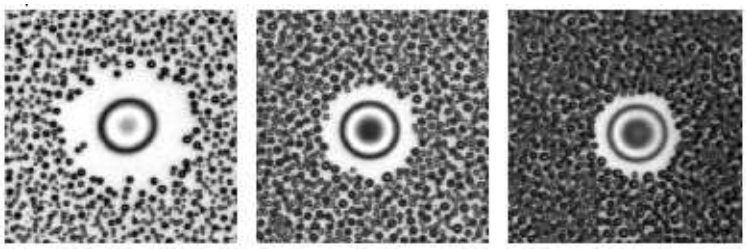


Abbildung 6.10: *Verarmungszonen (Experiment von C. Schäfle [37])*: Bei Kondensation von Diethylenglykol auf einem lyophoben Substrat mit einem lyophilen Kreis mit Durchmesser 10μm beobachtet man das Auftreten von Verarmungszonen (engl. *depletion zones*). Die Bilder zeigen den Endzustand von drei identischen Substraten, welche mit unterschiedlich starken Gasflüssen bedampft wurden, von links nach rechts wächst die Geschwindigkeit der Kondensation.

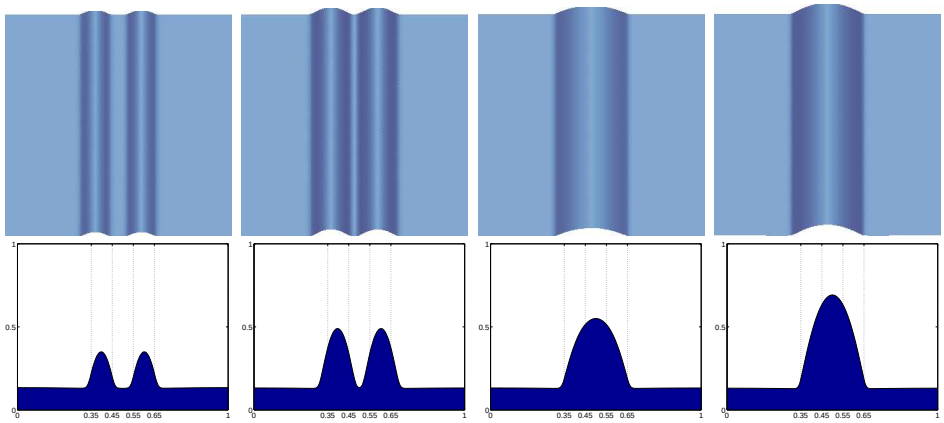


Abbildung 6.11: *Kondensation auf zwei parallelen, hydrophilen Streifen*: Numerische Simulation zum Substrat (6.20) und einen Potential (6.21) mit Anfangsdaten $u_0(x) = 0.125$. Kondensation wird modelliert durch $q(u) = \frac{0.2}{u+0.01}$. Die Bilder zeigen den Film zu den Zeiten 0.034, 0.070, 0.079, 0.101.

Abbildung 6.9 präsentiert numerische Ergebnisse zu diesem Problem. Dabei wurde der Einfluss der Kondensationsgeschwindigkeit genauer untersucht. Die Kondensationsgeschwindigkeit wird durch den Parameter C_1 festgelegt, in Abbildung 6.9 nimmt C_1 die Werte $\frac{1}{10}, \frac{1}{4}$ und 1 an. In allen drei Fällen bildet sich auf dem zentralen hydrophilen Gebiet ein Tropfen, um den herum eine Zone ohne Flüssigkeitstropfen entsteht. Je schneller die Kondensation, desto kleiner ist der Radius dieser Zone.

Der Mechanismus dahinter ist der folgende: Die in der Nähe des hydrophilen Gebietes kondensierte Flüssigkeit wird aufgrund des Potentialunterschiedes auf das hydrophile Gebiet gesogen, dort entsteht dadurch ein Tropfen. Da ein großer Tropfen energetisch günstiger ist als mehrere kleinere Tropfen, wächst der zentrale Tropfen auf Kosten der in der Umgebung kondensierten Masse weiter an - so entsteht die freie Zone. Im äußeren Gebiet findet aufgrund der auf die Anfangsdaten addierten Störung spinodale Entnetzung statt, welche wie im Falle eines homogenen Substrates (siehe Abb 6.8) einzelne Tropfen entstehen lässt.

Bei langsamer Kondensation haben die entstehenden Tropfen (sowohl der zentrale Tropfen als auch die äußeren Tropfen) mehr Zeit, die sie umgebende Masse aufzusammeln. Dadurch werden die um jeden Tropfen entstehenden freien Zonen größer.

Bei Experimenten mit kondensierendem Diethylenglykol konnten ähnliche Beobachtungen gemacht werden [39]. Abb 6.10 zeigt die Ergebnisse dieser Experimente – wie in Abbildung 6.9 entstehen bei Kondensation rund um den lyophilen Kreis Zonen ohne kondensierte Tropfen, sogenannte Verarmungszonen (*engl. depletion zones*). Leider sind die Experimente nicht direkt mit den numerischen Simulationen vergleichbar, da im Gegensatz zum Modellsystem Polystyrol auf Silizium die hier vorliegenden Grenzflächenpotentiale nicht exakt bekannt sind. Entsprechende Experimente mit Polystyrolfilmen wurden bisher nicht durchgeführt.

Das nächste Beispiel geht von einer etwas komplizierteren Geometrie aus. Hier ist das Substrat aus Streifen verschiedener Materialien aufgebaut:

$$\begin{aligned} &\Omega_1 = [0,1]^2 \setminus \Omega_2, \\ &\Omega_2 = \{x \in \mathbb{R}^2 : 0.35 \le x_1 \le 0.45\} \cup \{x \in \mathbb{R}^2 : 0.55 \le x_1 \le 0.65\}, \end{aligned} \tag{6.20}$$

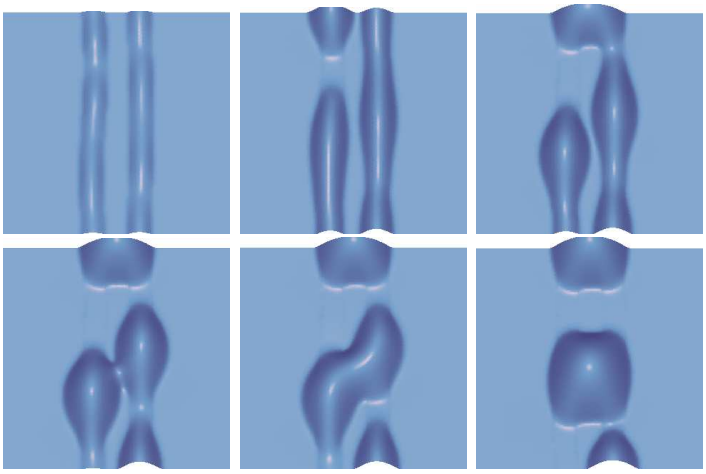


Abbildung 6.12: *Symmetriebruch bei gestörter paralleler Konfiguration*: Numerische Simulation zum Substrat (6.22), dem Potential (6.21) und den Anfangsdaten $u_0(x) = 0.125$. Kondensation wird modelliert durch $q(u) = \frac{0.2}{u+0.01}$. Die Bilder zeigen den Film zu den Zeiten 0.016, 0.038, 0.046, 0.052, 0.053, 0.061.

wobei Ω_2 im Vergleich zu Ω_1 hydrophil ist:

$$w(u,x) = \begin{cases} -u^{-2} + 10^{-6}u^{-8} & \text{falls } x \in \Omega_1, \\ -\frac{1}{2}u^{-2} + \frac{1}{2}\cdot 10^{-6}u^{-8} & \text{falls } x \in \Omega_2. \end{cases} \tag{6.21}$$

Die verwendete Näherung W von w ist durch (6.4) gegeben. Die numerische Simulation (Abb. 6.11) zeigt, dass sich die Flüssigkeit bei Kondensation auf diesem Substrat zunächst auf den beiden hydrophilen Streifen sammelt. Durch Kondensation weiterer Masse verbreitert sich der Flüssigkeits-Schlauch über die Grenze von Ω_2 hinaus bis schließlich beide Schläuche zusammengewachsen sind. Die Lösung bleibt dabei in y-Richtung konstant, da Anfangsdaten und Geometrie unabhängig von y sind.

Interessanter ist es daher, die Symmetrie des Substrates zu brechen, indem man eine kleine Störung addiert:

$$\begin{aligned} &\Omega_1 = [0,1]^2 \setminus \Omega_2, \\ &\Omega_2 = \{x \in \mathbb{R}^2 : 0.35 \le x_1 + S_1(x_2) \le 0.45\} \cup \{x \in R^2 : 0.55 \le x_1 + S_2(x_2) \le 0.65\}. \end{aligned} \tag{6.22}$$

Dabei ist

$$\begin{aligned} S_1(x_2) &= 0.01\sin(2\pi x_2), \\ S_2(x_2) &= 0.01\sin(\pi x_2 - \pi). \end{aligned} \tag{6.23}$$

Dadurch zerreißt die Flüssigkeit in mehrere Tropfen, welche versuchen, energetisch möglichst günstige Positionen auf den beiden Streifen einzunehmen. Wenn die angesammelte Masse groß genug wird, findet ein Benetzungsübergang statt und die auf die beiden Streifen verteilte Masse schließt sich zu einem großen Tropfen zusammen. Die dabei entstehenden Muster ähneln denen von kondensierenden Wassertropfen auf einem abwechselnd hydrophoben/hydrophilen Substrat [16] (siehe auch Abbildung 1.1).

Im Gegensatz zu dem in Abbildung 6.9 gezeigten Beispiel entstehen in den Abbildungen 6.11 und 6.12 keine Tropfen im äußeren Bereich des hydrophoben Gebietes. Dies liegt daran, dass der betrachtete Bereich so klein, bzw. die Kondensation so langsam gewählt wurde, dass die hier um das hydrophile Gebiet auftretenden Verarmungszonen größer sind als das betrachtete Gebiet Ω.

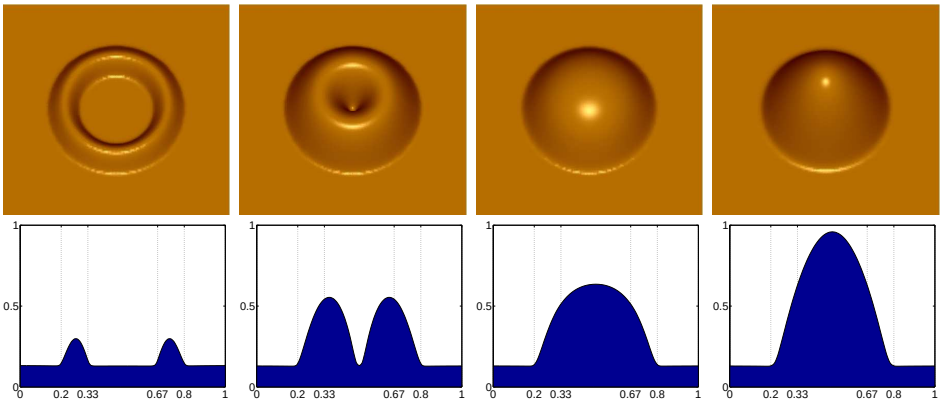


Abbildung 6.13: *Kondensation auf einem hydrophilen Kreisring*: Numerische Simulation zu einem durch (6.24) gegebenen Substrat, einem Grenzflächenpotential (6.21), $q(u) = \frac{0.2}{u+0.01}$ und Anfangswerten $u_0(x) = 0.1$. Die Bilder zeigen das Substrat $[0,1]^2$ zu den Zeiten $0.036, 0.078, 0.079, 0.096$. Die untere Zeile zeigt den Querschnitt entlang der Linie $x_2 = 0.5$.

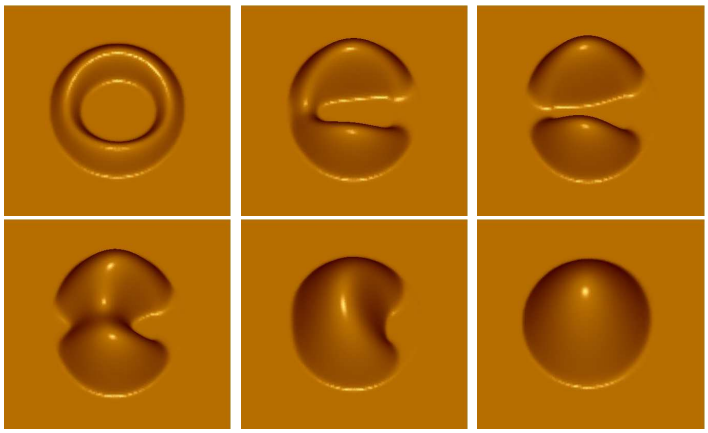


Abbildung 6.14: *Kondenstation auf einem hydrophilen, nahezu kreisförmigen Ring*: Numerische Simulation zu einem durch (6.25) gegebenen Substrat und einem Grenzflächenpotential (6.21). Es gilt $q(u) = \frac{0.2}{u+0.01}$ und $u_0(x) = 0.1$. Die Bilder zeigen das Gebiet $\Omega = [0,1]^2$ zu den Zeiten 0.044, 0.066, 0.073, 0.089, 0.091, 0.111.

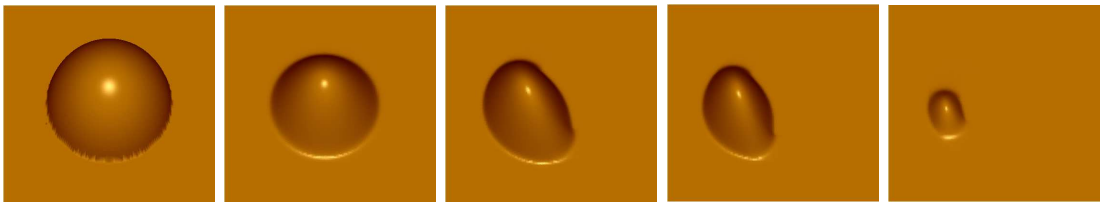


Abbildung 6.15: *Evaporation auf einem hydrophilen, nahezu kreisförmigen Ring*: Die Bilder zeigen die Evaporation eines Tropfens mit Radius 0.3 und Anfangshöhe 1.0, dabei ist $q(u) = -\frac{0.1}{u+0.01}$. Substrat und Potential sind wie in Abbildung 6.14 gewählt. Die Bilder zeigen den Tropfen zu den Zeiten $t = 0, 0.13, 0.15, 0.19, 0.24$.

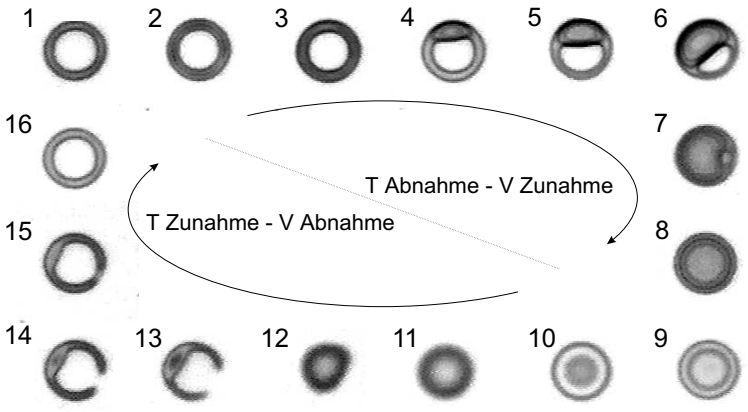


Abbildung 6.16: *Benetzungsübergänge bei Kondensation und Evaporation:* Bild 1-8: Bei Abnahme der Temperatur T nimmt das Volumen V der Flüssigkeit (hier Schwefelsäure) auf dem hydrophilen Ring (Außendurchmessser 7.5 μm) zu. Bild 9-16: Durch Heizen Volumenabnahme (Experimentelle Resultate von C. Schäfle [37]).

Ähnlich wie kondensierende Tropfen auf parallelen Streifen verhalten sich Tropfen, die auf hydrophilen Kreisringen kondensieren. In Abbildung 6.13 besteht das Substrat aus einem vergleichsweise hydrophilen Kreisring in einer hydrophoben Umgebung:

$$\begin{aligned} &\Omega_1 = [0,1]^2 \setminus \Omega_2, \\ &\Omega_2 = B_{0.3}(x_0) \setminus B_{0.17}(x_0), \quad x_0 = (\tfrac{1}{2}, \tfrac{1}{2}), \end{aligned} \tag{6.24}$$

und das Grenzflächenpotential ist wiederum durch (6.21) gegeben. Wie im Fall der parallelen Streifen sammelt sich die hinzukommende Flüssigkeit zunächst ausschließlich auf dem Kreisring. Der entstehende ringförmige Tropfen wächst dann aber über das hydrophile Gebiet hinaus. Aufgrund der Oberflächenspannung erfolgt das Wachstum zur Kreismitte hin, so dass sich schließlich ein einziger großer Tropfen bildet.

In Abbildung 6.14 sind die Grenzen des hydrophilen Gebietes Ω_2 nicht durch zwei Kreisringe, sondern lediglich durch zwei fast kreisförmige Ellipsen mit leicht verschobenen Brennpunkten gegeben:

$$\begin{aligned} &\Omega_1 = [0,1]^2 \setminus \Omega_2, \\ &\Omega_2 = \{x \in \mathbb{R}^2 : \left(\tfrac{x_1-0.505}{0.175}\right)^2 + \left(\tfrac{x_2-0.495}{0.17}\right)^2 \geq 1 \text{ und } \left(\tfrac{x_1-0.5}{0.295}\right)^2 + \left(\tfrac{x_2-0.505}{0.305}\right)^2 \leq 1\} \end{aligned} \tag{6.25}$$

In dieser Situation verliert die Lösung schnell ihre Symmetrie. Wie schon im Fall der leicht gewellten Streifen bildet die Flüssigkeit mehrere Tropfen aus, welche über Teilen des hydrophilen Gebietes entstehen. Mit der Zeit wachsen diese aber wieder zusammen, bis schließlich der gesamte Ring bedeckt ist.

Den umgekehrten Prozess zeigt Abbildung 6.15. Hier beginnt die Simulation mit einem Tropfen. Da die rechte Seite q negativ ist, nimmt die Masse des Tropfens ab. Trotzdem bleibt der Tropfen zunächst in einer zentralen Position. Sobald aber der Rand des kleiner werdenden Tropfens die innere Grenze des hydrophilen Gebietes erreicht hat, bewegt sich der Tropfen über das hydrophile Gebiet. Es bildet sich allerdings nicht wieder die Ringstruktur aus, die in den ersten Bildern der Abbildung 6.14 zu sehen ist.

Ähnliche Benetzungsmuster wurden in Experimenten mit Schwefelsäure-Tropfen auf hydrophilen Kreisringen beobachtet. Abbildung 6.16 zeigt diese Resultate.

Kapitel 7

Resümee

Ziel dieser Arbeit war es, ein konvergentes numerisches Verfahren zu entwickeln, welches die Dynamik dünner Flüssigkeitsfilme und die Ausbildung fluider Strukturen auf inhomogenen Oberflächen simuliert. In Kapitel 2 wurde gezeigt, dass diese Dynamik auf chemisch inhomogenen Substraten durch die Dünne-Filme-Gleichung

$$\partial_t u - \operatorname{div}\left(m(u)\nabla(-\Delta u + w_{,u}(u,x))\right) = q(u) \tag{7.1}$$

mathematisch beschrieben werden kann. Es wurde gezeigt, dass die Inhomogenität des Substrats durch ein x-abhängiges Grenzflächenpotential $w(u,x)$ beschrieben werden kann. Dieses ist in der Dünne-Filme-Näherung unstetig entlang der Grenze zwischen zwei verschiedenen Substratmaterialien.

Da man bei der Herleitung des Potentials w und des Quellterms q in Kapitel 2 davon ausging, dass das Substrat von einem Flüssigkeitsfilm bedeckt ist, sind $w(u,x)$ und $q(u)$ zunächst einmal nur für $u > 0$ definiert. Um numerische Stabilität des in 3.2.2 definierten Finite-Elemente-Verfahrens zu garantieren, mussten daher Funktionen W und Q definiert werden, die für alle $u \in \mathbb{R}$ definiert sind. Unter geeigneten Voraussetzungen an W und Q konnte dann die Existenz einer diskreten Lösung $U_{\tau h}$, welche Energie- und Entropieabschätzung erfüllt, gezeigt werden.

Um die diskrete Lösung des Verfahrens auch numerisch berechnen zu können, wurde das C++ Programmpaket EConLub2D erstellt und implementiert. Mit diesem lassen sich Dünne-Filme-Probleme in Raumdimension $d = 2$ sowohl für homogene als auch für inhomogene Oberflächen lösen.

In Kapitel 5 wurde gezeigt, dass die diskreten Lösungen $U_{\tau h}$ des Finite-Elemente-Verfahrens für $\tau, h \to 0$ gegen eine kontinuierliche Lösung u konvergieren, die die Dünne-Filme-Gleichung (7.1) im folgenden Sinn löst:

$$\int_\Omega \langle \partial_t u\,, \phi\rangle_{W^{1,r}(\Omega)' \times W^{1,r}(\Omega)} + \int_{[u>0]} m(u)\nabla p \nabla\phi = \int_\Omega Q(u)\phi \quad \forall \phi \in L^2(0,T,W^{1,r}(\Omega)). \tag{7.2}$$

Dabei ist $r = 2$ für $d = 1$ und $r > 2$ für $d = 2$. Der Druck $p \in L^2(\Omega_T)$ ist durch

$$p = -\Delta u + W_{,u}(u,x) \tag{7.3}$$

gegeben. Die Lösung u, deren Existenz damit für den inhomogenen Fall erstmalig bewiesen wurde, ist damit zunächst einmal eine Lösung zu dem Grenzflächenpotential W und der rechten Seite Q.

Es stellte sich also die Frage, wann die Funktionen W und Q darin durch w und q ersetzt werden können, oder aber zumindest so gewählt werden können, dass $w(\cdot, x) = W(\cdot, x)|_{\mathbb{R}^+}$ und $q = Q|_{\mathbb{R}^+}$ gilt.

Ist q durch (2.44) gegeben, so kann im Fall von Kondensation immer eine stetige Fortsetzung Q mit $q = Q|_{\mathbb{R}^+}$ gewählt werden, welche die geforderte Bedingung (Q1) erfüllt. Beschreibt (2.44) Evaporation, so ist es nicht möglich, Q so zu wählen, dass $q = Q|_{\mathbb{R}^+}$ gilt, da Bedingung (Q2) u.a. verlangt, dass $Q(u) = 0$ auf dem positiven Intervall $[0, \delta_q]$ gilt. Dies bedeutet, dass es nicht möglich ist, ein Verdunsten der gesamten Masse zu simulieren, es wird immer ein dünner Restfilm auf dem Substrat verbleiben. Andererseits wird das Verdunsten des Restfilms durch die Gleichungen auch gar nicht beschrieben, so dass es aus physikalischer Sicht wenig Sinn macht, die Simulation bis zu diesem Zeitpunkt durchführen zu wollen.

Ein Grenzflächenpotential w vom Typ (w2), z.B. ein vom Lennard-Jones-Potential abgeleitetes, ist singulär in $u = 0$. Daher kann für W lediglich für $u \geq \varepsilon_w > 0$ gelten: $W(u, x) = w(u, x)$. In Raumdimension $d = 1$ ließ sich W in Gleichung (7.2) aber trotzdem durch w ersetzen, da in Satz 5.1.10 gezeigt werden konnte, dass die Lösung größer als ε_w bleibt.

In Dimension $d = 2$ war dies nicht mehr möglich, was für die Praxis der numerischen Berechnung allerdings zweitrangig ist. Es konnte nämlich in Satz 6.1.1 gezeigt werden, dass ein $\varepsilon_w > 0$ in Abhängigkeit von der Gitterweite h so klein gewählt werden kann, dass die diskrete Lösung $U_{\tau h}$ größer als ε_w bleibt, also das diskrete Problem für $W = w$ löst. Für $h \to 0$ könnte das Minimum der diskreten Lösung aber gegen 0 konvergieren. In allen in Kapitel 6 untersuchten Beispielen deutet aber nichts auf ein solches Verhalten hin: Das Minimum der diskreten Lösung ist hier ausschließlich durch das Minimum des Grenzflächenpotentials w bestimmt und nicht durch die Wahl der Gitterweite h oder des Parameters ε_w.

Dadurch verbleibt bei allen Simulationen zu einem Potential zum Typ (w2) immer ein Restfilm auf dem Substrat, deren Dicke dem Minimum des effektiven Grenzflächenpotentials w entspricht. Es findet also keine komplette Entnetzung statt. Bei Polystyrolfilmen auf einem Siliziumsubstrat wird ein solcher Restfilm tatsächlich experimentell beobachtet. Das Lennard-Jones-Modell ist also, wie auch der in den Abbildungen (6.2) und (6.5)-(6.7) durchgeführte Vergleich zwischen numerischer Simulation und physikalischem Experiment zeigt, in dieser Situation ein geeignetes Modell.

In einigen anderen Experimenten ist ein solcher dünner Restfilm nicht zu beobachten, so zum Beispiel bei den in Abbildung 1.1 abgebildeten Wassertropfen. Hier stellt sich die Frage, ob solche Situationen durch ein Dünne-Filme-Modell beschrieben werden können. Voraussetzung dafür ist unter anderem, dass die beobachteten Tropfen noch einen *dünnen* Film bilden – und dies ist bei den nahezu runden Wassertropfen nicht mehr erfüllt.

In den in Abbildung 6.16 gezeigten Kondensations-Experimenten sind die beobachteten Kontaktwinkel zwar kleiner, aber auch bei diesen Experimenten kann die Existenz eines dünnen Restfilms auf dem hydrophoben Teil des Gebietes experimentell nicht bestätigt werden. Es bleibt also unklar, ob das Grenzflächenpotential hier mit einem Lennard-Jones-Ansatz modelliert werden kann. Ein Grenzflächenpotential, welches keinen Restfilm verursachen würde, ist aber nur schwer zu realisieren: ein rein destabilisierendes Potential der Art $w(u) = -\frac{A}{12\pi}u^{-2}$ würde nämlich bewirken, dass sich die gesamte Masse in einem Punkt

sammelt und für ein Potential mit einem Minimum bei $u = 0$ gibt es keine physikalische Herleitung.

Nichtsdestoweniger zeigen die hier durchgeführten numerischen Simulationen, dass das Lennard-Jones-Modell in der Lage ist, die in den Kondensations-Experimenten beobachteten Phänomene – von dem Vorhandensein eines Restfilms einmal abgesehen – zumindest qualitativ zu beschreiben.

Anhang A

Dokumentation des C++-Programmpaketes EConLub2D

Zum Lösen der Dünne-Filme-Gleichung mittels zweidimensionalen Finiten Elementen ist das C++-Programmpaket EConLub2D (**e***ntropy* **con***sistent solver for* **lub***rication-type equations*) entwickelt und implementiert worden. Es beinhaltet im Wesentlichen drei Komponenten: Erstens eine Sammlung von Klassen und Funktionen, welche die notwendigen Vektor- und Matrixstrukturen zur Verfügung stellen, zweitens die für Finite Elemente notwendigen Gitterstrukturen, und drittens einen Löser für die Dünne-Filme-Gleichung.

Alle Klassen sind im namespace `ecl` definiert.

Hinweis: Die hier abgedruckten Klassendefinitionen enthalten nicht immer alle im Code vorhandenen Daten, Funktionen und Operatoren – der Übersichtlichkeit halber werden lediglich die wichtigsten Elemente dargestellt.

A.1 Hilfsmittel der linearen Algebra

Das Lösen einer Differentialgleichung verlangt – früher oder später – das Lösen eines linearen Gleichungssystems. Daher braucht EConLub2D Klassen und Routinen, welche die notwendigen Elemente der linearen Algebra implementieren. Zum Rechnen mit `double` Vektoren und Matrizen sind im Headerfile matrix.h die Klassen `vector` und `matrix` definiert, dünn besetzte Matrizen können mit Hilfe der in sparsematrix.h definierten Klasse `SparseMatrix` dargestellt werden. Die in solver.h definierte Klasse `solver` schließlich stellt iterative Löser für lineare Gleichungssysteme bereit. In den folgenden Abschnitten sind diese Klassen genauer beschrieben.

A.1.1 Die Klasse `vector`

Die Klasse `vector` definiert einen Vektor mit `double` Einträgen. Ihre Deklaration lautet:

```
class vector{

 protected:
```

```
  double *data;                // Array der Einträge
  int dim;                     // Dimension

 public:
  vector(int dim=0);           // Konstruktor
  vector(const vector&);
  ~vector();                   // Destruktor

  void setDim(int dim);        // ändert Dimension (löscht Einträge)
  int getDim(){return dim;};   // gibt Dimension zurück
  void clear();                // setzt alle Einträge gleich 0

  double& operator[](int i) {return data[i];};
};
```

Es existieren zwei verschiedene Konstruktoren. Der erste erzeugt einen `vector` der Länge `dim` und setzt alle Einträge des Vektors auf 0. Der zweite Konstruktor wird mit `vector u=v` aufgerufen und erzeugt ein Kopie des Vektors `v`.

Das Überladen des Operators `[]` bewirkt, dass man auf den `i`-ten Eintrag eines Vektors `v` mit `v[i]` zugreifen kann. Des weiteren sind für `vector` unter anderem die Operatoren `*` und `=` überladen, welche das Skalarprodukt zweier Vektoren ausrechnen bzw. zwei Vektoren gleichsetzen.

A.1.2 Die Klassen `matrixBase` und `matrix`

Um Operationen mit Matrizen durchführen zu können, ohne zu wissen, wie diese Matrix im Detail abgespeichert ist, ist eine rein virtuelle Basisklasse `matrixBase` gegeben. Diese stellt genau die Operationen zur Verfügung, die ein iterativer Löser braucht: Man kann mit `getN()` bzw `getM()` die Dimension der Matrix erfragen und man kann eine Multiplikation mit einem `vector` ausführen.

```
class matrixBase{
 public:
  matrixBase(){};
  ~matrixBase(){};
  virtual vector& multiply(vector& x, vector& lsg)=0;
  virtual vector& multiply(vector& x, vector& lsg, list<int>&)=0;
  virtual int getN()=0;
  virtual int getM()=0;
};
```

Die Routine `multiply` ist in zwei Versionen vorhanden. Die erste erwartet zwei Parameter: den Vektor `x`, mit welchem die Matrix multipliziert werden soll, und den Vektor `lsg`, in welchen das Ergebnis der Multiplikation geschrieben wird. Zurückgegeben wird eine Referenz auf `lsg`. Die zweite Variante erhält zusätzlich noch eine Liste mit `int`-Werten übergeben. Diese Routine reduziert den Vektor und die Matrix auf die in der Liste aufgeführten Einträge und führt anschließend die Multiplikation durch. Enthält die Liste zum

Beispiel nur die Einträge (1,3,17), so wird lediglich die entsprechende 3×3 Matrix mit dem Vektor `(x[1],x[3],x[17])` multipliziert. Die Nützlichkeit einer solchen Multiplikationsroutine wird in den weiteren Abschnitten deutlich werden. Die Klasse `list` ist Teil der STL-Bibliothek , eine genaue Beschreibung findet sich z.B. in [26].

Eine einfache Implementation einer $m \times n$ Matrix mit `double`-Einträgen ist die folgende, von `matrixBase` abgeleitete Klasse:

```
class matrix : public matrixBase{
 protected:
  vector  *col;               // col[i] ist i-te Zeile
  int m;                      // Anzahl der Zeilen
  int n;                      // Anzahl der Spalten

 public:
  matrix(const int dim=0);          // Konstruktoren
  matrix(const int m,const int n);
  matrix(const matrix&);

  ~matrix();                        // Destruktor

  int getN(){return n;};
  int getM(){return m;};

  void setDim(int m,int n);     // ändert Dimension (löscht Einträge)
  void clear();                 // setzt alle Einträge auf 0

  vector& operator[](int i) {return col[i];};

  vector& multiply(vector& x, vector& lsg);
  vector& multiply(vector& x, vector& lsg, list<int>&);
};
```

Ein Aufruf des ersten Konstruktors erzeugt eine `dim`×`dim` Matrix, ein Aufruf des zweiten Konstruktors erzeugt eine `m`×`n` Matrix. Beide Konstruktoren setzen alle Einträge auf 0. Der dritte Konstruktor schließlich erzeugt eine Kopie der übergebenen Matrix. Es ist möglich, nachträglich mit `setDim` die Dimension der Matrix zu ändern, dabei gehen allerdings alle bisherigen Einträge verloren. `setDim` setzt alle Einträge wieder auf 0.

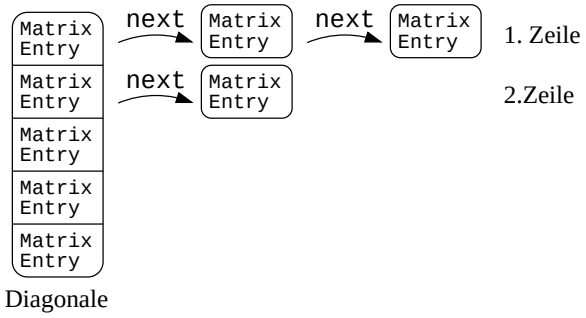
Wie schon bei der Klasse `vector` bewirkt das Überladen des Operators `[]`, dass man mit `M[i][j]` auf das ij-te Element einer Matrix `M` zugreifen kann.

A.1.3 Dünn besetzte Matrizen – die Klasse `SparseMatrix`

Ebenfalls von `matrixBase` abgeleitet ist die Klasse `SparseMatrix`, die eine dünn besetzte, quadratische Matrix implementiert. `SparseMatrix` speichert nur die Diagonale und die Einträge ungleich Null ab. Jeder Eintrag der dünn besetzten Matrix ist selbst ein Objekt der Klasse `MatrixEntry`, welche neben dem Wert des Eintrags noch die Spalte des Eintrags und einen Zeiger auf einen weiteren Eintrag enthält:

```
class MatrixEntry{
 protected:
   MatrixEntry *next;       // nächster Eintrag
   int         col;         // Spaltenindex des Eintrags
   double      value;       // Wert des Eintrags

   // [...]
};
```

Eine `SparseMatrix` besteht also aus einem Array von `MatrixEntry`, der Diagonalen. Hat die Matrix noch weitere Einträge, so werden diese mit Hilfe des `MatrixEntry::next` Zeigers an das Diagonalelement der entsprechenden Zeile angehängt. Es entsteht also folgende Struktur:



Für den Anwender ist es jedoch nicht notwendig, die genaue interne Struktur zu kennen, da geeignete Hilfsfunktionen zum Zugriff auf die einzelnen Einträge zur Verfügung stehen. Wichtige Funktionen der Klasse `SparseMatrix` sind neben den überladenen `multiply` Routinen der Konstruktor:

```
SparseMatrix::SparseMatrix(const int dim=0);
```

welcher eine quadratische Matrix der Dimension `dim`×`dim` erzeugt,

```
void SparseMatrix::clear();
```

welche alle Einträge wieder gleich Null setzt und

```
void SparseMatrix::set(int row, int col, double value);
void SparseMatrix::add(int row, int col, double value);
```

welche einen Wert `value` an die durch `row` und `col` definierte Stelle setzen (`set`) beziehungsweise hinzuaddieren (`add`). Damit nicht ungünstigstenfalls bei jedem Aufruf von `add` oder `set` eine neue Instanz von `MatrixEntry` erzeugt wird[1] sind die Operatoren

```
void* MatrixEntry::operator new(size_t);
void MatrixEntry::operator delete(void*, size_t);
```

überladen. Der überladene Operator holt den Speicher aus einem Puffer und stellt ihn bei Aufruf von `delete` wieder in den Puffer zurück.

[1] Das Beschaffen von Arbeitsspeicher ist eine vergleichsweise langsame Operation.

A.1.4 Iterative Löser – die Klasse `solver`

Zum Lösen eines linearen Gleichungssystems $Ax = b$ sind in der Klasse `solver` einige iterative Löser implementiert. Die Fehlerschranken für die absoluten und relativen Fehler und die maximale Anzahl der Iterationen können vor Aufruf des Lösers mit den Funktionen

```
void solver::setTolAbs(double x);
void solver::setTolRel(double x);
void solver::setMaxstep(int x);
```

gesetzt werden. `solver` enthält die beiden Konjugierte-Gradienten-Verfahren:

```
int solver::CG( matrixBase& A, vector& x, vector& b);
int solver::CG( matrixBase& A, vector& x, vector& b, std::list<int>&);
```

Das zweite Verfahren unterscheidet sich vom ersten lediglich dadurch, dass die Matrix `A` und die rechte Seite `b` auf die in der Liste angegebenen Einträge reduziert werden. Das CG-Verfahren terminiert, wenn der absolute Fehler klein genug ist oder die maximale Schrittweite erreicht wird. Rückgabewert des Verfahrens ist die Anzahl der benötigten Iterationsschritte. Der Rückgabewert ist 0, falls die Dimensionen von `A`, `x` und `b` nicht übereinstimmen. Außerdem ist ein BiCGstab-Verfahren (Definition siehe [45]) implementiert, sowohl ohne Vorkonditionierer

```
int solver::BiCGstab( matrixBase& A, vector& x, vector& b);
int solver::BiCGstab( matrixBase& A, vector& x, vector& b,
                      std::list<int>&);
```

als auch mit Vorkonditionierer:

```
int solver::PBiCGstab( matrixBase& A, vector& x, vector& b,
                       matrixBase& pre);
int solver::PBiCGstab( matrixBase& A, vector& x, vector& b,
                       matrixBase& pre, std::list<int>&);
```

Der Vorkonditionierer `pre` kann dabei eine in beliebiger Form abgespeicherte Matrix sein. Das BiCGstab-Verfahren und das vorkonditionierte BiCGstab-Verfahren werden beendet, falls der absolute *oder* der relative Fehler klein genug sind *oder* die maximale Anzahl an Iterationsschritten erreicht ist. Rückgabewert ist wiederum die Anzahl der benötigten Schritte. Wird ein Verfahren aus anderen Gründen (falsche Dimensionen der Vektoren und Matrizen oder ähnliches) beendet, so ist der Rückgabewert 0. Alle Verfahren schreiben die Lösung in den übergebenen Vektor `x`.

Von großem Vorteil ist, dass die Matrix `A` und der Vorkonditionierer `pre` lediglich als eine Referenz auf `matrixBase` übergeben werden. Dadurch funktionieren die Verfahren für jede Art Matrix und jeden Vorkonditionierer; zum Beispiel ist es nicht notwendig, zum Lösen der Gleichung $ABx = b$, wobei A und B dünn besetzte Matrizen sind, die Matrix AB zu berechnen. Dies wäre ungünstig, da es nicht möglich ist, zwei `SparseMatrix`-Objekte schnell zu multiplizieren, weil dies das Durchsuchen einer Spalte notwendig machen würde. Es genügt aber, eine von `matrixBase` abgeleitete Klasse zu implementieren, welche mittels der `multiply`-Routine $A(Bx)$ berechnen kann – und dies ist effizient und schnell. Ein Beispiel dieser Art ist die Klasse `NewtonMatrix`, welche im Abschnitt A.3.3.3 vorgestellt wird.

A.2 Finite Elemente

A.2.1 Datenstrukturen

Ein Finite-Elemente-Gitter besteht zunächst einmal aus Dreiecken. Die Ecken eines Dreiecks sind *lokal* mit 0,1,2 numeriert, ebenso die Kanten, wobei die Kante `i` der Ecke `i` gegenüberliegt. Jeder Eckpunkt eines Dreiecks ist aber gleichzeitig noch ein Knotenpunkt des Gitters, hat also auch noch einen *globalen* Index.

EConLub2D verwendet hierarchische Bisektionsgitter. Diese haben im Prinzip einen sehr einfachen Aufbau: Man beginnt mit einer einfachen Makrotriangulierung, die aus möglichst wenigen Dreiecken besteht. Feinere Gitter erzeugt man durch Unterteilen eines Dreiecks in zwei Teildreiecke. Dadurch entsteht eine Baumstruktur, in der jedes Dreieck (mit Ausnahme der Elemente der Makrotriangulierung) ein Vaterdreieck besitzt und eventuell zwei Kinderdreiecke.

A.2.1.1 Die Klasse `Element`

Herzstück der Gitterstruktur ist die Klasse `Element`, welche einem Dreieck der Triangulierung entspricht. Sie ist wie folgt definiert:

```
class Element{
 protected:
  double coord[3][2];
  int    node[3];
  int    nbEdgeIndex[3];

  Element *nb[3];       // Zeiger auf die Nachbar-Elemente (oder NULL)
  Element *child[2];    // Zeiger auf die Kinder-Elemente (oder NULL)
  Element *parent;      // Zeiger auf das Vater-Element

  int    level;
  int    flag;

 public:
  Element(double coord[3][2],int node[3], int nbEdgeIndex[3],
          Element *parent, int level);                  // Konstruktor
  ~Element();                                           // Destruktor

  // [...]
};
```

`coord` enthält also die Koordinaten der Eckpunkte an den lokalen Indizes $0, 1, 2$, `node[i]` speichert den globalen Index des Knotens mit dem lokalen Index `i`. Die Zeiger `child[i]` und `parent` erzeugen die Baumstruktur des hierarchischen Gitters, `level` zeigt die Gittertiefe eines Dreiecks an: Elemente der Makrotriangulierung haben Level 0, ihre Kinder Level 1, usw.

Die Nachbarschaftsbeziehungen innerhalb eines Gitters werden durch `nb` und `nbEdgeIndex` erfasst. `nb[i]` zeigt auf den Nachbarn jenseits der `i`-ten Kante. `nbEdgeIndex[i]` ist der lokale Index der `i`-ten Kante dieses Dreiecks im Nachbardreieck. Ist die `i`-te Kante eine Randkante, so ist `nb[i]=NULL`, und `nbEdgeIndex[i]` wird dazu benutzt um Informationen über die Art der Randwerte zu speichern (z.B `nbEdgeIndex[i]=-1` für Neumann-Randdaten). Die Informationen `nb[i]` und `nbEdgeIndex[i]` sind lediglich für die Elemente der Triangulierung[2] aktuell, für alle anderen Elemente ist `nb[i]=NULL`.

Der Parameter `flag` wird benutzt, um den Status eines Elementes anzuzeigen. Dabei können bitweise die folgenden Zustände an und ausgeschaltet werden:

`ACTIVE`	Das Element ist aktiv. Aktiv sind immer alle Elemente der Triangulierung und alle Eltern aktiver Elemente. Nicht-Aktiv sind Elemente, welche zwar noch im Speicher vorhanden sind, für alle Operationen auf dem Gitter aber unsichtbar sind.
`COARSE`	Das Element ist für ein Vergröbern markiert.
`REFINE`	Das Element ist zum Verfeinern vorgesehen.

Die Flags setzen die Bits `REFINE=1, COARSE=2` und `ACTIVE=4`. Ist ein Element aktiv und zum Verfeinern markiert, so ist also `flag=5`.

Der Konstruktor schließlich erzeugt einfach ein `Element` mit den übergebenen Werten und `flag=ACTIVE`. Der Destruktor löscht auch alle vorhandenen Kinder.

A.2.1.2 Die Klasse `MacroElement`

Ein `MacroElement` ist, wie sich durch die Namensgebung schon vermuten lässt, ein `Element`, welches Teil der Makrotriangulierung ist. Um gegebenenfalls mehrere Makroelemente miteinander verketten zu können, enthält die Klasse zusätzlich einen Zeiger `next`. Ansonsten ändert sich im Vergleich zur Klasse `Element` nichts:

```
class MacroElement : public Element {
  protected:
    MacroElement *next;

  public:
    MacroElement(double coord[3][2], int node[3], int nbEdgeIndex[3],
                 Element *parent, int level);
    ~MacroElement();
  };
```

A.2.1.3 Die Klasse `Mesh`

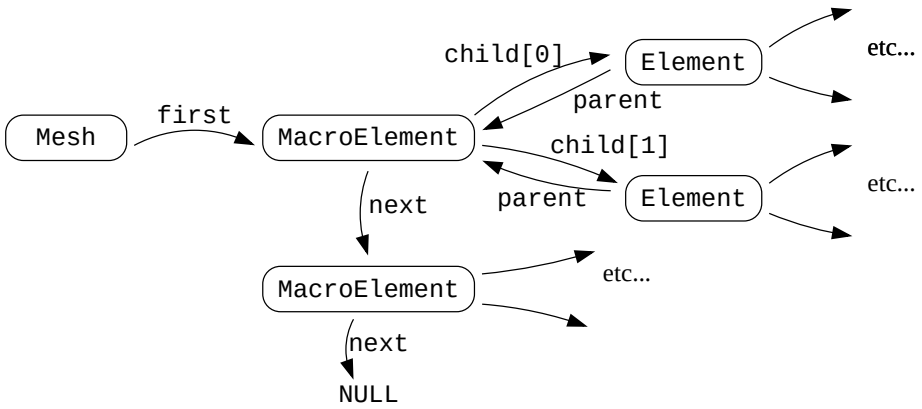
Nun fehlt also nur noch eine Klasse, welche die Organisation der verschiedenen `Element`- und `MacroElement`-Instanzen übernimmt. Dies ist die Klasse `Mesh`, welche wie folgt deklariert ist:

[2] Damit sind hier und später nur die Elemente der feinsten Zerlegung gemeint, nicht der ganze hierarchische Aufbau

```
class Mesh {

 protected:
   int maxNumberOfNodes;     // maximale Anzahl Gitterpunkte
   int maxDepth;             // maximale Gittertiefe
   int minDepth;             // minimale Gittertiefe

   MacroElement *first;      // erstes MacroElement

 public:
   Mesh(int maxNumberOfNodes, int minDepth, int maxDepth);
   virtual void createMacroTriangulation();

   std::list<int> active;
   std::list<int> sleeping;
   std::list<int> free;

   // [...]
};
```

`Mesh` legt also maximale und minimale Gittertiefe fest, ebenso die maximale Anzahl an Gitterpunkten. Letztere muß nicht unbedingt mit der auf dem Level `maxDepth` maximal möglichen Anzahl an Gitterpunkten übereinstimmen – für adaptiv verfeinerte Gitter machen auch andere Angaben Sinn.

`Mesh` speichert lediglich einen Zeiger `first` auf das erste Element der Makrotriangulierung. Besteht die Makrotriangulierung aus mehr als einem Makroelement, so werden die weiteren Makroelemente mit Hilfe des Zeigers `MacroElement::next` angehängt. Es entsteht also folgende Baumstruktur:



Der Konstruktor erhält die maximale Anzahl an Knoten, die minimale und die maximale Gittertiefe als Parameter übergeben. Er erzeugt die Makrotriangulierung durch Aufruf der Funktion `createMacroTriangulation()`. Ist diese Funktion nicht überladen, so wird das Einheitsquadrat $[0,1]^2 \subset \mathbb{R}^2$ durch zwei Dreiecke unterteilt. Anschließend wird das Gitter automatisch bis zum Level `minDepth` verfeinert.

Die Klasse `mesh` ermöglicht adaptives Verfeinern und Vergröbern des Gitters (mehr dazu im nächsten Abschnitt). Da beim Vergröbern eines Gitters natürlich nicht immer der Knoten mit der momentan höchsten Indexnummer entfernt wird, ist es nötig sich zu merken, welche Indizes gerade benutzt werden und welche nicht. Dies geschieht mit Hilfe der drei Listen `active, sleeping` und `free`. Die Liste `active` speichert die Indizes aller gerade von Elementen mit dem Flag `ACTIVE` benutzten Knoten (was allen Knoten der Triangulierung entspricht), `sleeping` sind alle darüberhinaus in nicht-`ACTIVE` Elementen vorhandenen Knoten, `free` enthält alle unbenutzten Indexnummern zwischen 0 und `maxNumberOfNodes-1`.

Dahinter verbirgt sich das folgende Konzept: Diskrete Funktionen $U \in V^h$ werden bekanntlich als Vektor dargestellt, jeder Eintrag U_i entspricht dem Wert der Funktion am Knotenpunkt mit dem Index i. Werden nun beim Vergröbern Knoten entfernt, so werden normalerweise die folgenden Anpassungen nötig: Erstens müssen die entsprechenden Einträge aus U gestrichen werden; aus einem 100-dimensionalen Vektor muss so zum Beispiel ein 95-dimensionaler Vektor gemacht werden. Zweitens muss die Indexnumerierung des Gitters dementsprechend angepasst werden. EConLub2D minimiert den notwendigen Zeitaufwand, indem auf diese Anpassungen verzichtet wird. Dies macht es allerdings notwendig, dass alle verwendeten Vektoren unabhängig von der momentanen Gittertiefe die Dimension `maxNumberOfNodes` haben. Gerechnet wird dann jedoch nur mit den sich gerade in der `active` Liste befindlichen Einträgen. Eine for-Schleife, welche alle aktiven Indizes und die Knotenwerte eines Vektors `u` ausgibt, sieht zum Beispiel wie folgt aus:

```
list<int>::const_iterator i;
for(i=active.begin();i!=active.end();++i){
    cout<<i<<u[*i];
};
```

Dies benötigt nicht mehr Zeitaufwand als eine normale for-Schleife. Lediglich der Speicheraufwand ist höher. Dies stellt aber in der Regel kein Problem dar.

A.2.2 Routinen der Klassen `Element` und `Mesh`

A.2.2.1 Die `traverse`–Routine

Um auf dem Gitter arbeiten zu können, ist eine Routine vonnöten, die es ermöglicht, auf einigen oder allen Elementen eine Aktion auszuführen. Dafür ist

```
void  Mesh::traverse(int tFlag, int elFlag, int maxlevel,
                     void (Element::*action)(void*), void* arg);
```

vorgesehen, welche auf allen Makroelementen die gleichnamige Funktion der Klasse `Element`

```
void Element::traverse(int tFlag, int elFlag, int maxlevel,
                       void (Element::*action)(void*), void* arg);
```

mit den gleichen Parametern aufruft. Diese wiederum ruft rekursiv die `traverse` Routinen ihrer Kinder auf und führt gegebenenfalls die übergebene Funktion `action` mit den Parametern `arg` aus. `action` kann eine beliebige, in der Klasse `Element` definierte Funktion sein, die einen Parameter vom Typ `void*` erwartet und `void` als Rückgabewert hat.

Die genaue Arbeitsweise der `Element::traverse` Routine wird durch die Parameter `tFlag`, `elFlag` und `maxlevel` bestimmt. Immer gilt, dass die Aktion `action` nur dann durchgeführt wird, wenn in dem entsprechenden Element der Flag `elFlag` gesetzt ist und das Element höchstens Gittertiefe `maxlevel` hat.

Der Parameter `tFlag` kann die Werte `PREFIX, POSTFIX, LEVEL` und `LEAVES` annehmen. Dabei bedeutet:

`PREFIX`:	`action` wird zuerst in einem Vaterelement ausgeführt, danach in den Kinderelementen
`POSTFIX`:	`action` wird zuerst in den Kinderelementen ausgeführt, danach im Vaterelement.
`LEVEL`:	`action` wird auf allen Elementen der Gittertiefe `maxlevel` durchgeführt.
`LEAVES`:	`action` wird auf allen Elementen durchgeführt, die keine aktiven Kinder haben (`child[i]` ist entweder `NULL` oder nicht `ACTIVE`) .

Als Beispiel sei hier die Implementation von `PREFIX` angegeben:

```
void Element::traverse(int tFlag, int elFlag, int maxlevel,
                       void (Element::*action)(void*), void* arg){
  switch(tFlag){
  case PREFIX:
    if (elFlag&flag){(this->*action)(arg);};
    if(child[0] && level<maxlevel){
      child[0]->traverse(tFlag,elFlag,maxlevel,action,arg);
      child[1]->traverse(tFlag,elFlag,maxlevel,action,arg);
    };
    break;
  // [...]
  };
};
```

A.2.2.2 Verfeinern und Vergröbern des Gitters

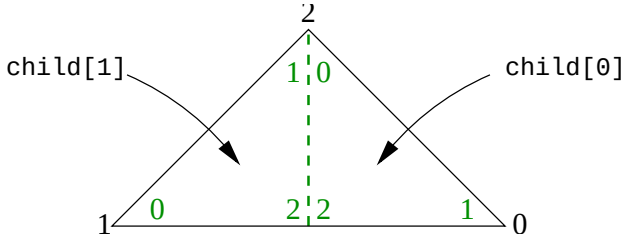
Um ein Dreieck durch Bisektion in zwei Kinderdreiecke zu unterteilen, wird der Mittelpunkt einer Kante als neuer Eckpunkt der beiden Kinder gewählt. Die solchermaßen unterteilte Kante wird auch *Verfeinerungskante* des Dreiecks genannt. Es stellt sich also die Frage, welche Kante eines Dreiecks als Verfeinerungskante zu wählen ist. Dabei muss verhindert werden, dass die gleiche Kante mehrfach hintereinander unterteilt wird, da so extrem spitze Dreiecke entstünden. Außerdem ist es a priori nicht gegeben, dass die Verfeinerungskante eines Dreiecks auch die Verfeinerungskante des ebenfalls zu unterteilenden Nachbardreiecks ist – eine Verfeinerung ist aber nur bei einer gemeinsamen Verfeinerungskante möglich.

Die von EConLub2D zum Unterteilen verwendete Methode `Element::recursiveRefine` geht deshalb wie folgt vor: sie testet zunächst, ob eine gemeinsame Verfeinerungskante mit dem Nachbarn vorliegt. Ist dies nicht der Fall, so wird zunächst das Nachbardreieck unterteilt. Anschließend wird das Dreieck selbst unterteilt.

Damit dies reibungslos funktioniert, ist es notwendig zu garantieren, dass nach nur einer Unterteilung des Nachbardreiecks eine gemeinsame Verfeinerungskante vorliegt. Dies wird durch die folgenden Regeln erreicht:

- Die Kante mit dem lokalen Index 2 ist die Verfeinerungskante.
- Ein neu eingefügter Knoten bekommt in den neuen Elementen den lokalen Index 2.

Die folgende Abbildung zeigt die Verteilung der lokalen Indizes beim Unterteilen eines Elementes. Die schwarzen Zahlen sind die lokalen Indizes des Elementes, die grünen Zahlen die lokalen Indizes der beiden Kinderelemente.



Die Funktion `recursiveRefine` lautet also wie folgt:

```
void Element::recursiveRefine(void *arg){
  if(nb[2] && nbEdgeIndex[2]!=2){          // gleiche Verfeinerungskante ?
    nb[2]->recursiveRefine(arg);
  };
  if(nb[2]){bisectInteriorEdge(arg);}      // Randelement ?
  else{bisectBorderEdge(arg);};
};
```

Das Unterteilen der Dreiecke selbst erledigen die Methoden `bisectInteriorEdge` (für Kanten im Inneren des Gebietes) und `bisectBorderEdge` (für Kanten am Rand des Gebietes).

Es ist theoretisch möglich, dass der rekursive Algorithmus nicht terminiert – und bei der Suche nach zwei Dreiecken mit gemeinsamer Verfeinerungskante in eine Schleife gerät (wenn z.B in einem Sechseck aus sechs Dreiecken jeweils die im Uhrzeigersinn liegende Kante Verfeinerungskante ist). Dies kann aber nur geschehen, wenn eine solche Schleife auch schon in der Makrotriangulierung möglich ist – und kann daher leicht vermieden werden.

Die Klasse `Mesh` stellt die folgenden Funktionen zur Verfügung, welche das Verfeinern und Vergröbern des Gitters managen. Zunächst einmal die beiden Routinen

```
void Mesh::markAllForRef();
void Mesh::markAllForCoa();
```

welche alle Elemente ohne Kinder (genauer: ohne Kinder, die als `ACTIVE` markiert sind) mit Hilfe der Flags `REFINE` und `COARSE` zum Verfeinern bzw. Vergröbern markieren. Zum Verfeinern gibt es die Routine

```
void Mesh::refineMesh(vector* v1=NULL, vector* v2=NULL);
```

welche alle mit `REFINE` markierten Elemente mit Hilfe der oben vorgestellten Funktion `recursiveRefine` verfeinert. Darüber hinaus können `refineMesh` noch bis zu zwei Vektoren übergeben werden, welche Funktionen aus dem Finite-Elemente-Raum V^h darstellen. Beim Verfeinern werden diese automatisch an das neue Gitter angepasst. Die Funktion `refineMesh` ist also ein Zweizeiler:

```
void Mesh::refineMesh(vector* v1,vector* v2){
  RefineStruct rs;rs.mesh=this;rs.v1=v1;rs.v2=v2;
  traverse(LEAVES, REFINE, maxDepth-1, &Element::recursiveRefine, &rs);
};
```

In der ersten Zeile werden die Parameter zu einem `struct` zusammengefasst, da sie der traverse-Routine als `void*` übergeben werden müssen. Der dafür nötige `RefineStruct` ist in structs.h definiert. Der Aufruf von `traverse` in der zweiten Zeile bewirkt, dass auf allen mit `REFINE` markierten Elementen ohne Kinder `recursiveRefine` aufgerufen wird. Ein Verfeinern über Level `maxDepth` hinaus wird verhindert, da die maximale Gittertiefe, auf der noch ein `recursiveRefine` durchgeführt wird, `maxDepth-1` ist. Nachdem ein Element verfeinert ist, werden die `REFINE` Flags automatisch entfernt.

Ähnlich funktioniert

```
void Mesh::refineToLevel(int level, vector* v=NULL, vector* v2=NULL);
```

Hier werden alle Teile des Gitters, welche noch nicht fein genug sind, bis zur Tiefe `level` verfeinert.

Für eine Vergröberung des Gitters ist die Funktion

```
void Mesh::coarseMesh();
```

zuständig. Vergröberung kann in einem hierarchischen Gitter wie `Mesh` ganz einfach stattfinden, indem die beiden Kinder eines Elements entfernt werden. Es können daher nur ganz bestimmte Gitterpunkte entfernt werden, da ein Zusammenfassen zweier beliebiger benachbarter Dreiecke zu einem neuen Dreieck nicht ohne Veränderung der gesamten Gitterstruktur möglich wäre. `coarseMesh` entfernt einen Gitterpunkt, falls alle davon betroffenen Dreiecke mit `COARSE` markiert sind und passt die Nachbarschaftsinformationen im Gitter dementsprechend an. Die dann überflüssigen Elemente werden aber nicht gelöscht, sondern sind lediglich nicht mehr `ACTIVE`, die Indexnummer des entfernten Knotens wandert also von der `active`-Liste in die `sleeping`-Liste. Wird eine vergröberte Stelle später wieder verfeinert, dann erzeugen die Routinen `bisectBorderEdge` und `bisectInteriorEdge` keine neuen Elemente, sondern wecken lediglich die schlafenden Kinder wieder auf.

Wie bei `REFINE` so gilt auch hier, dass alle `COARSE` Flags nach Aufruf von `coarseMesh` entfernt sind.

Endgültig entfernt werden alle schlafenden Elemente mit der Funktion

```
void Mesh::deleteSleepers();
```

Nach Aufruf dieser Funktion sind alle Elemente des Gitters, die vorher nicht `ACTIVE` waren, gelöscht. Die `sleeping`-Liste ist also leer, alle möglichen Indexnummern sind entweder in `free` oder in `active` enthalten.

A.3 Lösen der Dünne-Filme-Gleichung

A.3.1 Das Anfangs-Randwert-Problem

Die Klasse `Problem` stellt einen Entropie-konsistenten Löser für die Dünne-Filme-Gleichung zur Verfügung. Die Funktion `void Problem::solve()` löst auf $\Omega = [0,l]^2$, $I = (t_0, T)$ das kontinuierliche Anfangs-Randwertproblem:

$$\begin{aligned} \eta\partial_t u - \operatorname{div}(m(u)\nabla p) &= q(u) \text{ in } I \times \Omega, \\ p = -\varsigma\Delta u + w_{,u}(u,x) &\text{ in } I \times \Omega, \\ \frac{\partial u}{\partial \nu} = \frac{\partial p}{\partial \nu} &= 0 \text{ auf } \partial\Omega \times I, \\ u(0,.) &= u_0 \text{ in } \Omega. \end{aligned} \tag{A.1}$$

Dabei sind $\eta, \varsigma > 0$ beliebige positive Konstanten und die Mobilität $m(u)$ ist gegeben als

$$m(u) = cu^n, \quad n > 0. \tag{A.2}$$

Zur Vereinfachung der Rechnung und zur Stabilisierung löst `solve` allerdings nicht dieses Problem, sondern ein reskaliertes mit den dimensionslosen Größen:

$$\tilde{x} = \frac{x}{l_0}, \quad \tilde{u} = \frac{u}{h_0}, \quad \tilde{t} = \frac{V_0 t}{l_0}. \tag{A.3}$$

Die Skalierungsparameter l_0, h_0, V_0 werden dabei bestimmt durch

$$l_0 = l, \quad h_0 = \fint_\Omega u_0, \quad \frac{\eta V_0 h_0^{3-n}}{\varepsilon^3 \varsigma} = 1, \tag{A.4}$$

wobei $\varepsilon = \frac{h_0}{l_0}$ ist. Das reskalierte, dimensionslose Problem auf $\tilde{\Omega} = [0,1]^2$ und $\tilde{I} = (\frac{t_0 V_0}{l_0}, \frac{T V_0}{l_0})$ lautet also

$$\begin{aligned} \partial_{\tilde{t}}\tilde{u} - \tilde{\operatorname{div}}(c\,\tilde{u}^n \tilde{\nabla}\tilde{p}) &= \tilde{q}(\tilde{u}) \text{ in } \tilde{I} \times \tilde{\Omega}, \\ \tilde{p} = -\tilde{\Delta}\tilde{u} + \tilde{w}_{,\tilde{u}}(\tilde{u},\tilde{x}) &\text{ in } \tilde{I} \times \tilde{\Omega}, \\ \frac{\partial \tilde{u}}{\partial \tilde{\nu}} = \frac{\partial \tilde{p}}{\partial \tilde{\nu}} &= 0 \text{ auf } \partial\tilde{\Omega} \times \tilde{I}, \\ \tilde{u}(0,\tilde{x}) &= u_0(l_0\tilde{x}) \text{ in } \tilde{\Omega}, \end{aligned} \tag{A.5}$$

mit Funktionen

$$\tilde{q}(\tilde{u}) = \frac{h_0^{3-n}}{\varsigma\varepsilon^4} q(h_0\tilde{u}), \tag{A.6}$$

$$\tilde{w}(\tilde{u},\tilde{x}) = \frac{1}{\varsigma\varepsilon^2} w(h_0\tilde{u}, l_0\tilde{x}). \tag{A.7}$$

Dadurch ist es möglich alle Probleme auf dem Gebiet $[0,1]$ zu rechnen. Außerdem verschwinden die Konstanten η und ς aus der Rechnung. Es müssen lediglich die rechte Seite und das Potential entsprechend angepasst werden.

A.3.2 Parameter und Anfangsdaten

Die Parameter $t_0, T, l, \eta, \varsigma, n$ und c werden mit Hilfe der folgenden Funktionen der Klasse `Problem` gesetzt:

`void setStartT(double);`	Anfangszeit t_0
`void setT(double);`	Endzeit T
`void setLength(double);`	Gebietsgröße l
`void setEta(double);`	Viskosität η
`void setSigma(double);`	Oberflächenspannung ς
`void setMn(double);`	Mobilitäts-Exponent n
`void setMc(double);`	Mobilitäts-Koeffizient c

Für die Funktionen q, w und u_0 stehen eigene Basisklassen `rhs`, `potential` und `initial` zur Verfügung. Mit

```
void setInitialValue(initial*);
void setRHS(rhs*);
void setPotential(potential*);
```

werden in der Klasse `Problem` die entsprechenden Zeiger gesetzt. Diese Basisklassen sehen im einzelnen wie folgt aus:

A.3.2.1 Anfangswerte: `initial`

Die Deklaration der virtuellen Basisklasse `initial` lautet:

```
class initial{
 public:
   initial(){};
   virtual double f(double x1,double x2)=0;
 };
```

Um eigene Anfangsdaten zu implementieren, muss also eine von `initial` abgeleitete Klasse benutzt werden. Die folgende einfache Klasse `constant` erzeugt zum Beispiel einen Film mit der konstanten Anfangshöhe `height`:

```
class constant:public initial{
 protected:
   double height;
 public:
   constant(double h){height=h;};
   double f(double x1, double x2){return height;};
};
```

Um nun also eine Rechnung mit den Anfangsdaten $u_0(x) = 1$ durchzuführen, schreibt man im Hauptprogramm:

```
Problem p;
constant c(1.0);
p.setInitialValue(&c);
```

Die notwendige Skalierung der Anfangsdaten wird von `solve` automatisch vorgenommen, dies muss die Klasse `initial` nicht berücksichtigen.

A.3.2.2 Die Klasse `potential`

Die Basisklasse `potential`, die im Gegensatz zur Basisklasse `initial` nicht rein virtuell ist, sondern, falls sie nicht überladen ist, das Potential $w \equiv 0$ darstellt, ist wie folgt definiert:

```
class potential{

 public:
   potential(){};
   virtual void setScaling(double factor, double scalU, double scalX){};

   virtual double Wuplus(double u, double x, double y){return 0;};
   virtual double Wuminus(double u, double x, double y){return 0;};

   virtual double DWuplus(double u, double x, double y){return 0;};
};
```

Das Potential $w(u,x)$ wird dabei durch $w = w^+ + w^-$ in eine in u konvexe Funktion w^+ und eine in u konkave Funktion w^- aufgeteilt. Die in `potential` deklarierten Funktionen stellen allerdings nicht w dar, sondern die dimensionslose Version $\tilde{w}$. Genauer gilt:

$$\begin{aligned}
\texttt{Wuplus}(\tilde{u},\tilde{x}_1,\tilde{x}_2) &= \tilde{w}^+_{,\tilde{u}}(\tilde{u},\tilde{x}) = \frac{h_0}{\varepsilon^2\varsigma} w^+_{,u}(h_0\tilde{u}, l_0\tilde{x}),\\
\texttt{Wuminus}(\tilde{u},\tilde{x}_1,\tilde{x}_2) &= \tilde{w}^-_{,\tilde{u}}(\tilde{u},\tilde{x}) = \frac{h_0}{\varepsilon^2\varsigma} w^-_{,u}(h_0\tilde{u}, l_0\tilde{x}),\\
\texttt{DWuplus}(\tilde{u},\tilde{x}_1,\tilde{x}_2) &= \tilde{w}^+_{,\tilde{u}\tilde{u}}(\tilde{u},\tilde{x}) = \frac{h_0^2}{\varepsilon^2\varsigma} w^+_{,uu}(h_0\tilde{u}, l_0\tilde{x}).
\end{aligned} \tag{A.8}$$

Die Funktion `setScaling` wird von `solve()` nach Berechnung der Skalierung einmal aufgerufen und übergibt `potential` die Werte $(\frac{h_0}{\varsigma\varepsilon^2}, h_0, l_0)$.

Die Funktionsweise wird im folgenden Beispiel veranschaulicht: Sei w ein von x unabhängiges Lennard-Jones-Potential, gegeben durch

$$w(u) = -\frac{A}{12\pi}u^{-2} + Bu^{-8}.$$

Damit ist also

$$\begin{aligned}
w^+_{,u}(u) &= \frac{A}{6\pi}u^{-3},\\
w^-_{,u}(u) &= -8Bu^{-9}.
\end{aligned}$$

Dann muss gelten:

$$\begin{aligned} \texttt{Wuplus}(\tilde{u},\tilde{x}_1,\tilde{x}_2), &= -8B_s\tilde{u}^{-9} \\ \texttt{Wuminus}(\tilde{u},\tilde{x}_1,\tilde{x}_2), &= \frac{A_s}{6\pi}\tilde{u}^{-3} \\ \texttt{DWuplus}(\tilde{u},\tilde{x}_1,\tilde{x}_2), &= 72B_s\tilde{u}^{-10} \end{aligned} \tag{A.9}$$

und die Größen A_s und B_s werden durch `setScaling` wie folgt bestimmt:

$$\texttt{setScaling}(a,b,c)\{A_s = \frac{a}{b^3}A, \quad B_s = \frac{a}{b^9}B\}$$

Die Klasse `LJ` ist eine Implementation dieses Grenzflächenpotentials, ihr Konstruktor erhält die Werte von A und B übergeben:

```
LJ::LJ(double A, double B);
```

Um numerische Stabilität zu gewährleisten (w ist singulär in $u = 0$), ist nicht (A.9) implementiert, sondern eine Näherung $\tilde{W}$ von $\tilde{w}$:

$$\tilde{W}(\tilde{u}) = \begin{cases} \tilde{w}(\tilde{u}) & \text{falls } \tilde{u} \geq \varepsilon, \\ \tilde{w}(\varepsilon) + (\tilde{u}-\varepsilon)\tilde{w}_{,\tilde{u}}(\varepsilon) + \frac{1}{2}(\tilde{u}-\varepsilon)^2\tilde{w}_{,\tilde{u}\tilde{u}}(\varepsilon) & \text{falls } \tilde{u} < \varepsilon. \end{cases} \tag{A.10}$$

Damit ist sichergestellt, dass $\tilde{W}$ auf ganz $\mathbb{R}$ definiert ist. Der Parameter ε wird von der Klasse `LJ` selbst bestimmt und beträgt

$$\varepsilon = \frac{1}{4}\sqrt[6]{\frac{48\pi B_s}{A_s}} = \frac{1}{4}\operatorname*{argmin}_{\tilde{u}\in\mathbb{R}}\{\tilde{W}(\tilde{u})\}.$$

A.3.2.3 Inhomogene Substrate

Es gibt zwei Möglichkeiten, mit EConLub2D Potentiale für inhomogene Substrate zu implementieren. Die erste Möglichkeit ist, die oben beschriebene Klasse `potential` zu überladen und bei der Implementierung der Funktionen `Wuplus, Wuminus` und `DWuplus` die Inhomogenität mit zu berücksichtigen.
Für Gebiete $\Omega = \bigcup_{i=1}^{L}\Omega_i$ und ein Potential

$$w(u,x) = w^{(i)}(u), \quad \text{falls } x \in \Omega_i,$$

welches sich aus verschiedenen Potentialen $w^{(i)}(u)$ zusammensetzt, gibt es noch eine weitere Möglichkeit. Zunächst können mit Hilfe der Basisklasse `patch` die verschiedenen Teilgebiete Ω_i definiert werden. Dazu muss in der Klasse

```
class patch{
 public:
   patch(){};
   virtual bool isPatch(double x1,double x2){return true;};
};
```

lediglich die Funktion `isPatch` überladen werden. Diese gibt `true` zurück, falls der Punkt (`x1,x2`) zum durch `patch` dargestellten Gebiet gehören soll, ansonsten `false`. Wie aus der Definition ersichtlich, stellt die Basisklasse `patch` selbst das ganze Gebiet dar, da immer `true` zurückgegeben wird.

Die von `patch` abgeleitete Klasse `circle` stellt zum Beispiel einen Kreis mit Mittelpunkt (`Mx,My`) und Radius `R` dar:

```
class circle:public patch{
 protected:
   double R,Mx,My;
 public:
   circle(double r,double mx,double my){R=r;Mx=mx;My=my;};

   bool isPatch(double,double){
     if(pow(x-Mx,2)+pow(y-My,2)<=pow(R,2)) return true;
     return false;
   };
 };
```

Nun kann die von `potential` abgeleitet Klasse `inhomogen` genutzt werden, um aus den verschiedenen Flicken ein Potential zusammenzusetzen. Dazu steht die Funktion

```
void inhomogen::addPatch(patch* pat,potential* pot);
```

zur Verfügung, welche festsetzt, dass das Potential `inhomogen` auf dem durch `*pat` definiertem Gebiet das Potential `*pot` annimmt. Werden `inhomogen` durch `addPatch` mehrere sich überlagernde Flicken übergeben, so ist immer der zuletzt übergebene `patch` der entscheidende (er liegt sozusagen oben). Um also auf $\Omega = [0,1]^2$ das Potential

$$w(u,x) = \begin{cases} 0 & \text{falls } x \in B_{\frac{1}{4}}\left((\frac{1}{2},\frac{1}{2})\right), \\ -\frac{A}{12\pi}u^{-2} + Bu^{-8} & \text{sonst.} \end{cases}$$

zu definieren, schreibt man:

```
Problem p;
p.setLength(1.0);
inhomogen inh;

LJ pot1(A,B); patch pat1;
inh.addPatch(&pat1,&pot1);

potential pot2; circle pat2(0.25,0.5,0.5);
inh.addPatch(&pat2,&pot2);

p.setPotential(&inh);
```

Man sieht, dass es also nicht nötig ist, das Gebiet $\Omega \setminus B_{\frac{1}{4}}\left((\frac{1}{2},\frac{1}{2})\right)$ zu definieren. Es reicht

aus, zuerst auf dem ganzen Gebiet ein Potential festzulegen und dieses dann auf einem Teilgebiet zu überschreiben[3].

A.3.2.4 Die Klasse `rhs`

Ähnlich wie `potential` ist die Klasse `rhs` strukturiert, welche die rechte Seite der Differentialgleichung definiert:

```
class rhs{
 public:
   rhs(){};
   virtual void setScaling(double factor,double scalU){};
   virtual double f(double){return 0;};
};
```

Auch hier wird `setScaling` einmal von `solve` aufgerufen, diesmal mit den Parametern $(\frac{h_0^{3-n}}{\varsigma\varepsilon^4}, h_0)$. Als Rückgabe wird

$$\texttt{f}(\tilde{u}) := \tilde{q}(\tilde{u}) = \frac{h_0^{3-n}}{\varsigma\varepsilon^4} q(h_0\tilde{u})$$

erwartet. Der Stabilität des Verfahrens wegen darf die Funktion `rhs::f` keine Singularität in $\mathbb{R}$ haben. Falls die Funktion q eine solche hat, muss eine Näherung $\tilde{Q}$ für $\tilde{q}$ implementiert werden, welche ohne Singularität auskommt.

A.3.3 Die Funktion `solve()` im Detail

A.3.3.1 Beschreibung des Algorithmus – Verfahren und Fehlerschranken

`solve()` löst das Anfangs-Randwertproblem (A.1) mit Hilfe von Schema 3.2.2. In der Notation aus Kapitel 3 ist also in jedem Zeitschritt eine Nullstelle der Gleichung

$$U^{k+1} - U^k + \tau_{k+1} M_h^{-1} L_h^M(U^{k+1}) \Big(M_h^{-1} L_h U^{k+1} + \mathcal{I}_h W_{,u}^+(U^{k+1}) + \mathcal{I}_h W_{,u}^-(U^k) \Big) = \tau_{k+1} \mathcal{I}_h Q(U^{k+1}) \quad \text{(A.11)}$$

gesucht[4]. Diese bestimmt man wie folgt. Zunächst einmal setzt man $U_0^{k+1} = U^k$ und bestimmt mittels einer Fixpunktiteration U_{i+1}^{k+1} durch

$$\frac{1}{\tau_{k+1}}(U_{i+1}^{k+1} - U^k) + M_h^{-1} L_h^M(U_i^{k+1}) \Big(M_h^{-1} L_h U_{i+1}^{k+1} + \mathcal{I}_h W_{,u}^+(U_{i+1}^{k+1}) + \mathcal{I}_h W_{,u}^-(U^k) \Big) = \mathcal{I}_h Q(U_i^{k+1}). \quad \text{(A.12)}$$

[3] Intern regelt `inhomogen` dies dadurch, dass zuerst die `isPatch` Routine des zuletzt übergebenen `patch` überprüft wird.

[4] Um die Schreibweise übersichtlich zu halten, wird ab jetzt auf die Akzente $\sim$ verzichtet, obwohl (A.11) die Lösung des dimensionslosen Problems berechnet.

Man setzt $U^{k+1} = U_{i+1}^{k+1}$ als Lösung des nächsten Zeitschritts, falls

$$\|U_{i+1}^{k+1} - U_i^{k+1}\| \leq \texttt{tolS}\sqrt{D} \quad \text{oder} \quad i > \texttt{maxstepS}. \tag{A.13}$$

Die Lösung von (A.12) wird mit Hilfe eines Newton-Verfahrens mit Armijo-Schrittweitensteuerung (siehe [30]) bestimmt, das heißt man setzt $U_{i+1,0}^{k+1} = U_i^{k+1}$ und bestimmt die Newton-Suchrichtung Y_j durch Lösen des Gleichungssystems

$$DB(U_{i+1,j}^{k+1})Y_j = B(U_{i+1,j}^{k+1}). \tag{A.14}$$

Nun setzt man

$$U_{i+1,j+1}^{k+1} = U_{i+1,j}^{k+1} + \beta^m Y_j,$$

wobei $\beta = \frac{1}{2}$ gewählt wurde und $m \in \mathbb{N}_0$ die kleinste Zahl ist, so dass

$$\|B(U_{i+1,j+1}^{k+1})\| < (1 - \frac{1}{2}\beta^m)\|B(U_{i+1,j}^{k+1})\| \quad \text{und} \quad m < \texttt{maxstepNs} \tag{A.15}$$

gilt. Der Vektor B ist für $X \in \mathbb{R}^D$ definiert durch

$$\begin{aligned} B(X) := -\frac{1}{\tau_{k+1}}(X - U^k) + \mathcal{I}_h Q(U_i^{k+1}) \\ - M_h^{-1} L_h^M(U_i^{k+1}) \left(M_h^{-1} L_h X + \mathcal{I}_h W_{,u}^+(X) + \mathcal{I}_h W_{,u}^-(U^k) \right) \end{aligned} \tag{A.16}$$

und die Matrix $DB(X) = \frac{d}{dX}B(X)$ durch

$$DB(X) := \frac{1}{\tau_{k+1}} \operatorname{Id} + M_h^{-1} L_h^M(U_i^{k+1}) \left(M_h^{-1} L_h + \frac{d}{dX} \mathcal{I}_h W_{,u}^+(X) \right). \tag{A.17}$$

Das Newton-Verfahren wird beendet und $U_{i+1}^{k+1} = U_{i+1,j+1}^{k+1}$ gesetzt, falls

$$j \geq \texttt{maxstepN} \tag{A.18}$$

oder die beiden folgenden Bedingungen erfüllt sind:

$$\|Y_j\| < \texttt{tolNx}\sqrt{D}, \tag{A.19}$$

$$\|B(U_{i+1,j+1}^{k+1})\| < \texttt{tolNbAbs} \quad \text{oder} \quad \|B(U_{i+1,j+1}^{k+1})\| < \texttt{tolNbRel}\|B(U_{i+1,0}^{k+1})\|. \tag{A.20}$$

Die in (A.13), (A.15), (A.18), (A.19) und (A.20) verwendeten Parameter können im `main` Programm mit Hilfe der Funktionen

```
void Problem::setTolS(double);
void Problem::setTolNx(double);
void Problem::setTolNbRel(double);
void Problem::setTolNbAbs(double);
void Problem::setMaxstepN(int);
void Problem::setMaxstepNs(int);
void Problem::setMaxstepS(int);
```

festgesetzt werden. Zum Lösen des Gleichungssystems (A.14) benutzt `solve` das in der Klasse `solver` definierte `PBiCGstab`-Verfahren, der verwendete Vorkonditionierer ähnelt dem von Bramble, Pasciak und Xu [9] entwickelten und wird in Abschnitt A.3.3.4 genauer beschrieben. Mit

```
void Problem::setTolIAbs(double);
void Problem::setTolIRel(double);
void Problem::setMaxstepI(int);
```

können maximaler absoluter Fehler, maximaler relativer Fehler und die maximale Anzahl an Iterationsschritten für das `PBiCGstab`-Verfahren festgelegt werden.

A.3.3.2 Adaptivität in Raum und Zeit

`solve()` berechnet die Zeitschrittweite τ_{k+1} nach einer Idee von Grün und Rumpf [21] auf die folgende Art und Weise: Zunächst einmal wird der Druck

$$P^k = M_h^{-1} L_h U^k + \mathcal{I}_h W_{,u}^+(U^k) + \mathcal{I}_h W_{,u}^-(U^k) \tag{A.21}$$

bestimmt, anschließend wird für jedes $E \in \mathcal{T}_h$ der Wert

$$v(E) = \begin{cases} \frac{m(\overline{U^k}(E))}{\overline{U^k}(E)} |\nabla P^k(E)|_1 & \text{falls } \overline{U^k}(E) > 0, \\ 0 & \text{sonst} \end{cases} \tag{A.22}$$

berechnet. Dabei bezeichnet $\overline{U^k}(E)$ den Mittelwert von U^k auf E. Da P^k stückweise linear ist, ist ∇P^k konstant auf jedem Dreieck. Die $|.|_1$ Vektornorm ist für $x \in \mathbb{R}^2$ definiert durch $|x|_1 = \sum_{i=1}^2 |x_i|$. Die Zeitschrittweite τ_{k+1} berechnet sich nun für $k > 1$ als

$$\tau_{k+1} = \frac{h \ \texttt{Ctau}}{\texttt{vmin} + \min\{\max_{E \in \mathcal{T}_h} v(E), \texttt{vmax}\}}. \tag{A.23}$$

Dabei ist h die Gitterweite des feinsten Levels und die Werte von `Ctau`, `vmin`, `vmax` können mit Hilfe von

```
void Problem::setTStepControl(double Ctau, double vmin, double vmax);
```

gesetzt werden. Im ersten Zeitschritt wird immer $\tau_1 = \frac{h \ \texttt{Ctau}}{\texttt{vmax}}$ gewählt und `solve` rechnet auf einem bis zur maximal möglichen Gittertiefe bzw. Knotenanzahl verfeinerten Gitter. Die maximale Gittertiefe `maxdepth`, die minimale Gittertiefe `mindepth` und die maximale Knotenanzahl `maxdim` des von `solve` verwendeteten `Mesh` werden im Hauptprogramm mit Hilfe der Routine

```
void Problem::setMesh(int maxdim, int mindepth, int maxdepth);
```

vor dem Aufruf von `solve()` festgelegt. Falls `maxdepth`$\neq$`mindepth` ist, so wird das Gitter vor jedem weiteren Zeitschritt angepasst. Dazu wird auf jedem Dreieck der Triangulierung die Funktion

```
void Element::adaptMesh(void*);
```

ausgeführt, welche entscheidet, ob das Element mit Hilfe der Flags `REFINE` oder `COARSE` zum Verfeinern oder Vergröbern markiert wird. Dabei wird das folgende Kriterium benutzt:

- Falls ein Nachbardreieck $\tilde{E}$ von E existiert, so dass die euklidische Norm der Differenz der Gradienten
$$|\nabla U^k(E) - \nabla U^k(\tilde{E})| > \texttt{adapMax}$$
ist, so wird E zum Verfeinern markiert.
- Falls für alle Nachbardreiecke $\tilde{E}$ von E gilt, dass
$$|\nabla U^k(E) - \nabla U^k(\tilde{E})| < \texttt{adapMin}$$
ist, so wird E zum Vergröbern markiert.

Die Werte von `adapMin` und `adapMax` können mit

```
void Problem::setAdaptivity(double adapMin, double adapMax);
```

bestimmt werden. Nachdem so alle Elemente markiert sind, wird das Gitter durch Aufruf von `refineMesh` und `coarseMesh` an den entsprechenden Stellen verfeinert und vergröbert. Pro Zeitschritt wird ein Dreieck höchstens einmal verfeinert oder vergröbert. Anschließend wird die Lösung U^{k+1} des neuen Zeitschritts ausgerechnet.

A.3.3.3 Aufstellen des Linearen Gleichungssystems

Um die Matrix DB aus (A.17) berechnen zu können, benötigt `solve` Routinen, welche die auftretenden Matrizen M_h, L_h und $L_h^M(U)$ berechnen. Am Beispiel der Matrix

$$L_h = \left(\int_\Omega \nabla\phi_i \nabla\phi_j \right)_{i,j=1}^{D} \tag{A.24}$$

soll das generelle Vorgehen deutlich gemacht werden. Der ij-te Eintrag dieser Matrix (i und j sind hier die globalen Gitterpunkte der Knoten) berechnet sich durch

$$(L_h)_{ij} = \sum_{E \in \mathcal{T}_h} \int_E \nabla\phi_i(x) \nabla\phi_j(x) dx.$$

Um diese Matrix aufzustellen, müssen also auf jedem Element E der Traingulierung $\mathcal{T}_h$ die Integrale $\int_E \nabla\phi_i(x)\nabla\phi_j(x)dx$ berechnet und zum ij-ten Eintrag hinzuaddiert werden. Sei dazu $\hat{E}$ das Referenzdreieck mit den Eckpunkten $(1,0),(0,1),(0,0)$ und die Abbildung $f : \hat{E} \to E$ sei gegeben durch $f(\hat{x}) = A\hat{x} + b$. Dann lauten die Basisfunktionen auf $\hat{E}$:

$$\hat{\phi}_0(\hat{x}) = \hat{x}_1,$$
$$\hat{\phi}_1(\hat{x}) = \hat{x}_2,$$
$$\hat{\phi}_2(\hat{x}) = 1 - \hat{x}_1 - \hat{x}_2,$$

und die Einträge berechnen sich nun leicht aus

$$\int_E \nabla\phi_{\texttt{node[k]}}\nabla\phi_{\texttt{node[l]}}\,dx = \int_{\hat{E}} \langle A^{-T}\hat{\nabla}\hat{\phi}_k\,, A^{-T}\hat{\nabla}\hat{\phi}_l\rangle |\det A| d\hat{x}, \quad k,l\in\{0,1,2\}.$$

Für rechtwinklige, gleichschenklige Dreiecke vereinfacht sich die Berechnung nochmals deutlich, da dann $A^{-1}A^{-T}|\det A| = Id$ ist. Das generelle Vorgehen ist nun, mit Hilfe der `traverse`-Routine auf jedem Element eine Funktion auszuführen, welche die Einträge berechnet und zu einer mit Hilfe des `void*` Parameters übergebenen Matrix hinzuaddiert. Um die nötigen Matrizen, Vektoren u.ä. übergeben zu können, ist in structs.h der folgende `struct` definiert:

```
struct TraverseStruct {
   Problem* problem;
   void* arg1; void* arg2; void* arg3;
 };
```

welcher neben einem Zeiger auf die Klasse `Problem`, Zeiger auf bis zu drei weitere, beliebige Objekte enthalten kann. Die von traverse aufgerufenen Funktionen weisen diesen `void*` Zeigern dann wieder einen Typ zu. Es ist daher wichtig, dass die richtigen Argumente übergeben werden, da Fehler an dieser Stelle erst zur Laufzeit des Programms bemerkt werden.

Die folgende Tabelle gibt eine Übersicht über die vorhandenen Callbacks und die erwarteten Argumente. `arg1` enthält immer einen Zeiger auf das zu berechnende Objekt, `arg2` und `arg3` weitere zur Berechnung nötige Argumente.

Name der Funktion	berechnet	arg1	arg2	arg3
`void assembleMh(void*)`	M_h	`vector`	–	–
`void assembleMhLh(void*)`	$M_h^{-1}L_h$	`SparseMatrix`	`vector` M_h	–
`void assembleMhLhM(void*)`	$M_h^{-1}L_h^M(U)$	`SparseMatrix`	`vector` M_h	`vector` U
`void assembleWuplus(void*)`	$\mathcal{I}_h W_{,u}^+(U)$	`vector`	`vector` U	–
`void assembleWuminus(void*)`	$\mathcal{I}_h W_{,u}^-(U)$	`vector`	`vector` U	–
`void assembleDWuplus(void*)`	$\frac{d}{dU}\mathcal{I}_h W_{,u}^+(U)$	`vector`	`vector` U	–
`void setInitialValue(void*)`	U^0	`vector`	`double` l	–
`void timestepcontrol(void*)`	$\max_{E\in\mathcal{T}_h} v(E)$	`double`	`vector` U	`vector` P

Man beachte, dass die Matrizen M_h und $\frac{d}{dU}\mathcal{I}_h W_{,u}^+(U)$ als `vector` abgespeichert abgespeichert werden können, da sie Diagonalgestalt haben.

Beispiel: Der folgende Code berechnet die Matrix M_h auf einem Gitter der Tiefe 10 (Ein solches Gitter benötigt 1089 Stützstellen) und speichert die Diagonale im `vector v` ab.

```
Mesh m(1089,10,10)
vector v(1089);
TraverseStruct ts; ts.problem=&this; ts.arg1=&v;
m.traverse(LEAVES,ACTIVE,10,&Element::assembleMh,&ts);
```

Die auf diese Art und Weise berechneten Matrizen werden nun benutzt, um das Gleichungssystem (A.14) aufzustellen. Um die Matrix

$$DB(X) := \frac{1}{\tau_{k+1}}\,\mathrm{Id} + M_h^{-1}L_h^M(U_i^{k+1})\left(M_h^{-1}L_h + \frac{d}{dX}\mathcal{I}_h W_{,u}^+(X)\right) \tag{A.25}$$

nicht explizit ausrechnen zu müssen, was *sehr* zeitaufwendig wäre, da zwei dünn besetzte Matrizen multipliziert werden müssten, wird eine von `matrixBase` abgeleitete Klasse benutzt. Diese stellt allgemein Summen der Form

$$\frac{1}{\tau}\,\mathrm{Id} + A(B+D)$$

dar, wobei D eine Diagonalmatrix ist. Ihre Definition lautet

```
class NewtonMatrix:public matrixBase{

 protected:
  int dim;

  matrixBase *A, *B;
  double tau;
  vector *D;

 public:
  NewtonMatrix(int dim){this->dim=dim;};

  void set(matrixBase* A, matrixBase* B, vector* D, double tau){
    this->A=A; this->B=B; this->D=D; this->tau=tau;
  };

  vector& multiply(vector& x, vector& lsg, std::list<int> &nl);
  int getN(){return dim;};
};
```

Entscheidend ist, dass die Klasse lediglich Zeiger auf die anderen Matrizen abspeichert, dadurch also so gut wie kein zusätzlicher Speicheraufwand nötig ist. `multiply` nutzt die `multiply`-Routinen der Matrizen `A` und `B`, um den Vektor $\frac{1}{\tau}x + A(Bx + Dx)$ zu berechnen.

Damit hat die Klasse `NewtonMatrix` alle Fähigkeiten, die ein Löser der Klasse `solver` von einer Matrix verlangt, das Gleichungssystem kann also ohne weiteren Aufwand gelöst werden.

A.3.3.4 Effizientes Lösen des Gleichungssystems – BPX Vorkonditionierer

Die Idee des von Bramble, Pasciak und Xu [9] entwickelten Vorkonditionierers ist die folgende: Gegeben sei eine Triangulierung $\mathcal{T}_h$ mit Gitterweite h und ein dazugehöriger Finite-Elemente-Raum V^h. Gesucht ist nun eine Lösung $X \in V^h$ des Problems

$$AX = B \tag{A.26}$$

mit einem Operator $A : V^h \to V^h$ und einer rechten Seite $B \in V^h$. Nun konstruiert man eine Folge von Triangulierungen $\mathcal{T}_{h_l}$ zu verschiedenen Gitterweiten h_l mit dazugehörigen Finite-Elemente-Räumen

$$V^{h_0} \subset V^{h_1} \subset \cdots \subset V^{h_L} = V^h, \tag{A.27}$$

und bezeichnet mit $A_l : V^{h_l} \to V^{h_l}$ die Restriktion von A auf V^{h_l}. Ein geeigneter Vorkonditionierer C ist definiert durch

$$CX = \sum_{l=0}^{L} \frac{1}{h_l^2 \lambda_l} \sum_{i \in I_l} (X, \varphi_i^l) \varphi_i^l. \tag{A.28}$$

Dabei ist I_l die zu $\mathcal{T}_{h_l}$ gehörige Knotenindexmenge und $\{\varphi_i^l\}_{i \in I_l}$ eine Basis von V^{h_l}. λ_l ist eine Näherung für den Spektralradius $\varrho(A_l)$.

Die Anwendung dieses Vorkonditionierers bietet sich deshalb an, da durch den hierarchischen Aufbau der Klasse `Mesh` eine Folge von Triangulierungen gegeben ist, deren zugehörige Finite-Elemente-Räume (A.27) erfüllen. Der von EConLub2D zur Verfügung gestellte BPX-Vorkonditionierer konstruiert aus dem verwendeten Gitter eine Folge von Triangulierungen wie folgt: Zu $0 \leq l \leq$ `maxDepth` sei $\mathcal{T}_h^l$ die Triangulierung, welche man durch Weglassen aller Elemente von `Mesh` mit `level` $> l$ erhält. Dann formen die zu $\mathcal{T}_h^l$ gehörigen Finite-Elemente-Räume V_l^h eine Folge

$$V_0^h \subset V_1^h \subset \cdots \subset V_{\texttt{maxDepth}}^h = V^h. \tag{A.29}$$

Man beachte, dass für $l >$ `minDepth` nicht notwendigerweise alle in $\mathcal{T}_h^l$ enthaltenen Dreiecke Gittertiefe l haben, da das Gitter wegen der Adaptivität nicht an allen Stellen bis zur maximalen Gittertiefe verfeinert sein muss. Dies macht eine Anpassung von (A.28) notwendig. Dazu sei die Indexmenge I_l^* definiert als die Menge aller Knotenindizes i, welche in einem `Element` aus $\mathcal{T}_h^l$ mit `level`$=l$ im Gitter vorkommen. Damit enthält $\{\varphi_i^l\}_{i \in I_l^*}$ genau die Basisfunktionen φ_i^l, welche nicht schon in der Basis $\{\varphi_i^{l-1}\}_{i \in I_{l-1}}$ des gröberen Finite-Elemente-Raums enthalten sind. Der Vorkonditionierer ist nun definiert durch

$$CX = \sum_{l=0}^{\texttt{maxDepth}} \frac{1}{h_l^2 \lambda_l} \sum_{i \in I_l^*} (X, \varphi_i^l) \varphi_i^l. \tag{A.30}$$

Wie bereits in Abschnitt A.1.4 gesehen, sind Vorkonditionierer von der Klasse `matrixBase` abgeleitet. Die in `bpx.h` definierte Klasse `BPX` ist also ebenfalls vom Typ `matrixBase`. Es ist daher nicht nötig, explizit die Matrix C zu berechnen, die durch (A.30) gegeben ist. Es muss lediglich die `BPX::multiply` Routine entsprechend überladen werden. Diese arbeitet wie folgt:

In einem ersten Durchgang *(First Pass)* werden die L^2-Skalarprodukte (X, φ_i^l) berechnet. Auf dem feinsten Gitter $\mathcal{T}_h^L$ werden diese Werte näherungsweise mit Hilfe der Formel

$$(X, \varphi_i^L) = \sum_{j \in I_L} X_j (\varphi_j^L, \varphi_i^L) \approx \sum_{j \in I_L} X_j (\varphi_j^L, \varphi_i^L)_h = (M_h)_{ii} X_i \tag{A.31}$$

bestimmt. Die Werte auf den gröberen Gittern werden daraus rekursiv berechnet, sich die Basisfunktionen φ_i^{l-1} als Linearkombination von Basisfunktionen φ_j^l des feineren Gitters schreiben lassen.

In einem zweiten Durchgang *(Second Pass)* werden nun diese soeben berechneten Werte genutzt, um die Darstellung von CX in der Basis $\{\varphi_i^L\}_{i \in I_L}$ des feinsten Gitters zu berechnen (dies entspricht dem Vektor $CX \in V^h$). Begonnen wird auf dem gröbsten Gitter. Zunächst

werden die Vorfaktoren der φ_i^0 berechnet. Dann wird, wiederum unter Ausnutzung der Tatsache, dass sich die Basisfunktionen des gröberen Gitters durch eine Linearkombination von Basisfunktionen des feineren Gitters ausdrücken lassen, die Summe $\sum_{i \in I_0^*}$ aus (A.30) eliminiert, indem die Vorfaktoren der φ_i^1 entsprechend angepasst werden. Dies wird fortgesetzt, bis nur noch die Summe über I_L übrig bleibt. Die Vorfaktoren der φ_i^L bilden die gesuchte Darstellung. Für den Spektralradius λ_l wird dabei die Näherung

$$\lambda_l = \frac{1}{\tau} + \frac{1}{h_l^4}. \tag{A.32}$$

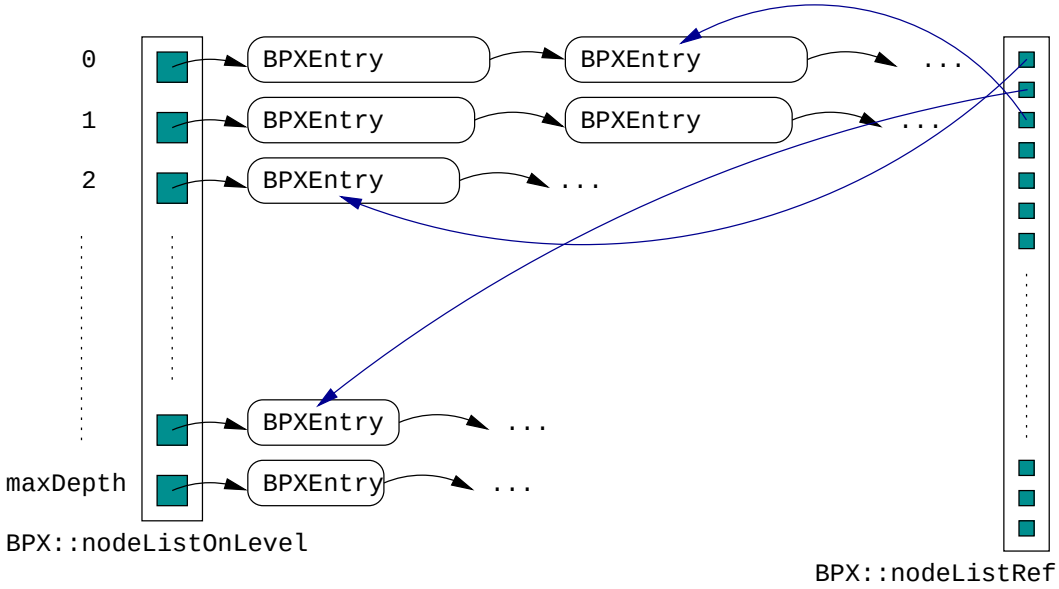
benutzt.

Um eine solche `multiply`-Funktion durchführen zu können, muss die Klasse `BPX` wissen, welche Knotenpunkte zu den Gittern $\mathcal{T}_h^l$ gehören. Außerdem muss `BPX` wissen, aus welchen Basisfunktionen des feinsten Gitters sich eine Basisfunktion des gröberen Gitters zusammensetzen lässt, also über Nachbarschaftsinformationen verfügen. Zusätzlich benötigt man genügend Speicherplatz, um alle Werte (X, φ_i^l) abspeichern zu können, also genau $\sum_{l=0}^{\texttt{maxDepth}} |I_l^*|$ Einträge. Dies alles muss bereits vor dem Aufruf von `multiply` bereitgestellt werden, eine Aufgabe, die vom Konstruktor übernommen wird.

Die Definition der Klasse `BPX` lautet:

```
class BPX : public matrixBase{

 protected:
   Mesh* mesh;
   Problem* pb;
   double tau;
   vector* mh;

   BPXEntry** nodeListRef;
   BPXEntry** nodeListOnLevel;

 public:
   BPX(Mesh* mesh,Problem* pb,double tau,vector* mh);
   ~BPX();

   vector& multiply(vector& x, vector& lsg, std::list<int> &nl);

   // [...]
 };
```

Der Konstruktor erhält Zeiger auf das zu verwendende Gitter, auf die Klasse `Problem`, von der aus `solve()` den Konstruktor aufruft, die Zeitschrittweite τ (nötig zum Berechnen der Eigenwerte), und auf die verdichtete-Massen-Matrix M_h übergeben. Außerdem baut der Konstruktor die folgende Datenstruktur auf:



Ein `BPXEntry` ist dabei eine Klasse, welche die folgenden Daten enthält:

```
class BPXEntry{

 protected:
   BPXEntry* next;
   int node;         // Indexnummer des Knotens
   int parent[2];    // Indexnummern der benachbarten Knoten
   double* value;    // Array
   int kmax;         // maximale zugehörige Gittertiefe

   // [...]
};
```

Für jeden Knoten des Gitters existiert in der obigen Struktur genau ein `BPXEntry`. Die Liste `nodeListRef` zeigt, welche Indexnummer wo zu finden ist, das heißt `nodeListRef[i]` zeigt auf das `BPXEntry` mit `node==i`. Sortiert sind die `BPXEntry`-Instanzen nach dem folgenden Prinzip: Ein Knoten `i`, der in allen Indexmengen I_l^* für `kmin`$\leq l \leq$`kmax` vorkommt, wird unter `nodeListOnLevel[kmin]` angehängt. Der Vektor `value` erhält dabei die Länge `kmax-kmin+1`. Auf diese Art und Weise erhält man insgesamt $\sum_{l=0}^{\texttt{maxDepth}} |I_l^*|$ freie Felder. Diese werden im *First Pass* genutzt, um die Werte der Skalarprodukte (X, φ_i^l) zu speichern und im *Second Pass* zum Speichern der Vorfaktoren der φ_i^l.

A.3.4 main.cpp – ein Beispiel

Das folgende Beispiel zeigt das Hauptprogramm der in Abbildung 6.7 gezeigten numerischen Simulation. In den ersten Zeilen wird ein Objekt der Klasse `Problem` erzeugt und die in der Rechnung verwendeten Parameter gesetzt:

```
int main(){
  Problem pb;

  pb.setEta(1200.);    pb.setSigma(0.0308);
  pb.setLength(4e-6);  pb.setT(1000);
  pb.setMn(3.);        pb.setMc(1./3.);
```

Dann werden Anfangswerte, Potential und rechte Seite gesetzt. Die dabei benutzte Klasse `perturbedFilm` ist eine von `initial` abgeleitete Klasse. Sie stellt einen flachen Film der angegebenen Höhe mit einer kleinen Störung dar.

```
  perturbedFilm init(4.9e-9); pb.setInitialValue(&init);

  inhomogen inh;
  LJ lj1(2.2e-20,6.25e-76);
  patch pat1;
  inh.addPatch(&pat1,&lj1);

  LJ lj2(2.5e-20,6.25e-76);
  circle pat2(.2e-6,2e-6,2e-6);
  inh.addPatch(&pat2,&lj2);

  pb.setPotential(&inh);

  rhs rs;  pb.setRHS(&rs);
```

Als nächstes werden nun die Abbruchkriterien für die verschiedenen Iterationsverfahren festgelegt. Das Newtonverfahren soll beendet werden, falls Bedingung (A.19) erfüllt ist, auf Bedingung (A.20) wird verzichtet. Dies wird realisiert, indem die entsprechenden Fehlertoleranzen auf 1.0 gesetzt werden:

```
  pb.setTolS(1e-7);
  pb.setTolIAbs(1e-10);  pb.setTolIRel(1e-6);
  pb.setTolNx(1e-12);    pb.setTolNbRel(1);    pb.setTolNbAbs(1);
  pb.setMaxstepS(25);    pb.setMaxstepI(2000);
  pb.setMaxstepN(25);    pb.setMaxstepNs(5);
```

Als letztes wird noch die Größe des Gitters festgelgt und die Funktion `solve()` aufgerufen:

```
  pb.setMesh(33025,8,15);

  pb.setOutputIntervall(10.0,50);
  pb.setFiledir("erg");   pb.setFilename("beispiel");

  pb.solve();
};
```

Die Ergebnisse der Rechnung werden in mehreren Dateien abgelegt. Pro Zeitschritt wird eine eigene Datei angelegt. Die Startwerte werden hier in die Datei erg/beispiel.0.gnu geschrieben,

die nächsten Ergebnisse in erg/beispiel.1.gnu, usw. Die angelegte Datei kann dann in gnuplot mit Hilfe des Befehls

```
splot "beispiel.0.gnu" with lines
```

angesehen werden. Um Speicherplatz zu sparen, werden nicht alle Zeitschritte ausgegeben. Die Funktion `setOutputIntervall` legt fest, an welchen Zeitpunkten eine Ausgabe erfolgen soll. Hier soll eine Ausgabe erfolgen, wenn seit der letzten Ausgabe mehr als 10.0 Zeiteinheiten vergangen sind, spätestens aber alle 50 Zeitschritte.

Zusätzlich zu den von gnuplot lesbaren Dateien werden auch noch Dateien mit den Namen erg/beispiel.#.grape angelegt. Diese können mit Hilfe eines geeigneten GRAPE-Interfaces visualisiert werden.

Anhang B

Notation

Mengen und Funktionen

$\partial\Omega$	Rand des Gebietes Ω
$\overline{\Omega}$	Abschluß des Gebietes Ω
Ω_T	$:= (0,T) \times \Omega$
$[u > \alpha]$	$:= \{(t,x) \in \Omega_T : u(t,x) > \alpha\}$
$[u(t) > \alpha]$	$:= \{x \in \Omega : u(t,x) > \alpha\}$
$B_r(x_0)$	$:= \{x \in R^d : \lvert x - x_0\rvert < r\}$
$f\vert_{[a,b]}$	Einschränkung der Funktion f auf $[a,b]$
$\chi_{[u>0]}$	charakteristische Funktion, $\chi(t,x) = 1$ falls $u(t,x) > 0$, $\chi(t,x) = 0$ sonst

Integration und Differentiation

$\fint$	Mittelwertintegral, $\fint_\Omega u := \frac{1}{\lvert\Omega\rvert}\int_\Omega u$
$w_{,u}(u,x)$	partielle Ableitung von $w(u,x)$ nach u
$\frac{\partial}{\partial\nu}$	Ableitung in Richtung der äußeren Einheitsnormalen an Ω
$\partial_\tau^- U$	Rückwärtsdifferenzenquotient, für $U \in S^{-1,0}(V^h)$ definiert durch $\partial_\tau^- U(t) := \frac{U^{k+1}-U^k}{\tau_{k+1}}$ für $\mathfrak{t}_k < t \le \mathfrak{t}_{k+1}$
Δ_h	diskreter Laplace-Operator, für $U \in V^h$ definiert durch $(\Delta_h U, \Psi)_h = (\nabla U, \nabla\Psi) \quad \forall\Psi \in V^h$

Funktionenräume und Normen

$C^k(\Omega)$	Menge der reellwertigen, k-fach stetig differenzierbaren Funktionen auf Ω
$C_0^k(\Omega)$	$:= \{u \in C^k(\Omega) : \mathrm{supp}(u)$ ist kompakte Teilmenge von $\Omega\}$
$C^k(X;Y)$	Menge der k-fach stetig differenzierbaren Funktionen $f : X \to Y$
$C^\beta(\Omega)$	Menge der hölderstetigen Funktionen zum Exponent β, $0 < \beta < 1$
$C^{\alpha,\beta}(\Omega_T)$	Menge aller Funktionen $f(t,x) \in C^0(\Omega_T)$, welche hölderstetig zum Exponent α bezüglich t und hölderstetig zum Exponent β bezüglich x sind
$\mathcal{P}^k$	Menge der Polynome vom Grad $\le k$
$L^p(\Omega)$	Raum der reellwertigen, Lebesgue-meßbaren Funktionen u für die $\lvert u\rvert^p$ integrierbar ist
$(\cdot,\cdot)$	Skalarprodukt in $L^2(\Omega)$, $(u,v) := \int_\Omega uv$

$W^{m,p}(\Omega)$	Sobolevraum der m-fach schwach differenzierbaren Funktionen mit schwachen Ableitungen in $L^p(\Omega)$
$H^m(\Omega)$	$:= W^{m,2}(\Omega)$
$W_0^{m,p}(\Omega)$	Abschluß von $C_0^\infty(\Omega)$ bzgl. der $W^{m,p}(\Omega)$-Norm
$H_0^m(\Omega)$	$:= W_0^{m,2}(\Omega)$
$\|.\|_{m,p}$	für $p < \infty$ Norm auf $W^{m,p}(\Omega)$, $\|u\|_{m,p} := \left(\sum_{\|\alpha\|\le m} \|\partial^\alpha u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}$
$\|u\|_{m,\infty}$	$:= \sup_{\|\alpha\|\le m} \|\partial^\alpha u\|_{L^\infty(\Omega)}$
$\|.\|_m$	$:= \|.\|_{m,2}$
$\|.\|_{m,p}$	für $p < \infty$ Halbnorm auf $W^{m,p}(\Omega)$, $\|u\|_{m,p} := \left(\sum_{\|\alpha\|= m} \|\partial^\alpha u\|^p_{L^p(\Omega)}\right)^{\frac{1}{p}}$
$\|u\|_{m,\infty}$	$:= \sup_{\|\alpha\|= m} \|\partial^\alpha u\|_{L^\infty(\Omega)}$
$\|.\|_m$	$:= \|.\|_{m,2}$
$\|.\|_X$	Norm des Banachraums X
$\|.\|$	euklidische Norm im $\mathbb{R}^D$
$C(0,T;X)$	Menge der stetigen Abbildungen von $[0,T]$ in einen Banachraum X
$L^p(0,T;X)$	Menge aller Bochner-meßbaren Funktionen $f : (0,T) \to X$, für die die Norm $\|f\|_{L^p(0,T;X)} := (\int_0^T \|f(t)\|^p_X dt)^{\frac{1}{p}}$, $p < \infty$ bzw. $\|f\|_{L^\infty(0,T;X)} :=$ esssup$_{0<t<T} \|f(t)\|_X$ beschränkt ist
X'	Dualraum des Banachraums X
$\langle u\,, v\rangle_{X'\times X}$	Duale Paarung zwischen $u \in X'$ und $v \in X$
$u_\varepsilon \to u$	starke Konvergenz von u_ε gegen u
$u_\varepsilon \rightharpoonup u$	schwache Konvergenz von u_ε gegen u

Finite-Elemente-Diskretisierung

d	Raumdimension, $\Omega \subset \mathbb{R}^d$
$\mathcal{T}_h$	Triangulierung von Ω (siehe Def. 3.1.1)
h	maximale Gitterweite der Triangulierung, $h := \max_{E\in\mathcal{T}_h} \mathrm{diam}(E)$
$\mathcal{N}_h$	Menge der Knotenpunkte der Triangulierung $\mathcal{T}_h$
$\mathfrak{x}_i$	Knotenpunkte von $\mathcal{T}_h$, $i = 1,\ldots,D$
V^h	$:= \{U \in C^0(\Omega) : U\|_E \in \mathcal{P}^1 \,\forall E \in \mathcal{T}_h\}$
φ_i	Basisfunktionen des Finite-Elemente-Raums V^h, definiert durch $\varphi_i(\mathfrak{x}_j) = \delta_{ij}$
D	Dimension des Finite-Elemente-Raums V^h
$\mathcal{I}_h$	nodaler Projektionsoperator, $\mathcal{I}_h u := \sum_{i=1}^D u(\mathfrak{x}_i)\varphi_i$
$(\cdot,\cdot)_h$	verdichtete Massen Skalarprodukt (siehe Def. 3.1.2)
$\mathfrak{t}_i$	diskrete Zeitgitterpunkte, $i = 0,\ldots,K$
τ_k	Zeitschrittweite, $\tau_k := \mathfrak{t}_k - \mathfrak{t}_{k-1}$
τ	maximale Zeitschrittweite, $\tau = \max_{1\le k\le K} \tau_k$
K	Anzahl der Zeitschritte
$S^{-1,0}(V^h)$	$:= \left\{U : [0,T] \to V^h \,\middle|\, U(t) = U(\mathfrak{t}_{k+1}) \text{ für } \mathfrak{t}_k < t \le \mathfrak{t}_{k+1}, k = 0,\ldots,K-1\right\}$

Entropie und Mobilität

$m(u)$	Mobilität, $m(u) = c\,u^n$
n	Mobilitäts-Exponent
$G(u)$	Entropie, gegeben durch $G(u) = \int_A^u \int_A^s m(r)^{-1} dr\, ds$, $A > 0$

$m_\sigma(u)$	Näherung von m (siehe Gleichung (3.21))
G_σ, M_σ	zulässiges Entropie-Mobilitäts-Paar (siehe Def. 3.2.1)

Diskrete Vektoren und Matrizen

$\langle \cdot, \cdot \rangle$	Euklidisches Skalarprodukt
$\mathbb{1}$	$:= (1, \dots, 1)^T$
M_h	$:= ((\varphi_i, \varphi_j)_h)_{i,j=1,\dots,D}$
L_h	$:= \left(\int_\Omega \nabla\varphi_i \nabla\varphi_j\right)_{i,j=1,\dots,D}$
$L_h^M(U)$	$:= \left(\int_\Omega M_\sigma(U)\nabla\varphi_i \nabla\varphi_j\right)_{i,j=1,\dots,D}$
$\mu(U)$	$:= \fint_\Omega U = \frac{\langle \mathbb{1}, M_h U\rangle}{\langle \mathbb{1}, M_h \mathbb{1}\rangle}$
U^-	um einen Zeitschritt verschobene Funktion, $U^-(t) := U^k$ falls $t \in (\mathfrak{t}_k, \mathfrak{t}_{k+1}]$

Literaturverzeichnis

[1] H.W. Alt. *Lineare Funktionalanalysis*. Springer-Verlag, Berlin-Heidelberg, 2002.

[2] C. Bauer and S. Dietrich. Quantitative study of laterally inhomogeneous wetting films. *Eur. Phys. J. B*, 10:767–779, 1999.

[3] J. Becker, G. Grün, M. Lenz, and M. Rumpf. Numerical methods for fourth order nonlinear diffusion problems. *Applications of Mathematics*, 47:517–543, 2002.

[4] 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*, 2:59–63, 2003.

[5] E. Beretta, M. Bertsch, and R. Dal Passo. Nonnegative solutions of a fourth order nonlinear degenerate parabolic equation. *Arch. Ration. Mech. Anal.*, 129:175–200, 1995.

[6] F. Bernis and A. Friedman. Higher order nonlinear degenerate parabolic equations. *J. Differential Equations*, 83:179–206, 1990.

[7] F. Bernis, L.A. Peletier, and S.M. Williams. Source-type solutions of a fourth order nonlinear degenerate parabolic equations. *Nonlinear Anal.*, 18:217–234, 1992.

[8] M. Bertsch, R. Dal Passo, H. Garcke, and G. Grün. The thin viscous flow equation in higher space dimensions. *Adv. Differential Equations*, 3:417–440, 1998.

[9] J.H. Bramble, J.E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55:1–22, 1990.

[10] L. Brusch, H. Kühne, U. Thiele, and M. Bär. Dewetting of thin films on heterogeneous substrates: Pinning versus coarsening. *Phys. Rev. E*, 66, 2002.

[11] Ph. G. Ciarlet. *The finite element method for elliptic problems*. North Holland, Amsterdam, 1978.

[12] R. Dal Passo, H. Garcke, and G. Grün. On a fourth order degenerate parabolic equation: global entropy estimates and qualitative behaviour of solutions. *SIAM J. Math. Anal.*, 29:321–342, 1998.

[13] A. Darhuber, S. Troian, and W. Reisner. Dynamics of capillary spreading along hydrophilic microstripes. *Phys. Rev. E*, 64:031603, 2001.

[14] S. Dietrich and M. Rauscher. Flow over a sharp chemical step in lubrication approximation. in preparation.

[15] R. Ferreira and F. Bernis. Source-type solutions to thin-film equations in higher space dimensions. *European J. Appl. Math.*, 8:507–524, 1997.

[16] H. Gau, S. Herminghaus, P. Lenz, and R. Lipowsky. Liquid morphologies on structured surfaces: from microchannels to microchips. *Science*, 283:46, 1999.

[17] G. Grün. Degenerate parabolic equations of fourth order and a plasticity model with nonlocal hardening. *Z. Anal. Anwendungen*, 14:541–573, 1995.

[18] G. Grün. *On free boundary problems arising in thin film flow.* 2001. Habilitation thesis, Universität Bonn.

[19] G. Grün. On the convergence of entropy consistent schemes for lubrication type equations in multiple space dimensions. *Math.Comp.*, 72:1251–1279, 2003.

[20] G. Grün, J. Becker, and M. Rumpf. On space-time adaptive convergent finite element schemes for a general class of lubrication-type equations. In H. A. Mang, F. G. Rammerstorfer, and J. Eberhardsteiner, editors, *Proceedings of the 5th World Congress on Computational Mechanics*. Vienna Technical University, ISBN 3-9501554-0-6, 2002.

[21] G. Grün and M. Rumpf. Nonnegativity preserving convergent schemes for the thin film equation. *Num. Math.*, 87:113–152, 2000.

[22] G. Grün and M. Rumpf. Simulation of singularities and instabilities in thin film flow. *Euro. J. Appl. Math.*, 12:293–320, 2001.

[23] S. Herminghaus, K. Jacobs, and R. Seemann. The glass transition of thin polymer films: Some questions and a possible answer. *Euro. Phys. J. E*, 5:531, 2001.

[24] J. Israelachvili. *Intermolecular and surface forces.* Academic Press, 1992.

[25] K. Jacobs. *Stabilität und Dynamik flüssiger Polymerfilme.* UFO-Verlag Allensbach, ISBN 3-930803-10-0, 1997.

[26] N.M. Josuttis. *The C++ Standard Library – A Tutorial and Reference.* Addison-Wesley, 1999.

[27] K. Kargupta and A. Sharma. Creation of ordered patterns by dewetting of thin films on homogeneous and heterogeneous substrates. *J. Colloid Int. Sci.*, 245:99–115, 2002.

[28] W. Koch, S. Dietrich, and M. Napiorkowski. Morphology and line tension of liquid films adsorbed on chemically structured substrates. *Phys. Rev. E*, 51:3300–3317, 1995.

[29] R. Konnur, K. Kargupta, and A. Sharma. Instability and morphology of thin liquid films on chemically heterogeneous substrates. *Phys. Rev. Lett.*, 84:931–934, 2000.

[30] P. Kosmol. *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben.* Teubner, Stuttgart, 1993.

[31] C. Neto, K. Jacobs, R. Seemann, R. Blossey, J. Becker, and G. Grün. Correlated dewetting patterns in thin polystyrene films. *J.Phys.:Cond.Mat.*, 15:421–426, 2003.

[32] C. Neto, K. Jacobs, R. Seemann, R. Blossey, J. Becker, and G. Grün. Satellite hole formation during dewetting: experiment and simulation. *J.Phys.:Cond.Mat.*, 15:3355–3366, 2003.

[33] S.M. Nikol'skii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer-Verlag, Berlin-Heidelberg, 1975.

[34] A. Oron, S.H. Davis, and S.G. Bankoff. Long-scale evolution of thin liquid films. *Reviews of Modern Physics*, 69:932–977, 1997.

[35] N. Rehse, C. Wang, M. Hund, M. Geoghegan, R. Magerle, and G. Krausch. Stability of thin polymer films on a corrugated substrate. *Eur. Phys. J. E*, 4:69–76, 2001.

[36] O. Reynolds. On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments, including an experimental determination of the viscosity of olive oil. *Philos. Trans. R. Soc. London*, 177:157–234, 1886.

[37] C. Schäfle. *Morphologie, Verdampfung und Kondensation von Flüssigkeiten auf benetzungsstrukturierten Oberflächen*. PhD thesis, Universität Konstanz, 2002.

[38] C. Schäfle, C. Bechinger, B. Rinn, C. David, and P. Leiderer. Cooperative evaporation in ordered arrays of volatile droplets. *Phys. Rev. Lett.*, 83:5302–5305, 1999.

[39] C. Schäfle, P. Leiderer, and C. Bechinger. Subpattern formation during condensation processes on structured substrates. *Europhys. Lett.*, 63:394–400, 2003.

[40] R. Seemann, S. Herminghaus, and K. Jacobs. Dewetting patterns and molecular forces: A reconciliation. *Phys. Rev. Lett.*, 86:5534–5537, 2001.

[41] R. Seemann, S. Herminghaus, and K. Jacobs. Gaining control of pattern formation of dewetting liquid films. *J. Phys.: Cond. Mat.*, 13:4952, 2001.

[42] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Annali di Matematica Pura ed Applicata*, 146:65–96, 1987.

[43] U. Thiele, L. Brusch, M. Bestehorn, and M. Bär. Modelling thin-film dewetting on structured substrates and templates: Bifurcation analysis and numerical simulation. *Eur. Phys. J. E*, 11:255–271, 2003.

[44] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer-Verlag, Berlin, 1997.

[45] H. A. van der Vorst. A fast and smoothly converging variant of bi–cg for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 13:73–86, 1992.

[46] T. Young. An essay on the cohesion of fluids. *Philos. Trans. Roy. Soc. London*, 95:65–87, 1805.

[47] E. Zeidler. *Linear monotone operators*, volume 2A of *Nonlinear Functional Analysis and its Applications*. Springer-Verlag, Berlin, 1990.

Ein herzliches Dankeschön

an all diejenigen, die mich beim Anfertigen dieser Arbeit unterstützt haben. Insbesondere möchte ich mich bedanken bei

... Günther Grün dafür, dass er mir die Gelegenheit zum Anfertigen dieser Arbeit gab, für sein Interesse am Fortschritt der Arbeit und für die zahlreichen Tipps, wenn mal wieder ein Beweis zu scheitern drohte.

... Prof. Frehse und allen aktuellen und ehemaligen Mitarbeitern der Abteilung Angewandte Analysis für die gute (Arbeits)atmosphäre.

... Martin Lenz, weil er für jede noch so kryptische Fehlermeldung des Compilers oder Linkers die Ursache kannte.

Wesentlich zum Gelingen der Arbeit haben auch die Kontakte beigetragen, die ich durch das DFG-Schwerpunktprogramm “Benetzung und Strukturbildung an Grenzflächen” gewonnen habe. Mein Dank gilt hier vor allem

... der (ehemals) Ulmer Gruppe, insbesondere Karin Jacobs, Ralf Seemann, Chiara Neto und Renate Konrad, für die gute Zusammenarbeit zum Thema “Entnetzung von Polymerfilmem” und für ihre Bereitschaft, mich jederzeit mit allen zu einem Vergleich von Simulation und Experiment nötigen Daten und Bildern auszustatten.

... Prof. Leiderer dafür, dass er mir das Bild- und Datenmaterial der Doktorarbeit von C. Schäfle zur Verfügung stellte.