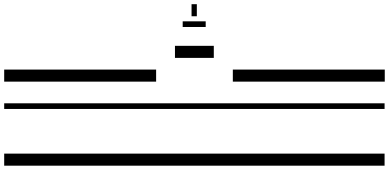


Westfälische
Wilhelms-Universität
Münster

Lösung eines Phasenfeld Modells für Rissentstehung mittels semiglatter Newton Methode

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science

Westfälische Wilhelms-Universität Münster
Fachbereich Mathematik und Informatik
Institut für Numerische und Angewandte Mathematik

Betreuung:
Prof. Dr. Benedikt Wirth

Eingereicht von:
Ines Ahrens

Münster, September 2015

Inhaltsverzeichnis

1. Einleitung

Wir wollen herausfinden, wie Risse entstehen, bzw wie sie sich fortsetzen,

Betrachten wir zunächst ein Beispiel. Ein Blatt Papier befindet sich in Ruheposition auf dem Gebiet $\Omega \subset \mathbb{R}^3$.Nun wird es an zwei gegenüberliegenden Seiten eingespannt und an diesen Seiten wird Kraft ausgeübt. Was passiert? Zunächst wird sich das Blatt Papier ein wenig dehnen, also deformieren. Die neue Position des Papiers beschreibt $\varphi(x)$. Also gilt $\varphi = u + id$ wobei u die Verschiebung des Körpers ist.

Das Ziel ist es, obwohl Kraft auf dem Körper ausgeübt wird, dass die Verzerrung des Papiers minimal ist. Dazu betrachten wir den Abstand zwischen $x \in \Omega$ und $x + h \in \Omega$ im aktuellen Zustand

$$\|\varphi(x+h) - \varphi(x)\|^2 = \|\nabla\varphi\|^2 + o(\|h\|^2) = h^T \nabla\varphi^T \nabla\varphi h + o(\|h\|^2)$$

Damit haben wir die Matrix $C := \nabla\varphi^T\nabla\varphi$ erhalten, die die lokale Änderung der Längen angibt. Diese nennt sich Cauchy-Grennscher Verzerrungstensor. Die Verzerrung E ist $\frac{1}{2}(C - I)$. Da $u = \phi - \mathrm{id}$ gilt, können wir die Ableitung von u einsetzen und erhalten

$$E_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) + \frac{1}{2}\sum_k \frac{\partial u_i}{\partial x_k}\frac{\partial u_j}{\partial x_k}$$

Da im linearen die quadratischen Terme vernachlässigt werden können, erhalten wir $\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) = \frac{\nabla u^T + \nabla u}{2}$.

Wenn kräftiger am Papier gezogen wird, reißt es. Dann ist die Formel für die Verzerrung nicht mehr gültig. Also brauchen wir einen Term, der angibt, wo und wie stark der Körper gerissen ist. $v : \Omega \to \mathbb{R}$ ist 0, falls der Körper vollständig gerissen ist und 1, wenn kein Riss vorhanden ist. Multiplizieren wir also $\frac{1}{2}(\nabla u^T + \nabla u)$ mit $v^2 + \epsilon_1$. Es

ergibt sich

$$\min_u \int_\Omega |\nabla u|^2(v^2+\epsilon_1)$$
$$u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2$$

Da wir u nur im zweidimensionalen betrachten, können wir $\frac{\nabla u+\nabla u^T}{2}^2$ vereinfachen zu $|\nabla u|^2$. Der Körper ist an Γ_1 und Γ_2 befestigt, was die Randbedingung u_0 angibt.

In der Modellierung wäre es nun ideal, falls der Körper nie reißt. Dies ist jedoch nicht möglich. Betrachten wir deswegen den Riss genauer. Einerseits wollen wir, dass der Riss möglichst klein ist, also die Oberfläche, auf der der Riss auftritt möglichst minimal ist. Anderersetits stellen wir fest, dass, wenn ein Riss vorhanden ist, immer ein Übergang zwischen dem Riss und dort wo es nicht gerissen ist, ist. Betrachten wir zunächst die Minimierung der Oberfläche. Wir wollen also, dass $\int_\Omega (1-v)^2$ minimal ist. Aber in der Realität stellt man fes, dass es Zwischenstück zwischen Riss un nicht Riss gibt. Damit dieses genügend glatt ist, addieren wir die Ableitung zum Quadrat mit ϵ_2^2 multipliziert hinzu. Daraus ergibt sich eine Darstellung vom Riss

$$\epsilon_2|\nabla v|^2 + \frac{1}{\epsilon_2}(1-v)^2$$

Zusammengefasst ist das gesamte Problem gegeben durch

$$\min_{u\in H^1(\Omega)^2, v\in H^1(\Omega)} \int_\Omega \left(v^2+\epsilon_1\right)|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \frac{1}{\epsilon_2}(1-v)^2\right)dx$$
$$\text{s.d. } 0 \le v \le v_0$$
$$u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2$$

dabei ist ν eine Konstante, $\epsilon_i > 0 \ \forall i \in \{1,2\}$ ein kleiner Parameter und $\Gamma_1, \Gamma_2 \subset \Omega$ der rechte bzw. linke Rand eines rechteckigen Gebietes $\Omega \subset \mathbb{R}^2$, d.h., $\Omega = [0,a]\times[0,b]$ $\Gamma_1 = \{0\}\times[0,b]$ $\Gamma_2 = \{a\}\times[0,b]$

$u : \Omega \to \mathbb{R}^2$ beschreibt die Verschiebung eines Körpers auf dem Gebiet Ω. Der Körper ist an Γ_1 und Γ_2 befestigt, was die Randbedingung u_0 angibt. $v : \Omega \to \mathbb{R}$ gibt an, wo und wie stark der Körper gerissen ist. 0 bedeutet, dass der Körper vollständig gerissen ist und 1, das kein Riss vorhanden ist.

Zur Lösung des Problems, werden analytische und numerische Grundlagen gebraucht. Für die analytische Betrachtung nutze ich partielle Differentialgleichungen und Opti-

mierung. Da die Lösung des Problems analytisch nicht zu finden ist, diskretisiere ich das Problem mit dreieckig lineare Lagrange Elemente und implementiere es mittels semiglatter Newton Methoden. Die Grundlagen dazu sind im zweiten Kapitel zu finden. Alle Themen sind sehr umfangreich und ich werde nur die wichtigsten Begriffe einführen können.

Die Untersuchung des Problems folgt in Kapitel drei. Beim ersten Betrachten fällt auf, dass die Optimierung nach u und v getrennt werden kann. Dieses wird im ersten Teil des dritten Kapitels erläutert. Im zweiten Teil wird die Optimierung nach u betrachtet. Zuerst wird die Existenz und Eindeutigkeit gesichert, um dann numerisch die Lösung mit dreieckig linearen Lagrange Elementen zu suchen. Die Optimierung nach v im dritten Teil hat den selben Aufbau. Nur ist hier das Problem komplizierter und die numerische Lösung erfolgt mit der semiglatten Newton Methode. Im letzten Kapitel werte ich die numerischen Resultate aus und ziehe Rückschlüsse für die Konvergenz der Methode aufgrund der Gitterweite.

2. Mathematische Grundlagen

2.1. Optimierung

Das Phasenfeldmodell für die Rissentstehung ist ein Optimierungsproblem mit Ungleichungsnebenbedingung. Um die Eindeutigkeit und Existenz einer Lösung zu sichern, werden Grundlagen in der Optimierung benötigt. Außerdem werden wir Bedingungen kennenlernen, mit denen sich das Optimierungsproblem in ein einfacheres Problem umschreiben lässt. Grundsätzlich lassen sich Optimierungsprobleme in Probleme mit und ohne Nebenbedingung aufteilen. Fangen wir zunächst mit der einfacheren Variante an.

2.1.1. Optimierungsproblem ohne Nebenbedingung

Optimierungsprobleme ohne Nebenbedingung kennt man im endlichdimensionalen bereits aus der Schule. Wir wollen ein Minimum oder Maximum finden und leiten dazu die zu optimierende Funktion ab und setzen die Ableitung gleich 0. Allerdings betrachten wir jetzt nicht mehr nur endlichdimensionale Probleme, sondern auch unendlichdimensionale. Sei also W ein Banachraum und $J : W \to \mathbb{R}$ ein Funktional. Das Optimierungsproblem ohne Nebenbedingung hat dann folgende Form

$$\min_{w \in W} J(w) \tag{2.1}$$

Um nun wieder die Ableitung 0 setzen zu können, muss erst der Ableitungsbegriff in Banachräumen definiert werden. Dies ist die Gâteaux-Ableitung. Die Definitionen stammen aus [1, S. 50]

Sei $F : U \subset X \to Y$ ein Operator zwischen Banachräumen und $U \neq \emptyset$ offen.
Definition 2.1.1 (Richtungsableitung)**.** *F heißt Richtungsableitbar in $x \in U$, falls*

$$\partial F(x)(h) = \lim_{t \to 0^+} \frac{F(x + th) - F(x)}{t} \in Y$$

für alle $h \in X$ existiert. Dann heißt $\delta F(x,h)$ Richtungsableitung von F in Richtung h.
Definition 2.1.2 (Gâteaux differenzierbar)**.** *F heißt Gâteaux differenzierbar in $x \in U$, falls F richtungsableitbar ist und die Richtungsableitung*

$$F'(x) : X \to Y$$
$$h \mapsto \partial F(x)(h)$$

beschränkt und linear ist d.h. $F'(x) \in L(X,Y)$
Definition 2.1.3 (Fréchet differenzierbar)**.** *F heißt Fréchet differenzierbar in $x \in U$, falls F Gâteaux differenzierbar ist und folgende Approximation gilt*

$$\|F(x+h) - F(x) - F'(x)h\|_Y = o\left(\|h\|_X\right) \text{ für } \|h\|_X \to 0$$

Nun können wir die Ableitung von J bestimmen und daraus resultierend das Optimeriungsproblem lösen. Das Theorem stammt aus der Vorlesung „Optimierung 2“, gelesen von Prof. B. Wirth [4].
Theorem 2.1.4. *Sei das Optimierungsproblem* (2.1) *gegeben. Sei $J : W \to \mathbb{R}$ Gâteaux differenzierbar in $\tilde{w} \in W$. Wenn $\tilde{w}$ das Optimierungsproblem löst, gilt*

$$\partial J(\tilde{w})(h) = 0 \quad \forall h \in W$$

Dabei ist h die Richtung der Ableitung.

Beweis. Für alle $h \in W$ muss $\alpha \mapsto J(\tilde{w} + \alpha h)$ minimal in $\alpha = 0$ sein. Daraus folgt

$$\frac{\partial}{\partial \alpha} f(x + \alpha h)|_{\alpha=0} = 0$$

□

Damit ist Optimierungsproblem zu einer Nullstellensuche geworden. Oftmals ist die Ableitung eine partielle Differentialgleichung. Für diese muss eine Lösung gefunden werden. Dies wird in den Grundlagen Partieller Differentialgleichungen 2.2 erklärt.

2.1.2. Optimierungsproblem mit Ungleichungsnebenbedingung

Oftmals tauchen als Nebenbedingungen Ungleichungsbedingungen wie $a \leq u \leq b$ auf, wobei $a, b, u \in X$ gilt und X ein Vektorraum ist. Dabei ist u die zu optimierende

Funktion, a die untere und b die obere Schranke. Damit überhaupt klar ist, wie das $\leq$ gemeint ist, wird ein positiver Kegel nach der Vorlesung „Optimierung 2“von Prof. Wirth [4] definiert.

Definition 2.1.5 (positiver Kegel)**.** *Sei X ein Vektorraum, $P \subset X$ ein konvexer Kegel. Für $x, y \in X$ schreiben wir $x \leq_P y$ oder $y \geq_p x$, falls $y - x \in P$. P heißt positiver Kegel.*

$x <_P y$ oder $y >_P x$ bedeutet $y - x \in \mathring{P}$

Wir werden Probleme der Form

$$\min_{w \in W} J(w) \quad s.d. \quad G(w) \leq_p 0$$

bearbeiten, wobei W, Z Banachräume sind, $J : W \to \mathbb{R}$ Gâteaux differenzierbar und $G : W \to Z$ die Nebenbedingung des Optimierungsproblems ist. $P \subset Z$ ist ein positiver Kegel. Die Nebenbedingung lässt sich in eine Raumnebenbedingung umschreiben, also $C := \{w \in W | G(w) \leq_P 0\}$. Dabei ist C nichtleer, abgeschlossen und konvex. Das Problem lautet

$$\min_{w \in W} J(w) \quad s.d. \quad w \in C \tag{2.2}$$

Je nachdem welche Notation grade praktischer ist, wird die eine oder andere benutzt. Bei Optimierungen dieser Art muss zunächst die Existenz und Eindeutigkeit der Lösung gesichert werden.

Theorem 2.1.6. *Sei*

1. *W reflexiver Banachraum*
2. *$C \subset W$ nichtleer, konvex und abgeschlossen*
3. *$J : W \to \mathbb{R}$ strikt konvex und stetig auf C*
4. *J Gâteaux differenzierbar*
5. $\lim\limits_{w \in C, \|w\|_W \to \infty} J(w) = \infty$

Dann existiert genau eine Lösung von (2.2)*.*

Beweis. Der Beweis und das Theorem sind in [1, S.66] zu finden □

Bei Optimierungsproblemen mit Nebenbedingung reicht als Bedingung für das Optimum nicht aus, dass die Ableitung 0 ist. Da das Optimum auf dem Rand des zulässigen Gebietes sein könnte, muss die Ableitung nicht zwingend 0 sein. Jedoch gibt es andere Bedingungen, die ausreichend für ein Optimum sind. Die Herleitung dieser Bedingungen, die wir im folgenden Karush-Kuhn-Tucker Bedingungen nennen werden, werde ich aufgrund des Umfanges hier nicht machen können. Ich werde sie nur angeben.

Theorem 2.1.7 (Lagrangefunktion)**.** *Seien X,Y normierte Räume, $P \subset Z$ ein positiver Kegel mit $\mathring{P} \neq \emptyset$. Sei $J : W \to \mathbb{R} \cup \{\infty\}$, $G : W \to Z$ konvex. Es existiert ein $\hat{w}$ im Bild(J), sodass $G(\hat{w}) <_P 0$. Außerdem gelte $\mu = \inf\{J(w) | G(w) \leq_P 0\} < \infty$.*

Dann existiert $z' \in Z^*$ mit $z' \geq_{P^*} 0$, sodass $\mu = \inf_{w \in W} J(w) + \langle G(w), z' \rangle_{Z,Z^*}$. Falls ein optimales $\overline{w}$ existiert, dann minimiert $\overline{w}$ $J(w) + \langle G(w), z' \rangle_{Z,Z^*}$.

Beweis. Der Beweis ist im Script zu der Vorlesung „Optimierung II “, gelesen von Prof. Wirth [4], zu finden. □

Nun haben wir die Bedingungen gegeben, sodass wir von (2.2) das KKT System aufstellen können. Dabei ist $\overline{w}$ die Lösung des Problems. μ und λ sind die Lagrange Multiplikatoren.

$$\begin{aligned} &\nabla J(\overline{w}) + \lambda - \mu = 0 \\ &\overline{w} \geq a \quad \mu \geq 0 \quad \mu(\overline{w} - a) = 0 \\ &\overline{w} \leq b \quad \lambda \geq 0 \quad \lambda(b - \overline{w}) = 0 \end{aligned}$$

Die unteren beiden Zeilen kann man als Projektion darstellen. Es gilt für alle $c > 0$

$$\begin{aligned} \mu &= \max\{0, \mu + c(\overline{w} - a)\} \\ \lambda &= \max\{0, \lambda + c(b - \overline{w})\} \end{aligned}$$

Daraus ergibt sich:

$$\begin{aligned} \mu - \lambda &= \max\{0, \mu + c(\overline{w} - a)\} - \max\{0, \lambda + c(b - \overline{w})\} \\ &= \max\{0, \mu + c(\overline{w} - a)\} + \min\{-\lambda + c(\overline{w} - b)\} \\ &= \max\{0, \mu - \lambda + c(\overline{w} - a)\} + \min\{\mu - \lambda + c(\overline{w} - b)\} \end{aligned} \tag{2.3}$$

Diese Darstellung stammt aus [2] und werde ich später nutzen, um das Problem über die Rissentstehung zu lösen.

2.2. Partielle Differentialgleichungen

Optimierungsprobleme kann man oft umschreiben, sodass, statt dem Optimierungsproblem, eine partielle Differentialgleichung gelöst wird. Dadurch kann man Rückschlüsse auf die Existenz und Eindeutigkeit von dem Optimierungsproblem ziehen. Die Theorie, die ich dazu verwende, ist aus der Vorlesung „partielle Differentialgleichungen“gelesen vom Professor B. Wirth [5].

Wir betrachten das elliptische Dirichlet-Problem auf einem beschränkten Gebiet $\Omega \subset \mathbb{R}^n$

$$\begin{aligned} Lu &= f \text{ auf } \Omega \\ u &= g \text{ auf } \partial\Omega \end{aligned} \tag{2.4}$$

mit $g \in H^1(\Omega)$, $f : \Omega \to \mathbb{R}$ und $Lu(x) := -\text{div}\,(A(x)\nabla u(x)) + b(x)\nabla u(x) + c(x)u(x)$, wobei $A : \Omega \to \mathbb{R}^{n\times n}$, $b : \Omega \to \mathbb{R}^n$ und $c : \Omega \to \mathbb{R}$

Definition 2.2.1 (schwache Lösung)**.** *$u \in g + H_0^1(\Omega)$ heißt schwache Lösung zu* (2.4)*, falls*

$$B(u,v) := \int_\Omega \nabla v^T A \nabla u + b\nabla u v + cuv \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega)$$

Damit eine schwache Lösung eindeutig ist, brauchen wir ein paar Voraussetzungen

Annahme 2.2.2. *Es existieren $\lambda, \Lambda, \nu > 0$, sodass $\forall x \in \Omega$, $\forall \xi, \zeta \in \mathbb{R}^n$ gilt*

1. $\xi^T A(x)\xi \geq \lambda|\xi|^2$
2. $|\xi^T A(x)\zeta| \leq \Lambda|\xi||\zeta|$
3. $\lambda^{-2}|b(x)|^2 + \lambda^{-1}|c(x)| \leq \nu^2$
4. $c(x) \geq 0$

Theorem 2.2.3 (Eindeutigkeit der schwachen Lösung)**.** *Seien die Annahmen 2.2.2 für das Problem* (2.4) *erfüllt. Falls eine schwache Lösung für* (2.4) *existiert, ist sie eindeutig.*

Beweis. Der Beweis wird im Script von Prof. B. Wirth zur Vorlesung „Partielle Differentialgleichungen“geführt. □

Theorem 2.2.4 (Existenz der schwachen Lösung)**.** *Sei Ω beschränkt mit Lipschitz*

Rand. A, b, c *seien beschränkt,* $f \in L^2(\Omega)$*. Dann existiert eine schwache Lösung* $u \in H^1(\Omega)$ *von* (2.4)*.*

Beweis. Der Beweis wird im Script von Prof. B. Wirth zur Vorlesung „Partielle Differentialgleichungen“geführt. □

2.3. Finite Elemente

Finite Elemente sind die Grundlage, um partielle Differentialgleichungen auf zweidimensionalen Gebieten numerisch darstellen zu können. Dazu wird zunächst das Gebiet in Dreiecke trianguliert. Dann werden Basisfunktionen auf diesen Dreiecken definiert, die sogenannten globalen Formfunktionen. Aus diesen ist die gesuchte Funktion zusammengesetzt und kann damit berechnet werden. DiesesVorgehen nennt man Galerkin-Ansatz. Die hier beschriebene Theorie richtet sich nach der Vorlesung „Numerik partieller Differentialgleichungen“gelesen von Dr. F. Wübbeling [3].

Es ist ein rechteckiges Gebiet in 2D gegeben. ObdA $\Omega = [0, a] \times [0, b]$. Auf diesem Gebiet legen wir ein äquidistantes Gitter G_h.

$$G_h := \left\{ (ih_1, jh_2) | i = 0, \cdots, \frac{a}{h_1}, j = 0, \cdots, \frac{b}{h_2} \right\}$$

$h = (h_1, h_2)$ ist die Schrittweite mit $a = (n+1)h_1$ und $b = (m+1)h_2$, $n+1$ die Anzahl der Stützpunkte in x-Richtung und $m+1$ die Anzahl der Stützpunkte in y-Richtung. Um ein sinnvolles Gitter zu erhalten, sollten h_1 und h_2 recht nahe beieinander gewählt werden. Nun wird durch die Gitterpunkte die Triangulierung gelegt. Diese nennen wir E_k und ist in 2.1 dargestellt.

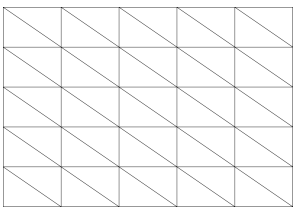


Abbildung 2.1.: Triangulierung eines rechteckigen Gebietes

Stellen wir das Referenzelement unserer Finiten Elemente auf. Wir benutzen dreieckig lineare Lagrange Elemente. Bei diesen sind die Funktionsauswertungen auf den Ecken der Dreiecke gegeben. Das Finite Element ist deswegen gegeben durch (E, P, Ψ), wobei

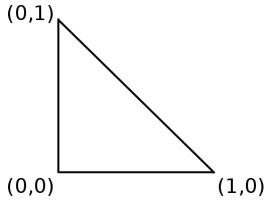


Abbildung 2.2.: Referenzdreieck

E das Referenzdreieck 2.2 ist, $P = \mathcal{P}_1$, sind Polynome auf $\mathbb{R}^2$ vom Grad 1 mit Basis $\{p_1, p_2, p_3\}$

$$p_1(x,y) := 1 \qquad p_2(x,y) := x \qquad p_3(x,y) := y$$

und $\Psi := \{\varphi_0, \varphi_1, \varphi_2\}$ sind Funktionale auf P und damit eine Basis von P^*. φ_i sind lokale Formfunktionen d.h. $\varphi_i(p_j) = \delta_{ij}$, $i, j \in \{1, 2, 3\}$. Dabei ist δ_{ij} das Kronecker-Delta. Außerdem soll gelten $\varphi_i(p_j) = p_j(a_i)$, wobei a_i eine Auswertung in einer Ecke des Dreiecks ist. Daraus ergibt sich, dass

$$\varphi_1 = 1 - x - y \qquad \varphi_2 = x \qquad \varphi_3 = y \tag{2.5}$$

Nun ist das Referenzelement gegeben. Jedes Element (E_k, P_k, Ψ_k) lässt sich mit der affin linearen Transformation

$$\begin{array}{rccc} T: & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & \begin{pmatrix} x \\ y \end{pmatrix} & \mapsto & \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \pm \begin{pmatrix} h_1 x \\ h_2 y \end{pmatrix} \end{array}$$

durch das Referenzelement darstellen. Dabei entspricht $(a_1, a_2)^t$ dem Eckpunkt mit dem 90° Winkel des Dreiecks und $(h_1, h_2)^t$ ist die Höhe bzw Breite des Dreiecks. Mit dem Transformationssatz können wir alle Berechnungen auf dem Referenzelement ausführen und dann auf das transformierte Element übertragen. Durch die Transformation muss dann zu allen Integralen $|\det DT(x,y)|^{-1}$ multipliziert werden. Das ergibt

$$|\det \mathrm{D}\ T(x,y)|^{-1} = |\det \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}|^{-1} = \frac{1}{h_1 h_2}$$

Die Familie $\{(E_k, P_k, \Psi_k)\}$ von Finiten Elementen, die durch unsere Triangulierung hervorgegangen ist, ist verträglich. Also können wir die globalen Formfunktionen aufstellen, die auf dem gesamten Gebiet Ω definiert sind. Die globale Formfunktion T_j ist

1 auf dem Gitterpunkt j und 0 sonst.

Für die Berechnung von linearen Funktionen auf dreieckig linearen Lagrange Elementen, brauchen wir oft eine explizite Darstellung. Durch die Triangulierung haben wir 2 Arten von Dreiecken. Dabei entspricht a^i der Wert der Funktion a an dem Eckpunkt i.

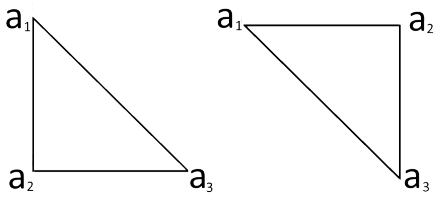


Abbildung 2.3.: gerade (links) und ungerade Dreiecke mit den Werten von a

$a(x, y)$ wird auf dem linken Dreieck von 2.3 dargestellt durch

$$a(x,y) = (a_3 - a_2)x + (a_1 - a_2)y + a_2 \quad \text{mit} \quad \nabla a(x,y) = \begin{pmatrix} a_3 - a_2 \\ a_1 - a_2 \end{pmatrix} \tag{2.6}$$

und auf dem rechten Dreieck von 2.3 wird $a(x, y)$ dargestellt durch

$$a(x,y) = (a_1 - a_2)x + (a_3 - a_2)y + a_2 \quad \text{mit} \quad \nabla a(x,y) = \begin{pmatrix} a_1 - a_2 \\ a_3 - a_2 \end{pmatrix} \tag{2.7}$$

2.4. Semidifferenzierbare Newton Methode

Semiglatte Newton Methoden werden gebraucht, um Nullstellen von nicht differenzierbaren Funktionen numerisch zu berechnen. Die Rissentstehung ist ein nicht differenzierbares Problem. Um die Idee der Newton Methoden zu verstehen, führe ich zunächst die einfache Newton Methode ohne Nebenbedingung und dann solche mit einfachen Nebenbedingungen ein. Um diese realisieren zu können, wird der Begriff der Semidifferenzierbarkeit benötigt. Das ist eine Mengenwertige Ableitung, mit der auch nicht differenzierbare, aber stetige Punkte in einer Funktion abgeleitet werden können. Damit kann die semidifferenzierbare Newton Methode eingeführt werden, von der wir auch die Konvergenz betrachten werden. Dieses Kapitel richtet sich nach [1, S. 115 ff].

2.4.1. Newton Methode ohne Nebenbedingung

Als erstes leiten wir uns zum Verständnis die einfache Newton Methode her. Dazu betrachten wir wie vorher das Minimierungsproblem

$$\min_{w \in \mathbb{R}^n} f(w) \qquad f : \mathbb{R}^n \to \mathbb{R} \tag{2.8}$$

Die Optimalbedingung zu diesem Problem lautet $\nabla f(w) = 0$. Nun wollen wir ein numerisches Verfahren für dieses Problem entwickeln. Dazu setzen wir $G := \nabla f$. Da wir ein diskretes Verfahren wollen, setzten wir $w_0, w_1, \cdots$ in G ein. Wir erhalten

$$G(w_{k+1}) = 0$$

Um ein iteratives Verfahren zu erhalten taylorn wir G in w_k. Das ergibt

Theorem 2.4.1 (einfaches Newtonverfahren)**.** *Der Algorithmus 1 löst das Optimierungsproblem* (2.8)*. Es konvergiert superlinear falls $G \in C^1$ und G' invertierbar ist.*

Data: w^0 *(möglichst Nah an der Lösung $\overline{w}$)*
for $k = 0, 1, \cdots$ **do**
  Löse $G'(w^k)s^k = -G(w^k)$;
  $w^{k+1} = w^k + s^k$;
end

Algorithm 1: einfache Newton Methode

2.4.2. Konvergenz der generalisierten Newton Methode

Nun möchten wir Aussagen über die Konvergenz der Newton Methode treffen können. Dazu definieren wir Konvergenzgeschwindigkeiten.

Definition 2.4.2 (Konvergenzgeschwindigkeit)**.** *Sei x_k eine Folge, die $\overline{x}$ approximiert.*

- *lineare Konvergenz:* $\|x_{k+1} - \overline{x}\| \leq c\|x_k - \overline{x}\| \quad \forall k > k_0$
- *superlineare Konvergenz: Sei c_k eine Nullfolge.* $\|x_{k+1} - \overline{x}\| \leq c_k\|x_k - \overline{x}\| \quad \forall k > k_0$
- *Konvergenz der Ordnung p:* $\|x_{k+1} - \overline{x}\| \leq c\|x_k - \overline{x}\|^p \quad \forall k > k_0$

Betrachte nun

$$G(x) = 0 \tag{2.9}$$

mit $G : X \to Y$, wobei X, Y Banachräume sind. Sei $\overline{x}$ die Lösung der Gleichung.

Um eine numerische Lösung von (2.9) zu erhalten, benutzen wir einen ähnlichen Algorithmus, wie den für das einfache Newtonverfahren, nur allgemeiner

Data: $x^0 \in X$ (möglichst Nah an der Lösung $\overline{x}$)
for $k = 0, 1, \cdots$ **do**
  Wähle invertierbaren Operator $M_k \in L(X, Y)$;
  Erhalte s_k *beim lösen von* $M_k s^k = -G(x^k)$;
  $x^{k+1} = x^k + s^k$;
end

Algorithm 2: Generalisierte Newton Methode

Bis jetzt war der Operator M_k die Ableitung von G. Dies ist jedoch nicht möglich, wenn G nicht differenzierbar ist. Wie der Operator M_k in diesem Fall sinnvoll zu wählen ist, wird später bestimmt.

Nun untersuchen wir die durch diesen Algorithmus gewonnene Folge x^k in einer Umgebung von $\overline{x}$. Sei $d^{k+1} = x^{k+1} - \overline{x}$ der Abstand zwischen dem Iterationsschritt und der Lösung. Dann gilt

$$\begin{aligned} M_k d^{k+1} &= M_k(x^{k+1} - \overline{x}) = M_k(x^k + s^k - \overline{x}) = M_k d^k - G(x^k) \\ &= G(\overline{x}) + M_k d^k - G(x^k) \end{aligned}$$

Wir erhalten

Theorem 2.4.3 (Konvrgenz der generalisierten Newton Methode)**.** *Betrachte* (2.9) *mit der Lösung* $\overline{x}$*. Sei* x^k *die Folge, die durch den Generalisierten Newton Algorithmus 2 erzeugt wurde. Sei* x^0 *nah genug an* $\overline{x}$ *gewählt*

1. *Falls* $\exists \gamma \in (0, 1)$ *mit*

$$\begin{aligned} &\|d^{k+1}\|_X = \|M_k^{-1}\left(G(\overline{x} + d^k) - G(\overline{x}) - M_k d^k\right)\|_X \leq \gamma \|d^k\|_X \\ &\forall k \text{ mit } \|d_k\|_X \text{ klein genug} \end{aligned}$$

   gilt, dann konvergiert $x^k \to \overline{x}$ *linear mit Konstante* γ

2. *Falls* $\forall \eta \in (0, 1) \quad \exists \delta_\eta > 0$*, sodass*

$$\begin{aligned} &\|d^{k+1}\|_X = \|M_k^{-1}\left(G(\overline{x} + d^k) - G(\overline{x}) - M_k d^k\right)\|_X \leq \eta \|d^{k+1}\|_X \\ &\text{für } \|d_k\|_X < \delta_\eta \end{aligned}$$

gilt, dann konvergiert $x^k \to \overline{x}$ super linear

3. *Falls $\exists \gamma \in (0,1)$ mit*

$$\|d^{k+1}\|_X = \|M_k^{-1}\left(G(\overline{x}+d^k) - G(\overline{x}) - M_k d^k\right)\|_X \leq C\|d^k\|_X^{1+\alpha}$$
$$\text{für } \|d_k\|_X \to 0$$

gilt, dann konvergiert $x^k \to \overline{x}$ super linear der Ordnung $\alpha + 1$

Beweis. Der Beweis ist in [1, S. 118] zu finden. □

Oft teilt man diese Kleinheitsannahmen in zwei Teile auf:
Definition 2.4.4 (Regularitätsannahme)**.** *Sei $M_k \in L(X,Y)$, wobei X, Y Banachräume sind. Dann ist die Regularitätsannahme gegeben durch:*

$$\|M_k^{-1}\|_{Y\to X} \leq C \quad \forall k \geq 0$$

Bemerkung (Operatornorm). Die Notation für die Operatornorm von einem linearen Operator $f : X \to Y$, wobei X, Y normierte Vektorräume sind lautet:

$$\|f\|_{X\to Y} := \sup_{\|x\|_X=1} \|f(x)\|_Y$$

Definition 2.4.5 (Approximationsannahme)**.** *Sei $M_k \in L(X,Y)$, wobei X, Y Banachräume sind, $\overline{x}$ die Lösung von $G(x) = 0$ und $d^k := x^k - \overline{x}$ Sei $\alpha + 1 > 1$ Dann ist die Approximationsannahme gegeben durch:*

$$\|G(\overline{x}+d^k) - G(\overline{x}) - M_k d^k\|_Y = o(\|d^k\|_X) \text{ für } \|d_k\|_X \to 0$$

oder

$$\|G(\overline{x}+d^k) - G(\overline{x}) - M_k d^k\|_Y = o(\|d^k\|_X^{1+\alpha}) \text{ für } \|d_k\|_X \to 0$$

Die geeignete Wahl von M_k ist das sogenannte Semidifferential. Was das genau ist und wie es gerechnet wird, klärt folgendes Kapitel.

2.4.3. Semidifferential

Definition 2.4.6 (verallgemeinerte Differentiale)**.** *Seien X, Y Banachräume und $G : X \to Y$ ein stetiger Operator. Dann ist die Menge der verallgemeinerten Differentiale definiert als*

$$\partial G : X \rightrightarrows L(X, Y)$$

Dabei meint $X \rightrightarrows L(X, Y)$, dass ein Punkt $x \in X$ auf eine Menge von linearen Operatoren abgebildet wird. Ein Beispiel für ein verallgemeinertes Differenzial ist das Clarke Differenzial. Dies ist jedoch nur für Vektorwertige Funktionen definiert.

Nun können wir, um unser Newtonverfahren umzugestalten $M_k \in \partial G(x^k)$ wählen. Damit unser Verfahren aber super linear konvergiert, muss gelten

$$\sup_{M \in \partial G(\overline{x}+d)} \|G(\overline{x} + d^k) - G(\overline{x}) - M_k d\|_Y = o\left(\|d\|_X\right) \text{ für } \|d\|_X \to 0$$

Dieses nennt sich semidifferenzierbar.

Definition 2.4.7 (semidifferenzierbar)**.** *Sei $G : X \to Y$ ein stetiger Operator zwischen Banachräumen. Sei $\partial G : X \rightrightarrows L(X, Y)$ mit nicht leeren Bildern gegeben wie oben.*

1. *G heißt ∂G semidifferenzierbar in $x \in X$, falls*

$$\sup_{M \in \partial G(x+d)} \|G(x + d^k) - G(x) - M_k d\|_Y = o\left(\|d\|_X\right) \text{ für } \|d\|_X \to 0 \qquad (2.10)$$

2. *G heißt ∂G semidifferenzierbar von der Ordnung $\alpha + 1 > 1$ in $x \in X$, falls*

$$\sup_{M \in \partial G(x+d)} \|G(x + d^k) - G(x) - M_k d\|_Y = \mathcal{O}\left(\|d\|_X^{\alpha+1}\right) \text{ für } \|d\|_X \to 0$$

Lemma 2.4.8. *Sei $G : X \to Y$ ein Operator zwischen Banachräumen und stetig Fréchet-differenzierbar in einer Umgebung von x. Dann ist G $\{G'\}$-semidifferenzierbar in x. Falls G' α-Hölderstetig in einer Umgebung von x ist, dann ist G $\{G'\}$-semidifferenzierbar in x von der Ordnung α.*

$\{G'\}$ beschreibt den Operator $\{G'\} : X \rightrightarrows L(X, Y)$ mit $\{G'\}(x) = \{G'(x)\}$

Beweis.

$$\begin{aligned}
&\|G(x+d^k) - G(x) - G'(x+d)d\|_Y \\
&\leq \|G(x+d^k) - G(x) - G'(x)d\|_Y + \|G'(x)d - G'(x+d)d\|_Y \\
&\leq o\left(\|d\|_X\right) + \|G'(x) - G'(x+d)\|_{X\to Y}\|d\|_X = o\left(\|d\|_X\right)
\end{aligned}$$

Der zweite Teil des Beweises erfolgt analog, siehe [1, S. 121] □

Theorem 2.4.9 (Rechenregeln semidifferenzierbare Funktionen)**.** *Seien X, Y, Z, X_i, Y_i Banachräume.*

1. *Falls die Operatoren $G_i : X_i \to Y_i$ ∂G_i-semidifferenzierbar in x sind, dann ist (G_1, G_2) $(\partial G_1, \partial G_2)$-semidifferenzierbar in x.*
2. *Falls die Operatoren $G_i : X \to Y$ ∂G_i-semidifferenzierbar in x sind, dann ist $G_1 + G_2$ $(\partial G_1 + \partial G_2)$-semidifferenzierbar in x.*
3. *Seien $G_1 : Y \to Z$ und $G_2 : X \to Y$ ∂G_i-semidifferenzierbar in $G_2(x)$ und in x. Sei außerdem ∂G_1 beschränkt in einer Umgebung von $x = G_2(x)$ und G_2 ist Lipschitzstetig in einer Umgebung von x. Dann ist $G = G_1 \circ G_2$ ∂G-semidifferenzierbar mit*

$$\partial G(x) = \{M_1 M_2 | M_1 \in \partial G_1\left(\partial G_2(x)\right), \quad M_2 \in \partial G_2(x)\}$$

Beweis. Der Beweis ist in [1, S. 122] zu finden. □

2.4.4. Semidifferenzierbare Newton Methode

Mit dem Semidifferential können wir nun die semidifferenzierbare Newton Methode definieren.

Data: $x^0 \in X$ (möglichst Nah an der Lösung $\overline{x}$)
for $k = 0, 1, \cdots$ **do**
  Wähle $M_k \in \partial G(x^k)$;
  Erhalte s_k beim lösen von $M_k s^k = -G(x^k)$;
  $x^{k+1} = x^k + s^k$;
end

Algorithm 3: semidifferenzierbare Newton Methode

Damit diese konvergiert, muss die Approximationsannahme und die Regularitätsannahme erfüllt sein. Die Approximationsannahme ist durch die Semidifferenzierbarkeit gegeben. Fehlt noch die Regularitätsannahme.

Definition 2.4.10 (Regularitätsannahme für das semidifferenzierbare Newton Verfahren)**.** *Betrachte* (2.9) *mit der Lösung* $\overline{x}$*. Dann lautet die Regularitätsannahme*

$$\exists C > 0, \quad \exists \delta > 0 : \|M^{-1}\|_{X\to Y} \leq C \quad \forall M \in \partial G(x) \quad \forall x \in X, \quad \|x - \overline{x}\|_X < \delta \tag{2.11}$$

Theorem 2.4.11 (Konvergenz das semidifferenzierbare Newton Verfahrens)**.** *Sei das Problem* (2.9) *gegeben mit der Lösung* $\overline{x}$*. Seien* X, Y *Banachräume,* $G : X \to Y$ *stetig und* ∂G *semidifferenzierbar und die Regularitätsannahme* (2.11) *sei gegeben. Dann existiert* $\delta > 0$*, sodass für alle* $x^0 \in X$ *mit* $\|x^0 - \overline{x}\|_X < \delta$ *die semidifferenzierbar Newton Methode super linear gegen* $\overline{x}$ *konvergiert.*

Falls G ∂G*-semidifferenzierbar der Odnung* $\alpha > 0$ *in* $\overline{x}$ *ist, dann ist die Konvergenzordnung* $1 + \alpha$

Beweis. 2.4.3 besagt, dass wenn ich ein Newtonverfahren der Form 2 habe, also $M_k \in \mathcal{L}(X, Y)$, M_k invertierbar ist und

$$\|M_k^{-1}\left(G(\overline{x} + d^k) - G(\overline{x}) - M_k d^k\right)\|_X = o(\|d^k\|_X)$$

gilt, dann konvergiert das Newtonverfahren super linear. Da $M_k \in \partial G$, ist $M_k \in \mathcal{L}(X, Y)$. M_k ist invertierbar, da die Regularitätsannahme gilt. Außerdem gilt mit der Regularitätsannahme und der Semidifferenzierbarkeit

$$\begin{aligned}
&\|M_k^{-1}\left(G(\overline{x} + d^k) - G(\overline{x}) - M_k d^k\right)\|_X \\
&\leq \|M_k^{-1}\|_X\| \left(G(\overline{x} + d^k) - G(\overline{x}) - M_k d^k\right)\|_X \\
&\leq Co(\|d^k\|_X) = o(\|d^k\|_X)
\end{aligned}$$

Also ist 2.4.3 anwendbar. □

Damit haben wir Bedingungen für die Konvergenz der semidifferenzierbaren Newton Methode gefunden. Diese können wir für den Beweis der Konvergenz bei unserer Newton Methode anwenden.

3. Anwendung auf das Phasenfeldmodell für Rissentstehung

Nachdem wir die mathematischen Grundlagen für die Betrachtung eines Optimierungsproblems kennengelernt haben, wollen wir diese anwenden. Zunächst teilen wir das Minimierungsproblem in zwei voneinander unabhängige Optimierungen auf: Die Optimierung nach u und die Optimierung nach v. Im Anschluss betrachten wir beide Optimierungen genauer, indem wir sie umschreiben und numerische Verfahren zur Lösung entwickeln. Zum Schluss fusionieren wir beide Verfahren.

3.1. Erste Betrachtung des Modells

Erinnern wir uns an die vorangegangene Problemstellung

$$\begin{aligned}
&\min_{u\in H^1(\Omega)^2, v\in H^1(\Omega)} \int_\Omega \left(v^2+\epsilon_1\right)|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \frac{1}{\epsilon_2}(1-v)^2\right)\mathrm{d}x\\
&\text{s.d. } 0 \le v \le v_0\\
&u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2
\end{aligned}$$

Beim genaueren Betrachten bemerkt man, dass die Ungleichungsnebenbedingung nur von v und die Randbedingung nur von u abhängt. Dies bietet die Möglichkeit das

Optimierungsproblem in zwei Teilprobleme aufzuteilen.

$$\min_{u \in H^1(\Omega)^2} \int_\Omega \left(v^2 + \epsilon_1\right) |\nabla u|^2 + \nu \left(\epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_2} (1-v)^2\right) \mathrm{d}x$$
$$u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2$$
$$\min_{v \in H^1(\Omega)} \int_\Omega \left(v^2 + \epsilon_1\right) |\nabla u|^2 + \nu \left(\epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_2} (1-v)^2\right) \mathrm{d}x$$
$$\text{s.d. } 0 \leq v \leq v_0$$

Wenn man beide Probleme implementiert, löst man zunächst die Optimierung nach u und setzt die Lösung dann in die Optimierung nach v ein. Danach setzt man die Lösung von v in die Optimierung nach u ein. Dieses Vorgehen wird mit der semidifferenzierbaren Newtonmethode wiederholt. Betrachten wir zuerst die Optimierung nach u.

3.2. Optimierung nach u

Das Kapitel ist in zwei Unterkapitel aufgeteilt: Die analytische und numerische Betrachtung.

Im analytischen Teil formulieren wir das Problem zu einem Problem ohne Nebenbedingung um. Dieses lässt sich dann als partielle Differentialgleichung schreiben. Die Existenz und Eindeutigkeit der schwachen Lösung sichern wir uns am Schluss des ersten Teils.

Für die numerische Betrachtung schreiben wir das Problem nochmal um und wenden dann Finite Elemente und den Galerkin Ansatz an. Das Resultat ist ein Gleichungssytem, das sich einfach lösen lässt.

3.2.1. Analytische Betrachtung

Erinnern wir uns an das Optimierungsproblem, das die Verschiebung des Körpers bei der Entstehung von einem Riss beschreibt.

$$\min_{u \in H^1(\Omega)^2} \int_\Omega \left(v^2 + \epsilon_1\right) |\nabla u|^2 + \nu \left(\epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_2} (1-v)^2\right) \mathrm{d}x$$
$$u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2$$

Es ist leichter ein Problem ohne Nebenbedingung zu betrachten, also nehmen wir die Nebenbedingung mit in dem Raum auf, über den wir optimieren. Also suchen wir statt $u \in H^1(\Omega)^2$

$$u \in u_0 + H_0^1(\Omega)^2 := u_0 + \{u \in H^1(\Omega)^2 | u = 0 \text{ auf } \Gamma_1 \cup \Gamma_2\}$$

Das Problem hat dann folgende Form

Da $u : \Omega \to \mathbb{R}^2$, müssen wir nach u_1 und nach u_2 minimieren. Es tauchen keine Mischung aus den Termen u_1 und u_2 auf, das heißt, dass wir die Optimierungen trennen können. Beide sind identisch, es müssen später nur unterschiedliche Werte eingesetzt werden. Betrachten wir oBdA die Optimierung nach u_1.

Theorem 3.2.1 (Bedingung für ein Minimum)**.** *Sei das Minimierungsproblem* (**??**) *gegeben und $\tilde{u_1}$ nimmt das Minimum an. Dann gilt*

$$\int_\Omega 2\left(v^2 + \epsilon_1\right) \nabla \tilde{u_1} \nabla \psi \, \mathrm{d}x = 0 \forall \psi \in u_0 + H_0^1(\Omega) \tag{3.1}$$

Beweis. Nach 2.1.4 muss nur überprüft werden, ob die Gâteaux-Ableitung von

$$\begin{aligned} J: \quad u_0 + H_0^1(\Omega)^2 &\to \mathbb{R} \\ u &\mapsto \int_\Omega \left(v^2 + \epsilon_1\right) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \tfrac{1}{\epsilon_2} (1-v)^2 \, \mathrm{d}x \end{aligned}$$

(3.1) ist. Leiten wir J ab:

$$\begin{aligned}
\partial J(u)(\psi) &= \lim_{t\to 0} \tfrac{1}{t}\Big(\quad J(u_1 + t\psi, u_2) - J(u_1, u_2)\Big) \\
&= \lim_{t\to 0} \tfrac{1}{t}\Big(\int_\Omega (v^2+\epsilon_1)\,|\nabla(u_1+t\psi)|^2 + |\nabla u_2|^2 + \epsilon_2|\nabla v|^2 + \tfrac{1}{\epsilon_2}\,(1-v)^2\,\mathrm{d}x \\
&\qquad\qquad - \int_\Omega (v^2+\epsilon_1)\,|\nabla u_1|^2 \qquad\quad + |\nabla u_2|^2 + \epsilon_2|\nabla v|^2 + \tfrac{1}{\epsilon_2}\,(1-v)^2\,\mathrm{d}x\Big) \\
&= \lim_{t\to 0} \tfrac{1}{t}\Big(\int_\Omega (v^2+\epsilon_1)\,(|\nabla(u_1+t\psi)|^2 - |\nabla u_1|^2)\,\mathrm{d}x\Big) \\
&= \lim_{t\to 0} \tfrac{1}{t}\Big(\int_\Omega (v^2+\epsilon_1)\,(|\nabla u_1 + t\nabla\psi|^2 - |\nabla u_1|^2)\,\mathrm{d}x\Big) \\
&= \lim_{t\to 0} \tfrac{1}{t}\Big(\int_\Omega (v^2+\epsilon_1)\,(|\nabla u_1|^2 + 2t\nabla u_1\nabla\psi + t^2|\nabla\psi|^2 - |\nabla u_1|^2)\,\mathrm{d}x\Big) \\
&= \lim_{t\to 0} \tfrac{1}{t}\Big(\int_\Omega (v^2+\epsilon_1)\,(2t\nabla u_1\nabla\psi + t^2|\nabla\psi|^2)\,\mathrm{d}x\Big) \\
&= \int_\Omega 2\,(v^2+\epsilon_1)\,\nabla u_1\nabla\psi\,\mathrm{d}x
\end{aligned}$$

Damit dies eine Gâteaux Ableitung ist, muss die Abbildung $J'(u_1) : \psi \mapsto \partial J(u_1, \psi) \in \mathbb{R}$ linear und beschränkt sein. Linearität ist einfach nachzurechnen. Beschränktheit lässt sich durch Cauchy-Schwarz zeigen. □

Also lautet unser analytisches Problem: Finde $u_1 \in u_0 + H_0^1(\Omega)$, sodass $\forall\psi \in u_0 + H_0^1(\Omega)$ gilt

$$0 = \int_\Omega 2\left(v^2+\epsilon_1\right)\nabla u_1\nabla\psi\,\mathrm{d}x \tag{3.2}$$

Nun ist noch interessant, ob eine Lösung existiert und ob sie eindeutig ist. Dieses hängt von u_0 und v_0 ab.

Theorem 3.2.2 (Existenz und Eindeutigkeit)**.** *Sei $u_0 \in H^1(\Omega)^2,, v_0 \in H^1(\Omega)$. Die schwache Lösung $u \in u_0 + H_0^1(\Omega)$ von (3.2) existiert und ist eindeutig.*

Beweis. Wir wenden 2.2.4 an. Dazu müssen wir die Bilinearform aufstellen und dann 2.2.2 zeigen. Die Bilinearform lautet

$$\begin{aligned}
B(u_1,\psi): \quad (u_0 + H^1(\Omega)0^2)^2 &\to \mathbb{R} \\
(u_1,\psi) &\mapsto \int_\Omega (v^2+\epsilon_1)\,2\nabla u_1\nabla\psi\,\mathrm{d}x
\end{aligned}$$

Mit den Bezeichnungen aus 2.2 ist $g = u_0,\ f, b, c = 0$ und

$$A(x) := \begin{pmatrix} v^2(x) + \epsilon_1 & 0 \\ 0 & v^2(x) + \epsilon_1 \end{pmatrix}$$

Aus 2.2.2 sind 3 und 4 bereits erfüllt, da $b, c = 0$ gilt. Beweisen wir 1.

Sei $\xi \in \mathbb{R}^n$. Dann gilt:

$$\begin{aligned} \xi^T A(x)\xi &= \xi^T \begin{pmatrix} v^2(x) + \epsilon_1 & 0 \\ 0 & v^2(x) + \epsilon_1 \end{pmatrix} \\ &= (v^2 + \epsilon_1)\xi \cdot \xi \\ &\geq \epsilon_1 |\xi|^2 \end{aligned}$$

Damit ist Annahme 1 erfüllt mit $\lambda = \epsilon_1$. Für Annahme 2 gilt

$$|\xi^T A(x)\zeta| = (v^2 + \epsilon_1)\xi \cdot \zeta \leq (v_0^2 + \epsilon_2)\xi \cdot \zeta \leq (\sup(v_0)^2 + \epsilon_2)|\xi||\zeta|$$

mit $\Lambda = \sup(v_0)^2 + \epsilon$ □

3.2.2. Numerische Betrachtung

Wir haben grade bewiesen, dass wir folgendes Problem lösen müssen:

Finde $u_1 \in u_0 + H_0^1(\Omega)$, sodass

$$0 = \int_\Omega \left(v^2 + \epsilon_1\right) \nabla u_1 \nabla \psi \, \mathrm{d}x \quad \forall \psi \in u_0 + H_0^1(\Omega)$$

Da die Nullstelle im Raum $H_0^1(\Omega)$ einfacher zu finden ist, als im Raum $u_0 + H_0^1(\Omega)$, stellen wir das Problem um. Dazu definieren wir $\tilde{u}_0 \in u_0 + H_0^1(\Omega)$, sodass $\tilde{u}_0\ u_{0_1}$ auf dem Rand $\Gamma_1 \cup \Gamma_2$ entspricht und sonst 0 ist. Definiere zusätzlich $\tilde{u} \in H_0^1(\Omega)$, sodass $u_1 = \tilde{u} + \tilde{u}_0$. Damit lässt sich das Problem umschreiben zu

Finde $\tilde{u} \in H_0^1(\Omega)$, sodass $\forall \psi \in H_0^1(\Omega)$

$$-\int_\Omega \left(v^2 + \epsilon_1\right) \nabla \tilde{u}_0 \nabla \psi \, \mathrm{d}x = \int_\Omega \left(v^2 + \epsilon_1\right) \nabla \tilde{u} \nabla \psi \, \mathrm{d}x$$

Zur numerischen Betrachtung bieten sich Finite Elemente, insbesondere die dreieckig linearen Lagrange Elemente an. Dafür triangulieren wir das Gebiet, wie in 2.3 dargestellt. Nun nutzen wir den Galerkin Ansatz. Dafür gilt ab jetzt $k := (m+1)(n+1)$

$$\tilde{u}(x,y) := \sum_{i=1}^{k} u_i^h T_i(x,y)$$

Dabei sind $T_i(x,y)$ die globalen Formfunktionen und u_i^h die gesuchten Konstanten. Setzt man die Definition von $\tilde{u}$ ein und ersetzt $\psi \in H_0^1(\Omega)$ durch die Basis von P^*, also den globalen Formfunktionen T_i, so gilt $\forall i \in \{1, \cdots, k\}$

$$\begin{aligned}
&-\int_\Omega \left(v^2+\epsilon_1\right) \nabla\tilde{u}_0 \nabla\psi \,\mathrm{d}x = \int_\Omega \left(v^2+\epsilon_1\right) \nabla\tilde{u} \nabla\psi \,\mathrm{d}x \\
\Leftrightarrow &-\int_\Omega \left(v^2+\epsilon_1\right) \sum_{i=1}^{k} u_{0\,i}^h \nabla T_i \nabla T_i \,\mathrm{d}x = \int_\Omega \left(v^2+\epsilon_1\right) \sum_{i=1}^{k} u_i^h \nabla T_i \nabla T_j \,\mathrm{d}x \\
\Leftrightarrow &-\sum_{i=1}^{k} u_{0\,i}^h \int_\Omega \left(v^2+\epsilon_1\right) \nabla T_j \nabla T_i \,\mathrm{d}x = \sum_{i=1}^{k} u_i^h \int_\Omega \left(v^2+\epsilon_1\right) \nabla T_i \nabla T_j \,\mathrm{d}x \\
\Leftrightarrow &L * u_0^h = L * u^h
\end{aligned}$$

wobei $u^h := (u_1^h, \cdots u_k^h)^T$, $u_0^h := (u_{0\,1}^h, \cdots u_{0\,k}^h)^T$ und $L := \left(\int_\Omega \left(v^2+\epsilon\right) \nabla T_i \nabla T_j \,\mathrm{d}x\right)_{ij}$

Also müssen wir L berechnen und dann das Gleichungssystem $L * u_0^h = L * u^h$ lösen.

Berechnung des u Integrals

Als erste Vereinfacherung betrachten wir nicht mehr das Integral über Ω, sondern über die einzelnen Dreiecke der Triangulierung. Desweiteren ist T_i linear, also ∇T_i konstant. Es gilt

$$\begin{aligned}
\int_\Omega \left(v^2+\epsilon_1\right) \nabla T_i \nabla T_j \,\mathrm{d}x &= \sum_{\tilde{E}\in E_k} \int_{\tilde{E}} \left(v^2+\epsilon_1\right) \nabla T_i \nabla T_j \,\mathrm{d}x \\
&= \sum_{\tilde{E}\in E_k} \nabla T_i \nabla T_j \int_{\tilde{E}} \left(v^2+\epsilon_1\right) \mathrm{d}x
\end{aligned}$$

Wir kennen $\nabla T_i \nabla T_j$ auf jedem Dreieck. Also muss nur noch $\int_E v^2 + \epsilon \,\mathrm{d}x$ berechnet werden. Es darf über das Referenzdreieck integriert werden, da durch den Transfor-

mationssatz das Integral über das transformierte Element gewonnen werden kann. Es gilt:

$$\int_E v^2 + \epsilon_1 \, \mathrm{d}x = \int_E v^2 \, \mathrm{d}x + \frac{1}{2}\epsilon_1$$

Da v bereits numerisch berechnet wurde, haben wir nur Funktionsauswertungen von v an den Ecken des Dreieckes gegeben und wir wissen, dass $v \in \mathcal{P}_1$. Also ist v eindeutig bestimmt und kann berechnet werden. Die Darstellung von v ist in (2.6) und (2.7) zu finden.

Die Berechnung von $\int_E v^2 \, \mathrm{d}x$ sieht wie folgt aus

$$\begin{aligned}\int_E v(x,y)^2 \, \mathrm{d}x \, \mathrm{d}y &= \int_0^1 \int_0^{1-y} \left((v_3 - v_1)x + (v_2 - v_1)y + v_1\right)^2 \mathrm{d}x \, \mathrm{d}y \\ &= \frac{1}{12}(v_1^2 + v_2^2 + v_3^2 + v_1v_2 + v_1v_3 + v_2v_3)\end{aligned}$$

Da die Berechnung über das transformierte Element durchgeführt wurde, muss noch der Multiplikator $\frac{1}{h_1 h_2}$ eingefügt werden.

Berechnen wir nun

$$L_{i,j} = \sum_{\tilde{E} \in E_k} \nabla T_i \nabla T_j \int_{\tilde{E}} \left(v^2 + \epsilon\right) \mathrm{d}x$$

T_i ist nur auf dem Gitterpunkt i 1 und sonst 0. Das heißt genauer, dass T_i nur auf sechs Dreiecken ungleich 0 ist. Um das Integral zu bestimmen braucht man also maximal sechs Dreiecke. Falls der Gitterpunkt am Rand liegen sollte, betrachtet man nur drei Dreiecke, an den Ecken entweder ein oder zwei Dreiecke.

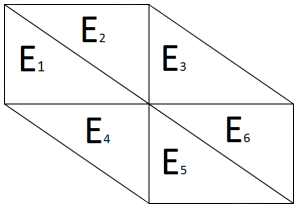


Abbildung 3.1.: Triangulierung im Inneren

Jetzt können wir L_{ij} für festes i, j berechnen. Falls i und j nicht adjazent sind, ist

$L_{ij} = 0$, da $T_i T_j = 0$ gilt. Seien nun T_I und T_j adjazent. Hier haben wir vier Fälle:

- i liegt auf j, also $j = i$
- j liegt rechts neben i also $j = i + 1$
- j liegt direkt unter i, $j = i + n + 1$
- j liegt schräg unter i, also $j = i + n + 2$

Betrachten wir für die einzelnen Berechnungen 3.1. T_i ist immer der Mittelpunkt dieser Zeichnung, T_j ist entsprechend des jeweiligen j positioniert. In den Berechnungen stimmen die Nummerierungen der Dreiecke mit den Nummerierungen in der Abbildung 3.1 überein und φ_k, $k \in \{1, 2, 3\}$ entspricht den φ_k in (2.5).

Betrachten wir nun die vier Fälle.

i und j sind gleich

$$\begin{aligned} L_{i,i} &= \sum_{E \in E_k} \nabla T_i \nabla T_i \int_E \left(v^2 + \epsilon\right) \mathrm{d}x \\ &= \nabla\varphi_1 \nabla\varphi_1 \int_{E_1} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_2 \nabla\varphi_2 \int_{E_2} \left(v^2 + \epsilon\right) \mathrm{d}x \\ &+ \nabla\varphi_0 \nabla\varphi_0 \int_{E_3} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_0 \nabla\varphi_0 \int_{E_4} \left(v^2 + \epsilon\right) \mathrm{d}x \\ &+ \nabla\varphi_2 \nabla\varphi_2 \int_{E_5} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_1 \nabla\varphi_1 \int_{E_6} \left(v^2 + \epsilon\right) \mathrm{d}x \end{aligned}$$

j liegt rechts neben i

$$\begin{aligned} L_{i,i+1} &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+1} \int_E \left(v^2 + \epsilon\right) \mathrm{d}x \\ &= \nabla\varphi_0 \nabla\varphi_1 \int_{E_3} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_0 \nabla\varphi_1 \int_{E_6} \left(v^2 + \epsilon\right) \mathrm{d}x \end{aligned}$$

j **liegt unter** i

$$\begin{aligned} L_{i,i+1+n} &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+1+n} \int_E \left(v^2 + \epsilon\right) \mathrm{d}x \\ &= \nabla\varphi_0 \nabla\varphi_1 \int_{E_4} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_0 \nabla\varphi_1 \int_{E_5} \left(v^2 + \epsilon\right) \mathrm{d}x \end{aligned}$$

j **liegt schräg unter** i

$$\begin{aligned} L_{i,i+2+n} &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+2+n} \int_E \left(v^2 + \epsilon\right) \mathrm{d}x \\ &= \nabla\varphi_1 \nabla\varphi_2 \int_{E_5} \left(v^2 + \epsilon\right) \mathrm{d}x + \nabla\varphi_1 \nabla\varphi_2 \int_{E_6} \left(v^2 + \epsilon\right) \mathrm{d}x \\ &= 0 \end{aligned}$$

Zusammenfassung Mit diesen Werten können wir nun die Matrix $\left(\int_\Omega (v^2+\epsilon)\nabla T_i \nabla T_j\right)_{ij}$ aufstellen:

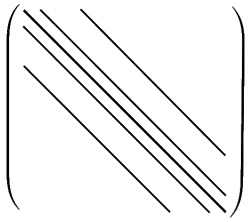


Abbildung 3.2.: Darstellung der Matrix $\left(\int_\Omega (v^2+\epsilon)\nabla T_i \nabla T_j\right)_{ij}$

Dabei sind auf der Diagonalen die Einträge $L_{i,i}$, auf der Nebendiagonalen die Einträge $L_{i,i+1}$ und auf der anderen Diagonale die Einträge $L_{i,i+n+1}$. Wir haben bis jetzt immer nur über das Referenzdreieck integriert. Da wir aber eigentlich über die transformierten Dreiecke integrieren, müssen wir zu der Matrix $1/(h_1 h_2)$ multiplizieren.

3.2.3. Aggregation

Nun haben wir die Matrix L und den Vektor u_0^h gegeben, um das Gleichungssystem $Lu^h = Lu_0^h$ zu berechnen. Allerdings haben wir noch nicht eingebracht, dass auf $\Gamma_1 \cup \Gamma_2$ $u = 0$ gilt. Eigentlich würden wir das im Vektor u mit aufnehmen, also die Zeilen 0

setzten, die den Rand repräsentieren und dann das Gleichungssystem lösen. Dies geht numerisch jedoch nicht so einfach. Die Information muss in L und in Lu_0^h codiert sein. Dazu setzt man die Zeilen in L 0, die zum Rand gehören. Die zugehörigen Diagonaleinträge werden 1 gesetzt. Diese Matrix nennen wir $\tilde{L}$. Die zugehörige Einträge in Lu_0^h, also die Einträge, die in der selben Zeile sind, setzt man 0. Den neuen Vektor nennen wir $\tilde{Lu_0^h}$ Dadurch erhält man, dass u^h an dieser Stelle 0 wird.

Damit haben wir beide Seiten diskretisiert und können das Gleichungssystem implementieren. Wir wollen

$$\frac{1}{h_1 h_2}\tilde{L}u = \frac{1}{h_1 h_2}\tilde{Lu_0^h} \Leftrightarrow \tilde{L}u = \tilde{Lu_0^h}$$

berechnen. Der Code dazu hat folgende Form

1. Berechne Matrix $\tilde{L}$
2. Berechne Vektor $\tilde{Lu_0^h}$
3. $u = \tilde{Lu_0^h} \backslash \tilde{L}$

Algorithm 4: Berechnung von u

Da sowohl $\tilde{L}$ als auch $\tilde{Lu_0^h}$ aus fast nur Nullen besteht, verwende ich in Matlab Sparse Matrizen. Dies führt zu einer wesentlich besseren Laufzeit.

3.3. Optimierung nach v

Bei der Optimierung nach v geht es um die Fortsetzung des Risses. Für dieses Optimierungsproblem mit Ungleichungsnebenbedingung sicheren wir zunächst die Existenz und Eindeutigkeit der Lösung. Danach stellen wir die Optimalitätsbedingungen auf. Das resultiert in ein Karush Kuhn Tucker System. Dieses wollen wir mittels semidifferenzierbarer Newton Methode lösen. Dazu müssen zunächst alle Funktionen der KKT Systems differenziert und danach diskretisiert werden. Dies geschieht wieder mit Finiten Elementen. Am Schluss führe beide Optimierungsprobleme zu einem Verfahren zusammen.

3.3.1. Analytische Betrachtung

Die Optimierung nach v hat folgende Form:

$$\min_{v\in H^1(\Omega)} \int_\Omega \left(v^2+\epsilon_1\right)|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \frac{1}{\epsilon_2}(1-v)^2\right)\mathrm{d}x \tag{3.3}$$
$$\text{s.d. } 0 \le v \le v_0$$

Das lässt sich allgemein als Optimierungsproblem mit Ungleichungsnebenbedingungen darstellen

$$\min_{w\in W} J(w) \quad \text{s.d.} \quad w \in C$$

wobei W ein Banachraum, $J : W \to \mathbb{R}$ G-diffbar und $C \subset W$.

In diesem Fall bedeutet das also, dass

$$\begin{aligned} J: \quad H^1(\Omega) &\to \mathbb{R} \\ v &\mapsto \int_\Omega \left(v^2+\epsilon_1\right)|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \tfrac{1}{\epsilon_2}(1-v)^2\right)\mathrm{d}x \\ C := \{v \in H^1(\Omega) | 0 \le v \le v_0\} & \end{aligned}$$

Zunächst wollen wir die Existenz und Eindeutigkeit der Lösung zeigen. Dafür brauchen wir, dass J Gâteaux differenzierbar ist.

Lemma 3.3.1. *J ist Gâteaux-Differenzierbar mit*

$$\begin{aligned} J'(v): \quad \overline{H^1}(\Omega) &\to \mathbb{R} \\ \psi_1 &\mapsto \left(2v|\nabla(u)|^2 - \nu\left(\epsilon_2 2\Delta v - \tfrac{2}{\epsilon_2}(1-v), \psi_1\right)\right)_{L^2(\Omega)} + (2\epsilon_2\nu\nabla v\nu, \psi_1)_{L^2(\partial\Omega)} \end{aligned}$$

Beweis. Zunächst müssen wir die Richtungsableitung bestimmen.

$$\begin{aligned}
\partial J(v)(\psi_1) &= \lim_{t\to 0}\tfrac{1}{t}\Big(\quad J(v+t\psi_1) - J(v)\Big) \\
&= \lim_{t\to 0}\tfrac{1}{t}\Big(\int_\Omega \left((v+t\psi_1)^2+\epsilon_1\right)|\nabla(u)|^2 + \nu\left(\epsilon_2|\nabla(v+t\psi_1)|^2 + \tfrac{1}{\epsilon_2}\left(1-(v+t\psi_1)\right)^2\right) \\
&\qquad - \int_\Omega \left(v^2+\epsilon_1\right)|\nabla u|^2 \qquad\qquad + \nu\left(\epsilon_2|\nabla v|^2 \qquad\quad + \tfrac{1}{\epsilon_2}\left(1-v\right)^2\right)\Big) \\
&= \lim_{t\to 0}\tfrac{1}{t}\Big(\int_\Omega \left(\left((v+t\psi_1)^2+\epsilon_1\right) - \left(v^2+\epsilon_1\right)\right)|\nabla(u)|^2 \\
&\qquad\qquad +\nu\left(\epsilon_2\left(|\nabla(v+t\psi_1)|^2 - |\nabla v|^2\right)\right. \\
&\qquad\qquad \left. +\tfrac{1}{\epsilon_2}\left((1-v-t\psi_1)^2-(1-v)^2\right)\right)\Big) \\
&= \lim_{t\to 0}\tfrac{1}{t}\Big(\int_\Omega \left(v^2+2vt\psi_1+t^2\psi_1^2-v^2\right)|\nabla u|^2 \\
&\qquad\qquad +\nu\left(\epsilon_2\left(|\nabla v|^2 + 2t\nabla v\nabla\psi_1 + t^2|\nabla\psi_1|^2 - |\nabla v|^2\right)\right. \\
&\qquad\qquad \left. +\tfrac{1}{\epsilon_2}\left((1-v)^2 - 2(1-v)t\psi_1 + t^2\psi_1^2 - (1-v)^2\right)\right)\Big) \\
&= \lim_{t\to 0}\Big(\int_\Omega \left(2v\psi_1 + t\psi_1^2\right)|\nabla u|^2 + \nu\left(\epsilon_2\left(2\nabla v\nabla\psi_1 + t|\nabla\psi_1|^2\right)\right. \\
&\qquad\qquad \left. -\tfrac{1}{\epsilon_2}\left(2(1-v)\psi_1 + t\psi_1^2\right)\right)\mathrm{d}x\Big) \\
&= \int_\Omega 2\psi_1 v|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla v\nabla\psi_1 - \tfrac{2}{\epsilon_2}(1-v)\psi_1\right)\mathrm{d}x \\
&= \int_\Omega 2v|\nabla u|^2\psi_1 - \nu\left(\epsilon_2 2\Delta v\psi_1 - \tfrac{2}{\epsilon_2}(1-v)\psi_1\right)\mathrm{d}x + \int_{\partial\Omega} 2\nu\epsilon_2\nabla v\nu\psi_1\,\mathrm{d}x
\end{aligned}$$

Damit es auch eine Gâteaux Ableitung ist, muss sie beschränkt und linear sein. Dies ist einfach zu sehen. □

Theorem 3.3.2. *Das Problem* (3.3) *besitzt genau eine Lösung, falls* v_0 *stetig ist.*

Beweis. Wir wollen 2.1.6 anwenden. Zunächst müssen wir alle Voraussetzungen prüfen.

1. $W = H^1(\Omega)$ ist ein Hilbertraum, also auch ein reflexiver Banachraum.

2. Nun muss gezeigt werden, dass C nichtleer, abgeschlossen und konvex ist. C ist nichtleer, da $0 \in C$.

   Sei v_n eine konvergente Folge in C. Dann gilt $0 \leq v_n \leq v_0 \ \forall n \in \mathbb{N}$. Es gilt auch $0 \leq \lim_{n\to\infty} u_n \leq v_0$. Also ist C abgeschlossen.

   Um Konvexität von C zu zeigen, sei $0 < \lambda < 1$ und $v, w \in C$. Dann gilt $0 \leq \lambda v + (1-\lambda)w$, da $\lambda > 0$. Außerdem gilt $\lambda v + (1-\lambda)w \leq \lambda v_0 + (1-\lambda)v_0 = v_0$. Also ist jede Konvexkombination in C enthalten, C ist konvex.

3. J ist strikt konvex. Der Beweis dazu kann durch einfaches nachrechnen geführt werden. Für Stetigkeit gilt dasselbe.

4. J ist Gâteaux differenzierbar nach (3.3.1)

5. Sei $w \in C$ mit $\|v\|_{H^1(\Omega)} \to \infty$. Dann gilt

$$\begin{aligned}
J(v) &= \int_\Omega (v^2+\epsilon_1)\,|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \tfrac{1}{\epsilon_2}(1-v)^2\right)\mathrm{d}x \\
&= \int_\Omega v^2|\nabla u|^2 + \epsilon_1|\nabla u|^2 + \nu\left(\epsilon_2|\nabla v|^2 + \tfrac{1}{\epsilon_2}(1-2v+v^2)\right)\mathrm{d}x \\
&= \int_\Omega v^2|\nabla u|^2 - \nu\left(\tfrac{2}{\epsilon_2}v + \tfrac{1}{\epsilon_2}v^2\right)\mathrm{d}x + \int_\Omega \epsilon_1|\nabla u|^2 + \nu\tfrac{1}{\epsilon_2}\,\mathrm{d}x + \int_\Omega \nu\epsilon_2|\nabla v|^2\,\mathrm{d}x \\
&\le \int_\Omega v^2\left(|\nabla u|^2 + \nu\tfrac{1}{\epsilon_2}\right)\mathrm{d}x + c + \epsilon_2\|\nabla v\|^2_{L^2(\Omega)} \\
&\le c'\|v\|^2_{L^2(\Omega)} + c + \epsilon_2\|\nabla v\|^2_{L^2(\Omega)} \\
&\le c''\left(\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)}\right) + c \\
&\le c''\|v\|^2_{H^1(\Omega)} + c \\
&\to \infty
\end{aligned}$$

   mit $c, c', c'' > 0$ passende Konstanten.

Alle Voraussetzungen aus 2.1.6 sind erfüllt, also existiert genau eine Lösung des Optimierungsproblems. □

Nachdem wir nun wissen, dass die Lösung existiert und eindeutig ist, wollen wir das Minimum finden. Dazu brauchen wir Optimalitätsbedingungen. Diese stellt das folgende Theorem auf

Theorem 3.3.3 (Optimalitätsbedingungen)**.** *Sei* $a := \inf\{J(w)|H(w) \le_P 0\}$. *Dann gilt:*

$$a = \inf_{v\in H^1(\Omega)} J(v) + \langle H(v), \begin{pmatrix}\lambda\\ \mu\end{pmatrix}\rangle_{H^1(\Omega),H^{-1}(\Omega)}$$

mit

$$\begin{aligned}
H : H^1(\Omega) &\to H^1(\Omega) \\
v &\mapsto \begin{pmatrix} -v \\ v - v_0 \end{pmatrix}
\end{aligned}$$

Beweis. Die Bedingungen aus 2.1.7 müssen gelten Sei $P := \{(v,w) \in H^1(\Omega)\times H^1(\Omega) | v \ge 0 \text{ und } w \ge 0\} \subset H^1(\Omega)\times H^1(\Omega)$. $\mathring{P} \neq \emptyset$, da $H^1(\Omega)$ nur stetige Funktionen enthält. Also ist P ein positiver Kegel.

$J : H^1(\Omega) \to \mathbb{R}$ sei wie oben definiert. H ist linear, also konvex.

Das Bild von J enthält ein $\hat{v}$, sodass $H(\hat{v}) <_P 0$ gilt, da es ein $v \in H^1(\Omega)$ geben muss, das echt zwischen 0 und v_0 liegt.

Außerdem ist $a := \inf\{J(w) | H(w) \leq_P 0\} < \infty$, da J stetig und beschränkt ist.

Also kann 2.1.7 angewendet werden. Damit existiert $(\mu, \lambda) \in H^{-1}(\Omega) \times H^{-1}(\Omega)$ mit $(\mu, \lambda) \geq 0$ Komponentenweise, sodass

$$a = \inf_{v \in H^1(\Omega)} J(v) + \langle H(v), \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \rangle_{H^1(\Omega), H^{-1}(\Omega)}$$

□

Damit müssen wir nur noch das Minimum der Lagrangefunktion suchen. Dies funktioniert, indem wir die Ableitung bestimmen und 0 setzen. Wir leiten die Lagragefunktion ab und erhalten $\nabla J(v) + \lambda - \mu = 0$. Ausformuliert sieht das so aus

$$\begin{aligned} &2v|\nabla u|^2 - \epsilon_2 2\Delta v - \frac{2}{\epsilon_2}(1 - v) + \lambda - \mu = 0 && \text{auf } \Omega \\ &2\epsilon_2 \nabla v \nu = 0 && \text{auf } \partial\Omega \end{aligned}$$

Nun ist alles gegeben, damit das KKT System aufgestellt werden kann.

$$\begin{aligned} &2\overline{v}|\nabla u|^2 - \epsilon_2 2\Delta\overline{v} - \tfrac{2}{\epsilon_2}(1 - \overline{v}) + \lambda - \mu = 0 \text{ auf } \Omega \\ &2\epsilon_2 \nabla\overline{v}\nu = 0 \text{ auf } \partial\Omega \\ &\overline{v} \geq a \quad \mu \geq 0 \quad \mu\overline{v} = 0 \\ &\overline{v} \leq b \quad \lambda \geq 0 \quad \lambda(v_0 - \overline{v}) = 0 \end{aligned}$$

Die Projektion für die Nebenbedingung lautet nach (2.3):

$$\mu - \lambda = \max\{0, \mu - \lambda + c(\overline{v} - v_0)\} + \min\{0, \mu - \lambda + c\overline{v}\} \forall c > 0$$

Daraus ergibt sich eine starke und schwache Formulierung. Die Starke lautet Suche $v \in H^1$, sodass

$$\begin{aligned} &2\overline{v}|\nabla u|^2 - \epsilon_2 2\Delta\overline{v} - \tfrac{2}{\epsilon_2}(1 - \overline{v}) + \eta = 0 \text{ auf } \Omega \\ &2\epsilon_2 \nabla\overline{v}\nu = 0 \text{ auf } \partial\Omega \\ &\eta = \max\{0, \eta + c(\overline{v} - v_0)\} + \min\{0, \eta + c\overline{v}\} \ \forall c > 0 \end{aligned}$$

Die schwache Formulierung ist dann

$$\int_\Omega 2\psi_1 v|\nabla u|^2 + \epsilon_2 2\nabla v \nabla \psi_1 - \frac{2}{\epsilon_2}(1-v)\psi_1 + \eta\psi_1 \,\mathrm{d}x = 0 \qquad \forall \psi_1 \in H^1(\Omega)$$
$$\int_\Omega \left(\eta - \max\{0, \eta + c(\overline{v} - v_0)\} - \min\{0, \eta + c\overline{v}\}\right)\psi_1 \,\mathrm{d}x = 0 \quad \forall c > 0, \forall \psi_1 \in H^1(\Omega)$$

mit $\eta = \mu - \lambda$

3.3.2. Anwendung auf semidifferenzierbare Newton Methode

Unser Ziel ist es, eine Methode zu finden, wie wir das KKT System lösen können. Betrachten wir also

$$\begin{aligned} G : H^1(\Omega) \times H^1(\Omega) &\to H^{-1}(\Omega)^2 \\ (v, \eta) &\mapsto \begin{pmatrix} \int_\Omega 2\psi_1 v|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla v \nabla \psi_1 - \frac{2}{\epsilon_2}(1-v)\psi_1 + \eta\psi_1\right)\mathrm{d}x \\ \int_\Omega \left(\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}\right)\psi_2 \,\mathrm{d}x \end{pmatrix} \end{aligned}$$

Wir wollen (v, η) finden, sodass $G = 0$. Direkt kann diese Formel nicht gelöst werden, da wir um G_2 zu berechnen (v, η) benötigen. Dies ist nicht gegeben. Also lösen wir das Problem mit einer Newton Methode. Für jeden Iterationsschritt ist (v, η) durch den vorherigen gegeben. Dafür brauchen wir aber die Ableitung von G. Da G_2 offensichtlich keine Gâteaux-Ableitung hat, brauchen wir das Semidifferenzial. Also wird die semidifferenzierbare Newton Methode gebraucht.

Sehen wir uns zunächst die $\partial G_1(v, \eta)(h)$ an. Es gilt:
Theorem 3.3.4. *$G_1(v, \eta)$ ist semidifferenzierbar mit*

$$\partial G_{1v}(v, \eta)(\phi_1) = \int_\Omega 2\psi_1\phi_1|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla\phi_1\nabla\psi_1 + \frac{2}{\epsilon_2}\phi_1\psi_1\right)\mathrm{d}x$$
$$\partial G_{1\eta}(v, \eta)(\phi_2) = \int_\Omega \nu\phi_2\psi_1 \,\mathrm{d}x$$

falls u fest gewählt ist, oder $u \in L^\infty$

Beweis. Nach **??** ist G_1 ∂G_1 semidifferenzierbar, falls G_1 stetig Fréchet differenzierbar ist. Bestimmen wir zunächst die Richtungsableitung in Richtung ϕ_1 bzw ϕ_2. Diese ist

gegeben durch

$$\partial G_{1v}(v,\eta)(\psi_1,\phi_1) = \int_\Omega 2\psi_1\phi_1|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla\phi_1\nabla\psi_1 + \frac{2}{\epsilon_2}\phi_1\psi_1\right)\mathrm{d}x$$
$$G_{1\eta}(v,\eta)(\phi_2) = \int_\Omega \nu\phi_2\psi_1\,\mathrm{d}x$$

Die Berechnung wird hier nicht weiter ausgeführt. Die Richtungsableitungen müssen linear und beschränkt sein. Linearität ist einfach zu sehen. Für Beschränktheit von $\partial G_{1v}(v,\eta)(\psi_1,\phi_1)$ gilt:

$$\begin{aligned}\|\partial G_{1v}(v,\eta)(\psi_1,\phi_1)\|_{H^{-1}} \leq = &\int_\Omega 2\psi_1\phi_1|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla\phi_1\nabla\psi_1 + \frac{2}{\epsilon_2}\phi_1\psi_1\right)\mathrm{d}x\\ &\leq 2\|\nabla u\|^2|(\psi_1,\phi_1)_{H^1}| + \nu\max\{2\epsilon_2,\frac{2}{\epsilon_2}\}|(\psi_1,\phi_1)_{H^1}|\\ &\leq \left(C + \tilde{C}\|\nabla u\|^2\right)\|\psi_1\|_{H^1}\|\phi_1\|_{H^1}\end{aligned}$$

Da $u \in L^\infty$, ist $\|\nabla u\|^2$ beschränkt. Damit ist G_{1v} beschränkt. Für Fréchet Differenzierbarkeit muss eine Abschätzung überprüft werden:

$$\begin{aligned}&\|G_1(v+h,\eta) - G_1(v,\eta) - \partial G_{1v}(v,\eta)(h)\|\\ &= \|\int_\Omega 2\psi_1(v+h)|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla(v+h)\nabla\psi_1 - \frac{2}{\epsilon_2}(1-(v+h))\psi_1 + \eta\psi_1\right)\mathrm{d}x\\ &\quad - \int_\Omega 2\psi_1 v|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla v\nabla\psi_1 - \frac{2}{\epsilon_2}(1-v)\psi_1 + \eta\psi_1\right)\mathrm{d}x\\ &\quad - \int_\Omega 2\psi_1 h|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla h\nabla\psi_1 + \frac{2}{\epsilon_2}h\psi_1\right)\mathrm{d}x\|\\ &= \|\int_\Omega 2\psi_1 h|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla h\nabla\psi_1 + \frac{2}{\epsilon_2}h\psi_1\right)\mathrm{d}x - \int_\Omega 2\psi_1 h|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla h\nabla\psi_1 + \frac{2}{\epsilon_2}h\psi_1\right)\mathrm{d}x\\ &= 0\end{aligned}$$

Da G_{1v} beschränkt und linear ist, ist G_{1v} auch stetig. □

Für das Semidifferenzial von G_2 beweisen wir zunächst ein Lemma

Lemma 3.3.5. *Betrachte $f : H^1(\Omega)^2 \to H^{-1}(\Omega)$ mit*

$$(\eta, v) \mapsto \eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\} \tag{3.4}$$

Dann ist f semidifferenzierbar mit

$$\frac{\partial f}{\partial \eta} = \begin{cases} \{0\} & \textit{falls } -c(v-v_0) < \eta \textit{ oder } \eta < -cv \\ \{1\} & \textit{falls } -cv < \eta < -c(v-v_0) \\ [0,1] & \textit{falls } -c(v-v_0) = \eta \textit{ oder } \eta = -cv \end{cases}$$

und

$$\frac{\partial f}{\partial v} = \begin{cases} \{-c\} & \textit{falls } -c(v-v_0) < \eta \textit{ oder } \eta < -cv \\ \{0\} & \textit{falls } -cv < \eta < -c(v-v_0) \\ [-c,0] & \textit{falls } -c(v-v_0) = \eta \textit{ oder } \eta = -cv \end{cases}$$

Beweis. f kann in einer anderen Form dargestellt werden

$$f(v,\eta) = \begin{cases} -c(v-v_0) & \text{falls } -c(v-v_0) \leq \eta \\ \eta & \text{falls } -cv < \eta < -c(v-v_0) \\ -cv & \text{falls } \eta \leq -cv \end{cases}$$

Die Äquivalenz von diese Form von f und (3.4), kann einfach nachgerechnet werden. Betrachten wir zunächst die Ableitung nach η. Es reicht, die Semidifferenzierbarkeit der einzelnen Abschnitte zu betrachten. Falls jeder Abschnitt semidifferenzierbar ist und die Übergänge auch, so ist f semidifferenzierbar.

Sei dazu $-c(v-v_0) < \eta$ oder $\eta < -cv$. Mit **??** gilt, dass, falls f stetig Fréchet differenzierbar ist, f ∂f semidifferenzierbar. Um Fréchet Differenzierbarkeit zu zeigen, bestimmen wir zunächst die Richtungsableitung. Diese ist offensichtlich 0. Dadurch folgt sofort die Fréchet differenzierbarkeit.

Sei nun $-cv < \eta < -c(v-v_0)$. Durch **??** müssen wir wieder die Fréchet Differenzierbarkeit überprüfen. Offensichtlich ist die Identität Fréchet differenzierbar. Das Differenzial ist hier 1.

Sei $\eta = -c(v-v_0)$. Sei zunächst $d > 0$. Die Abschätzung (**??**) muss gelten. Hier ist $\partial f(\eta + d, v) = \{0\}$ und damit

$$\begin{aligned} &\sup_{M \in \partial f(\eta+d,v)} \|f(\eta+d,v) - f(\eta) - Md\|_{H^{-1}(\Omega)} \\ &\quad = \| - c(v-v_0) + c(v-v_0)\|_{H^{-1}(\Omega)} = 0 = o\left(\|d\|_{H^1(\Omega)}\right) \text{ für } \|d\|_{H^1(\Omega)} \to 0 \end{aligned}$$

Sei nun $d < 0$. Da d nahe an 0 ist, gilt $d > -cv_0$ mit $v_0 > 0$. Es ist $\partial G_2^\eta(\eta + d) = \{1\}$

und damit

$$\begin{aligned}&\sup_{M\in\partial f(\eta+d,v)} \|f(\eta+d,v)-f(\eta,v)-Md\|_{H^{-1}(\Omega)}\\ &\quad= \|-c(v-v_0)+d+c(v-v_0)-d\|_{H^{-1}(\Omega)}=0=o\left(\|d\|_{H^1(\Omega)}\right) \text{ für } \|d\|_{H^1(\Omega)}\to 0\end{aligned}$$

Fehlt nur noch $\eta=-cv$. Sei zunächst $d>0$. Da d nahe an 0 ist, gilt auch $d<cv_0$. Es gilt $\partial f(\eta+d,v)=\{1\}$ und damit

$$\begin{aligned}&\sup_{M\in\partial G_2^\eta(\eta+d)} \|f(\eta+d,v)-f(\eta)-Md\|_{H^{-1}(\Omega)}\\ &\quad= \|-cv+d+cv-d\|_{H^{-1}(\Omega)}=0=o\left(\|d\|_{H^1(\Omega)}\right) \text{ für } \|d\|_{H^1(\Omega)}\to 0\end{aligned}$$

Sei nun $d<0$. Es gilt: Es gilt $\partial G_2^\eta(\eta+d)=\{0\}$ und damit

$$\begin{aligned}&\sup_{M\in\partial f(\eta+d,v)} \|f(\eta+d,v)-f(\eta,v)-Md\|_{H^{-1}(\Omega)}\\ &\quad= \|-cv+cv\|_{H^{-1}(\Omega)}=0=o\left(\|d\|_{H^1(\Omega)}\right) \text{ für } \|d\|_{H^1(\Omega)}\to 0\end{aligned}$$

Damit ist f semidifferenzierbar nach η. Für die Semidifferenzierbarkeit nach v gilt die Gleiche Rechnung. □

Das eigentliche Ziel war es, das Semidifferenzial von G_2 zu finden. Dieses können wir nun tun

Theorem 3.3.6. *$G_2 : H^1(\Omega)^2 \to H^{-1}(\Omega)$ mit*

$$(v,\eta)\mapsto \int_\Omega \left(\eta-\max\{0,\eta+c(v-v_0)\}-\min\{0,\eta+cv\}\right)\psi_2\,\mathrm{d}x$$

ist semidifferenzierbar mit

$$\partial G_{2\eta}(\eta,v)(\psi_2,\phi_2)=\int_\Omega \frac{\partial f}{\partial \eta}\psi_2\phi_2\,\mathrm{d}x$$

$$\partial G_{2v}(\eta,v)(\psi_2,\phi_1)=\int_\Omega \frac{\partial f}{\partial v}\psi_2\phi_1\,\mathrm{d}x$$

Beweis. Es gilt

$$\begin{aligned}
&\int_\Omega (\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\})\, \psi_2 \,\mathrm{d}x \\
&= \int_\Omega \eta\psi_2 \,\mathrm{d}x - \int_\Omega \max\{0, \eta + c(v - v_0)\}\psi_2 \,\mathrm{d}x - \int_\Omega \min\{0, \eta + cv\}\psi_2 \,\mathrm{d}x \\
&= F_1(v, \eta)(\psi_2) - F_2(v, \eta)(\psi_2) - F_3(v, \eta)(\psi_2)
\end{aligned}$$

Wir können nach **??** F_1, F_2, F_3 getrennt ableiten. Betrachten wir die Ableitung nach η. Die Ableitung von F_1 ist einfach

$$\partial F_{1\eta}(\psi_2, \phi_2) = \int_\Omega \psi_2\phi_2 \,\mathrm{d}x$$

F_2 und F_3 lassen sich mithilfe der Kettenregel aus **??** ableiten

$$\begin{aligned}
\partial F_{2\eta}(\psi_2, \phi_2) &= \int_\Omega \frac{\partial}{\partial\eta + c(v - v_0)} \max\{0, \eta + c(v - v_0)\} \frac{\partial\eta}{\partial\eta}\psi_2 \,\mathrm{d}x \\
&= \int_\Omega \frac{\partial f_2}{\partial\eta}\psi_2\phi_2 \,\mathrm{d}x \\
\partial F_{3\eta}(\psi_2, \phi_2) &= \int_\Omega \frac{\partial}{\partial\eta + cv} \min\{0, \eta + cv\} \frac{\partial\eta}{\partial\eta}\psi_2 \,\mathrm{d}x \\
&= \int_\Omega \frac{\partial f_3}{\partial\eta}\psi_2\phi_2 \,\mathrm{d}x
\end{aligned}$$

mit f_1 und f_2 der min bzw. der max Term. Die Ableitungen kann man davon einfach berechnen. Es ergibt sich:

$$\begin{aligned}
\partial G_{2\eta}(\eta, v)(\psi_2, \phi_2) &= \partial F_1(\eta, v)(\psi_2, \phi_2) + \partial F_2(\eta, v)(\psi_2, \phi_2) + \partial F_3(\eta, v)(\psi_2, \phi_2) \\
&= \int_\Omega \psi_2\phi_2 \,\mathrm{d}x + \int_\Omega \frac{\partial f_2}{\partial\eta}\psi_2\phi_2 \,\mathrm{d}x \int_\Omega \frac{\partial f_3}{\partial\eta}\psi_2\phi_2 \,\mathrm{d}x \\
&= \int_\Omega (id + \frac{\partial f_2}{\partial\eta} + \frac{\partial f_3}{\partial\eta})\phi_2\psi_2 \,\mathrm{d}x \\
&= \int_\Omega \frac{\partial f}{\partial\eta}\psi_2\phi_2 \,\mathrm{d}x
\end{aligned}$$

Es fügen sich die Ableitungen von f_2, f_3 und η wieder zu $\frac{\partial f}{\partial\eta}$ zusammen. □

Damit ergibt sich als Ableitung

$$G'(v,\eta) = \begin{pmatrix} G_{1v} & G_{1\eta} \\ G_{2v} & G_{2\eta} \end{pmatrix}$$

Also lautet das Gleichungssystem, das für das Newtonverfahren nach s gelöst werden muss

$$-\begin{pmatrix} G_1 \\ G_2 \end{pmatrix} = \begin{pmatrix} G_{1v} & G_{1\eta} \\ G_{2v} & G_{2\eta} \end{pmatrix} \begin{pmatrix} s^1 \\ s^2 \end{pmatrix}$$

Stellen wir dazu die Newton Methode auf

Data: v^0, η^0 möglichst Nah an der Lösung $\overline{v}, \overline{\eta}$
for $k = 0, 1, \cdots$ **do**
  Erhalte s_1^k *beim lösen von* $-\begin{pmatrix} G_1 \\ G_2 \end{pmatrix} = \begin{pmatrix} G_{1v} & G_{1\eta} \\ G_{2v} & G_{2\eta} \end{pmatrix} \begin{pmatrix} s^1 \\ s^2 \end{pmatrix}$;
  $v^{k+1} = v^k + s_1^k \quad \eta^{k+1} = \eta^k + s_2^k$;
end

Algorithm 5: semidiffbare Newton Methode

Nun wollen wir noch herausfinden, ob die Newton Methode konvergiert.

Theorem 3.3.7 (Konvergenz der Newton Methode)**.** *Der Algorithmus 6 konvergiert.*

Beweis. Wir wollen Theorem **??** anwenden. Dazu muss G stetig sein und ∂G semidifferenzierbar und die Regularitätsannahme muss erfüllt sein. Die Semidifferenzierbarkeit wurde in 3.3.4 schon gezeigt. Fehlt noch die Reguaritätsannahme. Dazu muss gezeigt werden, dass

$$\exists C > 0, \quad \exists \delta > 0 : \|M^{-1}\|_{X \to Y} \leq C \quad \forall M \in \partial G(x) \quad \forall x \in X, \quad \|x - \overline{x}\|_X < \delta$$

mit

$$M := \begin{pmatrix} \partial G_{1v} & \partial G_{1\eta} \\ \partial G_{2v} & \partial G_{2\eta} \end{pmatrix}$$

□

3.3.3. Numerische Betrachtung

Alle Funktionen aus dem Newtonsystem müssen numerisch dargestellt werden.

Für die Diskretisierung wird dasselbe Gitter und die Selben Elemente genommen wie bei der Optimierung nach u. Auch hier werden wir wieder mit dem Galerkin Ansatz arbeiten d.h.

$$v = \sum_{i=1}^{k} v_i^h T_i \qquad \eta = \sum_{i=1}^{k} \eta_i^h T_i \qquad v_0 = \sum_{i=1}^{k} {v_0}_i^h T_i$$

wobei die T_i wieder die globalen Formfunktionen sind.

Da u gegeben ist, ist u ein Vektor mit den Auswertungen an den Ecken der Dreiecke. Die Darstellung ist die gleiche wie in (2.7) und (2.7). Also gilt

$$|\nabla u|^2 = (u_{31} - u_{21})^2 + (u_{11} - u_{21})^2 + (u_{32} - u_{22})^2 + (u_{12} - u_{22})^2 =: u^{dis}$$

Numerische Darstellung von G_1

$$G_1(v,\eta) = \int_\Omega 2\psi_1 v|\nabla u|^2 + \nu\left(\epsilon_2 2\nabla v\nabla\psi_1 - \frac{2}{\epsilon_2}(1-v)\psi_1 +\right)\eta\psi_1 \,\mathrm{d}x$$

wird nun diskretisiert:

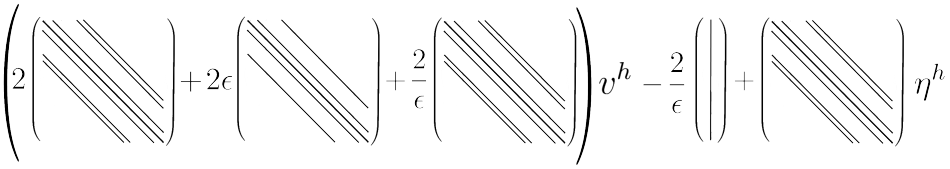
$$\begin{aligned}
&\int_\Omega 2\psi_1 v|\nabla u|^2 + \nu\left(2\epsilon_2\nabla v\nabla\psi_1 - \frac{2}{\epsilon_2}(1-v)\psi_1\right) + \eta\psi_1 \,\mathrm{d}x \\
&= \int_\Omega 2T_j \sum_{i=1}^{k} v_i^h T_i u^{dis} + \nu\left(\epsilon_2 2\sum_{i=1}^{k} v_i^h \nabla T_i \nabla T_j - \frac{2}{\epsilon_2}(1 - \sum_{i=1}^{k} v_i^h T_i)T_j\right) + \sum_{i=1}^{k}\eta_i^h T_i T_j \,\mathrm{d}x \\
&= \sum_{i=1}^{k} v_i^h \left(2\int_\Omega u^{dis} T_i T_j \,\mathrm{d}x + 2\epsilon_2\nu\int_\Omega \nabla T_i \nabla T_j \,\mathrm{d}x + \nu\frac{2}{\epsilon_2}\int_\Omega T_i T_j \,\mathrm{d}x\right) \\
&- \nu\frac{2}{\epsilon_2}\sum_{i=1}^{k}\int_\Omega T_j \,\mathrm{d}x + \sum_{i=1}^{k}\eta_i^h \int_\Omega T_i T_j \,\mathrm{d}x \\
&= (2A + \nu 2\epsilon_2 B + \nu\frac{2}{\epsilon_2}D)v^h - \nu\frac{2}{\epsilon_2}D * e + D\eta^h
\end{aligned}$$

mit $e := (1,\cdots,1)^T v^h := (v_1^h,\cdots v_k^h)^T$, $\eta^h := (\eta_1^h,\cdots \eta_k^h)^T$, $A_{ij} = \int_\Omega u^{dis} T_i T_j \,\mathrm{d}x$, $B_{ij} :=$

$\int_\Omega \nabla T_i \nabla T_j \,\mathrm{d}x$, $c_j := \int_\Omega T_j \,\mathrm{d}x$ und $D_{ij} := \int_\Omega T_i T_j \,\mathrm{d}x$.

Die Berechnung der Matrizen A, B, D und des Vektors c erfolgt im Anhang A.1. Es ergibt sich

$$(2A + 2\nu\epsilon_2 B + \frac{2\nu}{\epsilon_2} D)v^h - \frac{2\nu}{\epsilon_2} De + D\eta^h =$$



wobei bei A auf der Hauptdiagonalen $\frac{1}{12}\left(\sum_{i=1}^{6} u_{E_i}^{dis}\right)$, auf der Nebendiagonalen $\frac{1}{24}\left(u_{E_3}^{dis} + u_{E_6}^{dis}\right)$, auf der zweiten Nebendiagonalen $\frac{1}{24}\left(u_{E_4}^{dis} + u_{E_5}^{dis}\right)$ und auf der dritten Nebendiagonalen $\frac{1}{24}\left(u_{E_5}^{dis} + u_{E_6}^{dis}\right)$ steht.

Bei B steht auf der Hauptdiagonalen 4, auf der Nebendiagonalen und der zweiten Nebendiagonalen -1.

Bei D steht auf der Hauptdiagonalen $\frac{1}{2}$, auf der ersten, zweiten und dritten Nebendiagonalen $\frac{1}{12}$.

Durch die noch ausstehende Transformation der Dreiecke, muss der gesamte Term mit $1/h_1 h_2$ multipliziert werden.

Numerische Darstellung von G_2

$$\int_\Omega \left(\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}\right) \psi_2 \,\mathrm{d}x$$

Um dieses Funktional numerisch darzustellen, benutzen wir den Galerkin Ansatz mit

$$v = \sum_{i=1}^{k} v_i^h T_i(x,y) \qquad \eta = \sum_{i=1}^{k} \eta_i^h T_i(x,y) \qquad v_0 = \sum_{i=1}^{k} v_{0_i}^h T_i(x,y)$$

Daraus ergibt sich:

$$
\begin{aligned}
&\int_\Omega \left(\eta - \max\left\{0, \eta + c(v - v_0)\right\} - \min\left\{0, \eta + cv\right\}\right) \psi_2 \,\mathrm{d}x \\
&= \int_\Omega \left(\sum_{i=1}^{k} \eta_i^h T_i - \max\left\{ 0, \sum_{i=1}^{k} \eta_i^h T_i + c(\sum_{i=1}^{k} v_i^h T_i - \sum_{i=1}^{k} {v_0}_i^h T_i) \right\} \right. \\
&\qquad \left. - \min\left\{ 0, \sum_{i=1}^{k} \eta_i^h T_i + c \sum_{i=1}^{k} v_i^h T_i \right\} \right) T_j \,\mathrm{d}x \\
&= \int_\Omega \left(\sum_{i=1}^{k} \eta_i^h T_i - \sum_{i=1}^{k} \max\left\{0, \eta_i^h + c(v_i^h - {v_0}_i^h)\right\} T_i \right. \\
&\qquad \left. - \sum_{i=1}^{k} \min\left\{0, \eta_i^h + c v_i^h\right\} T_i \right) T_j \,\mathrm{d}x \\
&= \int_\Omega \sum_{i=1}^{k} \left(\eta_i^h - \max\left\{0, \eta_i^h + c(v_i^h - {v_0}_i^h)\right\} - \min\left\{0, \eta_i^h + c v_i^h\right\}\right) T_i T_j \,\mathrm{d}x \\
&= \left(\sum_{i=1}^{k} \eta_i^h - \max\left\{0, \eta_i^h + c(v_i^h - {v_0}_i^h)\right\} - \min\left\{0, \eta_i^h + c v_i^h\right\} \right) \int_\Omega T_i T_j \,\mathrm{d}x \\
&= D w_{v\eta}
\end{aligned}
$$

mit D aus der numerischen Darstellung von G_1 und
$(w_{v\eta})_i := \eta_i^h - \max\left\{0, \eta_i^h + c(v_i^h - {v_0}_i^h)\right\} - \min\left\{0, \eta_i^h + c v_i^h\right\}$. Die Summe und T_i darf aus dem max bzw. min herausgezogen werden, da T_i immer nur an einem Punkt ungleich 0 ist. Dadurch kommt niemals zustande, dass mehr als ein Term der Summe ungleich 0 ist. $w_{v\eta}$ kann auch explizit dargestellt werden:

$$
(w_{v\eta})_i = \begin{cases} -c(v_i^h - {v_0}_i^h) & \text{falls } -c(v_i^h - {v_0}_i^h) \leq \eta_i^h \\ \eta_i^h & \text{falls } -cv_i^h < \eta < -c(v_i^h - {v_0}_i^h) \\ -cv_i^h & \text{falls } \eta_i^h \leq -cv_i^h \end{cases} .
$$

Auch hier muss das Integral wieder transformiert werden, wodurch der Faktor $/h_1 h_2$ multipliziert wird.

Numerische Darstellung von G_{1v}

$$G_{1v}(v,\eta) = \int_\Omega 2\psi_1\phi_1|\nabla u|^2 + \nu 2\epsilon_2 \nabla\phi_1\nabla\psi_1 + \frac{2\nu}{\epsilon_2}\phi_1\psi_1 \,\mathrm{d}x$$

wird nun diskretisieren:

$$\begin{aligned}
&\int_\Omega 2\psi_1\phi_1|\nabla u|^2 + 2\nu\epsilon_2\nabla\phi_1\nabla\psi_1 + \frac{2\nu}{\epsilon_2}\phi_1\psi_1 \,\mathrm{d}x \\
&= \int_\Omega 2T_jT_iu^{dis} + 2\nu\epsilon_2\nabla T_i\nabla T_j + \frac{2\nu}{\epsilon_2}T_iT_j \,\mathrm{d}x \\
&= 2A + 2\nu\epsilon_2 B + \frac{2\nu}{\epsilon_2}D
\end{aligned}$$

Wir benutzen die gleichen Notationen, wie bei der numerischen Darstellung von G_1 und die gleiche Transformation.

Numerische Darstellung von $G_{1\eta}$

$$G_{1\eta}(v,\eta) = \int_\Omega \phi_2\psi_1 \,\mathrm{d}x$$

wird nun diskretisieren:

$$\int_\Omega \phi_2\psi_1 \,\mathrm{d}x = \int_\Omega T_jT_i = D$$

Auch hier muss wieder transformiert werden.

Numerische Darstellung von G_{2v}

Es soll

$$\partial G_{2v}(\eta, v)(\psi_2, \phi_1) = \int_\Omega \frac{\partial f}{\partial v} \psi_2 \phi_1 \, \mathrm{d}x$$

mit

$$\frac{\partial f}{\partial v} = \begin{cases} \{-c\} & \text{falls } -c(v - v_0) < \eta \text{ oder } \eta < -cv \\ \{0\} & \text{falls } -cv < \eta < -c(v - v_0) \\ [-c, 0] & \text{falls } -c(v - v_0) = \eta \text{ oder } \eta = -cv \end{cases} .$$

numerisch dargestellt werden. Statt $\frac{\partial f}{\partial v}$ implementieren wir eine Vereinfachung, die nicht mehr Mengenwertig ist. Dazu wählen wir statt $[-c, 0]$ einen Punkt aus dem Intervall z.B. $-c/2$. Nun kann $\frac{\partial f}{\partial v}$ diskretisiert werden zu f^h. Dies ist einfach die Funktion ausgewertet an den Gitterpunkten.

Nun wird $\partial G_{2v}(\eta, v)(\psi_2, \phi_1)$ diskretisiert. Hier wird wie immer ψ_2, ϕ_1 durch die globalen Formfunktionen T_i ersetzt und Ω durch die Vereinigung aller Dreiecke. Nun kann für jedes Dreieck $\int_E \frac{\partial f}{\partial v} T_i T_j \, \mathrm{d}x$ berechnet werden. Dies erfolgt im Anhang A.1.

Wir erhalten wieder eine Matrix folgender Form: Jeder Eintrag, der nicht 0 ist, besteht

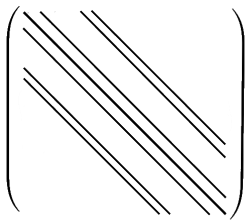


Abbildung 3.3.: Darstellung der Matrix $\left(\int_\Omega (v^2 + \epsilon) \nabla T_i \nabla T_j\right)_{ij}$

aus der Summe von unterschiedlichen Auswertungen der Funktion f auf den Dreiecken, über denen integriert wurde.

Numerische Darstellung von $G_{2\eta}$

Die numerische Darstellung ist genau die gleiche, wie bei G_{2v}, nur dass die Funktionsauswertungen von f andere sind.

3.3.4. Agregation

Nun sind alle Funktionen diskretisiert und das Problem kann implementiert werden. Stellen wir die genaue Newton Methode auf

Data: v^0, η^h (möglichst Nah an der Lösung $(\overline{v}, \overline{\eta})$)
for $k = 0, 1, \cdots$ **do**
Löse das Gleichungssytem $Lu^k = Lu_0^k$ *nach* u^k ;
Erhalte s_k *beim lösen von*

$$-\begin{pmatrix} G_1(v^k, \eta^k) \\ G_2(v^k, \eta^k) \end{pmatrix} = \begin{pmatrix} \partial G_{1v}(v^k, \eta^k)(T_i) & \partial G_{1\eta}(v^k, \eta^k)(T_i) \\ \partial G_{2v}(v^k, \eta^k)(T_i) & \partial G_{2\eta}(v^k, \eta^k)(T_i) \end{pmatrix} \begin{pmatrix} s_1^k \\ s_2^k \end{pmatrix};$$

$v^{k+1} = v^k + s_1^k \quad \eta^{k+1} = \eta^k + s_2^k$;
end

Algorithm 6: semidifferenzierbare Newton Methode

mit

$$\begin{aligned}
L &= \int_\Omega (v^2 + \epsilon_1) \nabla T_i \nabla T_j \, \mathrm{d}x \\
G_1 &= \left(2(\int_\Omega u^{dis} T_i T_j \, \mathrm{d}x)_{ij} + 2\nu\epsilon_2 (\int_\Omega \nabla T_i \nabla T_j \, \mathrm{d}x)_{ij} + \tfrac{2\nu}{\epsilon_2} (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij} \right) v^k \\
&\quad -\tfrac{2\nu}{\epsilon_2} (\int_\Omega T_j \, \mathrm{d}x)_j + (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij} \eta^k \\
G_2 &= (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij} (\eta_i^h - \max\left\{0, \eta_i^h + c(v_i^h - {v_0}_i^h)\right\} - \min\left\{0, \eta_i^h + c v_i^h\right\})_i \\
\partial G_{1v} &= 2(\int_\Omega u^{dis} T_i T_j \, \mathrm{d}x)_{ij} + 2\nu\epsilon_2 (\int_\Omega \nabla T_i \nabla T_j \, \mathrm{d}x)_{ij} + \tfrac{2\nu}{\epsilon_2} (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij} \\
\partial G_{1\eta} &= (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij} \\
\partial G_{2v} &= (\int_\Omega \tfrac{\partial f}{\partial v} T_i T_j)_{ij} \\
\partial G_{2\eta} &= (\int_\Omega \tfrac{\partial f}{\partial \eta} T_i T_j)_{ij} \\
f &= \eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}
\end{aligned}$$

$1/h_1 h_2$ kommt nicht in der Newton Methode vor, da es vorher gekürzt wurde.

3.4. Numerische Resultate

Dieses Kapitel behandelt die Implementation der Optimierung. Zunächst werde ich darauf eingehen, wie man die festen Parameter im Code A.2 wählen muss. Dabei meint feste Parameter, dass deren Veränderung zu keinen brauchbaren Ergebnissen mehr

führt. Hier sind das c, ϵ, η und v_0. Danach werde ich einige Beispiele für unterschiedliche Gitterweiten, verschiedene Risse und Möglichkeiten des Einspannens des Materials betrachten. Zum Schluss gehen wir kurz auf die Konvergenz der Methode ein.

Ganz am Anfang ist zu sagen, dass wir nur Risse betrachten, die mindestens zwei Gitterpunkte breit sind. Wenn ein Riss nur einen Gitterpunkt breit ist, wird v nicht konstant 0.

3.4.1. Justieren der festen Parameter

Zunächst ist es wichtig, dass u_0 an beiden Rändern unterschiedlich ist. Wenn u_0 an beiden Rändern gleich ist, ist die Verschiebung des Materials überall gleich, d.h. das Material verschiebt sich überall gleich, also das nur eine Translation stattgefunden hat. Selbst wenn wir einen Riss einfügen, werden wir diesen nicht sehen.

Als erstes treffen wir eine Wahl für ϵ_1 und ϵ_2. Für ϵ_1 muss gelten: $\epsilon_1 < \frac{1}{k}$, wobei k die Anzahl der Gitterpunkte ist. Für ϵ_2 muss gelten: $\epsilon_2 > \frac{1}{k}$,

Zur Wahl von c: Wählt man c zu groß und betrachtet

$$f(v, \eta) = \eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}$$

werden max und min 0, da $v - v_0$ negativ ist und v positiv ist. Es gilt $f(v, \eta) = \eta$. Nach dem Newtonverfahren ist $-G_2 = \partial G_{2v} s_1^k + \partial G_{2\eta} s_2^k$ mit

$$G_2 = (\int_\Omega T_i T_j \, \mathrm{d}_x)_{ij} \eta^h \qquad \partial G_{2v} = 0 \qquad \partial G_{2\eta} = (\int_\Omega T_i T_j \, \mathrm{d}x)_{ij}$$

Daraus folgt $\eta^h = -s^k$. Außerdem gilt $\eta^{k+1} = \eta^k + s^k = 0$. Dadurch fallen nach dem ersten Iterationsschritt alle Terme mit η weg. Also berücksichtigen wir die vorgegebene Bedingung von $0 < v < v_0$ nicht mehr. Diese soll aber explizit berücksichtigt werden. Also sollte man c nicht zu groß wählen. Auch wenn c zu klein gewählt wird, beachten wir die Schranke nicht d.h. sinnvollerweise sollte man ungefähr $c = 1$ wählen.

Die richtige Wahl von η am Anfang zu finden, ist gar nicht so einfach. η kann positiv und negativ sein, da $\eta = \lambda - \mu$ wobei $\lambda > 0, \mu > 0$ Lagrangemultiplikatoren sind. Experimentell findet man heraus, dass $\eta = 1$ kein Ergebnis liefert, bei $\eta \geq 2$ der Riss sehr steil wird, bei $.4 < \eta < 1$ ist das Papier nicht richtig gerissen, also v wird nie ganz 0 und/ oder das Papier ist nie ganz heile d.h. 1 wird nie angenommen. Bei $\eta = .34$ ist

es fr manch Probleme sehr gut, für andere nicht. Erst im minus Bereich, erhalten wir für alle Probleme gute Werte. Hier muss aber $\eta > -1$ gelten. Woher kommt das? Das Minimum in f ist immer 0, wenn $\eta > 0$ gilt.

Kommen wir zur Wahl von v_0 es sollte Sinn ergeben v_0 konstant 1 zu wählen, aber kleiner sollte es nicht gewählt werden, da sonst nicht zugelassen ist, dass das Gebiet nicht gerissen ist. Dies ist aber leider in der Implementation nicht zu sehen. Bei $v_0 >$ .7 kommen normale Ergebnisse raus, damit ist auch $v > v_0$. Das dürfte nicht sein. (Nochmal überprüfen). Eigentlich sollte v_0 auch recht nahe bei 1 gewählt sein, da $v : \Omega \to [0, 1]$ eigentlich gilt. Jedoch verändern auch große Werte von v_0 nichts am Ergebnis (auch nochmal testen). Dieses Verhalten lässt überlegen, ob es überhaupt Sinnvoll ist, diese Bedingung mit in das Problem zu nehmen. Man könnte probieren, di Rissentstehung auch ohne Ungleichungsnebenbedingungen zu testen.

3.4.2. Beispiele

Mit dem implementierten Algorithmus können wir nun einige Beispiele betrachten. Dazu betrachten wir zwei unterschiedliche Grafiken. In einer Grafik ist die Verschiebung des Gebietes dargestellt, in der anderen, ob und wie dolle ein Riss vorhanden ist. Bei beiden Grafiken ist jeweils das gesamte Gebiet, hier ein rechteckiges Gebiet in 2D auf der x und y Koordinate dargestellt. Die z Koordinate gibt bei der Verschiebung an, wie stark das Gebiet am Punkt (x, y) Verschoben ist und bei dem Riss, wie stark des jeweilige Gitterpunkt eingerissen ist. Ein Wert in der Nähe von 1 bedeutet, dass das Gebiet nicht gerissen ist und 0, dass es vollständig gerissen ist.

Für beide Grafiken betrachte ich je 8 Bilder. Diese geben an, wie viele Iterationsschritte der Newton Algorithmus durchgelaufen ist. Das erste Bild (also links oben) gibt das Ergebnis nach der ersten Iteration an. Das zweite Bild, rechts neben dem ersten Bild, gibt das Ergebnis nach 2 Schritten aus. Dann folgen 5, 10, 20, 50, 100 und zum Schluss, rechts unten 200 Iterationsschritte.

u_0 ist konstant und der Riss ist durchgängig

Beim ersten Beispiel ist das Gebiet an der rechten und linken Seite konstant eingespannt. Hier ist also u am linken Rand immer 1 und am rechten Rand immer 2. Am Anfang wurde ein Riss durch das gesamte Gebiet eingefügt, der parallel zum eingespannten Rand verläuft. Dieser befindet sich genau in der Mitte des Gebietes. Also bei einem

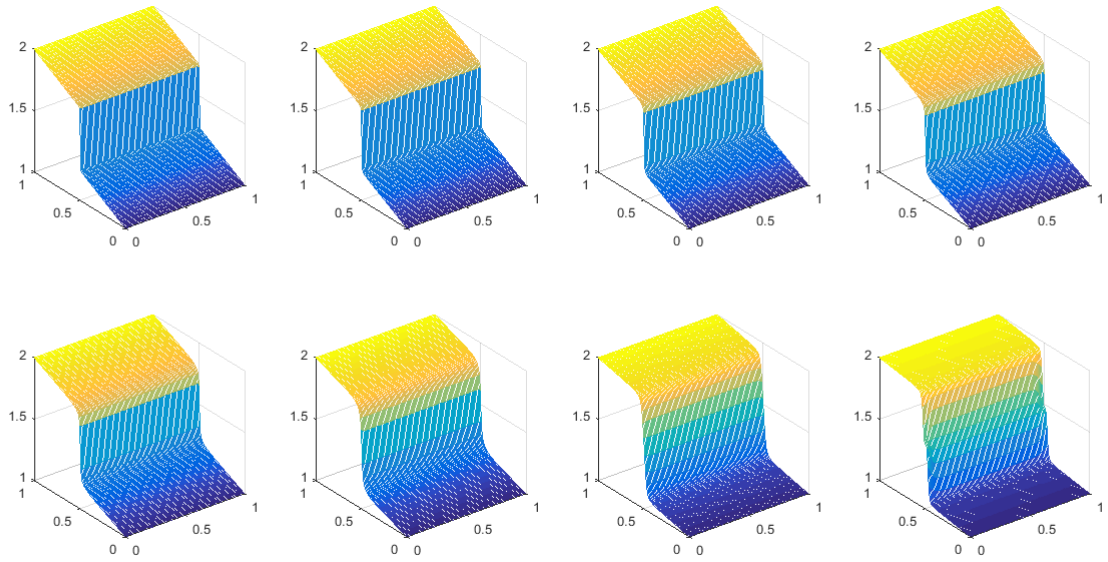


Abbildung 3.4.: Darstellung der Verschiebung des Gebietes bei konstantem u_0, einem Riss in der Mitte und $100x100$ Gitterpunkten

100x100 Gitter ist der Riss zwei Gitterpunkte Breit, also auf $\{(50, i*h) \times (51, i*h) | i \in \{0, \cdots 100\}\times$ mit $h = 1/100$. Dieses sieht man auch in dem ersten Bild von 3.5. Betrachtet man den Riss im Verlauf der Iterationen, sieht man, dass der Riss an der gleichen Stelle bleibt und geglättet wird. Manchmal schwankt das Gebiet, das nicht gerissen ist, zwischen .9 und 1.1. Die Verschiebung des Gebietes 3.4 sieht ähnlich aus. An der Stelle, wo der Riss ist, ist eine Unstetigkeitsstelle zu finden. Diese zieht sich durch alle Bilder durch. Die Veränderung im verlauf der Iterationen ist zum einen, dass die Unstetigkeit geglättet wird. Zum anderen ist die Verschiebung, die nicht beim Riss ist linear. Am Ende hingegen wird sie fast konstant. Die Glättung des Risses und damit auch die Glättung der Unstetigkeitsstelle bei der Verschiebung im Gebiet entsteht durch den $|\nabla v|^2$ Term im Optimierungsproblem, wie am Anfang erläutert. Dass die Verschiebung zunächst linear und dann konstant ist, hat die gleiche Begründung. Der Riss breitet sich nicht weiter aus, da das Gebiet nicht weiter verschoben wird. Also ergibt dieses Beispiel genau das, was wir erwarten.

Betrachten wir das gleiche Szenario bei einem $10x10$ Gitter wie in 3.7 und 3.6. Hier gibt es andere Ergebnisse. Der Riss verschiebt sich zum Rand.

Bei sehr kleinen Gitter (10x10) eher etwas komisch: Der Riss setzt sich zum Rand fort, wo $u_0 = 1$ gilt.

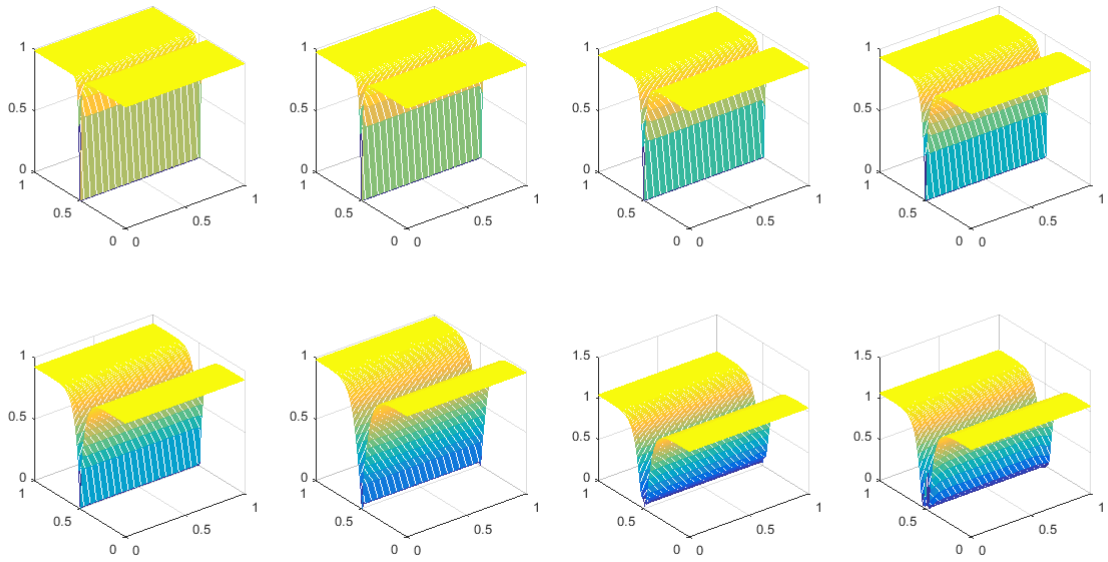


Abbildung 3.5.: Darstellung des Risses bei konstantem u_0, einem Riss in der Mitte und $100x100$ Gitterpunkten

u_0 ist konstant und das Material ist am Rand angerissen Beim nächsten Experiment betrachten wir wieder das gleiche u wir vorhin, nur reißen wir das Papier nur an einer Seite an. Dieser Riss ist 10 Gitterpunkte lang und wir betrachten wieder ein 100x100 Gitter. Betrachten wir zunächst die Verschiebung des Gebietes. Hier sieht man nur dort eine Unstetigkeitsstelle, wo am Anfang der Riss ist. Diese Unstetigkeitsstelle wird länger und breitet sich parallel zur Befestigung aus. Am Ende, bei 200 Iterationen, ist im gesamten Papier eine Unstetigkeit. Die Unstetigkeitsstelle wird wieder glatter, je mehr Iterationen das Programm durchgeführt hat.

Der Riss verhält sich fast identisch. Bei den ersten 20 Iterationen stellt man fast keine Ausbreitung des Risses fest. Bei 50 Iterationen ist der Wert von v dort, wo die Funktion nicht das Extremum annimmt, bei ca 0.7. Bei späteren Iterationen tritt das Phänomen nicht auf und der Riss ist durch das ganze Material zu sehen.

Das grobe Verhalten ist genau das erwartete Verhalten. Der Riss sollte sich parallel zur Befestigung fortsetzten und das tut er auch. Die Annormalität bei der 50. Iteration lässt sich darauf zurückführen, dass v nicht gerissen ist, heißt, dass v ungefähr 1 ist. Damit ist 0.7 noch im Rahmen.

Programm konvergiert Insgesamt lässt sich feststellen, dass das Programm bei vielen Gitterpunkten funktioniert. Mit einer kleineren Gitterweite treten Abweichungen auf.

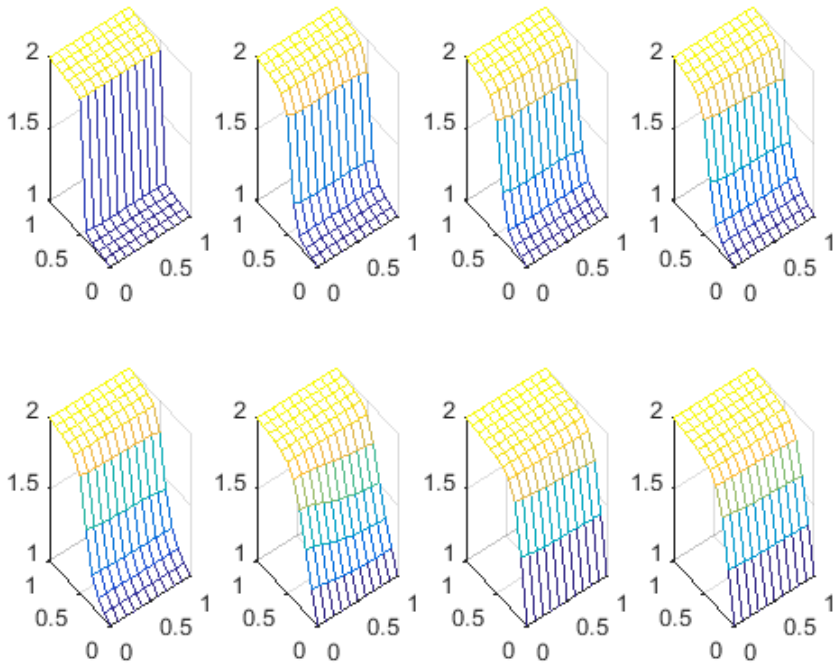


Abbildung 3.6.: Darstellung der Verschiebung des Gebietes bei konstantem u_0, einem Riss in der Mitte und $10x10$ Gitterpunkten

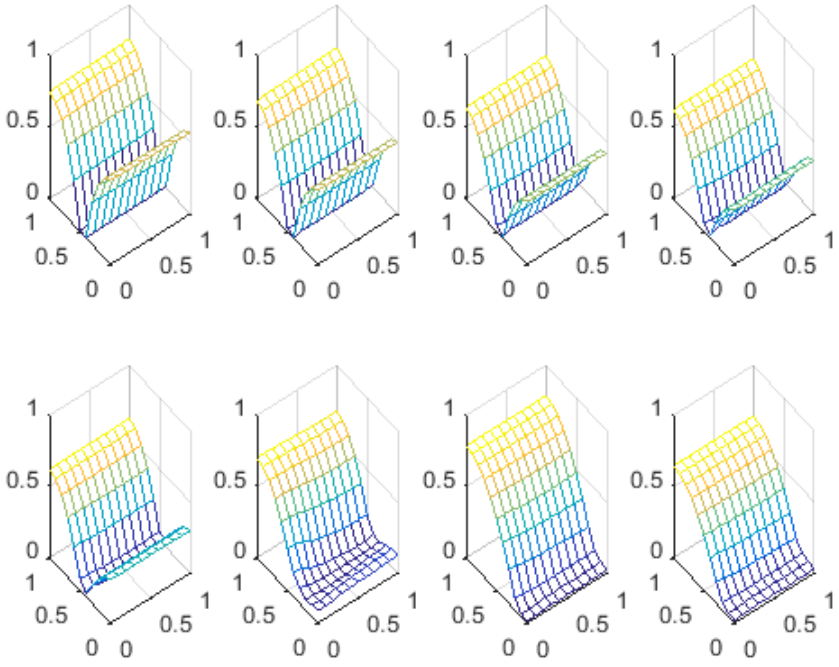


Abbildung 3.7.: Darstellung des Risses bei konstantem u_0, einem Riss in der Mitte und $10x10$ Gitterpunkten

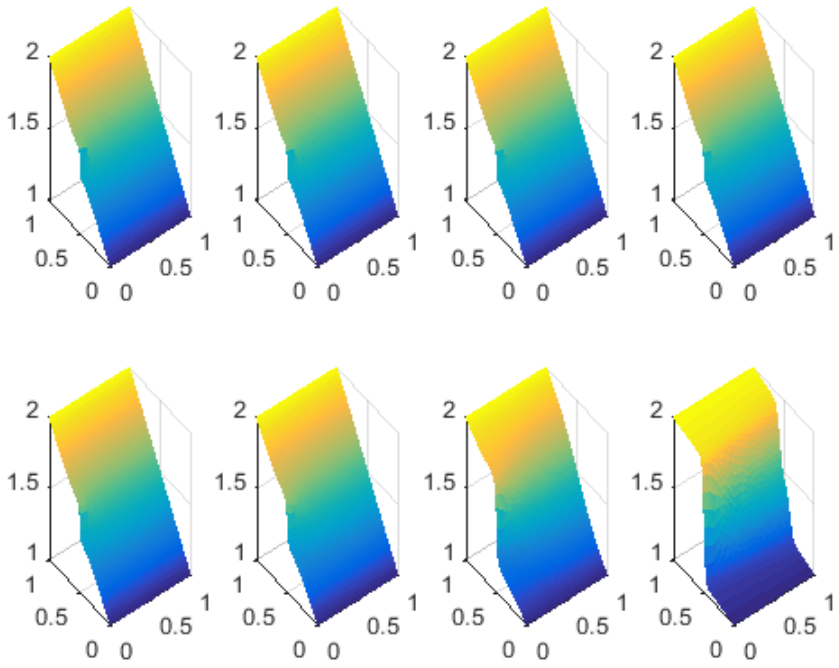


Abbildung 3.8.: Darstellung des Risses bei konstantem u_0, einem kleinen Riss am Rand und $100x100$ Gitterpunkten

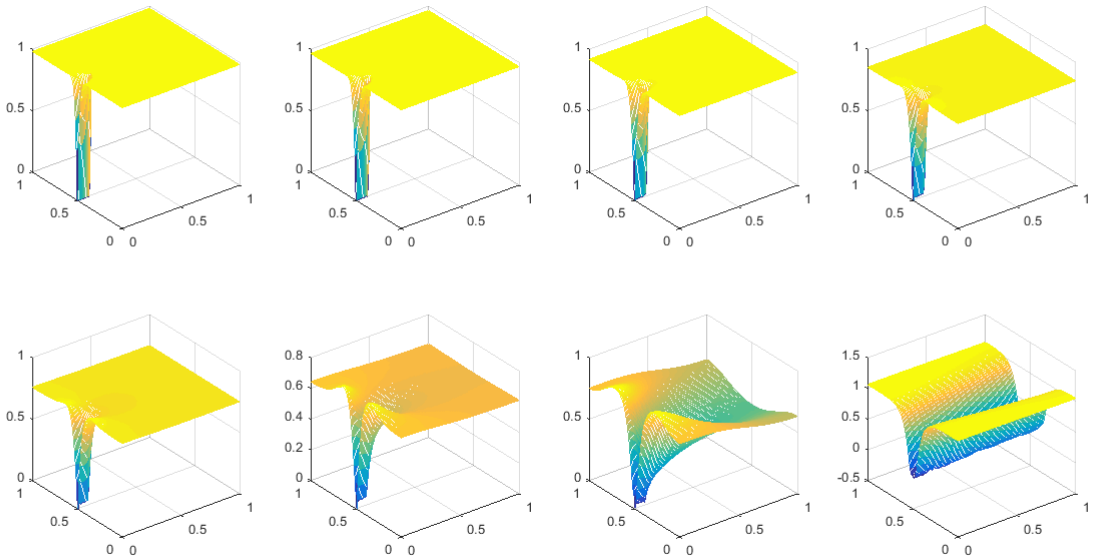


Abbildung 3.9.: Darstellung des Risses bei konstantem u_0, einem kleinen Riss am Rand und $100x100$ Gitterpunkten

4. Fazit und Ausblick

Ideen für Fazit, Ausblick, Auswertung

- bei semismoothness weniger vorr an $u \in L^\infty$

Optimierung nach u andere Arten zu implementieren, die optimierung nach u auch mit in NM rein mit NM Step und jedes einzelner NM step. Wäre bestimmt besser.

Jetzt könnte man noch untersuchen, ob auch mit weniger Voraussetzungen an u_0 und v_0 das Problem eine eindeutige Lösung hat. Dieses werde ich jedoch im Rahmen der Bachelorarbeit nicht untersuchen können.

- Diskussion über Einschränkungen oder Begrenzungen der Arbeit oder der Methoden, ohne die Sachen dabei negativ zu formulieren.
- Darstellung von offenen Problemen, Perspektiven und Ausblicke für zukünftige Arbeiten.

A. Allgemeine Informationen

A.1. Rechnungen

In diesem Kapitel werden alle Rechnungen vorgestellt, die zur numerischen Darstellung der Optimierung nach v notwendig sind.

A.1.1. Numerische Darstellung von G_1

Die Formel

$$\begin{aligned} G_1(v^h, \eta^h) = &\left(2(\int_\Omega u^{dis} T_i T_j \,\mathrm{d}x)_{ij} + 2\epsilon_2 (\int_\Omega \nabla T_i \nabla T_j \,\mathrm{d}x)_{ij} + \frac{2}{\epsilon_3} (\int_\Omega T_i T_j \,\mathrm{d}x)_{ij} \right) v^k \\ &- \frac{2}{\epsilon_3} (\int_\Omega T_j \,\mathrm{d}x)_j + (\int_\Omega T_i T_j \,\mathrm{d}x)_{ij} \eta^k \\ &= (2A + 2\epsilon_2 B + \frac{2}{\epsilon_3} D) v^h - \frac{2}{\epsilon_3} c + D\eta^h \end{aligned}$$

soll berechnet werden.

Um A, B, D zu berechnen, brauchen wir $\int_E \varphi_i \varphi_j$ bzw. $\int_E \nabla\varphi_i \nabla\varphi_j$, wobei E das Einheitsdreieck ist.

	$\int_E \varphi_i \varphi_j$	$\int_E \nabla\varphi_i \nabla\varphi_j$
$\varphi_0\varphi_0$	$\frac{1}{12}$	1
$\varphi_1\varphi_1$	$\frac{1}{12}$	$\frac{1}{2}$
$\varphi_2\varphi_2$	$\frac{1}{12}$	$\frac{1}{2}$
$\varphi_0\varphi_1$	$\frac{1}{24}$	$-\frac{1}{2}$
$\varphi_0\varphi_2$	$\frac{1}{24}$	$-\frac{1}{2}$
$\varphi_1\varphi_2$	$\frac{1}{24}$	0

Nun können wir die einzelnen Matrizen berechnen. Die Berechnung erfolgt analog zur Optimierung nach u. Bei den Matrizen gibt es immer die Fälle, dass i und j gleich sind, j rechts neben i ist, j direkt unter i liegt und j rechts unter i liegt. Für alle anderen i und j ist der Matrixeintrag immer 0. Die Bezeichnungen sind die Gleichen, wie bei u.

Berechnung der Matrix A

$$
\begin{aligned}
A_{i,i} &= \int_{\Omega} u^{dis} T_i T_i \,\mathrm{d}x \\
&= \int_{E_1} u^{dis}_{E_1} \varphi_1 \varphi_1 \,\mathrm{d}x + \int_{E_2} u^{dis}_{E_2} \varphi_2 \varphi_2 \,\mathrm{d}x + \int_{E_3} u^{dis}_{E_3} \varphi_0 \varphi_0 \,\mathrm{d}x \\
&\quad + \int_{E_4} u^{dis}_{E_4} \varphi_0 \varphi_0 \,\mathrm{d}x + \int_{E_5} u^{dis}_{E_5} \varphi_2 \varphi_2 \,\mathrm{d}x + \int_{E_6} u^{dis}_{E_6} \varphi_1 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{12} u^{dis}_{E_1} + \frac{1}{12} u^{dis}_{E_2} + \frac{1}{12} u^{dis}_{E_3} + \frac{1}{12} u^{dis}_{E_4} + \frac{1}{12} u^{dis}_{E_5} + \frac{1}{12} u^{dis}_{E_6} \\
&= \frac{1}{12} \left(\sum_{i=1}^{6} u^{dis}_{E_i} \right)
\end{aligned}
$$

$$
\begin{aligned}
A_{i,i+1} &= \int_{\Omega} u^{dis} T_i T_{i+1} \,\mathrm{d}x = \int_{E_3} u^{dis}_{E_3} \varphi_0 \varphi_1 \,\mathrm{d}x + \int_{E_6} u^{dis}_{E_6} \varphi_0 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{24} u^{dis}_{E_3} + \frac{1}{24} u^{dis}_{E_6} = \frac{1}{24} \left(u^{dis}_{E_3} + u^{dis}_{E_6} \right)
\end{aligned}
$$

$$
\begin{aligned}
A_{i,i+1+n} &= \int_{\Omega} u^{dis} T_i T_{i+1+n} \,\mathrm{d}x = \int_{E_4} u^{dis}_{E_4} \varphi_0 \varphi_1 \,\mathrm{d}x + \int_{E_5} u^{dis}_{E_5} \varphi_0 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{24} u^{dis}_{E_4} + \frac{1}{24} u^{dis}_{E_5} = \frac{1}{24} \left(u^{dis}_{E_4} + u^{dis}_{E_5} \right)
\end{aligned}
$$

$$
\begin{aligned}
A_{i,i+n+2} &= \int_{\Omega} u^{dis} T_i T_{i+2+n} \,\mathrm{d}x = \int_{E_5} u^{dis}_{E_5} \varphi_1 \varphi_2 \,\mathrm{d}x + \int_{E_6} u^{dis}_{E_6} \varphi_1 \varphi_2 \,\mathrm{d}x \\
&= \frac{1}{24} u^{dis}_{E_5} + \frac{1}{24} u^{dis}_{E_6} = \frac{1}{24} \left(u^{dis}_{E_5} + u^{dis}_{E_6} \right)
\end{aligned}
$$

Berechnung der Matrix B

$$
\begin{aligned}
B_{i,i} &= \int_{\Omega} \nabla T_i \nabla T_i \,\mathrm{d}x \\
&= \int_{E_1} \nabla\varphi_1 \nabla\varphi_1 \,\mathrm{d}x + \int_{E_2} \nabla\varphi_2 \nabla\varphi_2 \,\mathrm{d}x + \int_{E_3} \nabla\varphi_0 \nabla\varphi_0 \,\mathrm{d}x \\
&+ \int_{E_4} \nabla\varphi_0 \nabla\varphi_0 \,\mathrm{d}x + \int_{E_5} \nabla\varphi_2 \nabla\varphi_2 \,\mathrm{d}x + \int_{E_6} \nabla\varphi_1 \nabla\varphi_1 \,\mathrm{d}x \\
&= \frac{1}{2} + \frac{1}{2} + 1 + 1 + \frac{1}{2} + \frac{1}{2} = 4
\end{aligned}
$$

$$
\begin{aligned}
B_{i,i+1} &= \int_{\Omega} \nabla T_i \nabla T_{i+1} \,\mathrm{d}x = \int_{E_3} \nabla\varphi_0 \nabla\varphi_1 \,\mathrm{d}x + \int_{E_6} \nabla\varphi_0 \nabla\varphi_1 \,\mathrm{d}x \\
&= -\frac{1}{2} - \frac{1}{2} = -1
\end{aligned}
$$

$$
\begin{aligned}
B_{i,i+n+1} &= \int_{\Omega} \nabla T_i \nabla T_{i+1+n} \,\mathrm{d}x = \int_{E_4} \nabla\varphi_0 \nabla\varphi_1 \,\mathrm{d}x + \int_{E_5} \nabla\varphi_0 \nabla\varphi_1 \,\mathrm{d}x \\
&= -\frac{1}{2} - \frac{1}{2} = -1
\end{aligned}
$$

$$
\begin{aligned}
B_{i,i+n+2} &= \int_{\Omega} \nabla T_i \nabla T_{i+2+n} \,\mathrm{d}x \\
&= \int_{E_5} \nabla\varphi_1 \nabla\varphi_2 \,\mathrm{d}x + \int_{E_6} \nabla\varphi_1 \nabla\varphi_2 \,\mathrm{d}x = 0
\end{aligned}
$$

Berechnung der Matrix D

$$\begin{aligned} D_{i,i} &= \int_{\Omega} T_i T_i \,\mathrm{d}x \\ &= \int_{E_1} \varphi_1\varphi_1 \,\mathrm{d}x + \int_{E_2} \varphi_2\varphi_2 \,\mathrm{d}x + \int_{E_3} \varphi_0\varphi_0 \,\mathrm{d}x \\ &\quad + \int_{E_4} \varphi_0\varphi_0 \,\mathrm{d}x + \int_{E_5} \varphi_2\varphi_2 \,\mathrm{d}x + \int_{E_6} \varphi_1\varphi_1 \,\mathrm{d}x \\ &= \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{1}{2} \end{aligned}$$

$$\begin{aligned} D_{i,i+1} &= \int_{\Omega} T_i T_{i+1} \,\mathrm{d}x = \int_{E_3} \varphi_0\varphi_1 \,\mathrm{d}x + \int_{E_6} \varphi_0\varphi_1 \,\mathrm{d}x \\ &= \frac{1}{24} + \frac{1}{24} = \frac{1}{12} \end{aligned}$$

$$\begin{aligned} D_{i,i+n+1} &= \int_{\Omega} T_i T_{i+1+n} \,\mathrm{d}x = \int_{E_4} \varphi_0\varphi_1 \,\mathrm{d}x + \int_{E_5} \varphi_0\varphi_1 \,\mathrm{d}x \\ &= \frac{1}{24} + \frac{1}{24} = \frac{1}{12} \end{aligned}$$

$$\begin{aligned} D_{i,i+n+2} &= \int_{\Omega} T_i T_{i+2+n} \,\mathrm{d}x = \int_{E_5} \varphi_1\varphi_2 \,\mathrm{d}x + \int_{E_6} \varphi_1\varphi_2 \,\mathrm{d}x \\ &= \frac{1}{24} + \frac{1}{24} = \frac{1}{12} \end{aligned}$$

A.1.2. Berechnung von G_{2v}

Es muss $\int_{\Omega} \frac{\partial f}{\partial v} T_i T_j \,\mathrm{d}x$ berechnet werden. Dazu integriert man statt über Ω wieder über die einzelnen Dreiecke. Dabei ist zu beachten, dass für gerade und ungerade Dreiecke andere Ergebnisse zustande kommen:

i	j	$\int_E \frac{\partial f}{\partial v} T_i T_j \,\mathrm{d}x$ gerades Dreieck	$\int_E \frac{\partial f}{\partial v} T_i T_j \,\mathrm{d}x$ ungerades Dreieck
0	0	$\frac{1}{60}(f_1^h + 3f_2^h + f_3^h)$	$\frac{1}{60}(f_1^h + 3f_2^h + f_3^h)$
0	1	$\frac{1}{120}(f_1^h + 2f_2^h + 2f_3^h)$	$\frac{1}{120}(2f_1^h + 2f_2^h + f_3^h)$
0	2	$\frac{1}{120}(2f_1^h + 2f_2^h + f_3^h)$	$\frac{1}{120}(f_1^h + 2f_2^h + 2f_3^h)$
1	1	$\frac{1}{60}(f_1^h + f_2^h + 3f_3^h)$	$\frac{1}{60}(3f_1^h + f_2^h + f_3^h)$
1	2	$\frac{1}{120}(2f_1^h + f_2^h + 2f_3^h)$	$\frac{1}{120}(2f_1^h + f_2^h + 2f_3^h)$
2	2	$\frac{1}{60}(3f_1^h + f_2^h + f_3^h)$	$\frac{1}{60}(f_1^h + f_2^h + 3f_3^h)$

Dabei ist f_1^h bei einem geraden Dreieck die Auswertung von f^h an der oberen linken Ecke des Dreiecks. Die anderen Bezeichnungen sind darauf aufbauend.

Damit können wir ∂G_{2v} diskretisieren. Wir nennen die Diskretisierung F_{ij}. Hier hat man wieder die vier Fälle:

$$
\begin{aligned}
F_{i,i} &= \int_\Omega f^h T_i T_i \,\mathrm{d}x \\
&= \int_{E_1} f^h_{E_1} \varphi_1 \varphi_1 \,\mathrm{d}x + \int_{E_2} f^h_{E_2} \varphi_2 \varphi_2 \,\mathrm{d}x + \int_{E_3} f^h_{E_3} \varphi_0 \varphi_0 \,\mathrm{d}x \\
&\quad + \int_{E_4} f^h_{E_4} \varphi_0 \varphi_0 \,\mathrm{d}x + \int_{E_5} f^h_{E_5} \varphi_2 \varphi_2 \,\mathrm{d}x + \int_{E_6} f^h_{E_6} \varphi_1 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{60} \Big(\big(f_1^h + f_2^h + 3f_3^h\big)_{E_1} + \big(f_1^h + f_2^h + 3f_3^h\big)_{E_2} + \big(f_1^h + 3f_2^h + f_3^h\big)_{E_3} \\
&+ \big(f_1^h + 3f_2^h + f_3^h\big)_{E_4} + \big(3f_1^h + f_2^h + f_3^h\big)_{E_5} + \big(3f_1^h + f_2^h + f_3^h\big)_{E_6} \Big)
\end{aligned}
$$

$$
\begin{aligned}
A_{i,i+1} &= \int_\Omega f^h T_i T_{i+1} \,\mathrm{d}x = \int_{E_3} f^h_{E_3} \varphi_0 \varphi_1 \,\mathrm{d}x + \int_{E_6} f^h_{E_6} \varphi_0 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{120} \Big(\big(f_1^h + 2f_2^h + 2f_3^h\big)_{E_3} + \big(2f_1^h + 2f_2^h + f_3^h\big)_{E_6} \Big)
\end{aligned}
$$

$$
\begin{aligned}
A_{i,i+1+n} &= \int_\Omega f^h T_i T_{i+1+n} \,\mathrm{d}x = \int_{E_4} f^h_{E_4} \varphi_0 \varphi_1 \,\mathrm{d}x + \int_{E_5} f^h_{E_5} \varphi_0 \varphi_1 \,\mathrm{d}x \\
&= \frac{1}{120} \Big(\big(2f_1^h + 2f_2^h + f_3^h\big)_{E_4} + \big(f_1^h + 2f_2^h + 2f_3^h\big)_{E_5} \Big)
\end{aligned}
$$

$$A_{i,i+n+2} = \int_\Omega f^h T_i T_{i+2+n}\,\mathrm{d}x = \int_{E_5} f^h_{E_5}\varphi_1\varphi_2\,\mathrm{d}x + \int_{E_6} f^h_{E_6}\varphi_1\varphi_2\,\mathrm{d}x$$
$$= \frac{1}{120}\left(\left(2f_1^h + f_2^h + 2f_3^h\right)_{E_5} + \left(2f_1^h + f_2^h + 2f_3^h\right)_{E_6}\right)$$

hierbei bedeutet $\left(f_1^h + f_2^h + f_3^h\right)_{E_j}$, dass f_i^h f auf dem i-ten Gitterpunkt des Dreieck E_j ausgewertet wird. Die Transformation mit $1/h_1h_2$ wird auch hier am Schluss ausgeführt.

A.2. Code

first test

```
function result = calculate_u(edges, allSurroundingTriangles, u0, epsilon, v, m, n)
% CALCULATE_U berechnet u
%   Ziel ist es
%   - \int_Omega (v^2 + eps) \nabla u0 \nabla T_j
%   = \sum_i^{(n+1)(m+1)}  \int_Omega (v^2+eps) \nabla u \nabla T_i \nabla T_j
%   für alle j =1,... (m+1)(n+1) zu berechnen
%   Dafür muss die linke und rechte Seite numerisch berechnet werden.

    % Da für die Berechnung des Integrals für alle Dreiecke das Integral von
    % v^2+eps gebraucht wird, berechnen wir es hier.
    integral_v = integral_of_v_total(v, edges, epsilon,n,m);

%   rechte Seite
%   die rechte Seite ist 0 für alle Werte am Rand und sonst die Werte des
%   Integrals.
    % rechts muss an den rändern 0 gesetzt werden. Dadurch werden die 0
    % Randwerte eingebracht. Da die Matrix sonst singulär wird, muss der
    % Diagonaleintrag 1 gesetzt werden.
    rechts = integrate_u(allSurroundingTriangles,integral_v, m, n);
         for i=0:m
            rechts(i*(n+1)+1,: ) = 0;
            rechts(i*(n+1)+1+n,:) = 0;
            rechts(i*(n+1)+1,i*(n+1)+1) = 1;
            rechts(i*(n+1)+1+n,i*(n+1)+1+n ) = 1;
         end

%   linke Seite: - \int_Omega (v^2 + eps) \nabla u0 \nabla T_j = - \int_Omega (v^2 + e

    links = - rechts * u0;
```

```
    % links muss an den Rändern 0 gesetzt werden. Dadurch werden die 0 Randwerte einge
    for i=0:m
       links(i*(n+1)+1) = 0;
       links(i*(n+1)+1+n) = 0;
    end

    % hier wird x aus  Ax=b berechnet.

    result = rechts\links;

    result = result + u0;

end
```

Literaturverzeichnis

[1] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimazation with PDE Constraints*. Springer, 2009. 4, 6, 11, 14, 16

[2] Karl Kunisch. *Semi-smooth Newton Methods for Non-differentiable Optimazation Problems*. 2008. 7

[3] Dr. Frank WÃ¼bbeling. Vorlesung numerik partieller differentialgleichungen. 9

[4] Prof. Benedikt Wirth. Vorlesung optimierung ii. 5, 6, 7

[5] Prof. Benedikt Wirth. Vorlesung partielle differentialgleichungen. 8