Technische Universität Berlin
Fachgebiet Statik der Baukonstruktionen
Univ.-Prof. Habil. Dr.-Ing. R. Harbord
Betreuer: Dipl. -Ing. Boris Reyher

DIPLOMARBEIT

Untersuchungen zur Stabilität von geometrisch nichtlinearen Berechnungen

vorgelegt von:
cand.-Ing. Johannes Höhne
Matr.Nr.: 177281
September 2003

Ich widme diese Arbeit
meinem
Hauptschullehrer Herrn Schroeder
und meinem
Nachhilfelehrer Herrn Dänhardt
ohne die mein Traum vom Studium nie wahr geworden wäre.

Inhaltsverzeichnis

Kapitel 1
Einleitung und Motivation.....4
1.1 Beschreibung des Themas.....4
1.2 Art der Beschreibung.....4

Kapitel 2
Theorie.....6
2.1 Geometrische Nichtlinearität 6
2.1.1 Definitionen und Gedankenmodell.....6
2.1.2 Herleitung des Cauchy-Greenschen Verzerrungstensors.....8
2.1.3 Veranschaulichung und abgeleitete Größen.....13
2.2 Numerische Berechnungsverfahren.....23
2.2.1 Numerische Fehler und Pseudoarithmetik.....23
2.2.2 Konditionszahl von linearen Gleichungssystemen.....29
2.2.3 Bezug auf den Cauchy - Greenschen - Verzerrungstensor.....31
2.2.4 Cholesky Zerlegung.....33
2.2.5 Ausblick und Lösungsstrategie.....36

Kapitel 3
Bewertung der Ergebnisse.....39
3.1 Umfassendes Beispiel.....39
3.1.1 Zeitschrittintegration mit Prediktor Korrektor Verfahren.....40
3.1.2 Zugversagen.....46
3.1.3 Druckversagen.....50
3.2 Abschließende Bemerkung und Ausblick.....55

Kapitel 4
Verzeichnisse.....56
4.1 Gleichungen.....56
4.2 Abbildungen.....57
4.3 Literaturverzeichnis.....58

Kapitel 1

Einleitung und Motivation

1.1 Beschreibung des Themas

Um große Verformungen algorithmisch abbilden zu können, kann man auf die geometrische Nichtlinearität unmöglich verzichten. Der Cauchy-Greensche Verzerrungstensor scheint ein Mittel zu sein, um große Verzerrungen gut beschreiben zu können. Bei sehr großen Verformungen kann ein Zustand angenähert werden, bei dem Dimensionen verloren gehen würden. Man kann einen Würfel beispielsweise so stark zusammendrücken, dass er wie ein zweidimensionaler Gegenstand erscheint. Der Cauchy-Greensche Verzerrungstensor schließt diesen Zustand allerdings per Definition aus und reagiert mit exponentialen Verhalten. Dadurch erhält man Gleichungssysteme mit sehr großen Koeffizienten. Diese Arbeit soll beweisen, dass es dieses Verhalten geben muss und dass die Gruppe der direkten numerischen Gleichungslöser große Schwierigkeiten dabei hat, diese Systeme zu lösen.

1.2 Art der Beschreibung

Um sich einem neuen Thema in der Physik zu nähern, haben sich zwei gegensätzliche Wege über die Zeit herausgebildet. Der eine ist der Weg aus der physikalische Anschauung heraus. Beschreitet man ihn, so beobachtet und misst man, um daraus einen allgemeinen Schluss zu ziehen. Der andere ist der mathematische Weg. Verwendet man ihn, so versucht man physikalische Eigenschaften in Funktionen einzuprägen. Leider ist keiner der beiden Wege ideal. Gaspard Gustave de Coriolis beschritt vor etwa 180 Jahren den streng mathematischen Weg, als er die nach ihm benannte Corioliskraft durch eine saubere Ableitung der Bewegungsgleichungen entdeckte. Physikalisch offensichtlich ist die Corioliskraft sicherlich nicht, was ja für den mathematischen Weg sprechen würde. Andererseits wurden gerade in den letzten Jahren finite Elemente gebaut, die von der mathematischen Seite kommend, die Wirklichkeit nicht oder nur unzureichend beschreiben. Die Frage ist nun, wie man sich selbst entscheidet. Wie nähert man sich einem neuen Thema und wie vermittelt man es anderen? Klar ist nur folgendes: Die Mathematik kann durchaus physikalisch vollkommen unsinnige Dinge beschreiben. Die physikalisch sinnvollen Beschreibungen sind nur eine Teilmenge aller möglichen Beschreibungen. Wohlgemerkt, eine Teilmenge nicht aber eine Schnittmenge. Das heißt, dass man zum derzeitigen Stand der Forschung davon ausgeht, dass sich jeder physikalisch Sachverhalt mathematisch beschreiben lässt[1].

[1] Hierzu: "Sieben Experimente, die die Welt verändern können", Pupert Sheldrake, 1994, Scherz Verlag ISBN 3502-13653x

In dieser Arbeit soll der mathematische Weg beschritten werden. Allerdings wird vor jedem Schritt geklärt, wohin er gehen soll und nach jedem Schritt wird dieser aus physikalischer Sicht geprüft. Dadurch soll ein anschauliches, leicht verständliches und dennoch exaktes Bild erzeugt werden.

Da die FEM ein computerorientiertes Verfahren ist, kommt hier konsequent die Indexschreibweise zum Einsatz. Diese lässt sich leichter implementieren, da die Variablen der Schleifen bereits durch die Indizes vorgegeben sind. Lediglich die „berühmten“ Gleichungen werden auch in symbolischer Schreibweise angegeben, um sie besser wieder zuerkennen. Natürlich wird von der Einsteinschen Summenkonvention Gebrauch gemacht. Dies bedeutet, dass in jedem Term über gleiche Indizes über den gesamten Wertebereich summiert wird. Der Ausdruck

$$a_{ij} = b_{ik}\, c_{kj} + d_{ij}$$
$$k = \{1,2,3\}$$

ist also wie folgt zu verstehen:

$$a_{ij} = \left(\sum_{k=1}^{3} b_{ik}\, c_{kj} \right) + d_{ij}$$

Kapitel 2

Theorie

2.1 Geometrische Nichtlinearität

Wenn das Kapitel "geometrische Nichtlinearität" abgeschlossen ist, soll ein Werkzeug zur Verfügung stehen, das in der Lage ist, Eigenschaften von Verzerrungen zu beschreiben. Dieses Werkzeug wird benötigt, um die Verträglichkeitsaussage zu erfüllen. Im Endeffekt geht es ja darum, einer Verzerrung eine Kraft gegenüber zu stellen, die genau diese Verzerrung hervorruft. Eine analytische Funktion, die ganz allgemein jede beliebige Verzerrung beschreiben kann, gibt es natürlich nicht. Ansonsten würde sich die FEM ja ad absurdum führen.

2.1.1 Definitionen und Gedankenmodell

Zuerst muss definiert werden, was eigentlich verzerrt werden soll und was eine Verzerrung ist. Aus den Augen der Physik werden Materie behaftete Körper verzerrt. Diese Körper leben im dreidimensionalen Raum dem $\mathbb{R}^3$. Aus den Augen der Mathematik ist ein Körper nun nichts weiter als eine Menge $\overline{\Omega}$ aus dem $\mathbb{R}^3$:

$$\overline{\Omega} \subset \mathbb{R}^3$$

Allerdings lassen sich im $\mathbb{R}^3$ viele Mengen bilden, die die Kriterien eines physikalischen Körpers nicht erfüllen. So wäre zum Beispiel Regen so eine Menge. Regen besteht aber aus vielen Körpern und nicht aus einem. Es ist also nötig, $\overline{\Omega}$ als Gebiet zu definieren:

Def. i : $\overline{\Omega}$ ist eine offene, beschränkte und zusammenhängende Menge des $\mathbb{R}^3$ also ein Gebiet des $\mathbb{R}^3$

Manche gehen bei ihrer Definition noch weiter und fassen den Begriff „Gebiet“ noch enger. Das ist für diesen Fall jedoch nicht nötig. $\overline{\Omega}$ ist ein Körper mit einer Oberfläche (offen), er ist nicht unendlich groß (beschränkt) und er besteht aus einem Stück (zusammenhängend). Man sollte über die Oberfläche respektive den Rand $\partial\overline{\Omega}$ des Gebiets integrieren können. Es ist somit auch sinnvoll voraus zu setzen, dass die Oberfläche des Körpers zusammenhängend, geschlossen und begrenzt ist. In diesem Fall wird das Oberflächenintegral über die Flächenvektoren zu Null. Dies wird sich später bei der Untersuchung der Ränder als sehr nützlich erweisen.

Def. ii : Wird $\overline{\Omega}$ ohne jeden Index verwendet, so befindet sich das Gebiet in der Referenzkonfiguration.

Als Referenzkonfiguration bezeichnet man den Körper, wenn keine äusseren Kräfte auf ihn einwirken und er sich somit im unverformten Zustand befindet. Er ist also spannungsfrei. Von dieser Konfiguration ausgehend, werden nun die Verzerrungen beschrieben. Verzerrt ist ein Körper, wenn er nicht mehr das Gebiet $\overline{\Omega}$ einnimmt. Er bekommt den Index φ und wird zu $\overline{\Omega^{\varphi}}$.

Die Funktion, die eine Deformation des Körpers leisten kann, soll $\varphi_i(x_i)$ heißen. Sie bekommt als Funktionswerte $\overline{\Omega}$ übergeben und liefert $\overline{\Omega^{\varphi}}$ zurück.

$$\varphi_i(\overline{\Omega}) \rightarrow \overline{\Omega^{\varphi}}$$

Glg. 2.1.i

Auch hier bietet die Mathematik einen unendlich großen Satz von Funktionen an, die diese Deformation leisten können. Physikalisch sinnvoll sind allerdings nur wenige. Es folgen nun Forderungen an die Funktion, die sie erfüllen muss.

Def. iii : $\varphi_i(x_i)$ muss lokal injektiv sein

Injektiv sind Funktionen immer dann, wenn keine zwei Funktionswerte auf ein und denselben Bildwert abgebildet werden. Physikalisch bedeutet das, dass man ein Volumen nicht auf einen Punkt zusammendrücken kann. Dann würden nämlich Massepunkte aus der Referenzkonfiguration ein und den selben Ort in der deformierten Konfiguration einnehmen. Dadurch wird auch ein gegenseitiges Durchdringen ausgeschlossen. Interessant hierbei ist, dass sich bei der finiten Formulierung zwar ein und das selbe Element nicht selbst durchdringen kann, andere Elemente hingegen schon. Diese Problematik rief die Gruppe der Kontaktelemente ins Leben.

Forderte man Injektivität auch auf dem Rand $\partial\overline{\Omega}$, so schlösse man Verformungen, die zur Selbstberührung führen würden aus. Man könnt also zum Beispiel keinen Ring mehr formen.

Def. iv : $\varphi_i(x_i)$ ist orientierungserhaltend.

Das bedeutet, dass die Basis des betrachteten Gebiets zu jedem Zeitpunkt die gleiche Orientierung trägt (links oder rechts orientiert). Diese Aussage ist wichtig, da sich ein Orientierungsübergang mittels Deformation der Basis nur bewerkstelligen lässt, wenn zu

irgendeinem Zeitpunkt t mindestens zwei Basisvektoren zusammenfallen, was der Injektivität aber widerspricht.

$\varphi_i(x_i)$ ist eine Funktion, die als Ortsvektor vom Koordinatenursprung zur deformierten Lage des entsprechenden Funktionswertes zeigt. In der Statik ist diese Information aber von untergeordneter Bedeutung. Man benötigt vielmehr die Verschiebung als Vektor eines jeden Punktes in der Referenzkonfiguration zum entsprechenden Ort in der deformierten Konfiguration.

$\varphi_i(x_i)$ ist als die vektorielle Addition des Ortsvektors zum Funktionswert mit dem dazugehörigen Verschiebungsvektor definiert. Die Mathematik bezeichnet die Abbildung eines Funktionswertes auf sich selbst als *Identität* kurz *id* somit ergibt sich $\varphi_i(x_i)$ zu:

$$\varphi_i(x_i)=id_i+u_i(x_i), \qquad u_i\in\mathbb{R}^3$$

Glg. 2.1.ii

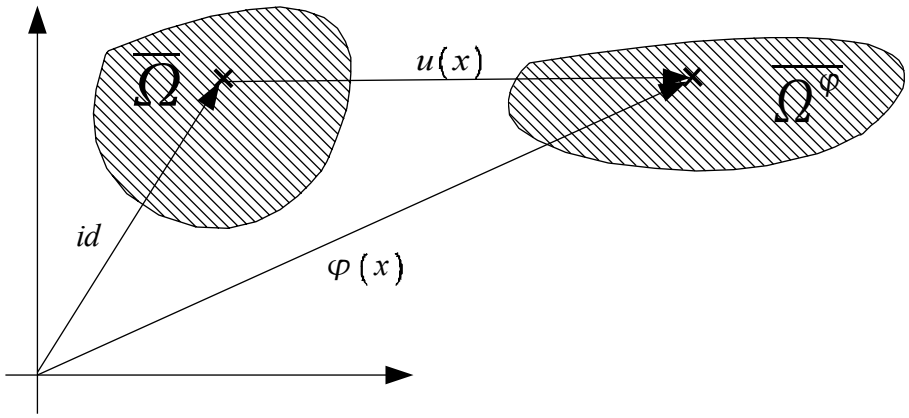

Abbildung 2.1: Deformation

2.1.2 Herleitung des Cauchy-Greenschen Verzerrungstensors

In der Theorie der linearen Geometrie wird die kinematische Verzerrung eines Gebiets mit dem Gradienten der Verschiebung angegeben.

$$\epsilon^K_{ji}=\frac{\partial u_i}{\partial x_j}$$

Glg. 2.1.iii

Diesen Ausdruck kann man sehr anschaulich in grafischer Art und Weise herleiten. Die Gefahr dabei ist allerdings, dass man leicht vergisst, dass er nicht vollständig und somit falsch ist. Der Zusammenhang ist nur für kleine Verformungen mit guter Genauigkeit verwendbar. Bevor nun

ein exaktes Maß für die Verzerrung hergeleitet wird, scheint es sinnvoll, den Fehler der linearen Theorie aufzudecken.

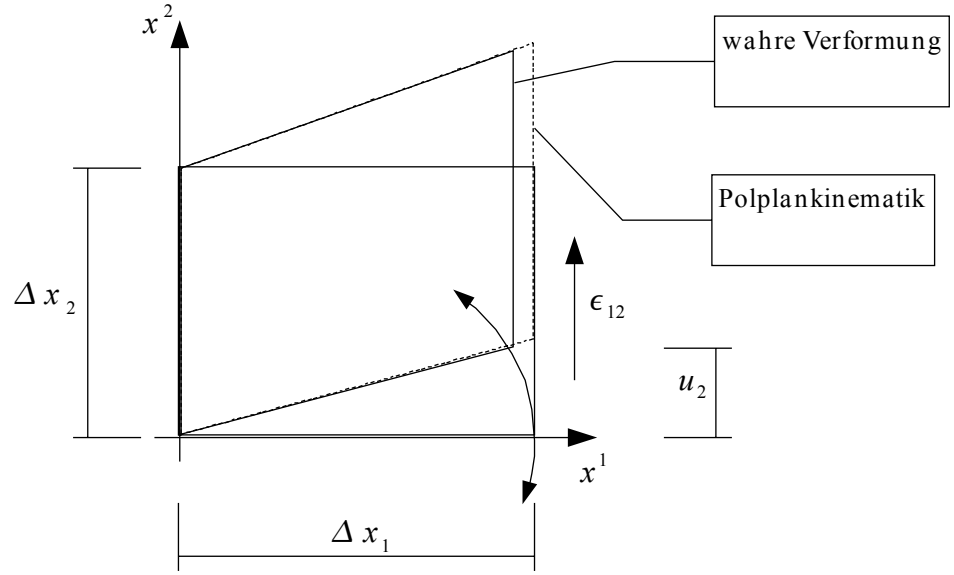

Abbildung 2.2 nach x^2 *Verschobenes Element*

In der linearen Theorie gilt, wie man aus der Zeichnung ablesen kann:

$$\epsilon_{12} = \frac{u_2}{\Delta x_1} \overset{\Delta x_1 \to dx_1}{\to} \frac{\partial u_2}{\partial x_1}$$

Glg. 2.1.iv

Ableitungen sind wie folgt definiert:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Glg. 2.1.v

Der Abstand h verändert sich aber nicht in Abhängigkeit zum Funktionswert. Die Verschiebung hingegen verkürzt den Abstand schon. Man kann sich in der Zeichnung davon überzeugen. Verwendet man die Ableitung der Verschiebung, um die Verzerrung darzustellen, so kann man ausschließlich Verformungen nach Polplankinematik darstellen, da die Verschiebung immer auf die unverformte Elementlänge bezogen wird. Im Laufe der finiten Approximation wird über die Verzerrung integriert, um die noch unbekannte Verschiebung zu erhalten. Da aber nun schon die Ableitung der Verschiebung die Verzerrung nicht richtig beschreibt, kann auch die Integration nicht richtig sein.

Um den Abstand zweier Punkte in einem Gebiet zueinander zu verändern, benötigt man eine entsprechende Kraft. Diesem Sachverhalt wird durch das Gleichsetzten der kinematischen und materiellen Verzerrung Sorge getragen:

$$\epsilon_{ij}^{K} = \epsilon_{ij}^{M}$$

Es gilt also ein Werkzeug zu entwickeln, das für jeden Punkt im Gebiet die Abstandsänderung zu seinen Nachbarn beschreibt. Hat man dieses Werkzeug, so ist die geometrische Seite der Gleichung exakt befriedigt. Denkt man sich nun einen beliebigen Punkt *x* im Gebiet und wählt einen willkürlichen *z* langen Vektor, so gilt:

$$\varphi_i(x_i + z_i) - \varphi_i(x_i) = \frac{\partial \varphi_i(x_i)}{\partial x_j} z_j + z_i$$

Glg. 2.1.vi

Der Ausdruck $\frac{\partial \varphi_i(x_i)}{\partial x_j}$ wird als der **Deformationsgradient** bezeichnet und hat in Matrixdarstellung folgendes Aussehen:

$$F_{ij} = \begin{Bmatrix} \frac{\partial \varphi_1(x_i)}{\partial x_1} & \frac{\partial \varphi_1(x_i)}{\partial x_2} & \frac{\partial \varphi_1(x_i)}{\partial x_3} \\ \frac{\partial \varphi_2(x_i)}{\partial x_1} & \frac{\partial \varphi_2(x_i)}{\partial x_2} & \frac{\partial \varphi_2(x_i)}{\partial x_3} \\ \frac{\partial \varphi_3(x_i)}{\partial x_1} & \frac{\partial \varphi_3(x_i)}{\partial x_2} & \frac{\partial \varphi_3(x_i)}{\partial x_3} \end{Bmatrix}$$

Glg. 2.1.vii

Ein Gradient ist eine Richtungsableitung. Er beschreibt also die Änderung einer Funktion in eine bestimmte Richtung. Multipliziert man den Gradienten mit einem Vektor, so erhält man die Änderung der Funktion in Richtung des Vektors. Die Glg. 2.1.vi kann grafisch dargestellt werden.

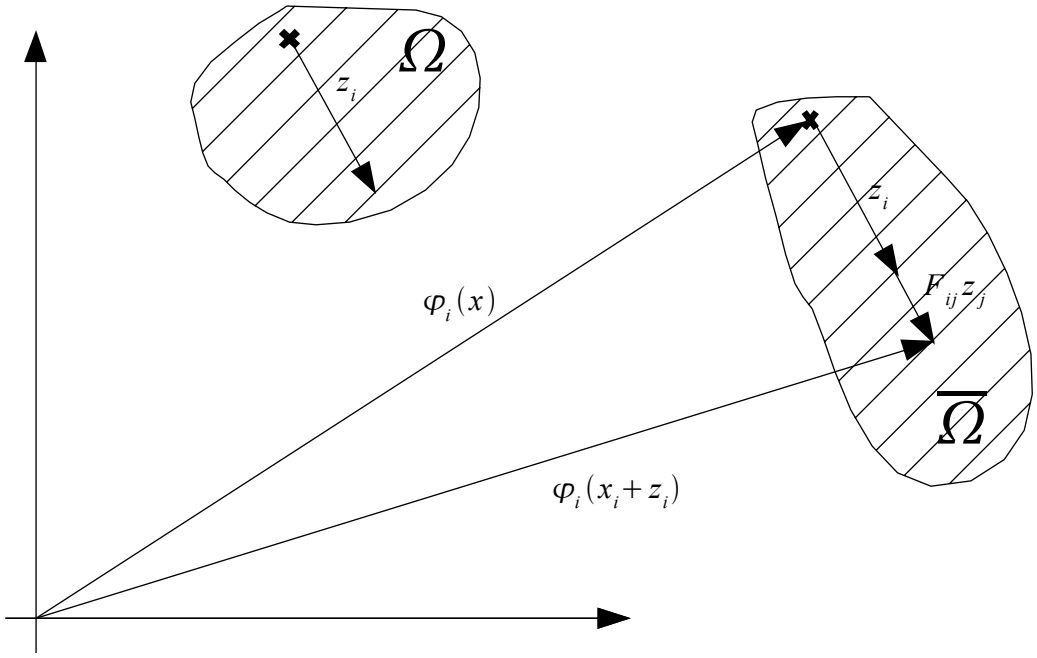

Abbildung 2.3 Deformation zwischen zwei Punkten

Für die Verträglichkeitsformulierung wird nur die betragsmäßige Längenänderung benötigt. Es muss ja nur die Frage beantwortet werden: Um wie viel wurden die beiden Punkte voneinander entfernt? Die Richtung ist ohne Belang. Somit ermittelt man den Euklidischen Abstand also die Länge des Vektors $F_{ij} z_i$:

$$\|\varphi_i(x_i + z_i) - \varphi_i(x_i)\| = \|F_{ij} z_j + z_j\| \quad \Rightarrow;$$

$$\|\varphi_i(x_i + z_i) - \varphi_i(x_i)\| = \|F_{ij} z_j\| + \|z_j\| \quad \Rightarrow;$$

$$\|\varphi_i(x_i + z_i) - \varphi_i(x_i)\| = \sqrt{(F_{ij} z_j)(F_{ik} z_k)} + \|z_j\|$$

Glg. 2.1.viii

Da die meiste Literatur leider die symbolische Schreibweise verwendet, sei hier dieser entscheidende Schritt noch einmal in symbolischer Schreibweise aufgeführt:

$$\|\underline{\varphi}(\underline{x} + \underline{z})\| - \|\underline{\varphi}(\underline{x})\| = \|\underline{\underline{F}}\,\underline{z} + \underline{z}\| \quad \Rightarrow;$$

$$\|\underline{\varphi}(\underline{x} + \underline{z})\| - \|\underline{\varphi}(\underline{x})\| = \|\underline{\underline{F}}\,\underline{z}\| + \|\underline{z}\| \quad \Rightarrow;$$

$$\|\underline{\varphi}(\underline{x} + \underline{z})\| - \|\underline{\varphi}(\underline{x})\| = \sqrt{(\underline{\underline{F}}\,\underline{z})^T (\underline{\underline{F}}\,\underline{z})} + \|\underline{z}\| \quad \Rightarrow;$$

$$\|\underline{\varphi}(\underline{x} + \underline{z})\| - \|\underline{\varphi}(\underline{x})\| = \sqrt{\underline{z}^T \underline{\underline{F}}^T \underline{\underline{F}}\,\underline{z}} + \|\underline{z}\|$$

Glg. 2.1.ix

Der Vektor z skaliert die Verzerrung auf eine bestimmte Länge, ist also veränderlich. Die Verzerrung für jeden Punkt im Gebiet wird aber durch das Produkt des Deformationsgradienten mit

sich selber angegeben. Der sich so ergebende Tensor ist ein Maß für die Verzerrung und heißt **rechter Cauchy-Greenscher Verzerrungstensor**. [2]

$$C_{jk} = F_{ij} F_{ik}$$

$$\underline{\underline{C}} = \underline{\underline{F}}^T \underline{\underline{F}}$$

Glg. 2.1.x

Setzt man die Glg. 2.1.ii nun die Glg. 2.1.x ein, so erhält man mit dem Wissen, dass der Gradient der Identität der Einheitstensor ist, folgenden Ausdruck:

$$C_{jk} = \frac{\partial \varphi_i}{\partial x_j} \frac{\partial \varphi_i}{\partial x_k} \qquad \Rightarrow;$$

$$C_{jk} = \left(\frac{\partial id_i}{\partial x_j} + \frac{\partial u_i}{\partial x_j} \right) \left(\frac{\partial id_i}{\partial x_k} + \frac{\partial u_i}{\partial x_k} \right) \qquad \Rightarrow;$$

$$C_{jk} = \left(I_{ij} + \frac{\partial u_i}{\partial x_j} \right) \left(I_{ik} + \frac{\partial u_i}{\partial x_k} \right) \qquad \Rightarrow;$$

$$C_{jk} = I_{jk} + \frac{\partial u_j}{\partial x_k} + \frac{\partial u_k}{\partial x_j} + \frac{\partial u_i}{\partial x_j} \frac{\partial u_i}{\partial x_k}$$

Glg. 2.1.xi

Natürlich kann man die Indexschreibweise nun auch wieder in symbolische Schreibweise umformen:

$$C_{jk} = I_{jk} + \underbrace{\frac{\partial u_j}{\partial x_k} + \frac{\partial u_j}{\partial x_k}^T + \frac{\partial u_j}{\partial x_i}^T \frac{\partial u_i}{\partial x_k}}_{2 E_{jk}} \qquad \Rightarrow;$$

$$\underline{\underline{C}} = \underline{\underline{I}} + \underbrace{\underline{\underline{\nabla u}}^T + \underline{\underline{\nabla u}} + \underline{\underline{\nabla u}}^T \underline{\underline{\nabla u}}}_{2 \underline{\underline{E}}}$$

Glg. 2.1.xii

Durch eine Umformung der Glg. 2.1.xii kann man die Abweichung von der Identität formulieren und erhält nun endgültig ein exaktes Maß für die Verzerrung:

---

[2] Die Symbolische Schreibweise definiert zwei verschiedene Produkte . Das rechte Produkt $A_{ik} A_{jk} = A_{ik} A_{kj}^T = A A^T$ und das linke Produkt $A_{ki} A_{kj} = A_{ik}^T A_{kj} = A^T A$. Man kann die Herleitung auch mit den linken Produkten durchführen, was aber hier keinen Vorteil bringt und nur der Vollständigkeit halber angemerkt wird.

$$E_{ik} = \frac{1}{2}(C_{jk} - I_{jk})$$

$$\underline{\underline{E}} = \frac{1}{2}(\underline{\underline{C}} - \underline{\underline{I}})$$

Glg. 2.1.xiii

Durch visuelle Inspektion von Glg. 2.1.xii mit der linearen Theorie wird ersichtlich, dass es sinnvoll ist, die Benennung 2E einzuführen. Der Term $\underline{\underline{\nabla u^T \nabla u}}$ beschreibt offenbar den nichtlinearen Anteil der Verzerrung. *E* wird als **Green-Lagrangescher-Verzerrungstensor** bezeichnet.

Im nächsten Gliederungspunkt soll gezeigt werden, wie der Deformationsgradient arbeitet. Dazu ist es aber nötig, auf die Dekomposition regulärer Tensoren einzugehen. Darunter versteht man die eindeutige, sprich invariante Zerlegung eines Tensors[3]. Dieser Tensor muss dazu regulär sein. Regulär ist ein Tensor, wenn sein Rang gleich dem Wertebereich seiner Indizes ist, sprich gleich der Anzahl seiner Zeilen oder Spalten. Tensoren deren Determinante ungleich Null ist, sind regulär. Somit ist auch der Deformationsgradient F_{ij} regulär und lässt sich in folgender Form zerlegen:

$$F_{ij} = R_{ik} U_{kj} \quad \begin{cases} R_{ik} : \textit{Orthogonale Rotationsmatrix} \\ U_{kj} : \textit{Symmetrische , positiv definite Skalierungsmatrix} \end{cases}$$

Glg. 2.1.xiv

Die Zerlegung ist, wie bereits erwähnt, eindeutig. Sie lässt sich wie folgt errechnen:

$$U_{kj} := \sqrt{F_{ki}^T F_{ij}}$$

$$R_{ik} := F_{ij} U_{jk}^{(-1)}$$

Glg. 2.1.xv

U_{kj} wird auch als der rechte Streckungstensor bezeichnet. Es gibt hier ebenfalls einen linken Streckungstensor $V = \sqrt{F F^T}$, der für die nun folgende Veranschaulichung keinen weiteren Vorteil bietet und somit nicht weiter verfolgt wird.

2.1.3 Veranschaulichung und abgeleitete Größen

Die nun folgenden Beispiele sollen auf graphische Art und Weise veranschaulichen, wie der Deformationsgradient arbeitet. Er mag ja analytisch exakt sein und die Verträglichkeit befriedigen, dennoch stößt man auch mit ihm an Grenzen des Darstellbaren. Die FEM ist ein Werkzeug, das nur

[3] Beweis: [2] Seite 188-192

mit dem entsprechenden Expertenwissen gewinnbringend und sicher eingesetzt werden kann. Es ist also nötig, eine klare Vorstellung und ein „Gefühl“ für den Deformationsgradienten zu entwickeln. Im ersten Teil dieses Absatzes soll der $\mathbb{R}^3$ auf den $\mathbb{R}^2$ reduziert werden. Es stehen somit drei Freiheitsgrade (zwei translatorische und ein rotatorischer) zur Verfügung. Die Referenzkonfiguration sei ein quadratisches Gebiet um den Koordinatenursprung mit der Kantenlänge 2:

$$\overline{\Omega}=\{x_i \mid i\in\{1,2\}\wedge x_i\in\mathbb{R}\wedge -1\le x_i\le 1\}$$

Glg. 2.1.xvi

- Reine Translation

  Die Deformation habe folgendes Aussehen:

$$\varphi_i(x_i)=\begin{Bmatrix}\varphi_1(x_i)=1\,x_1+0\,x_2+3\\ \varphi_2(x_i)=0\,x_1+1\,x_2+2\end{Bmatrix}$$

$$F_{ij}=\frac{\partial\varphi_i}{\partial x_j}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$detF_{ij}=1$$

Glg. 2.1.xvii

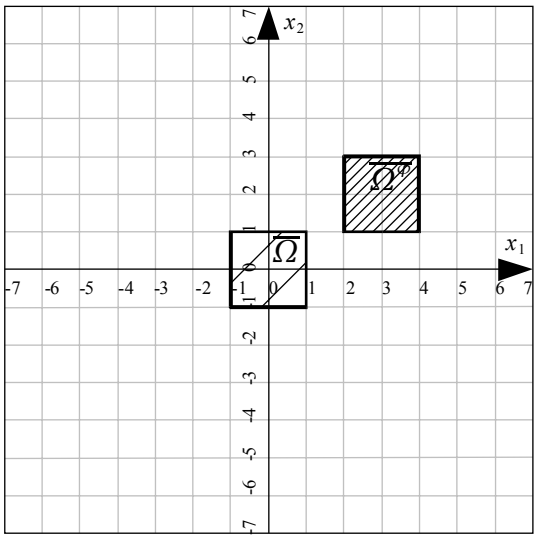

Abbildung 2.4 reine Translation

Bereits hier lässt sich erkennen, dass der Gradient anscheinend eine reine Verschiebung nicht bemerkt. Zu diesem Zeitpunkt soll nur festgehalten werden, dass die Determinante des Gradienten 1 ist.

- Reine Rotation

  Die Deformation habe folgendes Aussehen:

$$\varphi_i(x_i)=\begin{Bmatrix}\varphi_1(x_i)=\sin(45°)x_1-\cos(45°)x_2\\ \varphi_2(x_i)=\cos(45°)x_1+\sin(45°)x_2\end{Bmatrix}$$

$$F_{ij}=\frac{\partial\varphi_i}{\partial x_j}=\begin{pmatrix}\cos(45°) & \sin(45°)\\ -\sin(45°) & \cos(45°)\end{pmatrix}$$

$$F_{ij}=R_{ik}U_{kj}=\begin{pmatrix}\cos(45°) & \sin(45°)\\ -\sin(45°) & \cos(45°)\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$detF_{ij}=1$$

Glg. 2.1.xviii

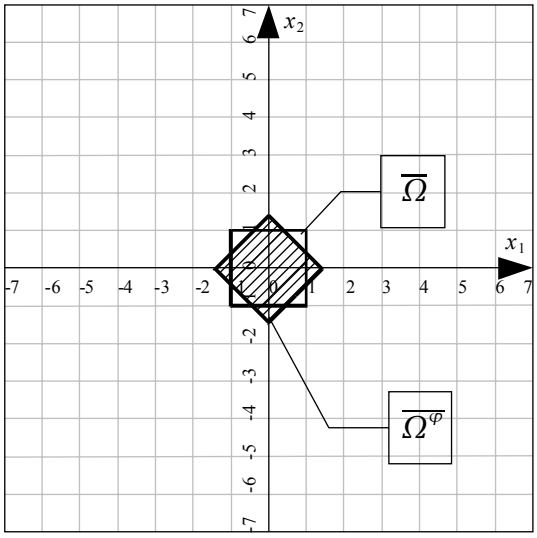

Abbildung 2.5 reine Rotation

Der Deformationsgradient nimmt die Form des Rotationstensors an[4], die Determinante bleibt weiterhin 1.

- Reine Skalierung

  Die Deformation habe folgendes Aussehen:

[4] In der Literatur wird er manchmal auch als Phiedreher bezeichnet, da er alles um den Winkel φ dreht.

$$\varphi_i(x_i)=\begin{Bmatrix}\varphi_1(x_i)=2\,x_1+0\,x_2\\ \varphi_2(x_i)=0\,x_1+4\,x_2\end{Bmatrix}$$

$$F_{ij}=\frac{\partial\varphi_i}{\partial x_j}=\begin{pmatrix}2 & 0\\ 0 & 4\end{pmatrix}$$

$$F_{ij}=R_{ik}\,U_{kj}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}2 & 0\\ 0 & 4\end{pmatrix}$$

$$detF_{ij}=8$$

Glg. 2.1.xix

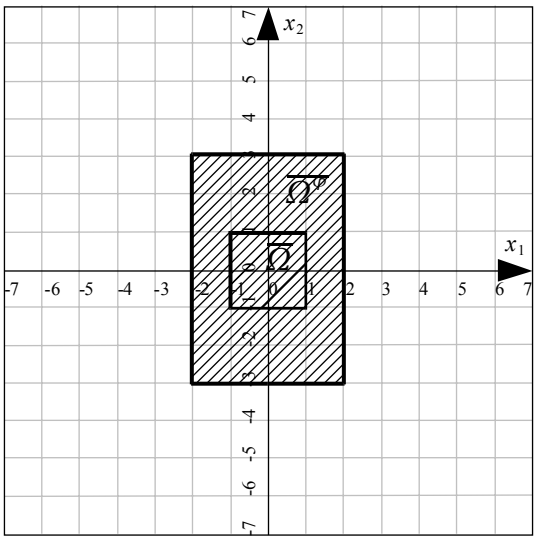

Abbildung 2.6 reine Skalierung

Die Determinante des Deformationsgradienten ist 8. Die Fläche der deformierten Konfiguration ist 32, dies ist das Achtfache der Referenzkonfiguration. Die Determinante scheint ein Maß für die Flächenänderung zu sein.

- Reine Scherung

  Die Deformation habe folgendes Aussehen:

$$\varphi_i(x_i)=\begin{Bmatrix}\varphi_1(x_i)=1\,x_1+2\,x_2\\ \varphi_2(x_i)=0\,x_1+1\,x_2\end{Bmatrix}$$

$$F_{ij}=\frac{\partial\varphi_i}{\partial x_j}=\begin{pmatrix}1 & 2\\ 0 & 1\end{pmatrix}$$

$$F_{ij}=R_{ik}\,U_{kj}=\begin{pmatrix}1+\dfrac{2(\sqrt{2}-1)}{2-\sqrt{5}} & \dfrac{\sqrt{2}-2}{2-\sqrt{5}}\\ \dfrac{\sqrt{2}}{2-\sqrt{5}} & -\dfrac{1}{2-\sqrt{5}}\end{pmatrix}\begin{pmatrix}1 & \sqrt{2}\\ \sqrt{2} & \sqrt{5}\end{pmatrix}$$

$$detF_{ij}=1$$

Glg. 2.1.xx

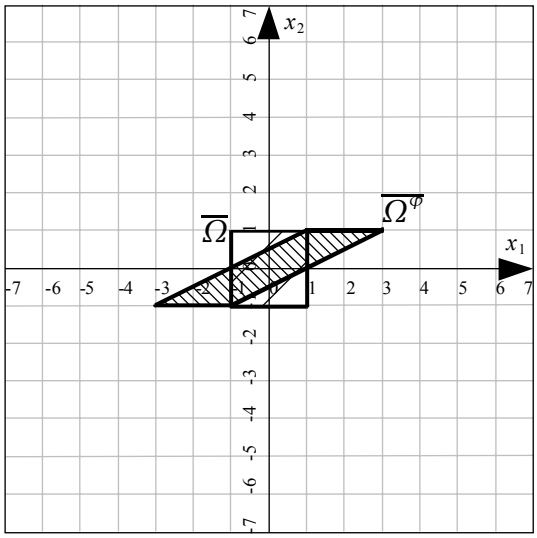

Abbildung 2.7 reine Scherung

Auch hier ist die Determinante wieder eins und die Fläche der verzerrten Geometrie hat sich nicht geändert. Es folgt also nun eine genauere Interpretation des Deformationsgradienten.

- Geometrische Interpretationen

Geometrisch kann man die Funktion $\varphi_i(x_i)$ auch anders deuten. Man kann sagen, dass die Funktion $\varphi_i(x_i)$ die Referenzkonfiguration $\overline{\Omega}$ in ein neues Koordinatensystem abbildet ohne sie zu verändern. Die folgende Abbildung soll diesen Sachverhalt illustrieren. Das Gebiet erstreckt sich wie auch in der Referenzkonfiguration von -1 bis 1 sowohl in g_1 als auch in g_2 Richtung. Das Koordinatensystem wurde hingegen verdreht, gestreckt und geschert.

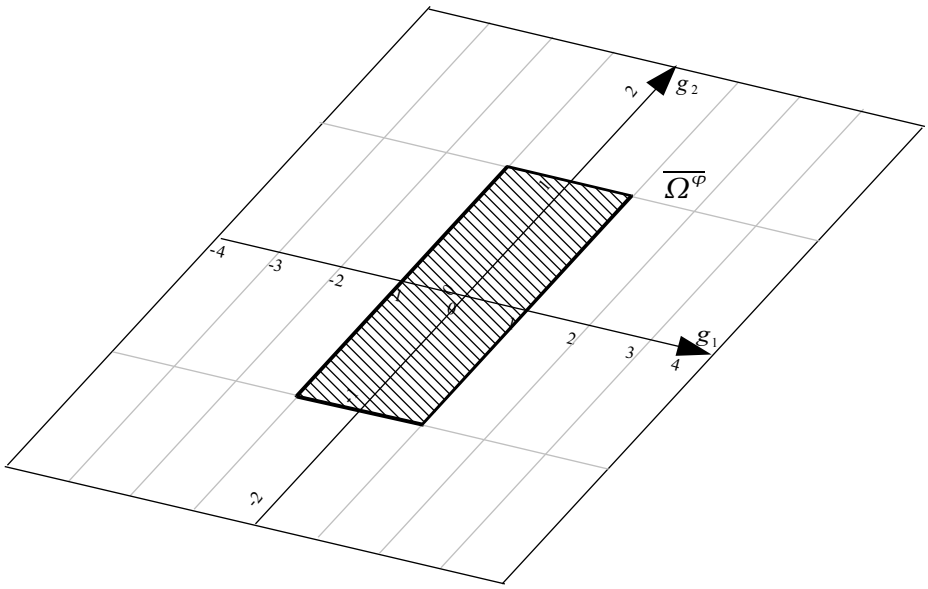

Abbildung 2.8 transformiertes Koordinatensystem

Ein Koordinatensystem wird anhand seiner Basen definiert. Es seien e_i die Basisvektoren des Referenzkoordinatensystems und $\tilde{e}_i$ die Basisvektoren der deformierten Konfiguration. So muss $\varphi_i(x_i)$ dazu in der Lage sein, die Basisvektoren zu transformieren:

$$\varphi_i(e_i)-\varphi_i(0_i)=\tilde{e}_i$$

Glg. 2.1.xxi

Am Beispiel der reinen Scherung wäre das also:

$$\varphi_i(e_1)=\begin{Bmatrix}\varphi_1(1,0)=1\\ \varphi_2(1,0)=0\end{Bmatrix}-\begin{Bmatrix}0\\0\end{Bmatrix} = \tilde{e}_1=(1,0)$$

$$\varphi_i(e_2)=\begin{Bmatrix}\varphi_1(0,1)=2\\ \varphi_2(0,1)=1\end{Bmatrix}-\begin{Bmatrix}0\\0\end{Bmatrix} = \tilde{e}_2=(2,1)$$

Glg. 2.1.xxii

Das Ergebnis entspricht genau der Erwartung, wie man sich anhand der Abbildung auch leicht überzeugen kann. Aus Glg. 2.1.viii ist ersichtlich, dass F_{ij} per Definition richtungskonstant ist. Das bedeutet, dass F_{ij} in Richtung z_j eine konstante relative Längenänderung aufweist. Bildlich gesprochen bedeutet dies, wenn das Gebiet ein Stab wäre und z_j in Richtung der Stabachse zeigen würde, dann wäre die Deformation über die gesamte Länge des Stabes konstant, was ja auch der Erfahrung entspricht. Es spielt also keine Rolle, wie lang z_j ist. Natürlich kann die Deformation in einer anderen Richtung anders sein. Aber auch für diese wäre sie wieder konstant. Beim Gradienten allgemein muss dieser Sachverhalt nicht zwangsläufig gelten. Im Falle des Deformationsgradienten ist es aber ein fester Bestandteil der Definition. Daraus folgt

unmittelbar, dass die Deformation höchstens linear sein kann. $\varphi_i(x_i)$ kann also nur eine Linearkombination der einzelnen Richtungen sein. Nur in diesem Fall ist der Gradient konstant. Die Ableitung einer linearen Funktion ist aber der Faktor der Variablen. Man leitet eine lineare Funktion einfach ab, indem man die Variable zu eins setzt und alle anderen Terme zu Null verschwinden lässt. Genau dies geschieht, wenn man die Basisvektoren kartesischer Koordinatensysteme in Glg. 2.1.xxi einsetzt. Somit lässt sich der Deformationsgradient auch derart interpretieren, dass die Basisvektoren des deformierten Systems spaltenweise nebeneinander stehen:

$$F_{ij} = \left(\overset{1}{\tilde{e}}_j \quad \overset{2}{\tilde{e}}_j \quad \overset{3}{\tilde{e}}_j \right) = \begin{pmatrix} \overset{1}{\tilde{e}}_1 & \overset{2}{\tilde{e}}_1 & \overset{3}{\tilde{e}}_1 \\ \overset{1}{\tilde{e}}_2 & \overset{2}{\tilde{e}}_2 & \overset{3}{\tilde{e}}_2 \\ \overset{1}{\tilde{e}}_3 & \overset{2}{\tilde{e}}_3 & \overset{3}{\tilde{e}}_3 \end{pmatrix} \quad [5]$$

Glg. 2.1.xxiii

Man kann sich in den Beispielen leicht von diesem Sachverhalt überzeugen. Nun wird auch offensichtlich, was geschieht, wenn die Determinante des Deformationsgradienten Null wird. Dann ist einer der Basisvektoren entweder selber Null lang oder fällt als Linearkombination in die anderen. Beide Fälle widersprechen aber der Injektivität der Deformation.

Das Spatprodukt ist definiert als $\varepsilon_{ijk} a_i b_j c_k$ [6], in symbolischer Schreibweise $(\underline{a} \times \underline{b}) \cdot \underline{c}$ .Die geometrische Interpretation des Spatprodukts ist das Volumen des von den drei Vektoren beschriebenen Parallelepipeds. Es errechnet sich über den Entwicklungssatz für das Kreuzprodukt, skalar mit dem dritten Vektor multipliziert. Im $\mathbb{R}^3$ entspricht es aber genau der Determinante über die drei Vektoren. $det\, F_{ij}$ ist also das Volumen, das von den drei deformierten Basisvektoren eingeschlossen wird. Da das Volumen, das von den undeformierten Basisvektoren eingeschlossen wird, immer 1 ist, stellt $det\, F_{ij}$ ein Maß für die Volumenänderung dar. Das deformierte Gebiet ist also das $det\, F_{ij}$ -fache der Referenzkonfiguration.

$$V^{\varphi} = det\,(F_{ij})\, V$$

Glg. 2.1.xxiv

Im Laufe der finiten Formulierung wird man gezwungen sein, über das deformierte Volumen zu integrieren. Integrale sind aus der Sicht der Mathematik differentielle Aufsummationen. Der Ausdruck $\int dV^{\varphi}$ bedeutet also, dass man alle „ dV^{φ} - Würfelchen“ aufsummieren soll. Über das deformierte Gebiet zu integrieren ist unschön und im Falle, dass man es noch nicht kennt,

[5] Der Index über dem $\tilde{e}$ bezeichnet den Basisvektor, der Index rechts davon die Richtung in kartesischen Koordinaten.

[6] Der ε Tensor ist ein erweitertes Kronecker-Symbol er kann die Werte -1 , 0 und +1 annehmen

unmöglich. Durch Glg. 2.1.xxiv ist aber nun ein Zusammenhang bekannt, mit dem man das Integral wie folgt überschreiben kann:

$$\int dV^{\varphi} = \int det\,(F_{ij})\,dV$$

Glg. 2.1.xxv

Bei der Herleitung eines finiten Elementes ist man oft auch dazu gezwungen, über die Fläche, also den Rand eines Volumens, zu integrieren. Es wird dadurch eine Transformation vom deformierten Rand zum nicht deformierten benötigt. Dies bewerkstelligt die **Piolatransformation für den Einheitstensor**. Die Herleitung ist leider nicht so übersichtlich wie für das Volumen und soll hier geometrisch anschaulich erfolgen.

Ein Flächenelement im Raum lässt sich durch seine betragsmäßige Fläche und einen senkrecht auf ihr stehenden Einheitsvektor beschreiben. Die hier vorkommenden Flächen sind immer Ränder von Volumen. Es ist also sinnvoll, den Einheitsvektor derart zu definieren, dass er immer aus dem Volumen heraus zeigt. Die folgende Zeichnung illustriert den Sachverhalt am Parallelepiped, das von der deformierten Basis aufgespannt wird.

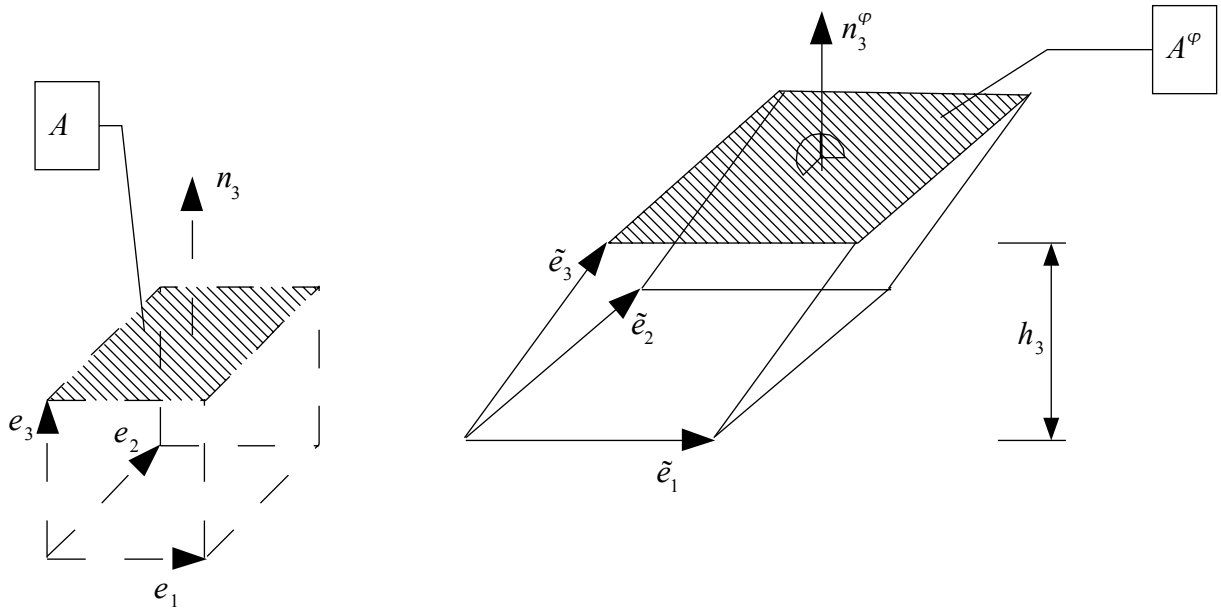

Abbildung 2.9 Flächenelement mit Flächenvektor

Die geometrische Bedeutung des Kreuzprodukts ist ein Vektor, der senkrecht auf den gekreuzten Vektoren steht und eine betragsmäßige Länge hat, die dem aufgespannten Parallelogramm entspricht. In diesem Fall wäre das also in Indexschreibweise:

$$n_k^{\varphi}\, A^{\varphi} = \tilde{e}_i \epsilon_{jki} \tilde{e}_j$$ [7]

Glg. 2.1.xxvi

[7] Die Definition des Kreuzprodukts ist $\underline{a} \times \underline{b} \simeq a_j \epsilon_{ikj} b_i$

In Glg. 2.1.xxiii wird gezeigt, dass man den Deformationsgradienten auch mittels der deformierten Basisvektoren darstellen kann. Das Skalarprodukt zweier normal aufeinander stehender Vektoren ist Null. Es beschreibt ja das Produkt der Beträge beider Vektoren mal dem Kosinus des eingeschlossenen Winkels $a_i b_i = |a||b|\cos(\varphi)$. Ist einer der beiden Vektoren aber eins lang, so wird ein rechtwinkliges Dreieck aufgespannt und das Skalarprodukt entspricht der Länge der Ankathete. Mit diesem Wissen kann man folgende Gleichung als die Höhe des Parallelepipeds in Bezug auf die betrachtete Fläche interpretieren:

$$F^T_{ji} n^\varphi_i = h^\varphi_j \Rightarrow h^\varphi_j = n^\varphi_j \left| h^\varphi_j \right|$$

Glg. 2.1.xxvii

Da der Normalvektor aus dem Kreuzprodukt zweier Basisvektoren hervorgegangen ist, muss die skalare Multiplikation mit diesen beiden Null werden. Somit ist der Ergebnisvektor an zwei Stellen Null und nur an einer Stelle verschieden von Null. Dieser Wert entspricht aber der Höhe des Parallelepipeds bezogen auf die betrachtete Fläche. Man kann sich von der Richtigkeit dieser Vorstellung anhand des folgenden Schemas überzeugen:

$$\begin{array}{c|c} & \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} = n^\varphi_i \\ \hline F^T_{ji} = \begin{pmatrix} \overset{1}{\tilde{e}}_1 & \overset{1}{\tilde{e}}_2 & \overset{1}{\tilde{e}}_3 \\ \overset{2}{\tilde{e}}_1 & \overset{2}{\tilde{e}}_2 & \overset{2}{\tilde{e}}_3 \\ \overset{3}{\tilde{e}}_1 & \overset{3}{\tilde{e}}_2 & \overset{3}{\tilde{e}}_3 \end{pmatrix} & \begin{pmatrix} n_i \overset{1}{\tilde{e}}_i \\ n_i \overset{2}{\tilde{e}}_i \\ n_i \overset{3}{\tilde{e}}_i \end{pmatrix} = h^\varphi_j \end{array}$$

Abbildung 2.10

Das Volumen eines Parallelepipeds errechnet sich aber aus Grundfläche mal der dazugehörigen Höhe. Für den deformierten Zustand kann also das Volumen wie folgt berechnet werden:

$$V^\varphi_j = F^T_{ji} n^\varphi_i A^\varphi$$

Glg. 2.1.xxviii

Das Volumen ist hier ein Vektor, der nur an einer Stelle besetzt ist. Das folgt direkt aus der Überlegung von Abbildung 2.10.

Für die Referenzkonfiguration gilt der gleiche Zusammenhang, nur dass hier der Deformationsgradient die Einheitsmatrix ist. Das Produkt aus Einheitstensor und Normalvektor ist natürlich wieder der Normalvektor selbst. Somit lässt sich folgendes festhalten:

$$V_j = I^T_{ji} n_i A = n_j A$$

Glg. 2.1.xxix

Der Vektor V_j ist natürlich an der gleichen Stelle besetzt wie V^{φ}_j

Der Zusammenhang zwischen den Volumen ist aber mit Glg. 2.1.xxiv bekannt. Die Piolatransformation ergibt sich also aus Glg. 2.1.xxviii und Glg. 2.1.xxix wie folgt:

$$det(F_{ij}) n_j A = F^T_{ji} n^{\varphi}_i A^{\varphi}$$

$$det(F_{ij}) F^{-T}_{ij} n_j A = n^{\varphi}_i A^{\varphi} \Rightarrow ;$$

$$n^{\varphi}_i dA^{\varphi} = det(F_{ij}) F^{-T}_{ij} n_j dA \Leftrightarrow ;$$

$$\boldsymbol{dA^{\varphi}_i = det(F_{ii}) F^{-T}_{ii} dA_i}$$

Glg. 2.1.xxx

Ist man nur an den betragsmäßigen Größen interessiert, so kann man die Euklidische Norm bilden, da $\|n^{\varphi}_i\| = 1$ erhält man:

$$n^{\varphi}_i dA^{\varphi} = det(F_{ij}) F^{-T}_{ij} n_j dA \Rightarrow ;$$

$$\|n^{\varphi}_i\| dA^{\varphi} = det(F_{ij}) \|F^{-T}_{ij} n_j\| dA \Rightarrow ;$$

$$\boldsymbol{dA^{\varphi} = det(F_{ii}) \|F^{-T}_{ii} n_i\| dA}$$

Glg. 2.1.xxxi

Ziel dieser Arbeit ist eine Untersuchung der geometrischen Nichtlinearität im Rahmen von numerischen Approximationsverfahren. Von diesem Punkt aus sind es noch einige Schritte bis zu einer vollständigen geometrisch nichtlinearen Formulierung, die hier aber nicht mehr aufgezeigt werden sollen, da sie den Rahmen verlassen würden. Die Piolatransformation ist die Schlüsselstelle für das Verständnis. In Glg. 2.1.xxx wird ein kleines Flächenelement, auf dem senkrecht ein Einheitsvektor, steht beschrieben. Weiterhin hat man ein Werkzeug zur Hand, das es einem ermöglicht, ein und dasselbe Element in zwei unterschiedlichen Konfigurationen (deformiert und undeformiert) zu betrachten. Erinnert man sich aber nun an die Definition der Normalspannung, so ist sie nichts weiter als ein kleines Flächenelement mit einer senkrecht auf ihr stehenden Kraft.

Skaliert man den Normalvektor des Flächenelementes, so erhält man ein Werkzeug für die Beschreibung von Normalspannungen. Die gleiche Überlegung gilt auch für die Schubspannungen. Die nichtlineare Theorie bezieht per Definition die Kräfte auf den deformierten Zustand. Am Anfang der Berechnung ist dieser Zustand aber noch unbekannt. Somit ist die Piolatransformation bildlich gesprochen das Bindeglied zwischen der undeformierten Gegenwart und der deformierten Zukunft. Den Deformationsgradienten zu finden, ist somit die zu lösende Aufgabe.

2.2 Numerische Berechnungsverfahren

Die bei einer finiten Formulierung entstehenden Gleichungssysteme werden im linearen Falle mittels Cholesky-Zerlegung mit guter Genauigkeit gelöst. Sobald man aber mit geometrisch nichtlinearen Systemen rechnet, liefert die Cholesky-Zerlegung ab einer Stauchung von etwa 20%[8] physikalisch nicht interpretierbare Lösungen. Bei einer Streckung tritt dieser Zustand auch ein, aber erst bei wesentlich größeren Deformationen. Dieser Abschnitt wird sich mit Fehlerverhalten von Computersystemen beschäftigen. Es soll gezeigt werden, wie ein Fehler entsteht und wie man diese Fehler messen kann. Weiterhin wird gezeigt, dass das System bei einer Stauchung zwar das analytisch Versagen weit vor dem numerischen eintritt, man diesen Effekt aber nummerisch relativ preisgünstig durch die nun vorgestellte Methode messen kann. Dadurch ließen sich einfache Abbruchschranken definieren.

2.2.1 Numerische Fehler und Pseudoarithmetik

Jede Zahl *r* zur Basis *p* lässt sich normalisiert darstellen:

$$r = \pm a\, p^{b}$$

$$a = (0, z_1 z_2 z_3 \cdots z_\infty)$$

$$r \neq 0$$

$$z_k \in \{0, \ldots, p-1\} \wedge z_1 \neq 0$$

$$p \in \mathbb{N} \wedge p > 1$$

$$b \in \mathbb{Z}$$

Glg. 2.2.i

[8] Das entspricht $det(F_{ij}) \sim -0.2$

a wird als Mantisse bezeichnet und *b* als der Exponent von *p*. Wobei *z* die Ziffern der Zahl sind und *b* im gleichen Zahlensystem liegen muss. Wenn *p* zum Beispiel 10 ist, gilt für folgende Zahlen:

$$1234{,}56789 \Leftrightarrow 0{,}123456789 \cdot 10^{4}$$
$$0{,}00012345 \Leftrightarrow 0{,}12345 \cdot 10^{-3}$$

In einem Rechner ist der Speicherplatz pro Zahl auf *n* Zeichen beschränkt. Es gibt mehrere Möglichkeiten, diese *n* Zeichen aufzuteilen. Man kann zum Beispiel zwei Zeichen dazu verwenden, sich die Vorzeichen zu merken, also zum einen das Vorzeichen der Mantisse und zum anderen das Vorzeichen des Exponenten. Ein weiter Teil muss dann für die Mantisse vorgesehen sein und den Rest der Zeichen verwendet man für die Stellen des Exponenten[9]. Für das Ziel, auf das hier zugesteuert wird, ist es nicht nötig, das Dezimalsystem zu verlassen. Es soll für die leichtere Lesbarkeit nur folgendes verabredet werden:

Def. v : $M(p, m, e)$ bezeichnet die Menge der in normalisierter Gleitkommadarstellung codierbarer Zahlen, mit Basis *p*, mit *m* Stellen für die Mantisse und mit *e* Stellen für den Exponenten (jeweils zur Basis p)

Spannt man beispielsweise *M(10,4,2)* auf, so ergeben sich folgende Darstellungen

Zahl in $\mathbb{R}$	***codierte Zahle in N***
$123456789 = 0{,}123456789 \cdot 10^{9}$	+1234/+09
$-15{,}976 = -0{,}15976 \cdot 10^{2}$	-1597/+02
$0{,}999999 \cdot 10^{100}$	+infinity
$0{,}13 \cdot 10^{-100}$	-infinity
$0{,}00005 = 0{,}5 \cdot 10^{4}$	0.0
größte darstellbare Zahl	+9999/+99
kleinste darstellbare positive Zahl	1000/-99

Da die Menge *M* endlich ist, können unendlich viele Elemente aus $\mathbb{R}$ auf ein und das selbe Element der Menge *M* abgebildet werden. Aus diesem Sachverhalt heraus ergeben sich die Fehler der Computerarithmetik.

Betrachtet man eine Funktion γ die reelle Zahlen in die Menge *M* abbildet:

[9] Das **I**nstitute of **E**lectrical and **E**lectronics **E**ngineers IEEE hat einen Standard verabschiedet (IEEE 754) wie Gleitkommazahlen mit einfacher und doppelter Genauigkeit abzuspeichern sind. Das Verfahren ist anders, als das hier vorgestellte.

$$\gamma : \mathbb{R} \rightarrow M(p, m, e)$$

Glg. 2.2.ii

und lässt nur Funktionswerte zwischen den Grenzen von *M* zu:

$$x_{min} \leq |x| \leq x_{max}$$

$$x = \pm (z_1 p^{-1} + z_2 p^{-2} + z_3 p^{-3} + \cdots) p^t$$

$$z_1 \neq 0$$

Glg. 2.2.iii

so ist γ gezwungen, Kommastellen ausserhalb von m zu kürzen. Dadurch entstehen die Rundungsfehler. γ Sieht also wie folgt aus:

$$\gamma(x) = \pm p^t \cdot \begin{cases} (z_1 p^{-1} + z_2 p^{-2} + \cdots + z_m p^{-m}) & falls\ z_{m+1} \leq \frac{p}{2} \\ (z_1 p^{-1} + z_2 p^{-2} + \cdots + z_m p^{-m} + p^{-m}) & falls\ z_{m+1} > \frac{p}{2} \end{cases}$$

Glg. 2.2.iv

Folgende Beispiele sollen Glg. 2.2.iv für *M(10,3,1)* veranschaulichen:

Zahl in $\mathbb{R}$	$\gamma(x)$	$\lvert x - \gamma(x) \rvert$	$\frac{\lvert x - \gamma(x) \rvert}{\lvert x \rvert}$
$12345 = 0{,}12345 \cdot 10^5$	$0{,}123 \cdot 10^5$	$0{,}45 \cdot 10^2$	3,645 ‰
$12355 = 0{,}12355 \cdot 10^5$	$0{,}124 \cdot 10^5$	$0{,}45 \cdot 10^2$	3,642 ‰

Aus dem Beispiel lässt sich auch bequem der größte **absolute Rundungsfehler** ableiten, der sich einstellen kann:

$$|\gamma(x) - x| \leq \frac{p^{-m}}{2} p^t$$

Glg. 2.2.v

Teilt man aus Glg. 2.2.v den Betrag von *x* heraus, so erhält man den **relativen Rundungsfehler**:

$$|x| \geq \frac{1}{p} p^t$$

$$\frac{|\gamma(x) - x|}{|x|} \leq \frac{p^{-m}}{2} p$$

Glg. 2.2.vi

Der relative Fehler ist nun genau das Maß, welches für die Beschreibung aller „Rechenfehler“ der Maschine benötigt wird. Man bezeichnet es als **Maschinengenauigkeit** oder als **roundoff unit**

$$eps := \frac{p}{2}\, p^{-m}$$

Glg. 2.2.vii

Bei binären Fließkommazahlen einfacher Genauigkeit verwendet man zum Beispiel 24 Bits für die Mantisse. Der Exponent ist 2. Somit ergibt sich eine Genauigkeit von $eps = 2^{-24} \approx 6 \cdot 10^{-8}$. Wenn man die Strecke von Berlin nach München (zirka 600km) mit einer ähnlichen Genauigkeit messen wollte, so würde der Fehler maximal 4cm betragen $600\,km \cdot 10^5 \frac{cm}{km} \cdot 2^{-24} \approx 3{,}6\,cm$. Das ist kein großer Fehler. Man muss sich aber zwei Dinge unbedingt vor Augen halten. Zum einen gilt dieser Fehler pro Operation (flop) die ausgeführt wird und bei durchschnittlichen Systemen kommt man leicht auf etliche Millionen Schritte. Zum Anderen haben Fehler die Eigenschaft sich aufzuschaukeln, was die folgende Betrachtung zeigen soll.

Betrachtet man zwei mit Fehlern behaftete Zahlen $\tilde{x} = x + \Delta x$ und $\tilde{y} = y + \Delta y$ wobei die Fehler sehr klein sein sollen, $\left(\left| \frac{\Delta x}{x} \right|, \left| \frac{\Delta y}{y} \right| \ll 1 \right)$ so folgt für Addition und Subtraktion der behafteten Größen:

$$(x + \Delta x) \pm (y + \Delta y) = x \pm y + \Delta x \pm \Delta y$$

Glg. 2.2.viii

Daraus folgt für den relativen Fehler:

$$\frac{\Delta x \pm \Delta y}{x \pm y} = \frac{\Delta x}{x \pm y} \pm \frac{\Delta y}{x \pm y}$$

Glg. 2.2.ix

Die Problematik wird allerdings erst sichtbar, wenn der Bruch wie folgt erweitert wird:

$$\frac{\Delta x}{x \pm y} \pm \frac{\Delta y}{x \pm y} = \frac{x}{x \pm y}\frac{\Delta x}{x} \pm \frac{y}{x \pm y}\frac{\Delta y}{y}$$

Glg. 2.2.x

Wenn $|x \pm y|$ sehr klein ist, verstärkt sich der Fehler ungemein. Es muss also bei der Konditionierung eines Problems dringend darauf geachtet werden, Differenzen etwa gleich großer Terme unbedingt zu vermeiden[10]. Multiplikation und Division verstärken den Fehler nicht. Anhand eines kleinen Zahlenbeispiels soll verdeutlicht werden, wie gefährlich die Fehlerverstärkung ist:

[10] Dieser Effekt wird als Auslöschung oder Scheingenauigkeit bezeichnet, da die bei der Normalisierung freiwerdenden Stellen der Mantisse willkürlich mit Nullen besetzt werden.

$$\left.\begin{array}{l} x=0{,}2345 \\ y=0{,}2344 \\ \Delta x=+5\cdot 10^{-5} \\ \Delta y=-5\cdot 10^{-5} \end{array}\right\} \Rightarrow \frac{\Delta x-\Delta y}{x-y}=\frac{10^{-4}}{10^{-4}}=1\simeq 100\,\%$$

Das bedeutet, dass das Ergebnis mit einem maximalen Fehler von 100% behaftet ist.

Die Numerik hat nun geeignete Verfahren entwickelt, um die größtmöglichen Fehler eines Algorithmus angeben zu können. Für das hier vorliegende Problem bietet sich die Rückwärtsanalyse an, die als Ergebnis das *c*-fache von eps als maximalen Fehler liefert. Der Vorteil dieses Verfahrens ist, dass die Konditionierung des Problems in *c* hervortritt und man somit ein Maß für die Konditionierung des konkreten Problems in den Händen hält.

Bei der Rückwärtsanalyse geht man davon aus, dass die Ergebnisse jeder Operation mit einem kleinen Fehler behaftet sind. Fasst man all die Fehler zusammen und konzentriert sie auf die Eingangsgrößen, so kann man den Algorithmus als eine korrekt ausgeführte Operation mit gestörten Eingangsdaten betrachten. Für die quadratische Gleichung $0=x^2-2a_1x+a_2$ bekommt man beispielsweise nach einer Verfolgung des Fehlers, wenn man die kleinere Lösung mittels „Mitternachtsformel“ $x_1=a_1-\sqrt{a_1^2-a_2}$ finden möchte[11]:

$$\left.\begin{array}{l} x^2-2xa_1(1+\epsilon_1)+a_2(1+\epsilon_2)=0 \\ \epsilon_1\le eps \\ \epsilon_2\le eps\cdot(5+4\frac{a_1^2}{|a_2|}) \end{array}\right\} \Rightarrow \begin{array}{l} |a_1|\ll|a_2|\Rightarrow \textit{gut konditioniert} \\ 0<|a_2|\ll 1<|a_1|\Rightarrow \textit{schlecht konditioniert} \end{array}$$

Für die meisten Fälle muss der Fehler aber über viele Schritte verfolgt werden. Dafür bietet sich die Taylor-Reihe an. Da die Fehler meist sehr klein sind, kann man auf Glieder höherer Ordnung bei der Entwicklung verzichten. Der Vollständigkeit halber wird der Rest mittels Landau-Notation abgeschätzt[12]. Hat man also eine Funktion $f(x)=y$ und will etwas über den Fehler des Ergebnisses wissen, so lässt sich leicht mittels Taylor-Reihe folgende Aussage treffen:

$$\begin{array}{l} y+\Delta y=f(x+\Delta x)=f(x)+f'(x)\Delta x+O(\Delta x^2) \\ \Delta y=f'(x)\Delta x+O(\Delta x^2) \\ \epsilon_y=\frac{\Delta y}{y}=\frac{f'(x)\Delta x}{y}=\frac{f'(x)\Delta x}{y}\frac{x}{x}=\frac{xf'(x)}{y}\epsilon_x \end{array}$$

Glg. 2.2.xi

[11] Das Beispiel ist aus [4]

[12] Diese Notation $O(\eta)$ besagt, dass die eigentliche Funktion für sehr große oder sehr kleine Werte nicht schneller steigt oder fällt als die Landau Funktion. Für $f(x)=x^3-x^2$ ist die Landau-Notation beispielsweise $O(x^3)$

Die Glieder höherer Ordnung werden beim relativen Fehler nicht mehr mit aufgeführt und spielen im Normalfall auch keine Rolle. Offensichtlich wurde in Glg. 2.2.xi ein Ausdruck für die Verstärkung des Fehlers gefunden. Dieser Einfluss wird **Konditionszahl** genannt

Def. vi : $cond_x = \left| \frac{x f'(x)}{y} \right|$ heißt die Konditionszahl bezüglich x der Berechnungsvorschrift $y = f(x)$ und lässt sich somit auch schreiben als

$$cond_x = \left| \frac{x f'(x)}{f(x)} \right|$$

Gut konditioniert ist ein Problem mit kleinen Konditionszahlen. Im Idealfall ist der relative Fehler des Ergebnisses so groß wie der Eingangsfehler. Die Konditionierung eines Problems hat nichts mit der Rechengenauigkeit zu tun. Auch mit unvorstellbar großen Mantissen kann ein schlecht konditioniertes Problem nicht verwertbare Lösungen erzeugen. Es fällt sofort ins Auge, dass ein Problem immer dann schlecht konditioniert ist, wenn entweder die Funktionswerte betragsmäßig sehr klein werden, sich also in der Nähe einer Nullstelle befinden, oder die Ableitung (die der Änderung zu den benachbarten Funktionswerten entspricht) sehr groß wird. Ganz schlecht ist es, wenn beides aufeinander trifft.

Man sollte gedanklich die Konditionierung eines Problems vom jeweiligen Lösungsverfahren trennen. Die Konditionierung ist quasi nur eine Art obere Grenze, die man mit dem besten aller Verfahren nicht überschreiten kann. Es führt kein Weg daran vorbei, die Stabilität des Lösungsverfahrens zu untersuchen. Ein stabiles Verfahren liefert für ein gut konditioniertes Problem gute Lösungen, für ein schlecht konditioniertes Problem höchstens annehmbare Ergebnisse. Kennt man allerdings die Konditionszahl quantitativ, so kann man bei einem stabilen Verfahren den maximal zu erwartenden Fehler angeben. Ist die Konditionszahl zum Beispiel 10^8 und die Maschinengenauigkeit 10^{-16} so kann man immer noch mit einer Genauigkeit von 10^{-8} rechnen, was für die Probleme der Statik durchaus akzeptabel ist.

2.2.2 Konditionszahl von linearen Gleichungssystemen

In diesem Abschnitt soll es darum gehen, wie man die Konditionszahl für Gleichungssysteme ermittelt.

Um ein griffiges skalares Ergebnis zu erhalten, muss die Norm von Vektoren auf Matrizen erweitert werden. Im Bereich der Matrizen ist das Errechnen von Normen allerdings nicht ganz so trivial wie bei Vektoren. Allgemein lässt sich sagen, dass folgende drei Bedingungen gelten müssen:

$$\|A\| \geq 0$$

$$\|\lambda A\| = |\lambda| \|A\|$$

$$\|A+B\| \leq \|A\| + \|B\|$$

Glg. 2.2.xii

Aus diesen Forderungen ergibt sich die Gruppe der Matrix p-Normen. Die euklidische Norm ist aus dieser Gruppe die Matrix 2-Norm. Sie ist wie folgt definiert:

Def. vii :Die Matrix 2-Norm ergibt sich zu

$$\|.\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

Sie erfüllt neben den Bedingungen aus Glg. 2.2.xii noch die Eigenschaft der Submultiplikativität

$$\|A \cdot B\|_2 \leq \|A\|_2 \cdot \|B\|_2$$

und ist mit der euklidischen Vektornorm verträglich:

$$\|A \cdot y\|_2 \leq \|A\|_2 \cdot \|y\|_2$$ [13]

Geometrisch lässt sich die Matrix 2-Norm als die größtmögliche Verlängerung des Vektors *x* durch die Matrix *A* deuten. Das ist nichts weiter als der größte Eigenwert.

Um nun die Kondition linearer Gleichungssysteme herzuleiten, geht man davon aus, dass die Matrix *A* mit Maschinenzahlen korrekt besetzt ist[14]. Der Vektor *x* ist von nun an eine Funktion über den Fehler und das Ergebnis *b* wird somit ebenfalls behaftet:

$$Ax(\epsilon_x) = b + \epsilon_b f$$

$$x(o) = x$$

Glg. 2.2.xiii

Mit dem Vektor *f*, der die gleiche Dimension hat wie *x* und *b*, kann man den Fehler verteilen. Im einfachsten Fall ist er ein Einheitsvektor. Stellt man Glg. 2.2.xiii nach *x* um und differenziert *x* einmal nach ε, so lässt sich die Taylor-Reihe entwickeln:

[13] Beweis siehe [3]

[14] Man kann *A* ebenso mit einem Fehler versehen, was zu einer genaueren Betrachtung führt, was hier aber nicht nötig ist, da der Fehler offensichtlich hervortreten wird.

$$x(\epsilon_x) = A^{-1}(b + f\ \epsilon_b)$$

$$x'(\epsilon_x) = A^{-1} f$$

$$x(\epsilon_x) = x(0) + \epsilon_x x'(0) + O(\epsilon_x^2) \Leftrightarrow$$

$$x(\epsilon_x) - x(0) = \epsilon_x x'(0) + O(\epsilon_x^2) \overset{x'(0)=A^{-1}f}{\Rightarrow}$$

$$x(\epsilon_x) - x(0) = \epsilon_x A^{-1} f + O(\epsilon_x^2)$$

Glg. 2.2.xiv

Nun macht man den Übergang zu den Beträgen und vom absoluten zum relativen Fehler. Durch die Submultiplikativität der Matrix p-Normen und deren Verträglichkeit mit Vektornormen wird aus der Gleichung eine Ungleichung, da das Produkt $A^{-1} f$ getrennt werden muss, um die Kondition deutlich zu machen:

$$x(\epsilon) - x(0) = \epsilon A^{-1} f + O(\epsilon^2) \overset{\|.\|}{\Rightarrow}$$

$$\|x(\epsilon) - x(0)\| = |\epsilon| \|A^{-1} f\| + O(\epsilon^2) \overset{\|A^{-1}f\| \le \|A^{-1}\| \|f\|}{\Rightarrow}$$

$$\|x(\epsilon) - x(0)\| \le |\epsilon| \|A^{-1}\| \|f\| + O(\epsilon^2) \overset{\frac{1}{\|x(0)\|}}{\Rightarrow}$$

$$\frac{\|x(\epsilon) - x(0)\|}{\|x(0)\|} \le |\epsilon| \|A^{-1}\| \frac{\|f\|}{\|x(0)\|} + O(\epsilon^2) \overset{\|x(0)\| \ge \frac{\|b\|}{\|A\|}}{\Rightarrow}$$

$$\frac{\|x(\epsilon) - x(0)\|}{\|x(0)\|} \le |\epsilon| \underbrace{\|A^{-1}\| \|A\|}_{cond_x} \frac{\|f\|}{\|b\|} + O(\epsilon^2) \overset{\|.\|_2}{\Rightarrow}$$

$$\boldsymbol{\epsilon_x \le |\epsilon_b| cond_x \frac{\|f\|_2}{\|b\|_2} + O(\epsilon^2)}$$

$$\boldsymbol{cond_x = \|A^{-1}\|_2 \|A\|_2}$$

Glg. 2.2.xv

In der fünften Zeile von Glg. 2.2.xv kann man erkennen, dass die Kondition nicht nur für die Matrix 2-Norm definiert ist, sondern für alle Matrix p-Normen. So lässt sich, auch wenn keine Eigenwerte vorliegen, für ein konkretes Problem schnell die Konditionszahl bestimmen. Dafür bietet sich entweder die Matrix 1-Norm oder die Matrix ∞-Norm an.

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$$
$$\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$$

Glg. 2.2.xvi

Das ist nichts weiter, als die betragsmäßig maximale Spalte beziehungsweise Zeile. Im Augenblick liegt nur ein Werkzeug vor, um die Kondition eines Gleichungssystems ausdrücken zu können.

2.2.3 Bezug auf den Cauchy - Greenschen - Verzerrungstensor

Nun soll gezeigt werden, wann die geometrische Nichtlinearität die Gleichungssysteme der FEM in Bereiche führt, in denen sie schlecht konditioniert sind. Ausgehend von folgender Darstellung, lässt sich eine Funktion für die Verschiebung jeder Stelle des Kragarms ableiten.

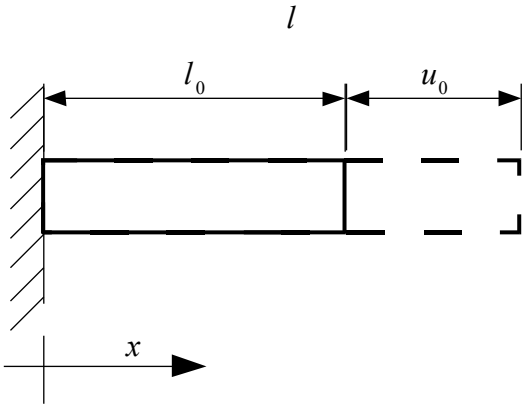

Abbildung 2.11 gedehnter Kragarm

$$u(x) = \frac{x}{l_0} u_0$$
$$x \in \{0 \le x \le l_0\}$$

Glg. 2.2.xvii

Aus Glg. 2.2.xvii kann man leicht die Eulersche und Greensche Lagrangesche Verzerrung herleiten.

$$\epsilon = u'(x) = \frac{u_0}{l_0}$$
$$E = \frac{1}{2}(2u'(x) + u'(x)^2) = \frac{1}{2}\frac{u_0}{l_0}(2 + \frac{u_0}{l_0})$$

Glg. 2.2.xviii

Zur Veranschaulichung der Ergebnisse, dient folgendes Diagramm, das die Verzerrung eines einen eins langen Kragarms über die eingeprägte Verschiebung u_0 darstellt. Mann kann deutlich erkennen, wie sich die Glieder höherer Ordnung, die bei der Polplankinematik vernachlässigt werden, auswirken. Bei einer Verschiebung von -1 (was einem totalen Zusammendrücken des Kragarms entspricht) nimmt die Eulersche Verzerrung den Wert -1 an, die Green Lagrangesche Verzerrung hingegen -0,5. Daraus lässt sich erkennen, dass der Verzerrungstensor sich offensichtlich analytisch nicht nicht gegen ein „Null-werden“ der Determinante wehrt. Dies ist somit Aufgabe des Materialgesetztes. Im nächsten Kapitel wird noch gezeigt, dass das Materialgesetzt diese Aufgabe auch erfüllt. Interessant ist hingegen der Bereich großer Dehnungen. Man beobachtet hier einen starken in diesem eindimensionalen Fall parabolischen Anstieg der Verzerrung. Wenn die hergeleitete Theorie der Kondition stimmt, müsste die Qualität des Gleichungssystems mit größer werdenden Dehnungen immer schlechter werden.

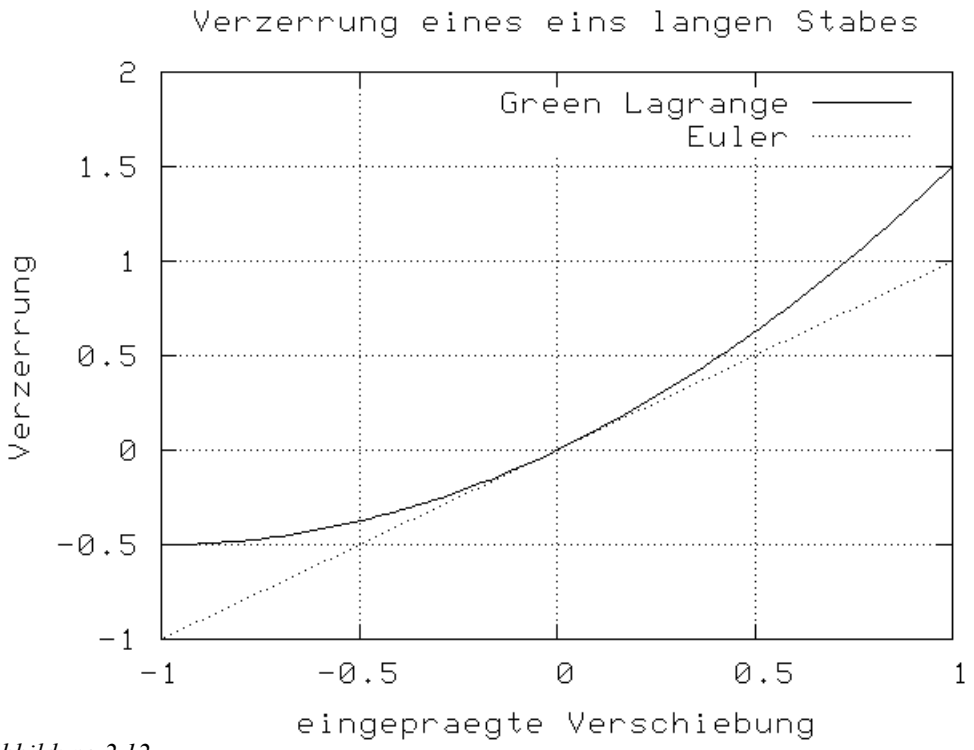

Abbildung 2.12

Für solche Aufgaben empfiehlt es sich, stabile Algorithmen einzusetzen. Das Cholesky-Verfahren ist sicherlich keine Zerlegung aus der stabilen Gruppe. Um aufzuzeigen warum das so ist, muss der Algorithmus komponentenweise abgeschätzt werden. Zunächst wird das Verfahren allerdings noch einmal genau vorgestellt, wobei sich bei der Herleitung der Algorithmen das Problem bereits deutlich zeigt.

2.2.4 Cholesky Zerlegung

Bei der Choleskyzerlegung handelt es sich um eine Verfahren, das lineare Gleichungssysteme $A_{ij} x_j = b_i$ deren Lösung eindeutig ist, in eine lösbare Form zu überführen. Dieses Verfahren eignet sich in $\mathbb{R}$ ausschließlich für symmetrische und positiv definite Matrizen. Das Lösen erfolgt in drei Schritten

1. Zerlegen der Matrix A

$$A_{ij} = G_{ik} G_{kj}^T$$

Glg. 2.2.xix

2. Vorwärtsauflösung der Substitution

$$G_{ik} y_k = b_i$$

Glg. 2.2.xx

3. Rückwertsauflöung der Desubstitution

$$G_{kj}^T x_j = y_k$$

Glg. 2.2.xxi

Die Zerlegung ist mit elementaren Rechenoperationen einfach herzuleiten

$$G_{ik} = \begin{pmatrix} \sqrt{a_{11}} & 0 & 0 & 0 & \cdots \\ \frac{a_{21}}{g_{11}} & \sqrt{a_{22} - g_{21}^2} & 0 & 0 & \cdots \\ \frac{a_{31}}{g_{11}} & \frac{a_{32} - g_{31} g_{21}}{g_{22}} & \sqrt{a_{33} - g_{31}^2 - g_{32}^2} & 0 & \cdots \\ \frac{a_{41}}{g_{11}} & \frac{a_{42} - g_{41} g_{21}}{g_{22}} & \frac{a_{43} - g_{41} g_{31} - g_{42} g_{32}}{g_{33}} & \sqrt{a_{44} - g_{41}^2 - g_{42}^2 - g_{43}^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Abbildung 2.13

Es empfiehlt sich spaltenweise vorzugehen. Für die Diagonalelemente gilt dann:

$$g_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{k<i} g_{ik}^2}$$

Glg. 2.2.xxii

Die unter dem entsprechenden Diagonalelementen liegenden Koeffizienten errechnen sich demnach zu:

$$g_{ij} = \frac{a_{ij} - \sum_{k=1}^{k<j} g_{ik} g_{jk}}{g_{ii}}$$

Glg. 2.2.xxiii

Hat man die Matrix zerlegt, erfolgt die Vorwärtsauflösung:

$$y_i = \frac{b_i - \sum_{k=1}^{k<i} y_k g_{ik}}{g_{ii}}$$

Glg. 2.2.xxiv

Die Rückwärtsauflösung errechnet sich demnach zu:

$$x_i = \frac{y_i - \sum_{k=n}^{k>i} x_k g_{ki}}{g_{ii}}$$

Glg. 2.2.xxv

Wobei die Matrix *A n* Spalten und Zeilen hat.

Wie aus Abbildung 2.13 ersichtlich wird, hat man bei der Zerlegung mit dem Phänomen der Auslöschung zu kämpfen. Wie in Glg. 2.2.x bereits hergeleitet, verstärkt sich ein Fehler, wenn zwei nahezu gleich große Terme voneinander abgezogen werden. Dadurch wird die Kommastelle beim Normalisieren des Ergebnisses nach rechts verschoben und die so entstehenden Stellen in der Mantisse werden mit Nullen besetzt. Diese Ziffer ist willkürlich und hat nichts mit dem konkreten Problem zu tun. Bei einer komponentenweisen Abschätzung kann man von folgender Gleichung ausgehen:

$$H_{ij} = A_{ij} - \tilde{G}_{ik} \tilde{G}_{kj}^{T}$$

Glg. 2.2.xxvi

Wobei die mit der Tilde versehenen Größen fehlerbehaftet sind. Unter Ausnutzung der Symmetrie von *A* benötigt man für die Berechnung jedes Elements von *H* respektive *A* genau *n* flops[15] wobei *n* der Spaltenindex ist. Daraus ergibt sich ein relativer Fehler von:

$$|H_{ij}| \leq eps \cdot n \left(|A_{ij}| + |\tilde{G}_{ik}| |\tilde{G}_{kj}^{T}| \right)$$

Glg. 2.2.xxvii

[15] ein flop ist eine Gleitkommaoperation

Woraus ersichtlich wird, dass die Cholesky Zerlegung bei betragsmäßig großen Koeffizienten schlechte Ergebnisse liefert, was ja auch der Erfahrung entspricht. Um den theoretischen Teil abzuschließen, soll hier noch ein kleines Zahlenbeispiel folgen, das das obige Problem illustriert. Dabei wird eine kleine Matrix mittels Cholesky zerlegt. Danach wird Glg. 2.2.xxvi ausgeführt. Zuerst wird die Matrix mit „normalen“ Zahlen besetzt. Erwartungsgemäß ist *H* nur mit Nullen besetzt.

$$A=\begin{bmatrix} 11 & & \\ 3 & 13 & \\ 5 & 7 & 17 \end{bmatrix} \qquad H=\begin{bmatrix} 0 & & \\ 0 & 0 & \\ 0 & 0 & 0 \end{bmatrix}$$

Danach werden die Koeffizienten der Matrix stark vergrößert und man erhält

$$A=\begin{bmatrix} 1.1E11 & & \\ 3 & 1.3E11 & \\ 5 & 7 & 1.7E11 \end{bmatrix} \qquad H=\begin{bmatrix} 1.5E\text{-}5 & & \\ 0 & 1.5E\text{-}5 & \\ 0 & 0 & 0 \end{bmatrix}$$

Der Fehler erscheint nicht sehr groß zu sein, man muss sich dabei aber vor Augen halten, dass es sich hierbei um eine 3x3 Matrix handelt. Normale Systeme haben etliche hundert Zeilen und Spalten. Man sollt dabei auch nicht vergessen, dass nicht lineare Systeme immer eine Iteration benötigen. Somit verstärkt sich der Fehler nochmal.

2.2.5 Ausblick und Lösungsstrategie

In modernen Programmsystemen wird meist ein kombiniertes Verfahren zum Lösen von Gleichungssystemen verwendet [3]. Dabei wird mit einem direkten Löser, wie zum Beispiel dem Cholesky Löser, ein Startvektor errechnet und diesem werden die Fehler in einem iterativen Verfahren ausgetrieben. Programmintern kann man bei jedem Rechenschritt die Konditionszahl des Gleichungssystems bestimmen und sich dann entscheiden, ob man den relativ teuren iterativen Löser einschaltet oder nicht. Die Literatur empfiehlt zum Beispiel das Gradientenverfahren als schnell konvergierende Strategie. Der Grundgedanke dieses Lösungsverfahrens ist einfach und genial.

Es fällt auf, dass ein n-dimensionaler Paraboloid folgender Darstellung genügt:

$$F(x_i)=\frac{1}{2}x_i A_{ij} x_j - b_i x_i$$

$$F(\underline{x})=\frac{1}{2}\underline{x}^T \underline{\underline{A}}\,\underline{x} - \underline{b}^T \underline{x}$$

Glg. 2.2.xxviii

Für den zweidimensionalen Fall erzeugt die Gleichung ein ähnliches Bild wie in Abbildung 2.14. Fordert man nun noch, dass A symmetrisch und positiv definit ist, so hat die Funktion keine Nullstelle, sonder nur ein Minimum. Das Minimum errechnet sich aber, indem man den Gradienten von $F(x_i)$ also $\nabla F(x_i)$ zu Null setzt,

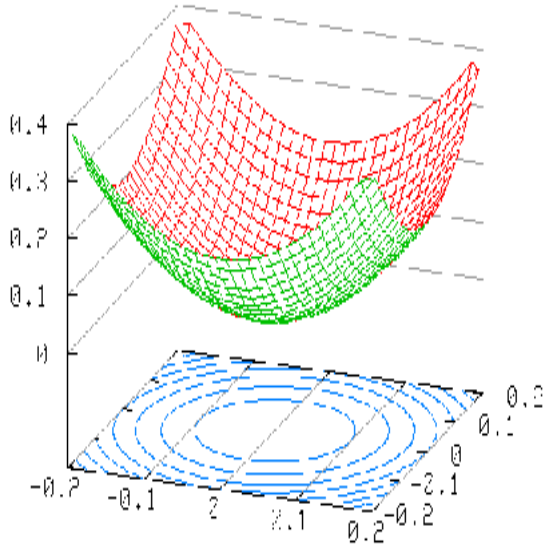

Abbildung 2.14

was einer waagrechten Tangentenfläche entspricht. Der Gradient von $F(x_i)$ ist genau $A_{ij} x_j - b_i$. Man kann sich von der Richtigkeit dieser Aussage überzeugen, wenn man die einzelnen Terme für einen kleinen Raum ausmultipliziert und dann ableitet. Das Minimum von *F* entspricht als genau der Lösung des Gleichungssystems. Man löst das Problem der linearen Gleichungssysteme also ab und ersetzt es durch ein Minimumproblem.

Für eine Iteration des Minimums verwendet man folgenden Ansatz:

$$x_i^{k+1} = x_i^k + \alpha^k d_i^k$$

Glg. 2.2.xxix

Dabei bezeichnet *k* den Iterationsschritt, d_i^k ist die Suchrichtung, die eine Verkleinerung des Funktionswertes von *F* birgt und α^k ist die Schrittweite, die man in Richtung d_i^k gehen muss. Bevor man nun den nächsten und entscheidenden Schritt machen will, sollte man sich folgende Kausalkette vor Augen halten.

Die Funktion *F* ist ein Skalar. Somit ist der Gradient von *F* ein Vektor. Multipliziert man den Gradienten mit einer Richtung auf der Oberfläche des Paraboloiden, so erhält man die Steigung dieser Richtung. Die größte positive Steigung ergibt sich aber immer für die Richtung, die auf einem Meridian vom Mittelpunkt weg liegt. Das Skalarprodukt zweier Vektoren entspricht dem Produkt der Beträge mal dem Kosinus des eingeschlossenen Winkels. Es wird maximal, wenn der Winkel Null wird, also deutet der Gradient immer genau von der Mitte weg zum Rand des Paraboloiden.

Somit ist es sinnvoll, als Vektor d_i^k den negativen Gradienten zu verwenden, da dieser immer zum Mittelpunkt deutet. Weiterhin ergibt sich aus der obigen Kausalkette, dass man jeden beliebigen Startvektor nehmen kann.

$$d_i^k = b_i - A_{ij} x_j$$

Glg. 2.2.xxx

Als letzter Schritt muss nun nur noch ein geeignetes α^k gewählt werden, damit der Ausdruck $F(x_i^k + \alpha^k (b_i - A_{ij} x_j^k))$ minimal wird. Dazu setzt man lediglich die Funktionsstelle in den Gradienten ein, löst das Gleichungssystem nach α^k auf und erhält:

$$\alpha^k = \frac{d_i^k d_i^k}{d_i^k A_{ij} d_j^k}$$

Glg. 2.2.xxxi

was in Zusammenhang mit Glg. 2.2.xxx ein Konvergenzbeweis ist. Somit ergibt sich für die Iteration folgende Gleichung:

$$x_i^{k+1} = x_i^k + \frac{\| b_i - A_{ij} x_j^k \|_2^2}{(b_i - A_{ij} x_j^k)^T \, Aij \, (b_j - A_{ji} x_i^k)} (b_i - A_{ij} x_j^k)$$

Glg. 2.2.xxxii

Kapitel 3

Bewertung der Ergebnisse

3.1 Umfassendes Beispiel

Das dritte Kapitel soll sich mit einem umfangreichen Beispiel aus dem Bereich der Statik beschäftigen. Dazu wurde ein Modell generiert, das der Geometrie von Abbildung 3.1 entspricht. Die Kantenlänge der einzelnen Elemente ist 1. Das Materialmodell ist linear elastisch, der E-Modul beträgt 1,0. Die Querdehnungszahl ist Null, somit ist nicht mit Querdehnung zu rechnen. Die Dichte ist ebenfalls auf 1,0 gesetzt. Damit ein numerischer Fehler bei der Berechnung der Verschiebung sich nicht als Stabilitätsversagen[16] tarnen kann, sind alle Knoten zwängungsfrei geführt (In der Abbildung wurden die Gleitlager an den Mittelknoten ausgeblendet, um das Bild übersichtlich zu halten).

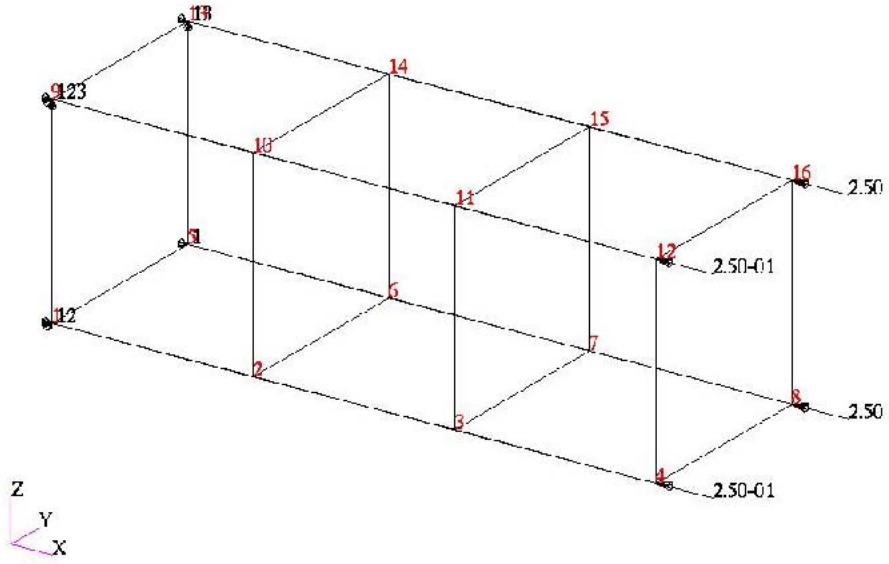

Abbildung 3.1

Das verwendete Volumenelement ist gemischt hybrid formuliert und am Fachgebiet für Baustatik an der technischen Universität Berlin entwickelt worden. Es hat alle üblichen Patch-Tests bestanden, unter andern auch den Durchschlagsversuch am Zweibock. Das Element entspricht, da es gemischt hybrid ist, genau den vorgestellten Gleichungssystemen der Form $Ax=b$, wobei A die Steifigkeitsmatrix, x die Verschiebung und b die eingeprägte Kraft ist. Der Löser ist ein aus der linearen Theorie kommendes Programmpaket mit dem Namen DAMAGE. In ihm ist ein

[16] Das System würde so einen Fehler als Imperfektion interpretieren.

spaltenweise pivotisierender[17] Cholesky-Löser eingebaut. Die Approximation des aus der Nichtlinearität entstehenden Anfangswertproblems wird mittels Zeitintegrationsverfahren bewerkstelligt. Das Materialmodell ist einfach linear elastisch. Es soll später gezeigt werden, dass sich das Element aus seiner Lagrangen Formulierung heraus nicht stark zusammendrücken lässt. Die Analytik des Elements versagt an einem vorher bestimmbaren Punkt. Dieser Punkt liegt weit vor dem numerischen Versagen, so dass dieses im Druckversuch nicht gezeigt werden kann.

3.1.1 Zeitschrittintegration mit Prediktor Korrektor Verfahren

Das im folgenden beschriebene Verfahren wurde Implementiert, um die geometrische Nichtlinearität zu bewerkstelligen. Die geometrische Nichtlinearität verbirgt sich dabei in der Operatorenmatrix, die nun aus einem linearen und einem nichtlinearen Anteil besteht.

$$\underline{\underline{D}} = \underline{\underline{D}}_{lin} + \underline{\underline{D}}_{nlin}$$

Glg. 3.1.i

Der lineare Anteil wird einmal am Anfang im Ort diskretisiert und aufgebaut. Danach ändert er sich nicht mehr. Der nichtlineare Anteil muss natürlich bei jedem Schritt neu aufgebaut werden, um die neu errechneten Verschiebungen in ihn einfließen zu lassen. In der Operatorenmatrix steckt also der ganze geometrisch nichtlineare Gedanke.

Bei dynamischen Formulierungen müssen die Größen nicht nur im Ort sondern auch in der Zeit diskretisiert werden. Für die Ortsdiskretisierung verwendet man keine Arbeitsgleichung (AGL), sondern eine Leistungsgleichung (LGL). Leistung ist die Änderung der Arbeit in der Zeit. Die Matrizendarstellung der Leistungsgleichung hat folgende Gestalt:

$$\boldsymbol{PvW}: \int \delta \underline{v}^T \left(\underline{\underline{D}}^T \underline{\sigma} - \underline{\bar{p}} + \underline{\underline{B}}\,\underline{u} + \underline{\underline{M}}\,\underline{\dot{v}} + \underline{\underline{C}}\,\underline{v} \right) dV = 0$$

$$\boldsymbol{PvK}: \int \delta \underline{\sigma}^T \left(\underline{\underline{D}}\,\underline{v} - \underline{\underline{E}}^{-1}\,\underline{\dot{\sigma}} \right) dV = 0$$

Glg. 3.1.ii

Dabei haben die Symbole folgende Bedeutung:

- D Operatorenmatrix (Kinematik)
- B Bettungsmatrix
- M Massenmatrix
- C Dämpfungsmatrix
- E^{-1} Nachgiebigkeitsmatrix
- v Geschwindigkeitsvektor

[17] Beim spaltenweisen Pivotisieren wird das betragsmäßig größte Element einer Spalte in die Diagonale getauscht. Das stabilisiert das Zerlegen der Matrizen etwas.

- σ Spannungsvektor
- p Vektor der eingeprägten Lasten
- u Verschiebungsvektor

Für die Diskretisierung im Ort verwendet man die entsprechenden Ansatzfunktionen:

$$\underline{u} = \sum_L \Phi(\xi^j)\,\underline{\hat{u}_L}$$
$$\underline{v} = \sum_L \Phi(\xi^j)\,\underline{\hat{v}_L}$$
$$\underline{\dot{v}} = \sum_L \Phi(\xi^j)\,\underline{\dot{\hat{v}}_L}$$
$$\underline{\bar{p}} = \sum_L \Phi(\xi^j)\,\underline{\hat{\bar{p}}_L}$$
$$\underline{\sigma} = \underline{\underline{T}}\,\underline{\underline{\psi}}\,\underline{\hat{\sigma}}$$

Glg. 3.1.iii

Dabei haben die Symbole folgende Bedeutung:

- $\Phi(\xi^j)$ Ansatzfunktion für die Diskretisierung im Ort
- j Elementkoordinaten (j = 1 , 2 , 3)
- L Elementknoten (L = A , B , ... , H)
- T Transformationsmatrix auf die Globalen Koordinaten
- ψ Ansatzmatrix der Spannungen
- (^) Kennzeichen für diskrete Werte im Ort

Man darf dabei nicht außer Acht lassen, dass alle Funktionen noch von der Zeit abhängen. Integriert man über die Ansatzfunktionen und konstanten Glieder LGL, so erhält man folgenden Ausdruck:

$$W^v = \left\{\delta\hat{\underline{v}}^T \,|\, \delta\hat{\underline{\sigma}}^T\right\} \left(\begin{bmatrix} \underline{\underline{\hat{M}}} & 0 \\ 0 & \underline{\underline{\hat{E}^{-1}}} \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}} \\ \dot{\hat{\underline{\sigma}}} \end{bmatrix} + \begin{bmatrix} \underline{\underline{\hat{C}}} & \underline{\underline{\hat{D}^T}} \\ \underline{\underline{\hat{D}}} & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{v}} \\ \hat{\underline{\sigma}} \end{bmatrix} + \begin{bmatrix} \underline{\underline{\hat{B}}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{u}} \\ 0 \end{bmatrix} - \begin{bmatrix} \hat{\underline{\bar{p}}} \\ 0 \end{bmatrix} \right) = 0$$

Glg. 3.1.iv

Es gilt also das Anfangswertproblem von Glg. 3.1.iv mittels Zeitschrittintegration zu lösen. Dazu bietet sich ein gemischtes Verfahren aus Euler vorwärts und Euler rückwärts an. Bei diesem werden die Raten ($\dot{\hat{\underline{v}}}$ und $\dot{\hat{\underline{\sigma}}}$) in der Intervallmitte ermittelt um dann auf die gesuchten

Größen am Intervallende zu Schließen. Nach dem Ermitteln der Raten wird allerdings die neue Geometrie angepasst. Man rechnet also nicht mit der geometrischen Konstellation am Anfang des Zeitschritts sondern mit der in der Mitte. Dadurch wird das Ergebnis verbessert. Folgende Zeichnung soll einen Zeitschritt illustrieren.

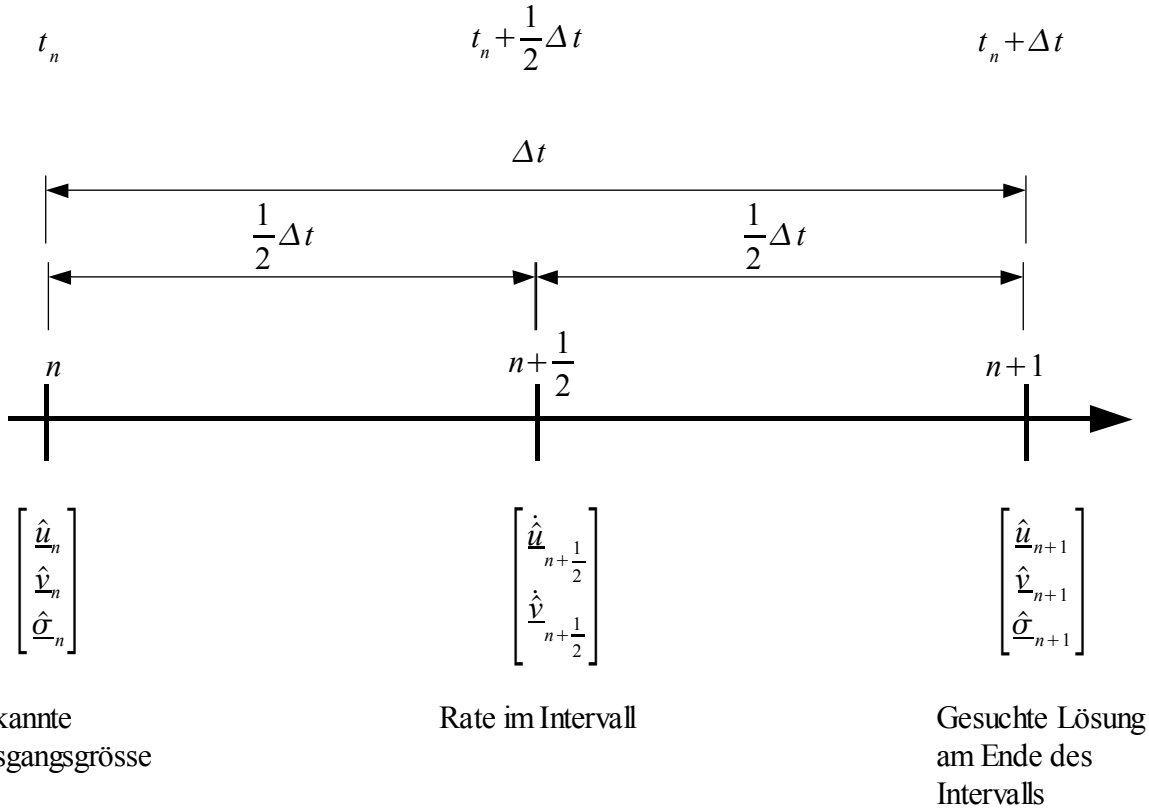

Abbildung 3.2 Zeitschritt mit Raten in der Intervallmitte

Anhand der Zeichnung lassen sich Ansätze für diskrete Größen in der Intervallmitte finden:

$$\hat{\underline{u}}_{n+\frac{1}{2}} = \hat{\underline{u}}_n + \frac{1}{2}\hat{\underline{v}}_n\Delta t + \frac{1}{4}\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\Delta t^2$$

$$\hat{\underline{v}}_{n+\frac{1}{2}} = \hat{\underline{v}}_n + \frac{1}{2}\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\Delta t$$

$$\hat{\underline{\sigma}}_{n+\frac{1}{2}} = \hat{\underline{\sigma}}_n + \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}}\Delta t$$

Glg. 3.1.v

Glg. 3.1.v kann nun in Glg. 3.1.iv eingesetzt werden:

$$\begin{bmatrix} \hat{\underline{\underline{M}}} & 0 \\ 0 & \hat{\underline{\underline{E}}}^{-1} \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}}_{n+\frac{1}{2}} \\ \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}} \end{bmatrix} + \begin{bmatrix} \hat{\underline{C}} & \hat{\underline{\underline{D}}}_n^T \\ \hat{\underline{\underline{D}}}_n & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{v}}_n + \frac{1}{2}\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\Delta t \\ \hat{\underline{\sigma}}_n + \frac{1}{2}\dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}}\Delta t \end{bmatrix} + \begin{bmatrix} \hat{\underline{\underline{B}}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{u}}_n + \frac{1}{2}\hat{\underline{v}}_n\Delta t + \frac{1}{4}\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\Delta t^2 \\ 0 \end{bmatrix} - \begin{bmatrix} \hat{\underline{\underline{p}}}_{n+\frac{1}{2}} \\ 0 \end{bmatrix} = \mathbf{0}$$

Glg. 3.1.vi

Spaltet man von die Vektoren der Raten ab, so erhält man:

$$\begin{bmatrix} \hat{\underline{\underline{M}}} & 0 \\ 0 & \hat{\underline{\underline{E}}}^{-1} \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}}_{n+\frac{1}{2}} \\ \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}} \end{bmatrix} + \begin{bmatrix} \hat{\underline{C}} & \hat{\underline{\underline{D}}}_n^T \\ \hat{\underline{\underline{D}}}_n & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{v}}_n \\ \hat{\underline{\sigma}}_n \end{bmatrix} + \begin{bmatrix} \frac{1}{2}\Delta t\,\hat{\underline{C}} & \frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n^T \\ \frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}}_{n+\frac{1}{2}} \\ \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}} \end{bmatrix} + \begin{bmatrix} \hat{\underline{\underline{B}}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\underline{u}}_n + \frac{1}{2}\hat{\underline{v}}_n\Delta t \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{4}\Delta t^2\,\hat{\underline{\underline{B}}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}}_{n+\frac{1}{2}} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \hat{\underline{\underline{p}}}_{n+\frac{1}{2}} \\ 0 \end{bmatrix} + = \mathbf{0}$$

Glg. 3.1.vii

Zuletzt wird das Gleichungssystem noch nach den Raten umgestellt.

$$\begin{bmatrix} \hat{\underline{\underline{M}}} + \frac{1}{2}\Delta t\,\hat{\underline{\underline{C}}} + \frac{1}{4}\Delta t^2\,\hat{\underline{\underline{B}}} & \frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n^T \\ \frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n & \hat{\underline{\underline{E}}}^{-1} \end{bmatrix} \begin{bmatrix} \dot{\hat{\underline{v}}}_{n+\frac{1}{2}} \\ \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}} \end{bmatrix} =$$

$$\begin{bmatrix} \hat{\underline{p}}_{n+\frac{1}{2}} - \hat{\underline{\underline{B}}}\left(\hat{\underline{u}}_n + \frac{1}{2}\Delta t\,\hat{\underline{v}}_n\right) - \hat{\underline{\underline{D}}}_n^T\,\hat{\underline{\sigma}}_n - \hat{\underline{\underline{C}}}\,\hat{\underline{v}}_n \\ -\hat{\underline{\underline{D}}}_n\,\hat{\underline{v}}_n \end{bmatrix}$$

Glg. 3.1.viii

Bei dem im Versuch verwendeten Element handelt es sich um eine gemischt hybride Formulierung, was impliziert, dass die Spannungen bereits auf Elementebene abgelöst und analytisch befriedigt werden. Dazu bietet sich das PvK (Glg. 3.1.ii) an:

$$\begin{aligned} &\underline{\underline{D}}_n\,\underline{v}_{n+\frac{1}{2}} - \underline{\underline{E}}^{-1}\,\dot{\underline{\sigma}}_{n+\frac{1}{2}} = 0 \\ &\dot{\underline{\sigma}}_{n+\frac{1}{2}} = \underline{\underline{E}}\,\underline{\underline{D}}_n\,\underline{v}_{n+\frac{1}{2}} \\ &\dot{\underline{\sigma}}_{n+\frac{1}{2}} = \underline{\underline{E}}\,\underline{\underline{D}}_n\left(\hat{\underline{v}}_n + \frac{1}{2}\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\Delta t\right) \\ &\dot{\underline{\sigma}}_{n+\frac{1}{2}} = \underline{\underline{E}}\,\underline{\underline{D}}_n\,\hat{\underline{v}}_n \;+\; \dot{\hat{\underline{\mathbf{v}}}}_{\boldsymbol{n+\frac{1}{2}}}\,\underline{\underline{E}}\,\underline{\underline{D}}_n\frac{1}{2}\Delta t \end{aligned}$$

Glg. 3.1.ix

Danach kann man Glg. 3.1.ix in Glg. 3.1.viii einsetzen und die Gleichung auf Systemebene nach der Beschleunigung umstellen:

$\hat{\underline{\underline{M}}} + \frac{1}{2}\Delta t\,\hat{\underline{\underline{C}}} + \frac{1}{4}\Delta t^2\,\hat{\underline{\underline{B}}}$	$\frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n^T$
$\frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n$	$\hat{\underline{\underline{E}}}^{-1}$

$\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}$
$\underline{\underline{E}}\,\underline{\underline{D}}_n\,\hat{\underline{v}}_n \;+\; \dot{\hat{\underline{\mathbf{v}}}}_{\boldsymbol{n+\frac{1}{2}}}\,\underline{\underline{E}}\,\underline{\underline{D}}_n\frac{1}{2}\Delta t$

=

$\hat{\underline{p}}_{n+\frac{1}{2}} - \hat{\underline{\underline{B}}}\left(\hat{\underline{u}}_n + \frac{1}{2}\Delta t\,\hat{\underline{v}}_n\right) - \hat{\underline{\underline{D}}}_n^T\,\hat{\underline{\sigma}}_n - \hat{\underline{\underline{C}}}\,\hat{\underline{v}}_n$
$-\hat{\underline{\underline{D}}}_n\,\hat{\underline{v}}_n$

Glg. 3.1.x Spannung analytisch befriedigt

$$\left[\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\right] \quad \left(\left[\hat{\underline{\underline{M}}}+\frac{1}{2}\Delta t\,\hat{\underline{\underline{C}}}+\frac{1}{4}\Delta t^2\,\hat{\underline{\underline{B}}}\right]+\left[\frac{1}{4}\Delta t^2\,\hat{\underline{\underline{D}}}_n^T\,\underline{\underline{E}}\,\hat{\underline{\underline{D}}}_n\right]\right) \quad =$$

$$\left[\hat{\underline{p}}_{n+\frac{1}{2}}-\hat{\underline{\underline{B}}}\left(\hat{\underline{u}}_n+\frac{1}{2}\Delta t\,\hat{\underline{v}}_n\right)-\hat{\underline{\underline{D}}}_n^T\,\hat{\underline{\sigma}}_n-\hat{\underline{\underline{C}}}\,\hat{\underline{v}}_n\right]-\left[\frac{1}{2}\Delta t\,\hat{\underline{\underline{D}}}_n^T\right]\left[\underline{\underline{E}}\,\underline{\underline{D}}_n\,\hat{\underline{v}}_n\right]$$

$$\dot{\hat{\underline{v}}}_{n+\frac{1}{2}}\overbrace{\left(\hat{\underline{\underline{M}}}+\frac{1}{2}\Delta t\,\hat{\underline{\underline{C}}}+\frac{1}{4}\Delta t^2\left(\hat{\underline{\underline{B}}}+\hat{\underline{\underline{D}}}_n^T\,\underline{\underline{E}}\,\hat{\underline{\underline{D}}}_n\right)\right)}^{\underline{\underline{k}}_e} \quad =$$

$$\underbrace{\hat{\underline{p}}_{n+\frac{1}{2}}-\hat{\underline{\underline{B}}}\,\hat{\underline{u}}_n-\hat{\underline{\underline{D}}}_n^T\,\hat{\underline{\sigma}}_n-\left(\hat{\underline{\underline{C}}}+\frac{1}{2}\Delta t\left(\hat{\underline{\underline{B}}}+\hat{\underline{\underline{D}}}_n^T\,\underline{\underline{E}}\,\hat{\underline{\underline{D}}}_n\right)\right)\hat{\underline{v}}_n}_{\underline{p}_e}$$

Glg. 3.1.xi Beschleunigung auf Systemebene

Danach werden die einzelnen Elemente über eine Belegungsmatrix $\underline{\underline{B}}_e$ zum kompletten System zusammen geschaltet:

$$\tilde{\underline{\underline{K}}}=\sum_e \underline{\underline{B}}_e^T\left(\hat{\underline{\underline{M}}}+\frac{1}{2}\Delta t\,\hat{\underline{\underline{C}}}+\frac{1}{4}\Delta t^2\left(\hat{\underline{\underline{B}}}+\hat{\underline{\underline{D}}}_n^T\,\underline{\underline{E}}\,\hat{\underline{\underline{D}}}_n\right)\right)\underline{\underline{B}}_e$$

Glg. 3.1.xii

$$\tilde{\underline{P}}_{n+\frac{1}{2}}=\sum_e \underline{\underline{B}}_e^T\left(\hat{\underline{p}}_{n+\frac{1}{2}}-\hat{\underline{\underline{B}}}\,\hat{\underline{u}}_n-\hat{\underline{\underline{D}}}_n^T\,\hat{\underline{\sigma}}_n-\left(\hat{\underline{\underline{C}}}+\frac{1}{2}\Delta t\left(\hat{\underline{\underline{B}}}+\hat{\underline{\underline{D}}}_n^T\,\underline{\underline{E}}\,\hat{\underline{\underline{D}}}_n\right)\right)\hat{\underline{v}}_n\right)\underline{\underline{B}}_e$$

Glg. 3.1.xiii

Die Geschwindigkeitsraten in Intervallmitte bestimmen sich somit über ein lineares Gleichungssystem:

$$\dot{\underline{\hat{V}}}_{n+\frac{1}{2}}=\tilde{\underline{\underline{K}}}^{-1}\,\tilde{\underline{P}}_{n+\frac{1}{2}}$$

Glg. 3.1.xiv

An dieser Stelle sollte man sich wieder ins Gedächtnis rufen, dass die Operatorenmatrix (Glg. 3.1.i) im geometrisch nichtlinearen Fall von der abhängt. Die Raten in Intervallmitte werden aber mit der Elementkonfiguration am Intervallanfang ermittelt $\hat{\underline{\underline{D}}}_n$. Man begeht also einen Fehler. Um diesen Fehler zu minimieren, schaltet man eine Prediktorschritt ein. Durch Glg. 3.1.v und die

ermittelten Raten kann man sich die Elementkonfiguration $\hat{\underline{\underline{D}}}_{n+\frac{1}{2}}$ in Intervallmitte bestimmen. Damit lässt sich der nichtlineare Teil der Operatorenmatrix für die Intervallmitte neu aufbauen. Natürlich begeht man dabei wieder einen Fehler, da die Raten ja mit der alten Operatorenmatrix $\hat{\underline{\underline{D}}}_n$ bestimmt wurden. Dennoch ist dieser Fehler nun kleiner als wenn man mit $\hat{\underline{\underline{D}}}_n$ rechnen würde. Mit der neuen Operatorenmatrix $\hat{\underline{\underline{D}}}_{n+\frac{1}{2}}$ errechnet man sich die Raten mittels Glg. 3.1.xii Glg. 3.1.xiii noch einmal. Dies bezeichnet man als Korrektorschritt. Um den Fehler zu minimieren kann man die beiden Schritte mehrmals hintereinander ausführen. Beispielsweise so oft, bis sich das Ergebnis stabilisiert. Es hat sich aber gezeigt, dass es effizienter ist lieber mit kleineren Zeitschritten zu rechnen und nur einen Prediktor- und Korrektorschritt durchzuführen.

Mit den neu ermittelten Raten bestimmt man sich die Unbekannten am Intervallende.

$$\hat{\underline{U}}_{n+1} = \hat{\underline{U}}_n + \hat{\underline{V}}_n \Delta t + \frac{1}{2} \dot{\hat{\underline{V}}}_{n+\frac{1}{2}} \Delta t^2$$

$$\hat{\underline{V}}_{n+1} = \hat{\underline{V}}_n + \dot{\hat{\underline{V}}}_{n+\frac{1}{2}} \Delta t$$

$$\hat{\underline{\sigma}}_{n+1} = \hat{\underline{\sigma}}_n + \dot{\hat{\underline{\sigma}}}_{n+\frac{1}{2}} \Delta t$$

Glg. 3.1.xv

Mit diesem eben beschriebenen verfahren wurden folgende versuche durchgeführt.

3.1.2 Zugversagen

Abbildung 2.12 zeigt bereits, dass die geometrisch nichtlineare Verzerrung bei großen Dehnungen stark zunimmt. Man sollte also ein numerisches Versagen um den Nullpunkt und sehr weit links vermuten (Dev. vi). Der Fehler, der beim lösen der Gleichungssysteme auftritt, sollte sich (wenn die Vermutung richtig ist) mit der Anzahl der Zeitschritte im negativen Sinne verstärken. Der Fehler wird ja von Zeitschritt zu Zeitschritt weitergegeben und vom Löser nicht beachtet. Weiterhin darf man auch nicht außer Acht lassen, dass die Konditionszahl nur den größtmöglichen Fehler angibt. Wenn man oft eine große Konditionszahl hat, bedeutet das natürlich auch, dass die Fehlerwahrscheinlichkeit steigt[18]. Um dieses Verhalten aufzudecken, wurde der Kragarm im ersten Versuch in fünfzig Zeitschritten stark gedehnt. In folgendem Diagramm sind die Verschiebung der

[18] Man kann sich das Bildlich vorstellen. Wenn man sein Auto einmal nicht abschließt, so ist es relativ unwahrscheinlich, dass etwas geklaut wird. Lässt man es hingegen immer unabgeschlossen, so ist es wesentlich wahrscheinlicher, dass etwas geklaut wird. Die Konditionalzahl entspricht bei diesem Beispiel dem Ort, wo man sein Auto abstellt. (St. Gallen, Berlin, Rio)

Knoten 4, 8, 12 und 16 über die eingeprägte Kraft angetragen. Alle drei Funktionsverläufe sind deckungsgleich.

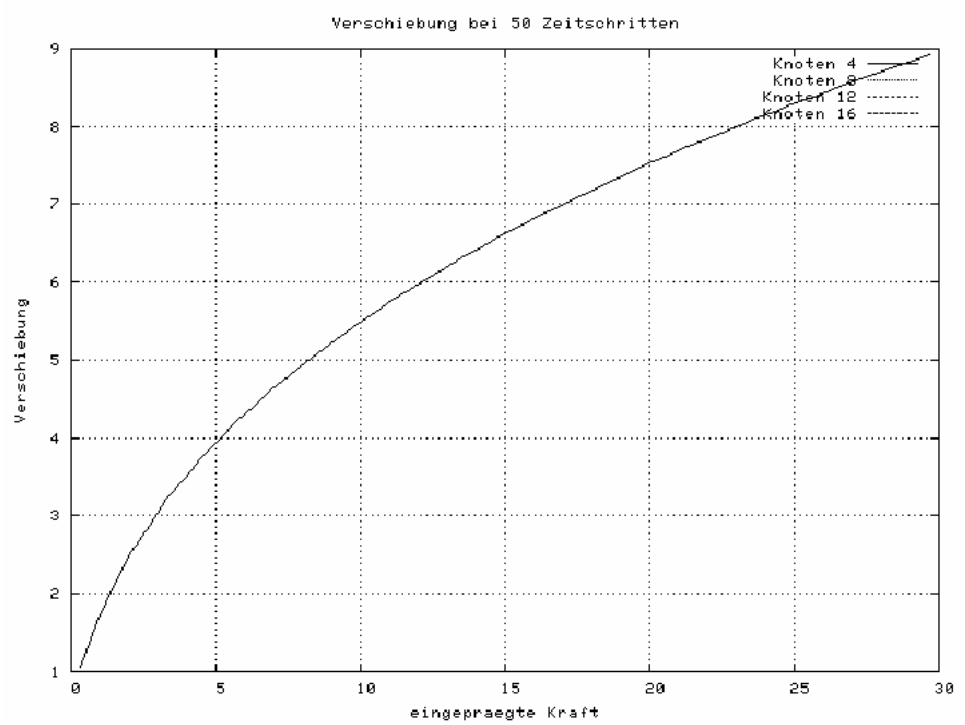

Abbildung 3.3

Für jeden Zeitschritt wurde die Konditionszahl bestimmt und in folgendem Diagramm angetragen. Die y-Achse ist dabei logarithmisch geteilt um die stark unterschiedlichen Konditionszahlen besser zu verdeutlichen.

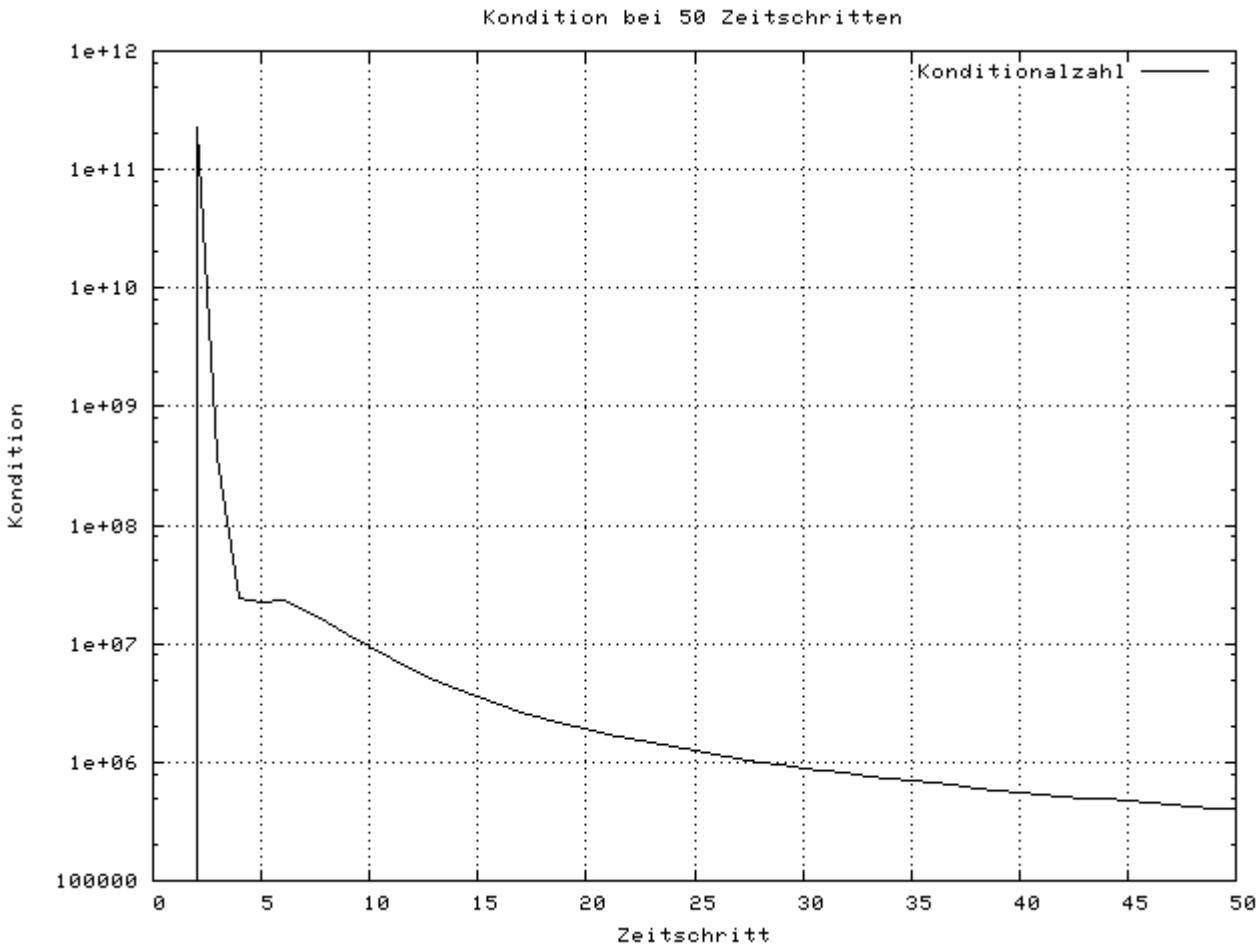

Abbildung 3.4

Es lässt sich sehr schön erkennen, dass das System mit der relativ großen Fehlerwahrscheinlichkeit am Anfang der Berechnung gut fertig wird.

Stimmt die Prognose, so müsste das System bei vielen Zeitschritten Versagen. Folgendes Diagramm verdeutlicht den Verlauf der Knotenverschiebung bei 500 Zeitschritten:

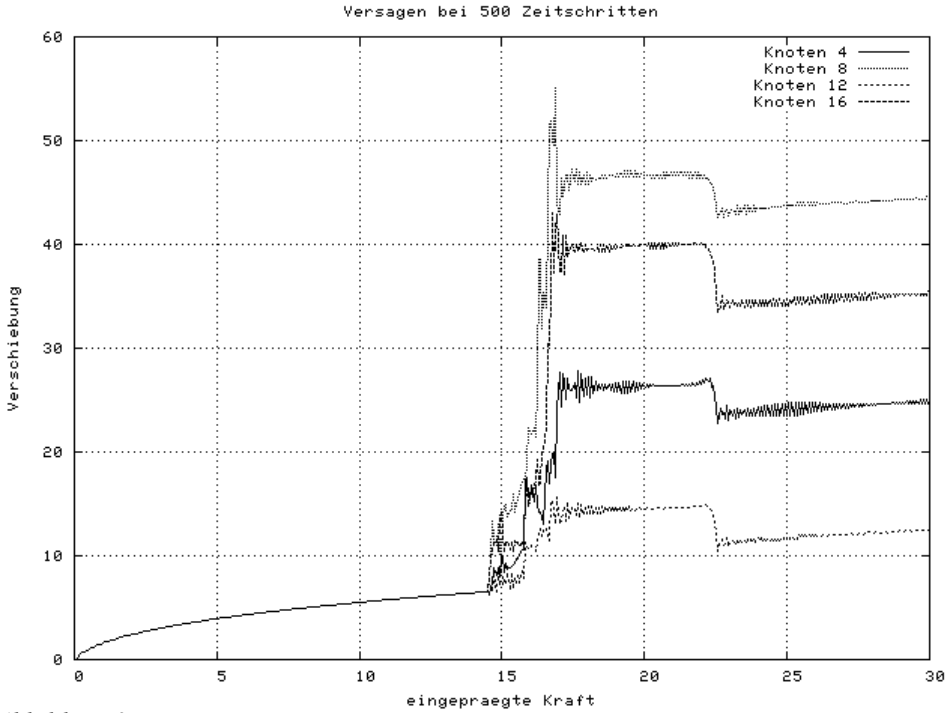

Abbildung 3.5

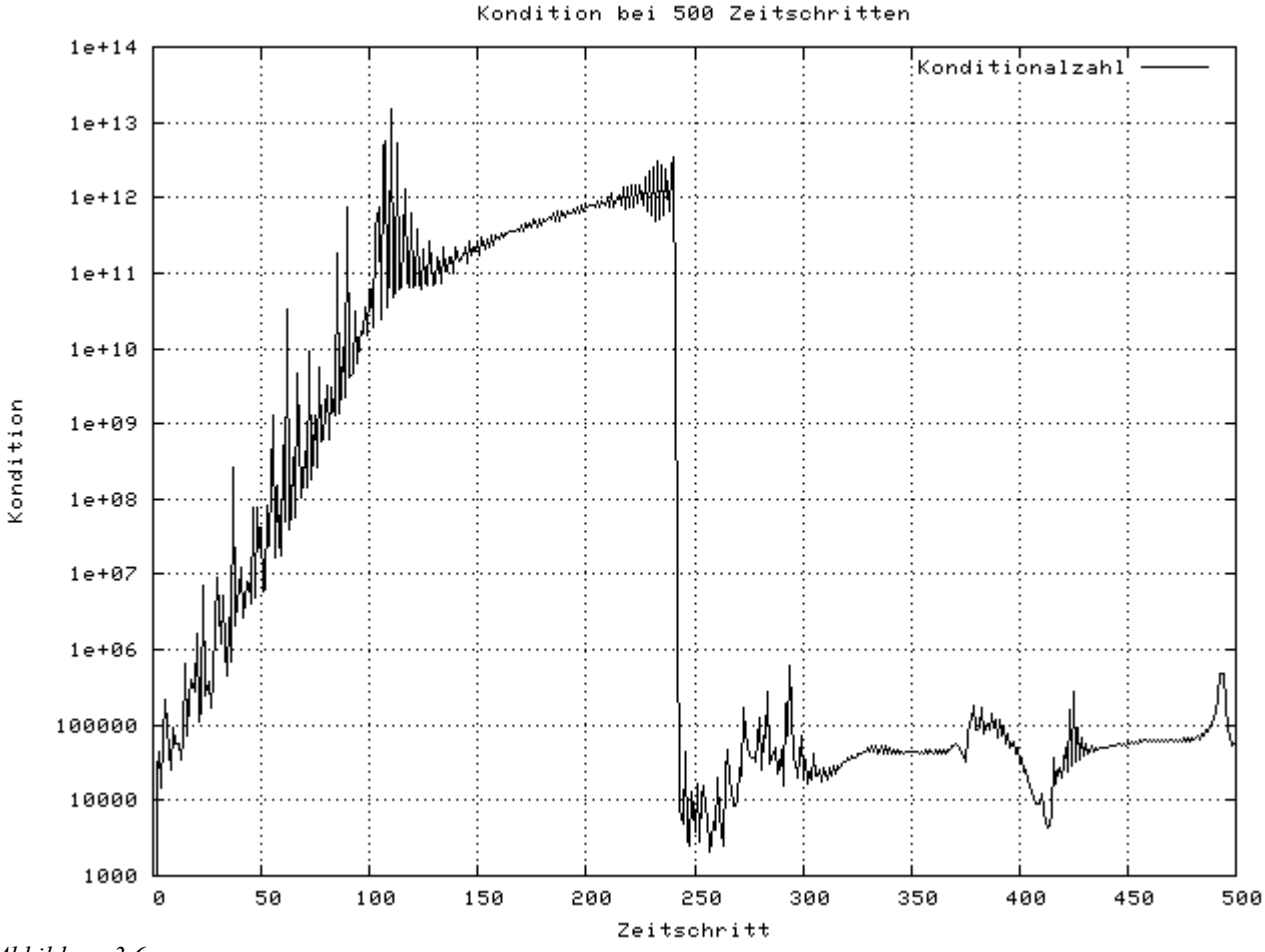

Abbildung 3.6

Das Systemversagen tritt deutlich hervor. Die Konditionszahlen zu diesem Verlauf untermauern die Prognose. Die Konditionszahlen steigen bis zum Systemversagen stetig an, die Fehler addieren sich auf, das System versagt numerisch und gleitet schließlich in einen neuen numerischen Gleichgewichtszustand. Die größte Konditionszahl liegt bei 10^{13}, die Maschinengenauigkeit hingegen bei $6 \cdot 10^{-8}$. Damit ergibt sich ein größtmöglicher Fehler von 600.000. Dass das Ergebnis nicht so falsch sein muss, hat das Beispiel mit 50 Zeitschritten gezeigt. Es geht um Wahrscheinlichkeiten, aber je öfter sich eine Gelegenheit bietet, desto wahrscheinlicher wird ein Systemversagen.

Wenn das Versagen analytischer Natur wäre, also aus dem Element selber käme, dürfte sich der Fehler bei mehr Zeitschritten nicht verstärken. Das ist ja eben die Natur der finiten Elemente, dass sie sich der analytisch exakten Lösung bei kleinerer elementierung im Ort oder in der Zeit nähern. Die Konditionszahlen eigenen sich aber nicht nur für eine Abschätzung der numerischen Qualität, sondern man kann mit ihnen auch indirekt die analytische Lösbarkeit eines Problems messen. Folgendes Kapitel soll einen Einblick geben.

3.1.3 Druckversagen

Die Last wurde in 500 Zeitschritten auf -0,3 erhöht und es ergab sich für die Knoten 4, 8, 12 und 16 folgender Verlauf:

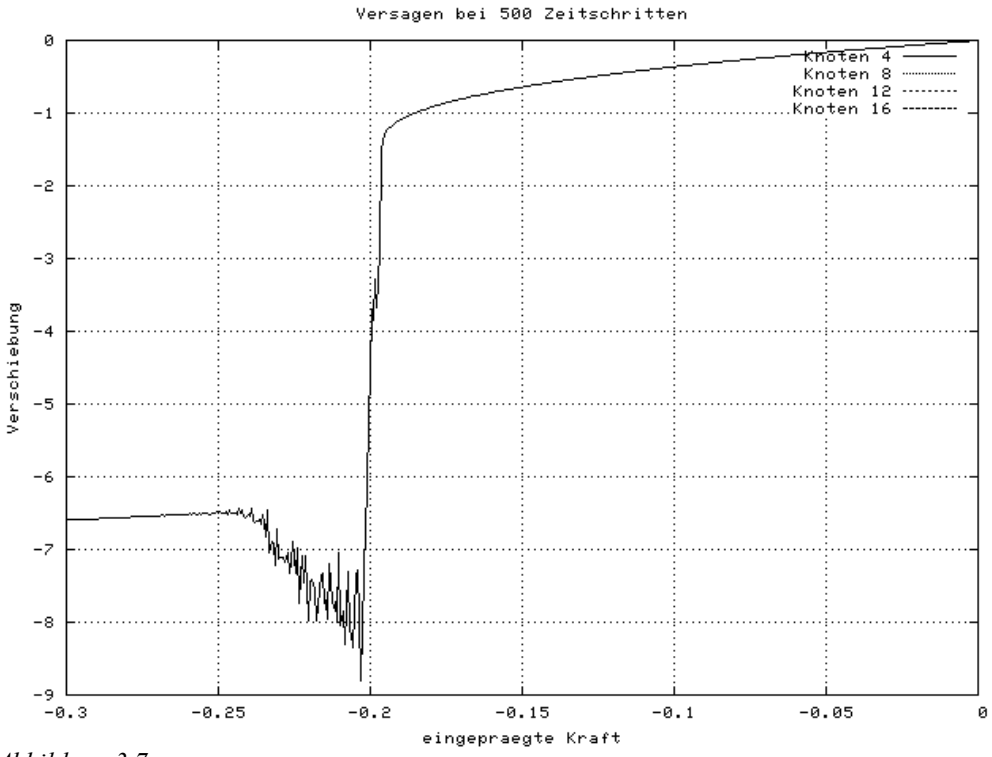

Abbildung 3.7

Man kann deutlich erkennen, dass sich das System bis zu einer Last von etwa 0,19 stabil verhält und den erwarteten Verschiebungen entspricht. Ab diesem Punkt versagt das System schlagartig und pendelt sich bei Verschiebungen ein, die auf der anderen Seite der Einspannung liegen. Das System hat die Injektivität verlassen und sich in einem linksorientierten Koordinatensystem eingependelt (Es hat sich „umgestülpt“).

Dieses Verhalten tritt immer auf. Es spielt dabei keine Rolle, ob man mit nur wenigen Zeitschritten rechnet oder mit sehr vielen. Sogar die Anzahl der Elemente ändert nichts am Ergebnis. Betrachtet man die Konditionszahlen, so ergibt sich folgendes Bild:

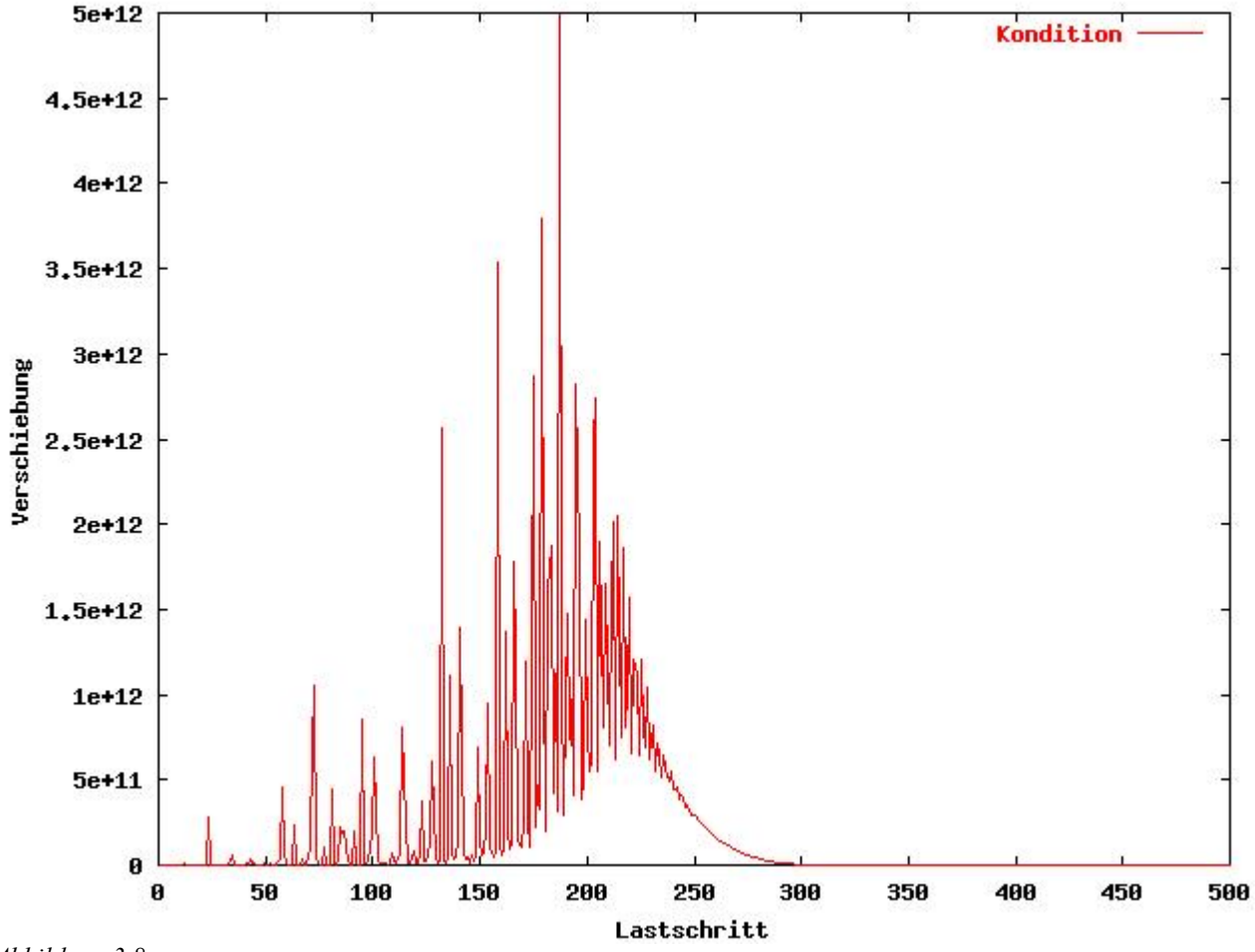

Abbildung 3.8

Man erkennt deutlich das exponentiale Ansteigen der Konditionszahlen bis zum Versagen des Systems. Danach nimmt die Kurve wieder logarithmisch ab, um sich in einem stabilen Zustand einzupendeln.

Ziel dieser Arbeit ist es numerische Stabilität zu untersuchen. Dennoch soll der Vollständigkeit halber der analytische Fehler aufgedeckt werden in dem das Tragverhalten des Kragarms analytisch hergeleitet wird. Allerdings ist diese Herleitung knapp gehalten.

Man gehe von folgender Geometrie aus:

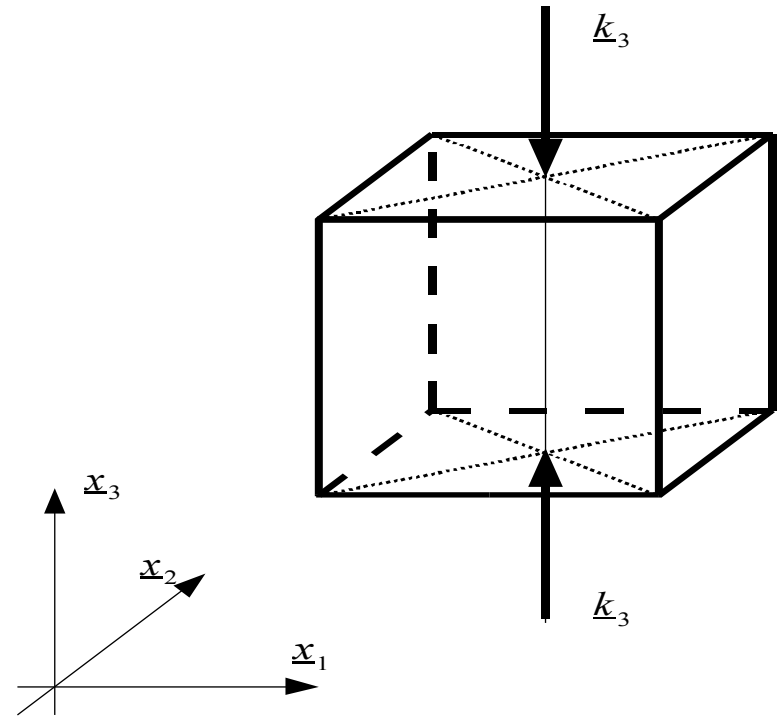

Abbildung 3.9

Sperrt man die Querkontraktion, so erhält man folgenden Deformationsgradient:

$$F_{ij}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}$$

Glg. 3.1.xvi

Daraus folgt der rechte Cauchy – Greensche – Verzerrungstensor:

$$C_{ij}=F_{ki}F_{kj}=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & (\lambda_3)^2\end{bmatrix}$$

Glg. 3.1.xvii

Und somit der Green – Lagrangesche – Verzerrungstensor

$$E_{ij}=\frac{1}{2}\left(C_{ij}-I_{ij}\right)=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \frac{1}{2}\left((\lambda_3)^2-1\right)\end{bmatrix}$$

Glg. 3.1.xviii

Für das Kräftegleichgewicht an der verformten Geometrie ergibt sich:

$$dk_i = \sigma_{ik} \cdot da_k$$
$$\sigma_{ik} = \frac{1}{\boldsymbol{J}} F_{il} S_{lm} F_{km}$$
$$da_k = \boldsymbol{J} (F^{-1})_{lk} dA_l$$

Glg. 3.1.xix

- Dabei ist a_k die Verformte und
- A_k die unverformte Konstellation
- $\boldsymbol{J}$ ist die Determinante der Jakobi Matrix
- S_{lm} ist der zweite Piola – Kirchhof – Tensor
- λ_{ik} ist der Cauchy – Spannungstensor

Setzt man die Bestimmungsgleichungen ein, so ergibt sich für das Gleichgewicht

$$d\,k_i = F_{il} S_{lm} dA_m$$

Für den hier behandelten Fall folgt damit:

$$k_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -k \end{bmatrix} = F_{il} S_{lm} A_m \Rightarrow ;$$
$$\lambda_1 S_{11} A_1 = 0 \Rightarrow S_{11} = 0$$
$$\lambda_2 S_{22} A_2 = 0 \Rightarrow S_{22} = 0$$
$$\lambda_3 S_{33} A_3 = -k$$

Glg. 3.1.xx

Für das linear elastische Materialgesetz ohne Querdehnung gilt:

$$S_{ij} = Dijkl\, E_{kl}$$
$$D_{ijkl} = \mu\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + \lambda\,\delta_{ij}\delta_{kl}$$
$$\mu = \frac{\boldsymbol{E}}{2}$$
$$\nu = 0$$
$$\lambda = 0$$

Glg. 3.1.xxi

An das obige Beispiel angepasst, folgt für den zweiten Piola - Kirchhof – Tensor:

$$S_{33} = D_{3311} E_{11} + D_{3322} E_{22} + D_{3333} E_{33} = \frac{-\boldsymbol{k}}{\lambda_3 A_3} \Rightarrow ;$$

$$2\mu \frac{1}{2}\left((\lambda_3)^2 - 1 \right) = \frac{-\boldsymbol{k}}{\lambda_3 A_3} \Leftrightarrow ;$$

$$(\lambda_3)^3 - (\lambda_3) + \frac{\boldsymbol{k}}{\mu A_3} = 0 \Leftrightarrow ;$$

$$\boldsymbol{k}(\lambda) = \mu A_3 \left(\lambda_3 - (\lambda_3)^3 \right)$$

Glg. 3.1.xxii

Man erhält somit eine Funktion für die eingeprägte Kraft k über λ. Folgende Abbildung verdeutlicht den Funktionsverlauf und dessen Ableitung

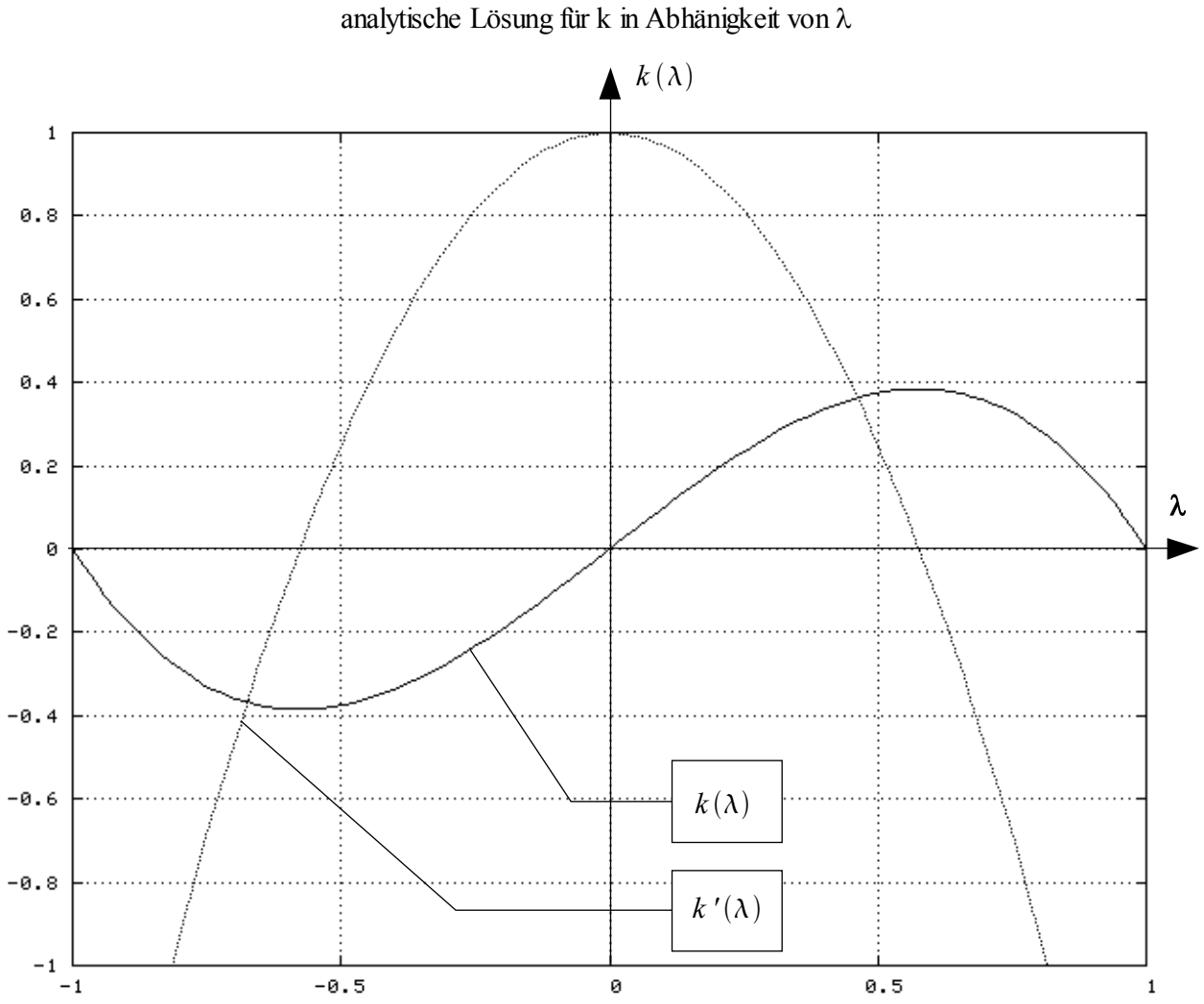

Abbildung 3.10

λ ist in diesem Fall ja gleichzusetzen mit der Determinante des Deformationsgradienten. Somit interessiert der negative Bereich nicht. Weiterhin gilt natürlich auch hier der Grundsatz der Statik,

dass in jedem Gleichgewichtszustand die Kraft injektiv auf die Verschiebung abgebildet wird, was nichts anderes heißt, als dass jedem Kraftzustand eindeutig ein Verschiebungszustand zugeordnet wird. Man kann sich aber davon überzeugen, dass hier den Verschiebungen symmetrisch um den Hochpunkt gleiche Kräfte zugeordnet werden. Zudem gilt in diesem Falle auch noch, dass größere Kräfte größere Verschiebungen hervorrufen. Somit ist der gesamte Bereich links vom Hochpunkt obsolet.

Die kritische Kraft errechnet sich für den hier behandelten Fall somit zu:

$$
\begin{aligned}
&k'(\lambda)=\mu A_3\left(1-3\left(\lambda_3\right)^2\right)\Rightarrow;\\
&k'(\lambda)=0\Rightarrow;\\
&\left(\lambda_3\right)_{01}=+\sqrt{\frac{1}{3}}\sim 0{,}5774\Rightarrow;\\
&u_{krit}=3*0{,}5774=1{,}732\\
&k_{krit}=\frac{1}{2}*1*(0{,}5773-0{,}5773^3)=0{,}1924
\end{aligned}
$$

Glg. 3.1.xxiii

Man kann sich in Abbildung 3.7 davon überzeugen, dass das System wirklich bei einer Verschiebung von 1,732 und einer Last von 0,1924 versagt.

3.2 Abschließende Bemerkung und Ausblick

Die hier vorliegende Arbeit hat deutlich gezeigt, dass sich die Gruppe der direkten Löser für Randwertprobleme nicht eignet. Die Integrationsverfahren, wie zum Beispiel die Zeitschrittintegration, erinnern sich nicht an die Fehler des vorangegangenen Zeitschritts. Dadurch kommt es zum Aufschaukeln der Fehler. Ob die gemischten Lösungsverfahren, bei denen ein Startvektor mittels direktem Löser gefunden wird und dessen Fehler mittels iterativen Verfahren ausgetrieben wird, das Problem beseitigen, sollte untersucht werden. Es fällt jedenfalls auf, dass kommerzielle Programmsysteme die Problematik des numerischen Versagens zumindest kennen. Ob sie das Problem allerdings beheben, oder nur verbergen, bleibt dahingestellt.

Kapitel 4

Verzeichnisse

4.1 Gleichungen

Glg. 2.1.i.....6
Glg. 2.1.ii.....7
Glg. 2.1.iii.....7
Glg. 2.1.iv.....8
Glg. 2.1.v.....8
Glg. 2.1.vi.....9
Glg. 2.1.vii.....9
Glg. 2.1.viii.....10
Glg. 2.1.ix.....10
Glg. 2.1.x.....11
Glg. 2.1.xi.....11
Glg. 2.1.xii.....11
Glg. 2.1.xiii.....12
Glg. 2.1.xiv.....12
Glg. 2.1.xv.....12
Glg. 2.1.xvi.....13
Glg. 2.1.xvii.....13
Glg. 2.1.xviii.....14
Glg. 2.1.xix.....15
Glg. 2.1.xx.....16
Glg. 2.1.xxi.....17
Glg. 2.1.xxii.....17
Glg. 2.1.xxiii.....18
Glg. 2.1.xxiv.....18
Glg. 2.1.xxv.....19
Glg. 2.1.xxvi.....20
Glg. 2.1.xxvii.....20
Glg. 2.1.xxviii.....20
Glg. 2.1.xxix.....21
Glg. 2.1.xxx.....21
Glg. 2.1.xxxi.....21
Glg. 2.2.i.....23
Glg. 2.2.ii.....24
Glg. 2.2.iii.....24
Glg. 2.2.iv.....24
Glg. 2.2.v.....25
Glg. 2.2.vi.....25
Glg. 2.2.vii.....25
Glg. 2.2.viii.....25
Glg. 2.2.ix.....26
Glg. 2.2.x.....26
Glg. 2.2.xi.....27
Glg. 2.2.xii.....28
Glg. 2.2.xiii.....29
Glg. 2.2.xiv.....29
Glg. 2.2.xv.....30
Glg. 2.2.xvi.....30
Glg. 2.2.xvii.....31
Glg. 2.2.xviii.....31
Glg. 2.2.xix.....32
Glg. 2.2.xx.....33
Glg. 2.2.xxi.....33
Glg. 2.2.xxii.....33
Glg. 2.2.xxiii.....33
Glg. 2.2.xxiv.....34
Glg. 2.2.xxv.....34
Glg. 2.2.xxvi.....34
Glg. 2.2.xxvii.....34
Glg. 2.2.xxviii.....35
Glg. 2.2.xxix.....36
Glg. 2.2.xxx.....37
Glg. 2.2.xxxi.....37
Glg. 2.2.xxxii.....37
Glg. 3.1.i.....39
Glg. 3.1.ii.....39
Glg. 3.1.iii.....40
Glg. 3.1.iv.....40
Glg. 3.1.v.....41
Glg. 3.1.vi.....42
Glg. 3.1.vii.....42
Glg. 3.1.viii.....43
Glg. 3.1.ix.....43
Glg. 3.1.x.....43
Glg. 3.1.xi.....44
Glg. 3.1.xii.....44
Glg. 3.1.xiii.....44
Glg. 3.1.xiv.....44
Glg. 3.1.xv.....45
Glg. 3.1.xvi.....51
Glg. 3.1.xvii.....51
Glg. 3.1.xviii.....51
Glg. 3.1.xix.....52
Glg. 3.1.xx.....52
Glg. 3.1.xxi.....52
Glg. 3.1.xxii.....53
Glg. 3.1.xxiii.....54

4.2 Abbildungen

Abbildung 2.1.....7
Abbildung 2.2.....8
Abbildung 2.3.....10
Abbildung 2.4.....13
Abbildung 2.5.....14
Abbildung 2.6.....15
Abbildung 2.7.....16
Abbildung 2.8.....17
Abbildung 2.9.....19
Abbildung 2.10.....20
Abbildung 2.11.....31
Abbildung 2.12.....32
Abbildung 2.13.....33
Abbildung 2.14.....36
Abbildung 3.1.....38
Abbildung 3.2.....41
Abbildung 3.3.....46
Abbildung 3.4.....47
Abbildung 3.5.....48
Abbildung 3.6.....48
Abbildung 3.7.....49
Abbildung 3.8.....50
Abbildung 3.9.....51
Abbildung 3.10.....53

4.3 Literaturverzeichnis

[1] Bronstein, Semendjajew, Musiol, Mühlig
Taschenbuch der Mathematik
Verlag Harri Deutsch
erschienen 1995
ISBN 3-8171-2002-8

[2] Heinz Schade
Tensoranalysis
Walter de Gruyter, Berlin New York
erschienen 1997
ISBN 3-11-014740-8

[3] Huckle Schneider
Numerik fürInformatiker
Springer Verlag
erschienen 2002
ISBN 3-540-42387-7

[4] V. Mehrmann, M. Bollhöfer
Numerische Mathematik für Ingenieure
http://www.moses.tu-berlin.de/mathematik/numerik1/ss_2003/ing_allg/nfing1.pdf

[5] Jörn Stypa
Existenz einer Lösung für die nichtlineare Elastizitätstheorie mit Anwendung in der Herzmechanik
http//www.math.uni-muenster.de/num/preprints/2000/stypa/paper.pdf

[6] Prof. Dr.-Ing. Rudolf Harbord
Lehrveranstaltung Statik der Baukonstruktionen IV -Vertiefung I -
ftp://statik.tu-berlin.de/pub/Lehre/Statik_VT_I/Skripte/Teil_5.pdf

Eidesstattliche Erklärung

Die selbständige und eigenhändige Anfertigung versichere ich an Eides statt.

Berlin, den