



# LÖSUNG EINES PHASENFELD MODELLS FÜR RISSENTSTEHUNG MITTELS SEMIGLATTER NEWTONMETHODE

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

<b>1</b>	<b>Einleitung</b>	<b>1</b>
<b>2</b>	<b>Mathematische Grundlagen</b>	<b>3</b>
2.1	Grundlagen Optimierung . . . . .	3
2.1.1	Optimierungsproblem ohne Nebenbedingung . . . . .	3
2.1.2	Optimierungsproblem mit Ungleichungsnebenbedingung . . . . .	5
2.2	Grundlagen pDGL . . . . .	6
2.3	Finite Elemente . . . . .	8
2.4	Semidifferenzierbare Newtonmethoden . . . . .	10
2.4.1	Newton Methoden mit einfachen Nebenbedingungen . . . . .	10
2.4.2	Konvergenz der generalisierten Newton Methode . . . . .	11
2.4.3	Semidifferential . . . . .	13
2.4.4	semidiffbare Newton Methoden . . . . .	15
<b>3</b>	<b>Anwendung auf das Phasenfeldmodell für Rissentstehung</b>	<b>17</b>
3.1	erste Betrachtung der Rissentstehung . . . . .	17
3.2	Optimierung nach $u$ . . . . .	18
3.2.1	Analytische Betrachtung . . . . .	18
3.2.2	Numerische Betrachtung . . . . .	20
3.2.3	Zusammenfassung . . . . .	31
3.3	Optimierung nach $v$ . . . . .	31
3.3.1	Newtonmethode . . . . .	35
3.3.2	numerische Betrachtung . . . . .	38
	<b>Literaturverzeichnis</b>	<b>51</b>

# 1 Einleitung

Ich befasse mich mit Rissen und der Darstellung des Problems als Newtonmethode. Dabei sind meine Aufgaben:

1. Optimalitätsbedingungen aufstellen (KKT)
2. Das Problem auf Semidifferenzierbarkeit untersuchen
3. Newtonmethode für das Problem aufstellen
4. Problem implementieren mithilfe von Finite Elemente
5. Konvergenz des Problems anhand der Implementation untersuchen: Hängt die Konvergenz vom Gitter ab?

Das Problem lautet:

$$\begin{aligned} \min_{u \in H^1(\Omega)^2, v \in H^1(\Omega)} & \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\ \text{s.d. } & 0 \leq v \leq v_0 \\ & u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2 \end{aligned}$$

**Notation 1.0.1.**  $|\nabla u|^2 := \frac{\partial u_1}{\partial x_1}^2 + \frac{\partial u_1}{\partial x_2}^2 + \frac{\partial u_2}{\partial x_1}^2 + \frac{\partial u_2}{\partial x_2}^2 = u_{1x_1}^2 + u_{1x_2}^2 + u_{2x_1}^2 + u_{2x_2}^2$

dabei ist  $\epsilon_i > 0 \forall i \in \{1, 2, 3\}$  ein kleiner Parameter und  $\Gamma_1, \Gamma_2 \subset \Omega$  ein der rechte bzw. linke Rand eines rechteckigen Gebietes  $\Omega \subset \mathbb{R}^2$ , das heißt,  $\Omega = [0, a] \times [0, b]$   $\Gamma_1 = \{0\} \times [0, b]$   $\Gamma_2 = \{a\} \times [0, b]$

$u : \Omega \rightarrow \mathbb{R}^2$  beschreibt die Verschiebung eines Körpers auf dem Gebiet  $\Omega$ , wenn ein Riss entsteht. Der Körper ist an  $\Gamma_1$  und  $\Gamma_2$  befestigt, was die Randbedingung  $u_0$  angibt.  $v : \Omega \rightarrow \mathbb{R}$  gibt an, wo und wie stark der Körper gerissen ist. 1 bedeutet, dass der Körper vollständig gerissen ist und 0, dass kein Riss vorhanden ist.

Analytische und numerische Grundlagen werden gebraucht, um die Verschiebung des Gebietes und den Riss zu finden. Für die analytische Betrachtung nutze ich partiell-

le Differentialgleichungen und Grundlagen der Optimierung. Da die Lösung des Problems analytisch nicht zu finden ist, diskretisiere ich das Problem mit dreieckig lineare Lagrange Elemente und implementiere es mittels semiglatte Newtonmethoden. Die Grundlagen dazu sind im zweiten Kapitel zu finden. Alle Themen sind sehr umfangreich und ich werde nur die wichtigsten Begriffe einführen können. Der Leser sollte schon Wissen über Finite Elemente, insbesondere die dreieckig linearen Lagrange Elemente mitbringen.

Die Untersuchung des Problems folgt in Kapitel drei. Beim ersten Betrachten fällt auf, dass man die Optimierung nach  $u$  und  $v$  trennen 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 Methoden. 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 Grundlagen Optimierung

Das Phasenfeldmodell für die Rissentstehung ist ein Optimierungsproblem mit Ungleichungsnebenbedingungen. 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  $J : W \rightarrow \mathbb{R}$  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 [?, S. 50]

Sei  $F : U \subset X \rightarrow Y$  ein Operator zwischen Banachräumen und  $U \neq \emptyset$  offen.

**Definition 2.1.1** (Richtungsableitung).  $F$  heißt *Richtungsableitbar* in  $x \in U$ , falls

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

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

$$F'(x) : X \rightarrow Y$$

$$h \mapsto \delta F(x, h)$$

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:

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

Nun können wir die Ableitung von  $J$  bestimmen und daraus resultierend das Optimierungsproblem lösen. Das Theorem stammt aus der Vorlesung „Optimierung 2“, gelesen von Prof. B. Wirth.

**Theorem 2.1.4.** Sei das Optimierungsproblem (2.1) gegeben. Sei  $J : W \rightarrow \mathbb{R}$  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 eine Bedingung für ein Optimum gegeben. Das Optimierungsproblem ist 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. Damit überhaupt klar ist, wie das  $\leq$  gemeint ist, wird ein positiver Kegel nach der Vorlesung „Optimierung 2“ von Prof. Wirth 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  $y - x \in \overset{\circ}{P}$

Wir werden Probleme der Form

$$\min_{w \in W} J(w) \quad \text{s.d.} \quad G(w) \leq_P 0 \quad (2.2)$$

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

$$\min_{w \in W} J(w) \quad \text{s.d.} \quad w \in C \quad (2.3)$$

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.  $J : W \rightarrow \mathbb{R}$  strikt konvex und stetig auf  $C$
4.  $J$  Gâteaux differenzierbar
5.  $\lim_{w \in C, \|w\|_W \rightarrow \infty} J(w) = \infty$

Dann existiert genau eine Lösung von (2.3).

*Beweis.* Der Beweis und das Theorem sind in [?, 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 Umfangs 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  $\dot{P} \neq \emptyset$ . Sei  $J : W \rightarrow \mathbb{R} \cup \{\infty\}$ ,  $G : W \rightarrow Z$  konvex. Es existiert ein  $\hat{w}$  im Bild( $J$ ), sodass  $G(\hat{w}) \leq_P 0$ . Außerdem gelte  $\mu = \inf\{J(w) | G(w) \leq_P 0\} < \infty$ .*

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

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

Nun haben wir die Bedingungen gegeben, sodass wir von (2.3) mit  $C$  wie oben das KKT System aufstellen können. Dabei ist  $\bar{w}$  die Lösung des Problems.  $\mu$  und  $\lambda$  heißen Lagrange Multiplikatoren.

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

Aus den letzten beiden Zeilen folgt, dass

$$\mu - \lambda = \max\{0, \mu - \lambda + c(\bar{w} - b)\} + \min\{0, \mu - \lambda + c(\bar{w} - a)\} \forall c > 0 \quad (2.4)$$

Diese Darstellung werde ich später nutzen, um das Problem über die Rissentstehung zu lösen.

## 2.2 Grundlagen pDGL

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.

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

$$\begin{aligned} Lu &= f \text{ auf } \Omega \\ u &= g \text{ auf } \partial\Omega \end{aligned} \tag{2.5}$$

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

**Definition 2.2.1** (schwache Lösung).  $u \in g + H_0^1(\Omega)$  heißt schwache Lösung zu (2.5), falls

$$B(u, v) := \int_{\Omega} \nabla v^T A \nabla u + b \nabla u v + c u v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega)$$

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 2.5 erfüllt. Falls eine schwache Lösung für 2.5 existiert, ist sie eindeutig.

*Beweis.* Der Beweis wird im Script von Prof. B. Wirth zur Vorlesung „Partielle Differentialgleichungen“ geführt.  $\square$

**Theorem 2.2.4** (Existenz der schwachen Lösung). Sei  $\Omega$  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 2.5.

*Beweis.* Der Beweis wird im Script von Prof. B. Wirth zur Vorlesung „Partielle Differentialgleichungen“ geführt.  $\square$

## 2.3 Finite Elemente

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

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

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

$h = (h_1, h_2)$  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  $m$  und  $n$  recht nahe beieinander gewählt werden. Nun wird durch die Gitterpunkte die Triangulierung gelegt. Diese nennen wir  $E_k$  und ist in 3.1 dargestellt.

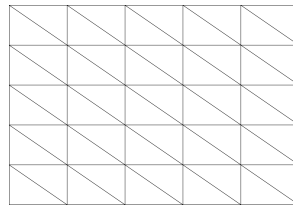


Abbildung 2.1: Triangulierung eines rechteckigen Gebietes

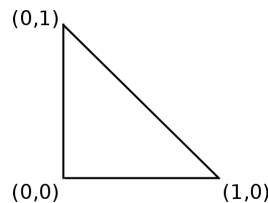


Abbildung 2.2: Referenzdreieck

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  $(E, P, \Psi)$ , wobei

$E$  das Referenzdreieck 3.2 ist,  $P = \mathcal{P}_1$ , sind Polynome auf  $\mathbb{R}^2$  vom Grad 1 mit Basis  $\{p_1, p_2, p_3\}$

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

und  $\Psi := \{\varphi_0, \varphi_1, \varphi_2\}$  sind Funktionale auf  $P$  und damit eine Basis von  $P^*$ .  $\varphi_i$  sind lokale Formfunktionen d.h.  $\varphi_i(p_j) = \delta_{ij}$ ,  $i, j \in \{0, 1, 2\}$ . Dabei ist  $\delta_{ij}$  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

$$\varphi_1 = 1 - x - y, \quad \varphi_2 = x, \quad \varphi_3 = y \quad (2.6)$$

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

$$T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

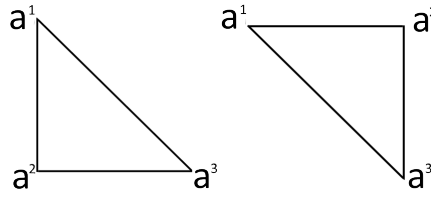
$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \pm \begin{pmatrix} h_1 x \\ h_2 y \end{pmatrix}$$

durch das Referenzelement darstellen. Dabei entspricht  $(a_1, a_2)^t$  dem Eckpunkt mit dem  $90^\circ$  Winkel des Rechteckes und  $(h_1, h_2)^t$  ist die Höhe 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

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

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  $\Omega$  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 und ungerade Dreiecke mit den Werten von  $a$ 

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

$$a(x, y) = (a^3 - a^2)x + (a^1 - a^2)y + a^2 \quad \text{mit} \quad \nabla a(x, y) = \begin{pmatrix} a^3 - a^2 \\ a^1 - a^2 \end{pmatrix} \quad (2.7)$$

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

$$a(x, y) = (a^1 - a^2)x + (a^3 - a^2)y + a^2 \quad \text{mit} \quad \nabla a(x, y) = \begin{pmatrix} a^1 - a^2 \\ a^3 - a^2 \end{pmatrix} \quad (2.8)$$

## 2.4 Semidifferenzierbare Newtonmethoden

Semiglatte Newtonmethoden werden gebraucht, um Nullstellen von nicht differenzierbaren Funktionen numerisch zu berechnen. Die Rissentstehung ist ein nicht differenzierbares Problem. Um die Idee der Newtonmethoden zu verstehen, führe ich zunächst einfache Newton Methoden 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 dann die semidifferenzierbare Newtonmethode eingeführt werden, von der wir auch die Konvergenz betrachten werden. Dieses Kapitel richtet sich nach [?, S. 115 ff].

### 2.4.1 Newton Methoden mit einfachen Nebenbedingungen

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

$$\min_{w \in \mathbb{R}^n} f(w) \quad f : \mathbb{R}^n \rightarrow \mathbb{R} \quad (2.9)$$

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, setzen wir  $w_0, w_1, \dots$  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). *Das Verfahren 1 löst das Optimierungsproblem (2.9). Es konvergiert superlinear falls  $G \in C^1$  und  $G'$  invertierbar ist.*

**Data:**  $w^0$  (möglichst Nah an der Lösung  $\bar{w}$ )

**for**  $k = 0, 1, \dots$  **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  $\bar{x}$  approximiert.*

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

Betrachte nun

$$G(x) = 0 \tag{2.10}$$

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

Um eine numerische Lösung von (2.10) zu erhalten, benutzen wir einen ähnlichen Al-

gorithmus, wie den für das einfache Newtonverfahren, nur allgemeiner:

**Data:**  $x^0 \in X$  (möglichst Nah an der Lösung  $\bar{x}$ )

**for**  $k = 0, 1, \dots$  **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  $\bar{x}$ . Sei  $d^{k+1} = x^{k+1} - \bar{x}$  der Abstand zwischen dem Iterationsschritt und der Lösung. Dann gilt:

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

Wir erhalten:

**Theorem 2.4.3.** *Betrachte (2.10) mit der Lösung  $\bar{x}$ . Sei  $x^k$  die Folge, die durch den Generalisierten Newton Algorithmus 2 erzeugt wurde. Sei  $x^0$  nah genug an  $\bar{x}$  gewählt*

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

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

*gilt, dann konvergiert  $x^k \rightarrow \bar{x}$  linear mit Konstante  $\gamma$*

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

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

*gilt, dann konvergiert  $x^k \rightarrow \bar{x}$  super linear*

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

$$\|d^{k+1}\|_X = \|M_k^{-1} (G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X \leq C \|d^k\|_X^{1+\alpha}$$

für  $\|d_k\|_X \rightarrow 0$

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

*Beweis.* Der Beweis ist in [?, 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:

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

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

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

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

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

oder

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

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  $G : X \rightarrow Y$  ein stetiger Operator. Dann ist die Menge der verallgemeinerten Differentiale

definiert als

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

Dabei meint  $\rightrightarrows L(X, Y)$ , dass ein Punkt  $x \in X$  auf eine Menge von linearen Operatoren abgebildet wird (und nicht nur auf einen Operator). 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

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

Dieses nennt sich semidiffbar.

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

1.  $G$  heißt  $\partial G$  semidiffbar in  $x \in X$ , falls

$$\sup_{M \in \partial G(x+d)} \|G(x + d^k) - G(x) - M_k d\|_Y = o(\|d\|_X) \text{ für } \|d\|_X \rightarrow 0 \quad (2.11)$$

2.  $G$  heißt  $\partial G$  semidiffbar von der Ordnung  $\alpha + 1 > 1$  in  $x \in X$ , falls

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

**Lemma 2.4.8.** Sei  $G : X \rightarrow Y$  ein Operator zwischen Banachräumen und stetig  $F$ -diffbar in einer Umgebung von  $x$ . Dann ist  $G$   $\{G'\}$ -semidiffbar in  $x$ . Falls  $G'$   $\alpha$ -Hölderstetig in einer Umgebung von  $x$  ist, dann ist  $G$   $\{G'\}$ -semidiffbar in  $x$  von der Ordnung  $\alpha$ .

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



*Beweis.*

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

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

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

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

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

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

## 2.4.4 semidiffbare Newton Methoden

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

**Data:**  $x^0 \in X$  (möglichst Nah an der Lösung  $\bar{x}$ )  
**for**  $k = 0, 1, \dots$  **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:** semidiffbare Newton Methode

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

noch die Regularitätsannahme.

**Definition 2.4.10** (Regularitätsannahme für semidiffbare Newton Verfahren). *Betrachte (2.10) mit der Lösung  $\bar{x}$ . Dann lautet die Regularitätsannahme*

$$\exists C > 0, \quad \exists \delta > 0 : \|M^{-1}\|_{X \rightarrow Y} \leq C \quad \forall M \in \partial G(x) \quad \forall x \in X, \quad \|x - \bar{x}\|_X < \delta \quad (2.12)$$

**Theorem 2.4.11** (Konvergenz des semidiffbaren Newton-Verfahrens). *Sei das Problem (2.10) gegeben mit der Lösung  $\bar{x}$ . Seien  $X, Y$  Banachräume,  $G : X \rightarrow Y$  stetig und  $\partial G$  semidiffbar und die Regularitätsannahme (2.12) sei gegeben. Dann existiert  $\delta > 0$ , sodass für alle  $x^0 \in X$  mit  $\|x^0 - \bar{x}\|_X < \delta$  die semidiffbare Newton Methode super linear gegen  $\bar{x}$  konvergiert.*

*Falls  $G$   $\partial G$ -semidiffbar der Ordnung  $\alpha > 0$  in  $\bar{x}$  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

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

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 Semidiffbarkeit:

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

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

### 3.1 erste Betrachtung der Rissentstehung

Erinnern wir uns an die vorangegangene Problemstellung:

$$\begin{aligned} & \min_{u \in H^1(\Omega)^2, v \in H^1(\Omega)} \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\ & \text{s.d. } 0 \leq v \leq v_0 \\ & u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2 \end{aligned}$$

Beim genaueren Betrachten bemerkt man, dass die zwei Nebenbedingungen entweder nur von  $u$  oder nur von  $v$  abhängen. Dies bietet die Möglichkeit das Optimierungsproblem in zwei Teilprobleme aufzuteilen. Zum einen die Optimierung nach  $u$  und zum anderen die Optimierung nach  $v$ .

$$\begin{aligned} & \min_{u \in H^1(\Omega)^2} \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\ & u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2 \\ & \min_{v \in H^1(\Omega)} \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\ & \text{s.t. } 0 \leq v \leq v_0 \end{aligned}$$

Wenn man beide Probleme implementiert, löst man zunächst ein Problem und setzt die Lösung dann in das andere Problem ein. Dieses Vorgehen wiederholt man. Betrachten wir zuerst die Optimierung nach  $u$ .

## 3.2 Optimierung nach $u$

### 3.2.1 Analytische Betrachtung

Die Optimierung nach  $u$  sieht wie folgt aus:

$$\min_{u \in H^1(\Omega)^2} \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx$$

$$u = u_0 \text{ auf } \Gamma_1 \cup \Gamma_2$$

Das Problem lässt sich vereinfachen, indem man

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

sucht, statt  $u \in H^1(\Omega)^2$ . Daraus ergibt sich ein Problem ohne Nebenbedingung.

$$\min_{u \in u_0 + H_0^1(\Omega)^2} \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \quad (3.1)$$

Da  $u : \Omega \rightarrow \mathbb{R}^2$ , müssen wir nach  $u_1$  und nach  $u_2$  minimieren. Dieses werden wir nicht gleichzeitig tun, sondern zunächst die Optimierung nach  $u_1$  und dann die Optimierung nach  $u_2$  betrachten. Die Optimierungen kann man trennen, da keine Mischung aus den Termen  $u_1$  und  $u_2$  auftauchen. Beide Optimierungen sind identisch, es müssen später nur unterschiedliche Werte eingesetzt werden. Betrachten wir also nur die Optimierung nach  $u_1$ .

**Theorem 3.2.1.** *Bedingung für Minimum Sei das Minimierungsproblem (3.1) gegeben. Sei  $\tilde{u}_1$  das Minimum. Dann gilt*

$$\int_{\Omega} 2(v^2 + \epsilon_1) \nabla \tilde{u}_1 \nabla \varphi dx = 0 \forall \varphi \in u_0 + H_0^1(\Omega) \quad (3.2)$$

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

$$J : u_0 + H_0^1(\Omega)^2 \rightarrow \mathbb{R}$$

$$u \mapsto \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx$$

(3.2) ist. Leiten wir  $J$  ab:

$$\begin{aligned}
\partial J(u, \varphi) &= \lim_{t \rightarrow 0} \frac{1}{t} \left( J(u_1 + t\varphi, u_2) - J(u_1, u_2) \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + \epsilon_1) |\nabla(u_1 + t\varphi)|^2 + |\nabla u_2|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \right. \\
&\quad \left. - \int_{\Omega} (v^2 + \epsilon_1) |\nabla u_1|^2 + |\nabla u_2|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + \epsilon_1) (|\nabla(u_1 + t\varphi)|^2 - |\nabla u_1|^2) dx \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + \epsilon_1) (|\nabla u_1 + t\nabla\varphi|^2 - |\nabla u_1|^2) dx \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + \epsilon_1) (|\nabla u_1|^2 + 2t\nabla u_1 \nabla\varphi + t^2 |\nabla\varphi|^2 - |\nabla u_1|^2) dx \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + \epsilon_1) (2t\nabla u_1 \nabla\varphi + t^2 |\nabla\varphi|^2) dx \right) \\
&= \int_{\Omega} 2(v^2 + \epsilon_1) \nabla u_1 \nabla\varphi dx
\end{aligned}$$

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

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

$$0 = \int_{\Omega} 2(v^2 + \epsilon_1) \nabla u_1 \nabla\varphi dx \quad (3.3)$$

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 (3.3) 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:

$$\begin{aligned}
B(u_1, \varphi) : (u_0 + H^1(\Omega))^2 &\rightarrow \mathbb{R} \\
(u_1, \varphi) &\mapsto \int_{\Omega} (v^2 + \epsilon_1) 2\nabla u_1 \nabla\varphi dx
\end{aligned}$$

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:

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

Damit ist Annahme 1 erfüllt mit  $\lambda = \epsilon_1$ . Zu Annahme 2:

$$|\xi^T A(x) \zeta| = (v^2 + \epsilon_1) \xi \cdot \zeta \leq (v_0^2 + \epsilon_2) \xi \cdot \zeta \leq (\sup(v_0)^2 + \epsilon_3) |\xi| |\zeta|$$

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

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.

### 3.2.2 Numerische Betrachtung

Zur Erinnerung: Folgendes Problem soll numerisch gelöst werden.

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

$$0 = \int_{\Omega} 2 (v^2 + \epsilon_1) \nabla u_1 \nabla \varphi \, dx$$

Die 2 kann gekürzt werden, da wir die Nullstelle dieser Funktion suchen. 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, sonst 0 ist. Der Übergang soll stetig sein. Desweiteren definieren wir  $\tilde{u} \in H_0^1(\Omega)$ , sodass  $u_1 = \tilde{u} + \tilde{u}_0$ . Damit lässt sich das Problem umschreiben zu

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

$$-\int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u}_0 \nabla \varphi \, dx = \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u} \nabla \varphi \, dx$$

gilt

Zur numerischen Betrachtung bieten sich Finite Elemente, insbesondere die dreieckig-linearen Lagrange Elemente an. Die Theorie wird als bekannt vorausgesetzt und hier nicht weiter besprochen. Zuerst triangulieren wir das Gebiet.

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

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

$h = (h_1, h_2)$  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  $m$  und  $n$  recht nahe beieinander gewählt werden. Die genaue Wahl wird nachher in der Implementierung vorgenommen. Nun wird durch die Gitterpunkte die Triangulierung gelegt. Diese nennen wir  $E_k$  und ist in 3.1 dargestellt.

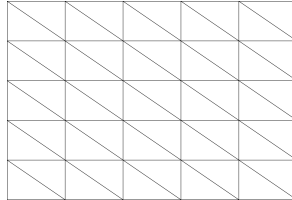


Abbildung 3.1: Triangulierung eines rechteckigen Gebietes

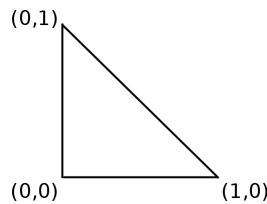


Abbildung 3.2: Referenzdreieck

Stellen wir das Referenzelement unserer Finiten Elemente auf. Es ist gegeben durch  $(E, P, \Psi)$ , wobei  $E$  das Referenzdreieck 3.2 ist.  $P = \mathcal{P}_1$  sind Polynome auf  $\mathbb{R}^2$  vom

Grad 1 mit Basis  $\{p_0, p_1, p_2\}$

$$p_0(x, y) := 1 \quad p_1(x, y) := x \quad p_2(x, y) := y$$

$\Psi := \{\varphi_0, \varphi_1, \varphi_2\}$  sind Funktionale auf  $P$  und damit eine Basis von  $P^*$ .  $\varphi_i$  sind lokale Formfunktionen d.h.  $\varphi_i(p_j) = \delta_{ij}$ ,  $i, j \in \{0, 1, 2\}$ . Dabei ist  $\delta_{ij}$  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

$$\varphi_0 = 1 - x - y, \quad \varphi_1 = x, \quad \varphi_2 = y$$

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

$$T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \pm \begin{pmatrix} h_1 x \\ h_2 y \end{pmatrix}$$

durch das Referenzelement darstellen. Dabei entspricht  $(a_1, a_2)^t$  dem Eckpunkt mit dem  $90^\circ$  Winkel des Rechteckes und  $(h_1, h_2)^t$  ist die Höhe des Dreiecks. Mit dem Transformationssatz können wir alle Berechnungen auf dem Referenzelement ausführen und dann auf das transformierte Element übertragen.

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  $\Omega$  definiert sind. Die globale Formfunktion  $T_j$  ist 1 auf dem dem Gitterpunkt  $j$  und 0 sonst.

Kommen wir wieder zu unserem eigentlichen Zielfunktional zurück. Wir wollten  $\tilde{u} \in H_0^1(\Omega)$  finden, sodass  $\forall \varphi \in H_0^1(\Omega)$

$$-\int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u}_0 \nabla \varphi \, dx = \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u} \nabla \varphi \, dx$$

gilt.

Wir benutzen den Galerkin Ansatz. Dafür gilt ab jetzt  $k := (m + 1)(n + 1)$

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



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

$$\begin{aligned}
& - \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u}_0 \nabla \varphi \, dx = \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u} \nabla \varphi \, dx \\
& \Leftrightarrow - \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u}_0 \nabla T_i \, dx = \int_{\Omega} (v^2 + \epsilon_1) \sum_{i=1}^k u_i^h \nabla T_i \nabla T_j \, dx \\
& \Leftrightarrow - \int_{\Omega} (v^2 + \epsilon_1) \nabla \tilde{u}_0 \nabla T_i \, dx = \sum_{i=1}^k u_i^h \int_{\Omega} (v^2 + \epsilon_1) \nabla T_i \nabla T_j \, dx \\
& \Leftrightarrow b = A * u^h
\end{aligned}$$

wobei  $u^h := (u_1^h, \dots, u_k^h)^T$ ,  $A := \left( \int_{\Omega} (v^2 + \epsilon) \nabla T_i \nabla T_j \, dx \right)_{ij}$  und  $b := \left( \int_{\Omega} (v^2 + \epsilon) \nabla \tilde{u}_0 \nabla T_i \, dx \right)_i$

Also gilt  $u^h = b \setminus A$ . Daraus folgt, dass wir  $A$  und  $b$  berechnen müssen. Betrachten wir zunächst die Matrix  $A$

### Berechnung des $u$ Integrals

Als erste Vereinfacherung betrachten wir nicht mehr das Integral über  $\Omega$ , sondern über die einzelnen Dreiecke der Triangulierung. Desweiteren ist  $T_i$  linear, also  $\nabla T_i$  Konstant. Es gilt:

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

Wir kennen  $\nabla T_i \nabla T_j$  auf jedem Dreieck. Also muss nur noch  $\int_E v^2 + \epsilon \, dx$  berechnet werden. Es darf über das Referenzdreieck integriert werden, da durch den Transformationssatz das Integral über das transformierte Element gewonnen werden kann. Es

gilt:

$$\int_E v^2 + \epsilon_1 \, dx = \int_E v^2 \, dx + \frac{1}{2} \epsilon_1$$

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 Berechnung ist in ?? mit  $v = a$  zu finden.

Nun können wir das Integral berechnen. Dies ist einfach und wird hier nicht weiter ausgeführt.

$$\begin{aligned} \int_E v(x, y)^2 \, dx \, dy &= \int_0^1 \int_0^{1-y} ((v_3 - v_1)x + (v_2 - v_1)y + v_1)^2 \, dx \, dy \\ &= \frac{1}{12} (v_1^2 + v_2^2 + v_3^2 + v_1 v_2 + v_1 v_3 + v_2 v_3) \end{aligned}$$

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

Berechnen wir nun

$$B(T_i, T_j) = \sum_{\tilde{E} \in E_k} \nabla T_i \nabla T_j \int_{\tilde{E}} (v^2 + \epsilon) \, dx$$

$T_i$  ist nur auf einem Gitterpunkt 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 einfach nur drei Dreiecke, an den Ecken entweder ein oder zwei Dreiecke. Für den Rand setzt man einfach die Dreiecke, die nicht existieren gleich 0.  $B(T_i, T_j)$  nimmt für unterschiedliche  $i$  und  $j$  unterschiedliche Werte an.

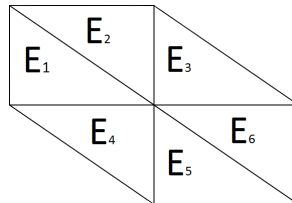


Abbildung 3.3: Triangulierung im Inneren

Jetzt können wir  $B(T_i, T_j)$  berechnen. Für unterschiedliche  $i$  und  $j$  kommen natürlich unterschiedliche Werte heraus. Betrachten wir zunächst die Situation, dass  $i$  und  $j$  nicht benachbart sind. Dann ist  $B(T_i, T_j) = 0$ , da  $\nabla T_i \nabla T_j$  nur auf benachbarten Dreiecken ungleich 0 ist. Daraus ergeben sich vier Fälle:  $i$  und  $j$  sind dasselbe, also  $j = i$ .  $j$  liegt rechts neben  $i$  also  $j = i + 1$ ,  $j$  liegt direkt unter  $i$ ,  $j = i + n + 1$  und  $j$  liegt schräg unter  $i$ , also  $j = i + n + 2$ . Betrachten wir für die einzelnen Berechnungen 3.4.  $T_i$  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 3.4 überein und  $T^k, k \in \{1, 2, 3\}$  ist das Gleiche  $T^k$  wie in (??).

### $i$ und $j$ sind gleich

$$\begin{aligned}
 B(T_i, T_i) &= \sum_{E \in E_k} \nabla T_i \nabla T_i \int_E (v^2 + \epsilon) \, dx \\
 &= \nabla T^0 \nabla T^0 \int_{E_3} (v^2 + \epsilon) \, dx + \nabla T^0 \nabla T^0 \int_{E_4} (v^2 + \epsilon) \, dx \\
 &\quad + \nabla T^1 \nabla T^1 \int_{E_1} (v^2 + \epsilon) \, dx + \nabla T^1 \nabla T^1 \int_{E_6} (v^2 + \epsilon) \, dx \\
 &\quad + \nabla T^2 \nabla T^2 \int_{E_2} (v^2 + \epsilon) \, dx + \nabla T^2 \nabla T^2 \int_{E_5} (v^2 + \epsilon) \, dx
 \end{aligned}$$

### $j$ liegt rechts neben $i$

$$\begin{aligned}
 B(T_i, T_{i+1}) &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+1} \int_E (v^2 + \epsilon) \, dx \\
 &= \nabla T^0 \nabla T^1 \int_{E_3} (v^2 + \epsilon) \, dx + \nabla T^0 \nabla T^1 \int_{E_6} (v^2 + \epsilon) \, dx
 \end{aligned}$$

### $j$ liegt unter $i$

$$\begin{aligned}
 B(T_i, T_{i+1+n}) &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+1+n} \int_E (v^2 + \epsilon) \, dx \\
 &= \nabla T^0 \nabla T^1 \int_{E_4} (v^2 + \epsilon) \, dx + \nabla T^0 \nabla T^1 \int_{E_5} (v^2 + \epsilon) \, dx
 \end{aligned}$$

$j$  liegt schräg unter  $i$

$$\begin{aligned}
 B(T_i, T_{i+2+n}) &= \sum_{E \in E_k} \nabla T_i \nabla T_{i+2+n} \int_E (v^2 + \epsilon) \, dx \\
 &= \nabla T^1 \nabla T^2 \int_{E_5} (v^2 + \epsilon) \, dx + \nabla T^1 \nabla T^2 \int_{E_6} (v^2 + \epsilon) \, dx \\
 &= 0
 \end{aligned}$$

**Zusammenfassung** Mit diesen Werten können wir nun die Matrix  $\int_{\Omega} (v^2 + \epsilon) \nabla u \nabla \varphi$  aufstellen:

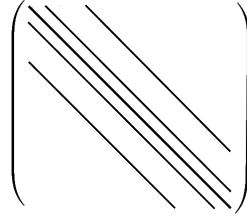


Abbildung 3.4: Triangulierung im Inneren

Dabei sind auf der Diagonalen die Einträge  $B(T_i, T_i)$ , auf der Nebendiagonalen die Einträge  $B(T_i, T_{i+1})$  und auf der anderen Diagonale die Einträge  $B(T_i, T_{i+n+1})$ . 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.

Wir wissen noch, dass auf  $\Gamma_1 \cup \Gamma_2$   $u = 0$  gilt. Deshalb ist auf den Rändern  $B(T_i, T_j) = 0$ . Also muss in der Implementierung für  $i = 1$  und  $i = m$  oder  $j = 1$  und  $j = m$  der entsprechende Eintrag 0 gesetzt werden. Damit entsteht eine singuläre Matrix. Da wir das nicht wollen, setzten wir die Einträge auf der Diagonalen die durch 0 ersetzt worden sind, auf 1. Wenn wir nachher bei der linken Seite die entsprechenden Einträge auch 0 setzen, erhalten wir die gewünschten Randterme.

### Betrachtung des $u_0$ Integrals

Berechnen wir nun numerisch

$$\int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \forall i \in \{1, \dots, k\}$$

Es gilt mit den gleichen Begründungen wie bei der rechten Seite:

$$\begin{aligned} \int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx &= \sum_{E \in E_k} \int_E (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\ &= \sum_{\substack{E \in E_k \text{ und} \\ E \text{ liegt am Rand}}} \int_E (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \end{aligned}$$

da  $u_0$  überall, außer auf den äußeren Dreiecken 0 ist. Damit müssen nur die  $T_i$  betrachtet werden, die am oder neben dem Rand liegen. Alle anderen Einträge sind 0, wie in Grafik 3.5 dargestellt. Damit haben wir 4 Fälle für die Berechnung von dem Integral.  $T_i$  kann auf dem linken oder rechten Rand 1 sein, oder auf einem Dreieck neben dem linken oder rechten Rand 1 sein.


 images/u0\_matrix.png

Abbildung 3.5: Veranschaulichung des Vektors  $\int_{\Omega} (v^2 + \epsilon) \nabla u_0 \nabla T_i$

**$T_i$  ist auf dem Rand 1** Bis jetzt haben wir uns immer die sechs Dreiecke um  $T_i$  angeschaut. Nun reicht es, die drei Dreiecke am Rand zu betrachten, da  $T_i$  am Rand ist, wie in Zeichnung 3.6 dargestellt.

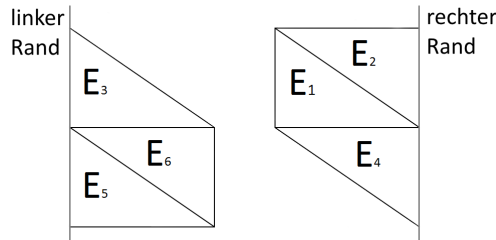


Abbildung 3.6: umliegende Dreiecke von  $T_i$  mit rechten bzw. linken Rand

**linker Rand** Wir erhalten hier

	$T_i$	$\nabla T_i$	$\nabla u_0$	$\nabla T_i \nabla u_0$
$E_3$	$-x - y - 1$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^1 - u_0^2 \end{pmatrix}$	$2u_0^2 - u_0^1$
$E_5$	$y$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^1 - u_0^2 \end{pmatrix}$	$u_0^1 - u_0^2$
$E_6$	$x$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} u_0^1 \\ 0 \end{pmatrix}$	$u_0^1$

Daraus ergibt sich:

$$\begin{aligned}
 & \int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_3} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_5} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_6} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_3} (v^2 + \epsilon_1) (2u_0^2 - u_0^1) \, dx + \int_{E_5} (v^2 + \epsilon_1) (u_0^1 - u_0^2) \, dx + \int_{E_6} (v^2 + \epsilon_1) u_0^1 \, dx
 \end{aligned}$$

**rechter Rand** Es ergibt sich folgende Tabelle

	$T_i$	$\nabla T_i$	$\nabla u_0$	$\nabla T_i \nabla u_0$
$E_1$	$x$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} u_0^3 \\ 0 \end{pmatrix}$	$u_0^3$
$E_2$	$y$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^3 - u_0^2 \end{pmatrix}$	$u_0^3 - u_0^2$
$E_4$	$-x - y - 1$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^3 - u_0^2 \end{pmatrix}$	$2u_0^2 - u_0^3$

Es folgt:

$$\begin{aligned}
 & \int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_1} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_2} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_4} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_1} (v^2 + \epsilon_1) u_0^3 \, dx + \int_{E_2} (v^2 + \epsilon_1) (u_0^3 - u_0^2) \, dx + \int_{E_4} (v^2 + \epsilon_1) (2u_0^2 - u_0^3) \, dx
 \end{aligned}$$

**$T_i$  ist neben dem Rand 1** Bei den  $T_i$  neben dem linken bzw. rechten Rand gehen wir analog vor. Wir betrachten jetzt 3.7.

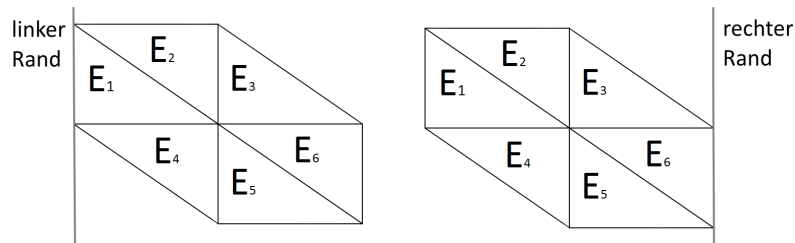


Abbildung 3.7: umliegende Dreiecke von  $T_i$  neben dem rechten bzw. linken Rand

#### linker Rand

	$T_i$	$\nabla T_i$	$\nabla u_0$	$\nabla T_i \nabla u_0$
$E_1$	$x$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^1 - u_0^2 \end{pmatrix}$	$-u_0^2$
$E_2$	$y$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} u_0^1 \\ 0 \end{pmatrix}$	$0$
$E_4$	$-x - y - 1$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} -u_0^1 \\ 0 \end{pmatrix}$	$-u_0^1$

Es folgt:

$$\begin{aligned}
 & \int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_1} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_2} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_4} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= - \int_{E_1} (v^2 + \epsilon_1) u_0^2 \, dx + 0 - \int_{E_4} (v^2 + \epsilon_1) u_0^1 \, dx
 \end{aligned}$$

### rechter Rand

	$T_i$	$\nabla T_i$	$\nabla u_0$	$\nabla T_i \nabla u_0$
$E_3$	$-x - y - 1$	$\begin{pmatrix} -1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} u_0^3 \\ 0 \end{pmatrix}$	$-u_0^3$
$E_5$	$y$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} -u_0^3 \\ 0 \end{pmatrix}$	$0$
$E_6$	$x$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -u_0^2 \\ u_0^3 - u_0^2 \end{pmatrix}$	$-u_0^2$

Es ergibt sich

$$\begin{aligned}
 & \int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= \int_{E_3} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_5} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx + \int_{E_6} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i \, dx \\
 &= - \int_{E_3} (v^2 + \epsilon_1) u_0^3 \, dx + 0 - \int_{E_6} (v^2 + \epsilon_1) u_0^2 \, dx
 \end{aligned}$$

**Zusammenfassung** Durch die vorherigen Berechnungen kommen wir auf den Vektor 3.8, der fast nur aus Nullen besteht. Nur die Einträge, die zu dem Rand des Gitters oder direkt neben dem Rand des Gitters liegen, sind ungleich 0. Dieses zeigt 3.5


 images/u0\_vektor.png

Abbildung 3.8: Vektor  $\int_{\Omega} (v^2 + \epsilon_1) \nabla u_0 \nabla T_i$



Es fehlt noch die Transformation der Dreiecke, über die wir integriert haben. Also multiplizieren wir hier auch wieder den Vektor mit  $1/h_1h_2$ .

### 3.2.3 Zusammenfassung

Damit haben wir beide Seiten diskretisiert und können nun die Matrizen implementieren. Wir wollen

$$\frac{1}{h_1h_2}Au = \frac{1}{h_1h_2}b \Leftrightarrow Au = b$$

berechnen. Der Code dazu hat folgende Form:

1. Berechne Matrix  $A$
2. Berechne Vektor  $b$
3.  $u = b \setminus A$

**Algorithm 4:** Berechnung von  $u$

Da sowohl  $A$  als auch  $b$  aus fast nur Nullen besteht, verwende ich in Matlab Sparse Matrizen. Dies führt zu einer wesentlich kürzeren Laufzeit.

## 3.3 Optimierung nach $v$

Die Optimierung nach  $v$  ist ein Optimierungsproblem mit einfacher Nebenbedingung und lässt sich in folgender Form schreiben:

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

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

In diesem Fall bedeutet das also, dass

$$\begin{aligned} J : H^1(\Omega) &\rightarrow \mathbb{R} \\ v &\mapsto \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\ C &:= \{v \in H^1(\Omega) | 0 \leq v \leq v_0\} \end{aligned}$$

**Theorem 3.3.1.** *Das Problem (??) besitzt genau eine Lösung, falls  $v_0$  stetig ist.*

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

- $W = H^1(\Omega)$  ist ein Hilbertraum, also ist er ein reflexiver Banachraum.
- 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  $0 \leq v_n \leq v_0 \quad \forall n \in \mathbb{N}$ . Es gilt auch  $0 \leq \lim_{n \rightarrow \infty} v_n \leq v_0$ . Also ist  $C$  abgeschlossen.

Für Konvexität 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.

- $J$  ist strikt konvex. Der Beweis dazu kann durch einfaches nachrechnen geführt werden. Für Stetigkeit gilt dasselbe.
- Sei  $w \in C$  mit  $\|v\|_{H^1(\Omega)} \rightarrow \infty$ . Dann gilt

$$\begin{aligned}
 J(v) &= \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 dx \\
 &= \int_{\Omega} v^2 |\nabla u|^2 + \epsilon_1 |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - 2v + v^2) dx \\
 &= \int_{\Omega} v^2 |\nabla u|^2 - \frac{2}{\epsilon_3} v + \frac{1}{\epsilon_3} v^2 dx + \int_{\Omega} \epsilon_1 |\nabla u|^2 + \frac{1}{\epsilon_3} dx + \int_{\Omega} \epsilon_2 |\nabla v|^2 dx \\
 &\leq \int_{\Omega} v^2 \left( |\nabla u|^2 + \frac{1}{\epsilon_3} \right) dx + c + \epsilon_2 \|\nabla v\|_{L^2(\Omega)}^2 \\
 &\leq c' \|v\|_{L^2(\Omega)}^2 + c + \epsilon_2 \|\nabla v\|_{L^2(\Omega)}^2 \\
 &\leq c'' \left( \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \right) + c \\
 &\leq c'' \|v\|_{H^1(\Omega)}^2 + c \\
 &\rightarrow \infty
 \end{aligned}$$

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

Alle Voraussetzungen aus 2.1.6 sind erfüllt, also existiert genau eine Lösung des Optimierungproblems.  $\square$

Nun stellen wir das KKT-System auf. .

**Theorem 3.3.2.** Sei  $a := \inf \{J(w) | G(w) \leq_P 0\}$ . Dann gilt:

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

*Beweis.* Die Bedingungen aus 2.1.7 müssen gelten: Sei  $P := \{(v, w) \in H^1(\Omega) \times$

$H^1(\Omega)|v \geq 0 \text{ und } w \geq 0\} \subset H^1(\Omega) \times H^1(\Omega)$ .  $\overset{\circ}{P} \neq \emptyset$ , da  $H^1(\Omega)$  nur stetige Funktionen enthält. Also ist  $P$  ein positiver Kegel.

$J : H^1(\Omega) \rightarrow \mathbb{R}$  sei wie oben definiert.

$$G : H^1(\Omega) \rightarrow H^1(\Omega)$$

$$v \mapsto \begin{pmatrix} -v \\ v - v_0 \end{pmatrix}$$

$G$  ist linear, also konvex.

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

Außerdem ist  $a := \inf\{J(w) | G(w) \leq_P 0\} < \infty$ , 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

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

□

Um das KKT System konkret angeben brauchen wir die Gâteaux Ableitung von  $J$ .

**Theorem 3.3.3.**  *$J$  ist Gâteaux-Differenzierbar mit*

$$J'(v) : \overline{H^1}(\Omega) \rightarrow \mathbb{R}$$

$$s \mapsto \left( 2v|\nabla(u)|^2 - \epsilon_2 2\Delta v - \frac{2}{\epsilon_3}(1-v), s \right)_{L^2(\Omega)} + (2\epsilon_2 \nabla v \nu, s)_{L^2(\partial\Omega)}$$

*Beweis.* Zunächst kommt die Richtungsableitung:

$$\begin{aligned}
\partial J(v, s) &= \lim_{t \rightarrow 0} \frac{1}{t} \left( J(v + ts) - J(v) \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} ((v + ts)^2 + \epsilon_1) |\nabla(u)|^2 + \epsilon_2 |\nabla(v + ts)|^2 + \frac{1}{\epsilon_3} (1 - (v + ts))^2 \right. \\
&\quad \left. - \int_{\Omega} (v^2 + \epsilon_1) |\nabla u|^2 + \epsilon_2 |\nabla v|^2 + \frac{1}{\epsilon_3} (1 - v)^2 \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (((v + ts)^2 + \epsilon_1) - (v^2 + \epsilon_1)) |\nabla(u)|^2 \right. \\
&\quad \left. + \epsilon_2 (|\nabla(v + ts)|^2 - |\nabla v|^2) \right. \\
&\quad \left. + \frac{1}{\epsilon_3} ((1 - v - ts)^2 - (1 - v)^2) \right) \\
&= \lim_{t \rightarrow 0} \frac{1}{t} \left( \int_{\Omega} (v^2 + 2vts + t^2 s^2 - v^2) |\nabla u|^2 \right. \\
&\quad \left. + \epsilon_2 (|\nabla v|^2 + 2t \nabla v \nabla s + t^2 |\nabla s|^2 - |\nabla v|^2) \right. \\
&\quad \left. + \frac{1}{\epsilon_3} ((1 - v)^2 - 2(1 - v)ts + t^2 s^2 - (1 - v)^2) \right) \\
&= \lim_{t \rightarrow 0} \left( \int_{\Omega} (2vs + ts^2) |\nabla u|^2 + \epsilon_2 (2 \nabla v \nabla s + t |\nabla s|^2) \right. \\
&\quad \left. - \frac{1}{\epsilon_3} (2(1 - v)s + ts^2) dx \right) \\
&= \int_{\Omega} 2sv |\nabla u|^2 + \epsilon_2 2 \nabla v \nabla s - \frac{2}{\epsilon_3} (1 - v)s dx \\
&= \int_{\Omega} 2v |\nabla u|^2 s - \epsilon_2 2 \Delta v s - \frac{2}{\epsilon_3} (1 - v)s dx + \int_{\partial \Omega} 2\epsilon_2 \nabla v \nu s dx
\end{aligned}$$

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

Aus  $\nabla J(v) + \lambda - \mu = 0$  folgt, dass

$$\begin{aligned}
2v |\nabla u|^2 - \epsilon_2 2 \Delta v - \frac{2}{\epsilon_3} (1 - v) + \lambda - \mu &= 0 && \text{auf } \Omega \\
2\epsilon_2 \nabla v \nu &= 0 && \text{auf } \partial \Omega
\end{aligned}$$

Damit lautet das KKT System:

$$\begin{aligned}
2\bar{v} |\nabla u|^2 - \epsilon_2 2 \Delta \bar{v} - \frac{2}{\epsilon_3} (1 - \bar{v}) + \lambda - \mu &= 0 \text{ auf } \Omega \\
2\epsilon_2 \nabla \bar{v} \nu &= 0 \text{ auf } \partial \Omega \\
\bar{v} &\geq a \quad \mu \geq 0 \quad \mu \bar{v} = 0 \\
\bar{v} &\leq b \quad \lambda \geq 0 \quad \lambda (v_0 - \bar{v}) = 0
\end{aligned}$$

Die Projektion für die Nebenbedingung lautet nach (??):

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

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

$$\begin{aligned} 2\bar{v}|\nabla u|^2 - \epsilon_2 2\Delta \bar{v} - \frac{2}{\epsilon_3}(1 - \bar{v}) + \eta &= 0 \text{ auf } \Omega \\ 2\epsilon_2 \nabla \bar{v} \nu &= 0 \text{ auf } \partial\Omega \\ \eta &= \max\{0, \eta + c(\bar{v} - v_0)\} + \min\{0, \eta + c\bar{v}\} \quad \forall c > 0 \end{aligned}$$

gilt. Die schwache Formulierung ist dann

$$\begin{aligned} \int_{\Omega} 2\varphi v |\nabla u|^2 + \epsilon_2 2\nabla v \nabla \varphi - \frac{2}{\epsilon_3}(1 - v)\varphi + \eta \varphi \, dx &= 0 \quad \forall \varphi \in H^1(\Omega) \\ \int_{\Omega} (\eta - \max\{0, \eta + c(\bar{v} - v_0)\} - \min\{0, \eta + c\bar{v}\}) \varphi \, dx &= 0 \quad \forall c > 0, \forall \varphi \in H^1(\Omega) \end{aligned}$$

mit  $\eta = \mu - \lambda$

### 3.3.1 Newtonmethode

Wir wollen das Problem implementieren, indem wir die Newton Methode anwenden. Diese sieht wie folgt aus:

**Data:**  $u^0$  (möglichst nah an der Lösung  $\bar{w}$ )  
**for**  $k = 0, 1, \dots$  **do**  
    | Löse  $G'(w^k)s^k = -G(w^k);$   
    |  $w^{k+1} = w^k + s^k;$   
**end**

**Algorithm 5:** Newton Methode

Hier betrachten wir die Funktion

$$G : H^1(\Omega) \times H^1(\Omega) \rightarrow H^{-1}(\Omega)^2$$

$$(v, \eta) \mapsto \begin{pmatrix} \int_{\Omega} 2\varphi v |\nabla u|^2 + \epsilon_2 2\nabla v \nabla \varphi - \frac{2}{\epsilon_3} (1-v)\varphi + \eta \varphi \, dx \\ \int_{\Omega} (\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}) \varphi \, dx \end{pmatrix}$$

Nun brauchen wir die Ableitung. Da  $G_2$  offensichtlich keine Gâteaux-Ableitung hat, brauchen wir das Semidifferenzial. Bei  $G_1$  entspricht das Semidifferenzial der Gâteaux-Ableitung. Diese lässt sich einfach hinschreiben mit der Richtung  $\phi$ .

$$G_{1v}(v, \eta) = \int_{\Omega} 2\varphi \phi |\nabla u|^2 + \epsilon_2 2\nabla \phi \nabla \varphi + \frac{2}{\epsilon_3} \phi \varphi \, dx$$

$$G_{1\eta}(v, \eta) = \int_{\Omega} \phi \varphi \, dx$$

Das Semidifferenzial von  $G_2$  ist nicht ganz so einfach. Beweisen wir zunächst ein Lemma

**Lemma 3.3.4.** *Betrachte  $f : H^1(\Omega)^2 \rightarrow H^{-1}(\Omega)$  mit*

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

*Dann ist  $f$  semidifferenzierbar mit*

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

*und*

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

*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  $\eta$ . 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 2.4.8 gilt, dass, falls  $f$  stetig Fréchet-Differenzierbar ist,  $f$   $\partial f$  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 2.4.8 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 2.11 muss gelten. Hier ist  $\partial f(\eta + d, v) = \{0\}$  und damit

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

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

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

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

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

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

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

Damit ist  $f$  semidifferenzierbar nach  $\eta$ . Für die semidifferenzierbarkeit nach  $v$  gilt die gleiche Rechnung.  $\square$

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

**Theorem 3.3.5.**  $G_2 : H^1(\Omega)^2 \rightarrow H^{-1}(\Omega)$  mit

$$(v, \eta) \mapsto \int_{\Omega} (\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}) \varphi \, dx$$

ist semidifferenzierbar mit

$$\partial G_{2\eta}(\eta, v)(\varphi, \phi) = \int_{\Omega} \frac{\partial f}{\partial \eta} \varphi \phi \, dx$$

$$\partial G_{2v}(\eta, v)(\varphi, \phi) = \int_{\Omega} \frac{\partial f}{\partial v} \varphi \phi \, dx$$

*Beweis.*  $\square$

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

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

### 3.3.2 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.

$$v = \sum_{i=1}^k v_i^h T_i$$

wobei die  $T_i$  wieder die globalen Formfunktionen sind.

Da  $u$  schon durch den vorherigen Iterationsschritt gegeben ist, ist  $u$  ein Vektor mit den Auswertungen an den Ecken der Dreiecke. Die Darstellung ist die gleiche wie in ??  
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 $G_1$

$$G_1(v, \eta) = \int_{\Omega} 2\varphi v |\nabla u|^2 + \epsilon_2 2\nabla v \nabla \varphi - \frac{2}{\epsilon_3} (1 - v) \varphi + \eta \varphi \, dx$$

wird nun diskretisiert:

$$\begin{aligned} & \int_{\Omega} 2\varphi v |\nabla u|^2 + 2\epsilon_2 \nabla v \nabla \varphi - \frac{2}{\epsilon_3} (1 - v) \varphi + \eta \varphi \, dx \\ &= \int_{\Omega} 2T_j \sum_{i=1}^k v_i^h T_i u^{dis} + \epsilon_2 2\nabla \left( \sum_{i=1}^k v_i^h T_i \right) \nabla T_j - \frac{2}{\epsilon_3} \left( 1 - \sum_{i=1}^k v_i^h T_i \right) T_j + \eta T_j \, dx \\ &= 2 \sum_{i=1}^k v_i^h \int_{\Omega} u^{dis} T_i T_j \, dx + 2\epsilon_2 \sum_{i=1}^k v_i^h \int_{\Omega} \nabla T_i \nabla T_j \, dx - \frac{2}{\epsilon_3} \sum_{i=1}^k \int_{\Omega} T_j \, dx \\ & \quad + \frac{2}{\epsilon_3} \sum_{i=1}^k \alpha_i \int_{\Omega} T_i T_j \, dx + \int_{\Omega} \eta T_j \, dx \\ &= 2Av^h + 2\epsilon_2 v^h B - \frac{2}{\epsilon_3} c + \frac{2}{\epsilon_3} Dv^h \\ &= (2A + 2\epsilon_2 B + \frac{2}{\epsilon_3} D)v^h - \frac{2}{\epsilon_3} c + e \end{aligned}$$

mit  $v^h := (v_1^h, \dots, v_k^h)^T$ ,  $A_{ij} := \int_{\Omega} u^{dis} T_i T_j \, dx$ ,  $B_{ij} := \int_{\Omega} \nabla T_i \nabla T_j \, dx$ ,  $c_j := \int_{\Omega} T_j \, dx$ ,  $D_{ij} := \int_{\Omega} T_i T_j \, dx$  und  $e_j := \int_{\Omega} \eta T_j \, dx$ .

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 Ein-

heitsdreieck 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**  $A_{ij} = \int_{\Omega} u^{dis} T_i T_j$

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

$$\begin{aligned}
 A_{i,i+1} &= \int_{\Omega} u^{dis} T_i T_{i+1} = \int_{E_3} u_{E_3}^{dis} \varphi_0 \varphi_1 \, dx + \int_{E_6} u_{E_6}^{dis} \varphi_0 \varphi_1 \, dx \\
 &= \frac{1}{24} u_{E_3}^{dis} + \frac{1}{24} u_{E_6}^{dis} = \frac{1}{24} (u_{E_3}^{dis} + u_{E_6}^{dis})
 \end{aligned}$$

$$\begin{aligned}
A_{i,i+1+n} &= \int_{\Omega} u^{dis} T_i T_{i+1+n} = \int_{E_4} u_{E_4}^{dis} \varphi_0 \varphi_1 \, dx + \int_{E_5} u_{E_5}^{dis} \varphi_0 \varphi_1 \, dx \\
&= \frac{1}{24} u_{E_4}^{dis} + \frac{1}{24} u_{E_5}^{dis} = \frac{1}{24} (u_{E_4}^{dis} + u_{E_5}^{dis})
\end{aligned}$$

$$\begin{aligned}
A_{i,i+n+2} &= \int_{\Omega} u^{dis} T_i T_{i+2+n} = \int_{E_5} u_{E_5}^{dis} \varphi_1 \varphi_2 \, dx + \int_{E_6} u_{E_6}^{dis} \varphi_1 \varphi_2 \, dx \\
&= \frac{1}{24} u_{E_5}^{dis} + \frac{1}{24} u_{E_6}^{dis} = \frac{1}{24} (u_{E_5}^{dis} + u_{E_6}^{dis})
\end{aligned}$$

**Berechnung der Matrix**  $B_{ij} = \int_{\Omega} \nabla T_i \nabla T_j$

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

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

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



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

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

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

### Berechnung des Vektors $\mathbf{e}$

$$B(T_i) = \int_{\Omega} \eta T_i$$

$\eta$  ist auf dem ungeraden Dreieck gegeben durch

$$\eta(x, y) = (\eta_3 - \eta_2)x + (\eta_1 - \eta_2)y + \eta_2$$

und auf dem geraden gegeben durch

$$\eta(x, y) = (\eta_1 - \eta_2)x + (\eta_3 - \eta_2)y + \eta_2$$

Dadurch ergibt sich

	$T_i$	$\int_{E_i} \eta T_i$
$E_1$	$x$	$\eta_1 + \eta_2 + 2\eta_3$
$E_2$	$y$	$\eta_1 + \eta_2 + 2\eta_3$
$E_3$	$1 - x - y$	$\eta_1 + 2\eta_2 + \eta_3$
$E_4$	$1 - x - y$	$\eta_1 + 2\eta_2 + \eta_3$
$E_5$	$y$	$2\eta_1 + \eta_2 + \eta_3$
$E_6$	$x$	$2\eta_1 + \eta_2 + \eta_3$

In der Tabelle entspricht  $\eta_1$  in der Spalte für  $E_i$  nicht  $\eta_1$  in der Spalte für  $E_j$ , da die

1 für das jeweilige Dreieck angibt, dass der Wert oben rechts in der Ecke steht. Das selbe gilt auch für  $\eta_2$  und  $\eta_3$ .

Den Vektor  $e$  kann man berechnen, indem man die Werte addiert. Dabei gilt wieder, dass, falls  $i$  am Rand liegt, die nicht existenten Dreiecke ausgelassen werden.

**Zusammenfassung** Also ergibt sich

$$(2A + 2\epsilon_2 B + \frac{2}{\epsilon_3} D)\alpha - \frac{2}{\epsilon_3} c + e =$$

$$\left(2 \begin{pmatrix} \diagup \\ \diagup \\ \diagup \\ \diagup \end{pmatrix} + 2\epsilon \begin{pmatrix} \diagup \\ \diagup \\ \diagup \end{pmatrix} + \frac{2}{\epsilon} \begin{pmatrix} \diagup \\ \diagup \\ \diagup \\ \diagup \end{pmatrix}\right) \alpha - \frac{2}{\epsilon} |\Omega| \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

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

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

Der Vektor  $c$  hat bei allen Einträgen, die nicht zu einem Randpunkt gehören eine 1, bei Einträgen am Rand und nicht in einer Ecke, eine  $\frac{1}{2}$ , an der linken oberen und der rechten unteren Ecke eine  $\frac{1}{3}$  und der Eintrag auf den anderen beiden Ecken ist  $\frac{1}{6}$ .

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

Der Vektor  $e$  hat überall Einträge. Da diese von  $\eta$  abhängen, haben sie keine erkennbare Form.

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

**numerische Darstellung von  $G_2$** 

$$\int_{\Omega} (\eta - \max\{0, \eta + c(v - v_0)\} - \min\{0, \eta + cv\}) \varphi \, dx$$

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

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

Daraus ergibt sich:

$$\begin{aligned}
& \int_{\Omega} (\eta - \max \{0, \eta + c(v - v_0)\} - \min \{0, \eta + cv\}) \varphi \, dx \\
&= \int_{\Omega} \left( \sum_{i=1}^k \eta_i^h T_i - \max \left\{ 0, \sum_{i=1}^k \eta_i^h T_i + c \left( \sum_{i=1}^k v_i^h T_i - \sum_{i=1}^k v_{0i}^h T_i \right) \right\} \right. \\
&\quad \left. - \min \left\{ 0, \sum_{i=1}^k \eta_i^h T_i + c \sum_{i=1}^k v_i^h T_i \right\} \right) T_j \, dx \\
&= \int_{\Omega} \left( \sum_{i=1}^k \eta_i^h T_i - \max \left\{ 0, \sum_{i=1}^k (\eta_i^h + c(v_i^h - v_{0i}^h)) T_i \right\} \right. \\
&\quad \left. - \min \left\{ 0, \sum_{i=1}^k (\eta_i^h + cv_i^h) T_i \right\} \right) T_j \, dx \\
&= \int_{\Omega} \left( \sum_{i=1}^k \eta_i^h T_i - \sum_{i=1}^k \max \{0, \eta_i^h + c(v_i^h - v_{0i}^h)\} T_i \right. \\
&\quad \left. - \sum_{i=1}^k \min \{0, \eta_i^h + cv_i^h\} T_i \right) T_j \, dx \\
&= \int_{\Omega} \sum_{i=1}^k (\eta_i^h - \max \{0, \eta_i^h + c(v_i^h - v_{0i}^h)\} - \min \{0, \eta_i^h + cv_i^h\}) T_i T_j \, dx \\
&= \left( \sum_{i=1}^k \eta_i^h - \max \{0, \eta_i^h + c(v_i^h - v_{0i}^h)\} - \min \{0, \eta_i^h + cv_i^h\} \right) \int_{\Omega} T_i T_j \, dx \\
&= Dw_{v\eta}
\end{aligned}$$

mit  $D$  aus der numerischen Darstellung von  $G_1$  und

$(w_{v\eta})_i := \eta_i^h - \max \{0, \eta_i^h + c(v_i^h - v_{0i}^h)\} - \min \{0, \eta_i^h + cv_i^h\}$ .  $w_{v\eta}$  kann auch explizit dargestellt werden:

$$(w_{v\eta})_i = \begin{cases} -c(v_i^h - v_{0i}^h) & \text{falls } -c(v_i^h - v_{0i}^h) \leq \eta_i^h \\ \eta_i^h & \text{falls } -cv_i^h < \eta < -c(v_i^h - v_{0i}^h) \\ -cv_i^h & \text{falls } \eta_i^h \leq -cv_i^h \end{cases}$$



**numerische Darstellung von  $G_{1v}$** 

$$G_{1v}(v, \eta) = \int_{\Omega} 2\varphi\phi|\nabla u|^2 + \epsilon_2 2\nabla\phi\nabla\varphi + \frac{2}{\epsilon_3}\phi\varphi \, dx$$

wird nun diskretisieren:

$$\begin{aligned} & \int_{\Omega} 2\varphi\phi|\nabla u|^2 + \epsilon_2 2\nabla\phi\nabla\varphi + \frac{2}{\epsilon_3}\phi\varphi \, dx \\ &= \int_{\Omega} 2 \sum_{j=1}^k T_j \sum_{i=1}^k T_i b + \epsilon_2 2\nabla\left(\sum_{i=1}^k T_i\right)\nabla\left(\sum_{j=1}^k T_j\right) + \frac{2}{\epsilon_3} \sum_{i=1}^k T_i \sum_{j=1}^k T_j \, dx \\ &= 2 \sum_{i,j=1}^k \int_{\Omega} b T_i T_j \, dx + 2\epsilon_2 \sum_{i,j=1}^k \int_{\Omega} \nabla T_i \nabla T_j \, dx + \frac{2}{\epsilon_3} \sum_{i,j=1}^k \int_{\Omega} T_i T_j \, dx \\ &= 2A + 2\epsilon_2 B + \frac{2}{\epsilon_3} D \end{aligned}$$

Wir benutzen die gleichen Notationen, wie bei der numerischen Darstellung von  $G_1$ .

**numerische Darstellung von  $G_{1\eta}$** 

$$G_{1\eta}(v, \eta) = \int_{\Omega} \phi\varphi \, dx$$

wird nun diskretisieren:

$$\int_{\Omega} \phi\varphi \, dx = \int_{\Omega} \sum_{j=1}^k T_j \sum_{i=1}^k T_i = D$$

Wir benutzen die gleichen Notationen, wie bei der numerischen Darstellung von  $G_1$ .

**numerische Darstellung von  $G_{2v}$** 

Es soll

$$\partial G_{2v}(\eta, v)(\varphi, \phi) = \int_{\Omega} \frac{\partial f}{\partial v} \varphi \phi \, dx$$

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. Diese diskrete Funktion hat eine explizite Darstellung auf den Dreiecken. Diese ist in ?? dargestellt.

Nun wird  $\partial G_{2v}(\eta, v)(\varphi, \phi)$  diskretisiert. Hier wird wie immer  $\varphi, \phi$  durch die globalen Formfunktionen  $T_i$  ersetzt und  $\Omega$  durch die Vereinigung aller Dreiecke. Nun kann für jedes Dreieck  $\int_E \frac{\partial f}{\partial v} T_i T_j \, dx$  berechnet werden. Dabei ist wieder zu beachten, dass für gerade und ungerade Dreiecke andere Ergebnisse zustande kommen:

$ij$	$\int_E \frac{\partial f}{\partial v} T_i T_j \, dx$ gerades Dreieck	$\int_E \frac{\partial f}{\partial v} T_i T_j \, dx$ ungerades Dreieck
00	$\frac{1}{60}(f_1^h + 3f_2^h + f_3^h)$	$\frac{1}{60}(f_1^h + 3f_2^h + f_3^h)$
11	$\frac{1}{60}(f_1^h + f_2^h + 3f_3^h)$	$\frac{1}{60}(3f_1^h + f_2^h + f_3^h)$
22	$\frac{1}{60}(3f_1^h + f_2^h + f_3^h)$	$\frac{1}{60}(f_1^h + f_2^h + 3f_3^h)$
01	$\frac{1}{120}(f_1^h + 2f_2^h + 2f_3^h)$	$\frac{1}{120}(2f_1^h + 2f_2^h + f_3^h)$
02	$\frac{1}{120}(2f_1^h + 2f_2^h + f_3^h)$	$\frac{1}{120}(f_1^h + 2f_2^h + 2f_3^h)$
12	$\frac{1}{120}(2f_1^h + f_2^h + 2f_3^h)$	$\frac{1}{120}(2f_1^h + f_2^h + 2f_3^h)$

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  $\partial G_{2v}$  diskretisieren. Wir nennen die Diskretisierung  $F_{ij}$ . Hier hat man wieder die vier Fälle:

$$\begin{aligned}
F_{i,i} &= \int_{\Omega} f^h T_i T_i \\
&= \int_{E_3} f_{E_3}^h \varphi_0 \varphi_0 \, dx + \int_{E_6} f_{E_6}^h \varphi_1 \varphi_1 \, dx + \int_{E_5} f_{E_5}^h \varphi_2 \varphi_2 \, dx \\
&\quad + \int_{E_4} f_{E_4}^h \varphi_0 \varphi_0 \, dx + \int_{E_1} f_{E_1}^h \varphi_1 \varphi_1 \, dx + \int_{E_2} f_{E_2}^h \varphi_2 \varphi_2 \, dx \\
&= \frac{1}{60} \left( (f_1^h + f_2^h + 3f_3^h)_{E_1} + (f_1^h + f_2^h + 3f_3^h)_{E_2} + (f_1^h + 3f_2^h + f_3^h)_{E_3} \right. \\
&\quad \left. + (f_1^h + 3f_2^h + f_3^h)_{E_4} + (3f_1^h + f_2^h + f_3^h)_{E_5} + (3f_1^h + f_2^h + f_3^h)_{E_6} \right)
\end{aligned}$$

$$\begin{aligned}
A_{i,i+1} &= \int_{\Omega} f^h T_i T_{i+1} = \int_{E_3} f_{E_3}^h \varphi_0 \varphi_1 \, dx + \int_{E_6} f_{E_6}^h \varphi_0 \varphi_1 \, dx \\
&= \frac{1}{120} \left( (f_1^h + 2f_2^h + 2f_3^h)_{E_3} + (2f_1^h + 2f_2^h + f_3^h)_{E_6} \right)
\end{aligned}$$

$$\begin{aligned}
A_{i,i+1+n} &= \int_{\Omega} f^h T_i T_{i+1+n} = \int_{E_4} f_{E_4}^h \varphi_0 \varphi_1 \, dx + \int_{E_5} f_{E_5}^h \varphi_0 \varphi_1 \, dx \\
&= \frac{1}{120} \left( (2f_1^h + 2f_2^h + f_3^h)_{E_4} + (f_1^h + 2f_2^h + 2f_3^h)_{E_5} \right)
\end{aligned}$$

$$\begin{aligned}
A_{i,i+n+2} &= \int_{\Omega} f^h T_i T_{i+2+n} = \int_{E_5} f_{E_5}^h \varphi_1 \varphi_2 \, dx + \int_{E_6} f_{E_6}^h \varphi_1 \varphi_2 \, dx \\
&= \frac{1}{120} \left( (2f_1^h + f_2^h + 2f_3^h)_{E_5} + (2f_1^h + f_2^h + 2f_3^h)_{E_6} \right)
\end{aligned}$$

hierbei bedeutet  $(f_1^h + f_2^h + f_3^h)_{E_j}$ , dass  $f_i^h$   $f$  auf dem  $i$ -ten Gitterpunkt des Dreieck  $E_j$  ausgewertet wird.

### 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.

### **Zusammenfassung**

Nun sind alle Funktionen diskretisiert und das Problem kann implementiert werden.

# Literaturverzeichnis

[opt] Optimazation with PDE Constraints