

# TENSORSTRUKTUR DER ZELLMATRIZEN BEI FINITEN ELEMENTEN

Bachelorarbeit

eingereicht von

Enes Witwit

betreut von

Prof. Dr. Kanschat

**Fakultät für Mathematik und Informatik**

**Universität Heidelberg**

---

Hiermit versichere ich, Enes Witwit, dass ich die vorliegende Bachelorarbeit selbstständig verfasst habe. Ich versichere, dass ich keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe, und dass die eingereichte Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist.

16. Mai 2017

Heidelberg

Unterschrift

### **Zusammenfassung**

Die folgende Arbeit konzentriert sich auf die Berechnung der Pseudoinversen von der Masse Matrix und der Steifigkeitsmatrix der Laplace Bilinearform, insbesondere die effiziente Berechnung des Matrix-Vektor Produkts mit der Pseudoinversen der genannten Matrizen. Insgesamt werden zwei verschiedene Ansätze hergeleitet. Der erste versucht das Ziel durch eine Singulärwertzerlegung höherer Ordnung zu erreichen. Der zweite Ansatz nutzt die Struktur der genannten Matrizen um eine einfache Berechnung der Pseudoinversen über eine einfache Singulärwertzerlegung. Mit der zweiten Methode erreichen wir insgesamt für die Berechnung der Pseudoinversen und dem Matrix-Vektor-Produkt eine Komplexität von  $O(n^3)$ , was erstaunlich ist. Hingegen stoßen wir beim ersten Ansatz auf diverse Probleme bezüglich der einfachen Berechnung des Produktes und der Komplexität.

# Inhaltsverzeichnis

|  |           |
|--|-----------|
| Notation   | I         |
| Abbildungsverzeichnis  | II        |
| Tabellenverzeichnis  | III       |
| <b>1 Einführung</b>  | <b>1</b>  |
| <b>2 Theorie</b>   | <b>3</b>  |
| 2.1 Numerische Behandlung partieller Differentialgleichungen . . . . . | 3         |
| 2.2 Tensor Dekomposition . . . . .                                     | 16        |
| <b>3 Tensorstruktur und Pseudoinverse der Zellmatrizen</b>             | <b>21</b> |
| 3.1 Summenfaktorisierung . . . . .                                     | 21        |
| 3.2 Singulärwertzerlegung höherer Ordnung . . . . .                    | 27        |
| <b>4 Effiziente Implementierung</b>                                    | <b>30</b> |
| <b>5 Resultate</b>   | <b>36</b> |
| Literatur  | 37        |

## Notation

|                                     |  |
|-------------------------------------|--|
| $a(\cdot, \cdot)$                   | Bilinearform   |
| $D^\alpha$                          | Ableitung $ \alpha $ -ter Ordnung zum Multiindex $\alpha$                      |
| $I$                                 | Identität  |
| $J(\cdot)$                          | Funktional   |
| $O(\cdot), o(\cdot)$                | Landau Symbole   |
| $P_l$                               | Menge aller Polynome vom Höchstgrad $l$  |
| $\mathbb{R}, \mathbb{N}$            | reelle Zahlen, natürliche Zahlen   |
| $V$                                 | Banachraum   |
| $\dim(V)$                           | Dimension von $V$  |
| $V_h$                               | endlichdimensionaler Finite-Elemente-Raum                                      |
| $\ \cdot\ _V$                       | Norm von $V$   |
| $(\cdot, \cdot)$                    | Skalarprodukt in $V$ , wenn $V$ Hilbertraum                                    |
| $f(v)$ oder $\langle f, v \rangle$  | Wert des Funktionals $f \in V^*$ bei Anwendung auf $v \in V$                   |
| $\Omega$                            | Grundgebiet bezüglich der räumlichen Veränderlichen                            |
| $\delta\Omega$                      | Rand von $\Omega$  |
| $\text{int}\Omega$                  | Inneres von $\Omega$   |
| $C^l(\Omega), C^{l,\alpha}(\Omega)$ | Räume differenzierbarer bzw. Hölder-stetiger Funktionen                        |
| $L_p(\Omega)$                       | Räume zur $p$ -ten Potenz integrierbarer Funktionen ( $1 \leq p \leq \infty$ ) |
| $W_p^l(\Omega)$                     | Sobolev-Raum   |
| $H^l(\Omega), H_0^l(\Omega)$        | Sobolev-Räume für $p=2$  |
| $\text{supp}(v)$                    | Träger der Funktion $v$  |
| $\Delta$                            | Gradient   |

|                    |  |
|--------------------|--|
| $div$              | Divergenz  |
| $\nabla$           | Laplace-Operator   |
| $\nabla_h$         | Diskreter Laplace-Operator                                     |
| $h_i$ , $h$        | Diskretierungsparamter bezüglich der räumlichen Veränderlichen |
| $\mathfrak{X}$     | Skript Buchstaben für Höher-dimensionale Tensoren              |
| $\mathbf{A}^{(n)}$ | bezeichnet die n-te Matrix einer Matrixfolge                   |

# Abbildungsverzeichnis

|   |  |    |
|---|--|----|
| 1 | Ansatzfunktionen $\phi_i$ [CG05, 184]  | 10 |
| 2 | Tensor dritter Ordnung $\mathfrak{X} \in \mathbb{R}^{I \times J \times K}$ [TK09, 456] | 16 |
| 3 | HOSVD eines Tensors dritter Ordnung [TK09, 475]  | 19 |
| 4 | [Tea, 3]   | 21 |

## Tabellenverzeichnis



# 1 Einführung

Hochleistungsrechnen ist eine neue Disziplin, die mit dem Drang entstand, immer komplexere Probleme zu lösen. Diese neue Art an Probleme anzugehen, löst sie nicht. Doch sie eröffnet uns eine neue Sichtweise auf dasselbe Probleme und gibt uns am Ende womöglich neue Lösungsansätze. Parallelisierung ist ein großer Zweig des Hochleistungsrechnens. Das Grundgerüst dieser Methodik besteht darin komplexe Probleme modular zu lösen, das heißt sie in viele wenig komplexere Subprobleme zu unterteilen, die unabhängig voneinander berechenbar sind. Am Ende fasst man die Ergebnisse der Subprobleme zu einem Ergebnis zusammen, welches die Lösung des komplexen Problems darstellt.

$$v = A(u) = \sum_{k=1}^{n_{cells}} C^T P_k^T A_k (P_k C u) \quad (1.1)$$

ist die Gleichung, welche wir mit Hilfe der finite Elemente Methode lösen wollen.  $A$  ist ein möglicherweise nichtlinearer Operator, der Vektor  $u$  als Input nimmt und das Integral vom Operator multipliziert mit Testfunktionen  $\phi_i$  mit  $i = 1, \dots, n$ , berechnet [MK12, 136]. Damit wir diese Gleichung besser nachvollziehen können, werden wir sie im nächsten Kapitel herleiten. Doch um diese Arbeit zu motivieren kann man sich diese Gleichung so vorstellen, dass man ein großes Problem in eine Summe von kleineren Problemen unterteilt.

Nun die Matrix  $A_k$  kann durchaus, wie wir nachher erkennen werden, sehr groß werden. Wenn sie so groß wird, dass sie nicht mehr im Cache liegt, bekommen wir das Problem, dass der Zugriff auf Elemente der Matrix sehr teuer wird. Dementsprechend wollen wir uns vor allem bei der Herleitung unserer Methoden um diese Subprobleme zu lösen auf sogenannten matrix-freie Heransgehensweisen konzentrieren. Das bedeutet wir wollen nicht explizit die Matrix  $A_k$  ausrechnen, sondern stattdessen ihre Elemente dort berechnen wo sie auch gebraucht werden. Ob dies möglich ist, ist für diese Arbeit von großer Bedeutung.

Der Sinn dieser Arbeit ist einen effiziente Ansatz herzuleiten, der uns

$$A_k^+ v_k \quad (1.2)$$

berechnet, wobei  $A_k^+$  eine Pseudoinverse darstellt. Dies kann als Präkonditionierer genutzt werden, um (1.1) Gleichung zu lösen. Das heißt die Resultate dieser Arbeit werden benutzt, um damit einen Präkonditionierer zu bauen. Daraus folgern wir,

dass die Exaktheit der Lösung von (2.15) eine redundante Rolle spielt, vielmehr sollten wir eine optimale Lösung im Trade-Off zwischen Genauigkeit und Komplexität anstreben.

Im zweiten Kapitel werden wir erstmal ein theoretisches Grundgerüst schaffen, um die beiden Gleichungen besser zu durchleuchten und nachvollziehen zu können. Wir werden uns die Grundlagen der numerischen Behandlung von partiellen Differentialgleichungen anschauen und versuchen die oberen Gleichungen herzuleiten. Im zweiten Teil des zweiten Kapitels werden wir uns mit Tensor Dekomposition beschäftigen. Dies liegt daran, dass wir uns den Operator  $A$  als Tensor undefinieren können und die Pseudoinverse mit Hilfe der sogenannten Singulärwertzerlegung höherer Ordnung berechnen können.

Im dritten Kapitel werden wir uns zwei Methoden anschauen die Pseudoinverse zu berechnen. In der ersten Methode die Struktur des Operators  $A$  zu nutzen und leiten eine Repräsentation von  $A$  her die uns die Pseudoinverse mit wenig Aufwand gibt. In der zweiten Methode werden wir uns wie bereits erwähnt,  $A$  als Tensor undefinieren und versuchen die Singulärwertzerlegung höherer Ordnung effizient zu berechnen und uns dort auch Strukturen vom Tensor zu nutzen zu machen.

Im vierten Kapitel sprechen wir über die effiziente Implementierung beider Algorithmen und im fünften Kapitel fassen wir alle Resultate zusammen und schließen mit einem Fazit ab.

## 2 Theorie

Das Ziel dieses Kapitels ist es die Verständigung über Notation zu klären und eine theoretische Grundlage für das dritte Kapitel, das Herz dieser Arbeit, zu schaffen.

Zu erst behandeln wir einen funktionalanalytischen Ansatz für elliptische Probleme. Begriffe wie schwache Lösung, klassische Lösung, Variationsgleichung und Sobolev Raum werden geklärt. Dies liefert uns das Basiswissen für die Galerkin Methode. Im dritten Unterkapitel *Methode der Finiten Elemente* ist das Ziel die Gleichung  $Au = f$  herzuleiten mit  $A$  als globale Steifigkeitsmatrix. Im darauffolgenden Kapitel wollen wir uns eine flexiblere Methode anschauen, als die Methode der Finiten Elemente und die Wichtigkeit zu dieser Arbeit soll besonders hervorgehoben werden.

Dann kommen wir zu einem ganz anderen Thema, nämlich dem der Tensor Dekomposition. Erstmal schauen wir uns grundlegende Definitionen an um mit diesem Basiswissen die Singulärwertzerlegung höherer Ordnung einzuführen. Mit dieser werden wir uns gewisse Finite Elemente Operatoren als Tensoren umdefinieren und uns deren Zerlegung anschauen. Wir werden diese Untersuchungen in Kapitel drei und vier durchführen. In Hoffnung natürlich, dass die Zerlegung einfach ist und daraus die Pseudoinverse herleitbar ist.

### 2.1 Numerische Behandlung partieller Differentialgleichungen

#### 2.1.1 Schwache Lösungen

Gegeben sei folgendes Randwertproblem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ in } \delta\Omega \end{aligned} \tag{2.1}$$

mit  $\Omega = [0, 1]^2$  und  $f : \bar{\Omega} \rightarrow \mathbb{R}$  zwei mal stetig partiell differenzierbar. Der funktionalanalytische Ansatz schlägt vor, dass wir uns zu erst mit notwendigen Funktionenräumen beschäftigen und uns auf analytischer Ebene eine Umformulierung der Differentialgleichung zu nutze machen, welche uns letztlich die Grundlage für die finiten Elementen Methode liefert.

Wir sehen von der Form der Differentialgleichung, dass die gesuchte Lösung bestimmte Differenzierbarkeits- und Stetigkeitsbedingungen zu erfüllen hat. Nämlich, dass  $u$  zweimal partiell stetig differenzierbar sein sollte. Die Lösung sollte also in

dem Raum der zweimal partiell stetig differenzierbaren Funktionen liegen.

Nun kann es aber sein, dass eine Lösung für dieses Problem nicht in diesem Raum existiert. Wir können uns dafür als Beispiel die Betragsfunktion  $b : \mathbb{R} \rightarrow \mathbb{R}$  anschauen.

$$b(x) = |x| \quad (2.2)$$

Offensichtlich ist diese Funktion in Null nicht stetig differenzierbar. Trotzdem können wir eine Ableitung finden, die wir *schwache Ableitung* nennen.

$$b'(x) = \begin{cases} -1 & \text{für } x < 0 \\ 0 & \text{für } x = 0 \\ 1 & \text{für } x > 0 \end{cases} \quad (2.3)$$

Dies zeigt uns, dass wir potentiell auch Lösungen finden können, in Räumen die nicht alle notwendigen Regularitätsbedingungen erfüllen. Diese Lösungen nennen wir dann *schwache Lösungen*. Doch wie finden wir schwache Lösungen? Dafür nutzen wir eine funktionalanalytische Idee, die aus der Distributionstheorie stammt. Wir multiplizieren mit einer Testfunktion  $v \in C(\bar{\Omega})$  und integrieren über das Gebiet.

$$\int_{\Omega} -\Delta u v \, dx = \int_{\Omega} f v \, dx \quad (2.4)$$

Der nächste Schritt, welcher durchwegs fundamental für die Herleitung ist, ist die intuitive Nutzung der Struktur und Integrationswerkzeuge um zu erreichen, dass an  $u$  weniger Differenzierbarkeitsanforderungen gebunden sind. Für diesen Schritt ist der Satz von Green und die partielle Integration von Wichtigkeit.

**Lemma 2.1.** (*Partielle Integration*)

Es sei  $\Omega \subset \mathbb{R}^n$ ,  $n = 2, 3$  ein beschränktes Lipschitz-Gebiet. Dann gilt

$$\int_{\Omega} \frac{\delta u}{\delta x_i} v \, dx = \int_{\partial\Omega} u v \cos(n, e^i) \, ds - \int_{\Omega} u \frac{\delta v}{\delta x_i} \, dx$$

für beliebige  $u, v \in C^1(\bar{\Omega})$  und  $n$  bezeichne die äußere Normale im jeweiligen Randpunkt [CG05, 139].

Für unsere Differentialgleichung (2.4) nutzen wir die partielle Integration und erhalten.

$$\int_{\Omega} -\nabla u \nabla v \, dx = \int_{\Omega} f v \, dx \quad (2.5)$$

Dies ist die so genannte *Variationsgleichung*. Sie ist ein erster Indiz für die später gewünschte Bilinearform der zugrundeliegenden Topologie. Die Lösung  $u$  von (2.5) nennt man *schwache Lösung* für das Problem (2.1). Die Lösung  $u \in C_0^2$  von (2.1) nennt man *klassische Lösung*. Nun wissen wir in welchem Raum die klassische Lösung liegt, aber welche Topologie ist für die schwache Lösung sinnvoll? Ein funktionalanalytischer Ansatz versucht nun die Räume zu definieren, in der die Lösung  $u$  für (2.5) liegt. In unserem Fall wäre folgender Raum ergiebig

$$H_0^1(\Omega) = \{ v \in L_2(\Omega) : \frac{\delta v}{\delta x_i} \in L_2(\Omega), v = 0 \text{ in } \delta\Omega, i = 1, 2 \}$$

Diesen Raum nennt man Sobolev Raum. Allgemein sind Sobolev Räume definiert durch

**Definition 2.2.** (*Sobolev Raum*)

Es sei  $\alpha = (\alpha_1, \dots, \alpha_n)$  mit  $\alpha_i \geq 0$  ganzzahlig. Weiterhin sei  $|\alpha| = \sum_i \alpha_i$  und

$$D^\alpha u := \frac{\delta^{|\alpha|}}{\delta x_1^{\alpha_1} \dots \delta x_n^{\alpha_n}} u$$

Für  $k = 1, 2, \dots$  definieren wir nun den Sobolev Raum

$$W^{k,p}(\Omega) = \{ v \in L_p(\Omega) : D^\alpha v \in L_p(\Omega), |\alpha| \leq k \}$$

Das heißt Sobolev Räume sind eine Teilmenge von den  $L_p$  Räumen. Von der analytischen Perspektive ist die Wahl des Funktionenraumes essentiell für den Nachweis der Existenz der Lösung. Von der Perspektive der finiten Elementen Methode ist dies für die Fehlerabschätzung wichtig, da wir dann die induzierte Norm des Funktionenraumes benutzen [Joh08, 36]. Beide genannten Themen würden den Rahmen dieser Bachelorarbeit sprengen, daher verweise ich an gegebenen Stellen an weiterführende Literatur.

Die Sobolev Räume wurden mit Skalarprodukten ausgestattet, sodass unsere Variationsgleichung intuitiv als Skalarprodukt der zugrundeliegenden Sobolev Räume geschrieben werden kann.

**Definition 2.3.** (*Sobolev Norm*)

$$\|v\|_{W^{k,p}(\Omega)} = \begin{cases} \left( \sum_{|\alpha| \leq k} \|D^\alpha v\|_{L^p(\Omega)}^{1/p} dx \right)^{1/2} \text{ für } p < \infty \\ \max_{|\alpha| \leq k} \|D^\alpha v\|_{L^\infty(\Omega)} dx \text{ für } p = \infty \end{cases}$$

$W^{k,p}(\Omega)$  ist mit der Norm  $\|\cdot\|_{W^{k,p}(\Omega)}$  ein vollständiger Vektorraum, somit also ein Banachraum. Für  $p = 2$  wird die Norm durch das Skalarprodukt induziert.

**Definition 2.4.** (*Sobolev Skalarprodukt*)

$$(u, v)_{W^{k,2}(\Omega)} := \sum_{|\alpha| \leq k} (\partial^\alpha u, \partial^\alpha v)_{L^2(\Omega)}$$

$W^{k,2}(\Omega)$  ist daher ein Hilbertraum, und wir schreiben  $H^k(\Omega) := W^{k,2}(\Omega)$ . Für die Untersuchung elliptischer Randwertaufgaben zweiter bzw. vierter Ordnung sind vor allem Sobolev-Räume  $H^1(\Omega)$  bzw.  $H^2(\Omega)$  und deren Unterräume von Bedeutung [CG05, 134].

Wir wollen noch ein wichtiges Resultat erwähnen, das Relevanz für die Herleitung der Variationsgleichung hat. Durch wiederholte Anwendung und Beachtung von Grenzübergängen von Lemma (2.1) erhält man folgendes Resultat

**Satz 2.5.** (*Green*)

Es sei  $\Omega \subset \mathbb{R}^n$ ,  $n = 2, 3$  beschränktes Lipschitz-Gebiet. Dann gilt

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} \frac{\delta u}{\delta n} v \, ds - \int_{\Omega} \Delta u \Delta v \, dx$$

für beliebige  $u, v \in C^1(\bar{\Omega})$  und  $n$  bezeichne die äußere Normale im jeweiligen Randpunkt [CG05, 140].

Der Satz von Green liefert uns ein mächtiges Werkzeug zur Herleitung der Variationsgleichung.

Nun kommen wir wieder zurück zu unserem ursprünglichen Problem.

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ in } \partial\Omega \end{aligned}$$

Wir wissen jede Lösung dieses Problems erfüllt die Variationsgleichung

$$\int_{\Omega} -\nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \text{ für alle } v \in H_0^1(\Omega) \quad (2.6)$$

Wir definieren uns  $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$  mit

$$a(u, v) := \int_{\Omega} \nabla u \nabla v \, dx$$

Wir können zeigen, dass die Lösung des Problems (2.6) unter gewissen Bedingungen äquivalent ist, zum Lösen des folgenden Minimierungsproblems. Das Minimierungsproblem nennen wir auch Variationsaufgabe oder Variationsproblem.

**Satz 2.6.** *(Verbindung zwischen Variationsaufgabe und Variationsgleichung)*

*Es sei  $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$  eine Bilinearform. Sei ferner  $f \in V^*$ . Falls die Bilinearform  $a(\cdot, \cdot)$  symmetrisch ist, ist  $u \in V$  ein Minimierer der Gleichung*

$$J(u) = \min_{v \in V} J(v) = \frac{1}{2}a(u, u) - f(u) \quad (2.7)$$

*genau dann wenn  $u$  die schwache Formulierung löst.*

### 2.1.2 Galerkin Verfahren

Unser Ausgangspunkt ist nun die Lösung des Problems

$$\text{Finde } u \in V : a(u, v) = f(v) \quad \forall v \in V \quad (2.8)$$

wobei  $V$  ein Banachraum,  $a(\cdot, \cdot)$  eine Bilinearform und  $f$  ein Funktional auf  $V$ .  $V$  ist in unserem Fall ein Sobolev Raum. Da Sobolev Räume unendlich dimensional sind, ist es schwer sich mit der Konstruktion einer Lösung vertraut zu machen. Daher ist die Kernidee des Galerkin Verfahrens unseren Banachraum zu diskretisieren. Dazu führen wir eine sogenannte konforme Approximation durch. Wir wählen  $V_n \subset V$  mit  $\dim(V_n) = n < \infty$ . Wir erhalten ein neues diskretes Problem und haben mit (2.7) nun zwei Probleme.

$$u = \arg \min_{v \in V} J(v) \quad (\text{stetig})$$

$$u_n = \arg \min_{v_n \in V_n} J(v_n) \quad (\text{diskret})$$

Daraus folgt

$$J(u_n) \geq J(u)$$

Dies folgt direkt aus der Wahl des Raumes als Teilraum des ursprünglichen Raumes. Man nennt diese Methode *konform*, da der diskrete Raum ein Teilraum von dem

ursprünglichen Raum ist und das Funktional  $J$  gleich bleibt. Was hat uns das Ganze gebracht? Da  $\dim(V_n) = n < \infty$  können wir eine Basis für  $V_n$ , welche wir später Ansatzfunktionen nennen, bestimmen. Das bringt uns den Vorteil, dass wir  $u_n$  als lineare Kombination der Basiselemente von  $V_n$  approximieren können mit endlich vielen Parametern.

Die diskrete schwache Formulierung sieht jetzt wie folgt aus

$$\text{Finde } u_n \in V_n : a(u_n, v_n) = f(v_n) \quad \forall v_n \in V_n \quad (2.9)$$

Sei nun  $e_1, \dots, e_n$  eine Basis von  $V_n$ . Es ist ausreichend nur die Basis zum Testen zu nutzen. Das obere Problem (2.9) reduziert sich auf

$$\text{Finde } u_n \in V_n : a(u_n, e_i) = f(e_i) \quad \forall i \in \{1, \dots, n\} \quad (2.10)$$

Da  $u_n \in V_n$  können wir  $u_n$  als Linearkombination der Basiselemente von  $V_n$  schreiben.

$$u_n = \sum_{j=1}^n u_j e_j \quad (2.11)$$

Setzen wir dies in Gleichung (2.10) erhalten wir eine neue Darstellung des diskreten Problems.

$$\sum_{j=1}^n u_j a(e_j, e_i) = f(e_i) \quad (2.12)$$

Das können wir in einem linearen Gleichungssystem mit  $A_{ij} = a(e_j, e_i)$  und  $u = (u_1, \dots, u_n)^T$  zusammenfassen. Insgesamt erhalten wir:

$$Au = f \quad (2.13)$$

Das heißt, es gilt einen Vektor  $u$  zu finden, der diese Gleichheit erfüllt um das diskrete Variationsproblem (2.9) zu lösen und damit eine Approximation für unser stetiges Problem (2.8) zu erhalten. Man kann allgemein davon ausgehen, dass für größeres  $n$  die Approximation besser wird. Das Verhalten des Fehlers im Bezug zum Diskretisierungsparameters  $n$  wird in [CG05, 154] ausgiebig untersucht.



**Bemerkung 2.7.** (*Galerkin Eigenschaften*)

1. *Galerkin Orthogonalität*

*Der Fehler liegt orthogonal auf dem Teilraum von  $V$ .*

2. *Symmetrie*

*Die Matrix  $A$  ist genau dann symmetrisch, wenn die Bilinearform symmetrisch ist.*

### 2.1.3 Methode der Finiten Elemente

Im vorherigen Kapitel haben wir das Ritz-Galerkin Verfahren kennengelernt. Der Kernaspekt dieser konformen Approximation war eine diskretisierung des Raumes und damit einhergehend global einheitlich definierte Basisfunktionen des diskreten Raumes. Nun öffnen wir die zuletzt genannte Einschränkung und fordern nur noch stückweise definierte Funktionen. Wo genau eine Funktion, in der Regel ein Polynom, definiert ist, hängt von unserer Gebietszerlegung ab. Das heißt für die *Finite Elemente Methode* (FEM) ist es zu erst notwendig

1. Das Grundgebiet in geometrisch einfache Teilgebiete  $\Omega_h = \{\Omega_k\}_{k=1, \dots, N}$  z.B. Dreiecke und Rechtecke bei Problemen in der Ebene oder Tetraeder und Quader bei Problemen im dreidimensionalen Raum.
2. Definition von Ansatz- und Testfunktionen über Teilgebieten
3. Da wir zwischen den Teilgebieten eine Stetigkeit fordern, definiert man Übergangsbedingungen, die uns globale Stetigkeit sichern

Die Stetigkeit der globale Lösung wird gefordert, damit wir  $V_n \subset V$  bekommen mit  $V$  ein Sobolev Raum. [CG05, 175].

**Bemerkung 2.8.** (*Voraussetzungen an Zerlegung*) [CG05, 176]

*Die Voraussetzungen an die Zerlegung  $Z = \{\Omega_j\}_{j=1}^m$  sind*

1.  $\bar{\Omega} = \bigcup_{j=1}^m \bar{\Omega}_j$

2.  $\text{int}\Omega_i \cap \text{int}\Omega_j = \emptyset$  , falls  $i \neq j$

**Beispiel 2.9.** *E sei  $\Omega = [a, b]$ . Wir definieren Gitterpunkte  $\{x_i\}_{i=0}^N$  über  $\bar{\Omega}$  beschrieben wie folgt*

$$a = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = b$$

*und eine Zerlegung  $Z = \{\Omega_j\}_{j=1}^m$  mit  $\Omega_i := (x_{i-1}, x_i)$  für  $i = 1, \dots, N$ . Ferner sei  $h_i := x_i - x_{i-1}$ ,  $i = 1, \dots, N$ . Wir wählen lineare Ansatzfunktionen  $V_h = \text{lin}\{\phi_i\}_{i=0}^N$ , wobei die Ansatzfunktionen sind durch*

$$\phi_i(x) = \begin{cases} \frac{1}{h_i} (x - x_{i-1}) & \text{für } x \in \Omega_i \\ \frac{1}{h_{i+1}} (x_{i+1} - x) & \text{für } x \in \Omega_{i+1} \\ 0 & \text{sonst} \end{cases}$$

*definiert. Es gilt nach Konstruktion  $\phi \in C(\bar{\Omega})$  sowie  $\phi_i|_{\Omega_j} \in C^1(\bar{\Omega}_j)$ , somit hat man insgesamt  $\phi_i \in H^1(\Omega)$  [CG05, 184]. Die folgende Abbildung stellt die Graphen von Ansatzfunktionen  $\phi_i$  dar.*

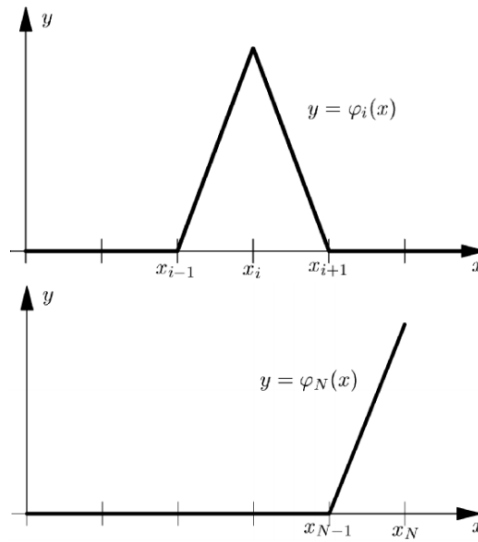


Abbildung 1: Ansatzfunktionen  $\phi_i$  [CG05, 184]

*Es gilt demnach  $\phi_i(x_k) = \delta_{ik}$  mit  $i, k = 0, 1, \dots, N$ .*

Nun haben wir eine Idee davon, wie man ein Gebiet zerlegt und wie Ansatzfunktionen aussehen könnten und welche Eigenschaften sie zu erfüllen haben, doch wie genau sieht nun das diskrete Problem bei der *Finite Elemente Methode* nun aus?

Dazu müssen wir uns die sogenannte *Assemblierung* der *Steifigkeitsmatrix*  $A_n$  anschauen. Wir werden uns die Teilelemente der Zerlegung dazu einzeln anschauen und sogenannte *Elementsteifigkeitsmatrizen* ausrechnen. Im *Assemblierungsschritt* werden wir dann die *Elementsteifigkeitsmatrizen* zu der *globalen Steifigkeitsmatrix* zusammen setzen. Wir werden hier nodale Basisfunktionen benutzen. Diese sind durch  $\varphi_k(x_l) = \delta_{kl}$  definiert. Weiterhin sei  $\hat{N}$  die Zahl der Freiheitsgrade. Rekapitulieren wir das zu lösende Problem

1. Finde ein  $u_n \in V_n : a(u_n, v_n) = f(v_n)$  für alle  $v_n \in V_n$
2. Sei  $\{\varphi_i\}_{i=1}^{\hat{N}}$  die Basis von  $V_n$
3. Definiere  $A_n = (a(\varphi_k, \varphi_i))_{i,k=1}^{\hat{N}}$  und  $f_n = (f(\varphi_i))_{i=1}^{\hat{N}}$
4. Löse lineares Gleichungssystem  $A_n u_n = f_n$  zur Bestimmung der Koeffizienten  $u_i$  der Darstellung  $u_n(x) = \sum_{i=1}^{\hat{N}} u_i \varphi_i(x)$

In unserem Fall war  $a(u, v) = \int_{\Omega} \Delta u \Delta v dx$  bzw.  $f(v) = \int_{\Omega} f v dx$ . Die zu  $\Omega_j$  gehörige Elementsteifigkeitsmatrix besitzt die Form

$$\begin{aligned} A_h^j &= (a_{ik}^j)_{i,k \in I_j} \\ a_{ik}^j &= \int_{\Omega_j} \Delta \varphi_i \Delta \varphi_k dx \quad \text{mit} \quad I_j = \{i : \text{supp } \varphi_i \cap \Omega_j \neq \emptyset\} \end{aligned} \quad (2.14)$$

Analog dazu die elementweise rechte Seite

$$f^j = (f_i^j)_{i \in I_j} \quad \text{mit} \quad f_i^j = \int_{\Omega} f \varphi_i dx$$

Die globale Steifigkeitstmatrix  $A_n$  und die rechte Seite  $f_n$  ergeben sich dann wegen der Additivität des Integrals als Summen

$$A_n = (a_{ik})_{i,k=1}^{\hat{N}} \quad \text{mit} \quad a_{ik} = \sum_{j=1, i \in I_j, k \in I_j}^M a_{ik}^j$$

und

$$f_h = (f_i)_{i=1}^{\hat{N}} \quad \text{mit} \quad f_i = \sum_{j=1, i \in I_j}^M f_i^j$$

$I_j$  ist gerade die List der Eckpunkte der Elemente. Wir sehen für die Elemente der Elementsteifigkeitsmatrix in (2.14) ein Integral über ein Teilgebiet  $\Omega_j$ . Hier liegt eine gute Chance viele Operationen zu sparen, indem wir uns die Struktur vom Integral zu nutze machen und mit Hilfe vom Transformationssatz für Integrale eine einheitlichere Form herleiten.

Dazu definieren uns sogenannten *Referenzelemente*. Beispielsweise wenn wir eine Zerlegung in Rechtecken wählen, würden wir uns ein Referenzrechteck definieren oder im Falle von Dreiecken ein Referenzdreieck. Weiterhin definieren wir uns eine lineare Transformation die gerade die Ecken unseres Teilgebiets auf die Ecken des korrespondierenden Referenzelements abbildet. Mit Hilfe vom Transformationssatz formen wir das Integral um und erhalten plötzlich für alle Elementsteifigkeitsmatrizen dieselben Integrale bloß mit verschiedenen Konstanten multipliziert. Die Konstanten sind gerade die Beträge der Determinanten der linearen Transformationen.

Um dann die die Elemente der Elementsteifigkeitsmatrizen in die globale Steifigkeitsmatrix einzubetten, ist dies eine Frage des Umdenkens von einer lokalen Struktur in die globale Struktur. Dieses Umdenken ist ausschlaggebend für das Verstehen vom Finite Elemente Ansatz und auch der späteren Arbeit zur Herleitung der Pseudoinversen.

Wir rekapitulieren die erste Gleichung dieser Arbeit

$$v = A(u) = \sum_{k=1}^{n_{cells}} C^T P_k^T A_k (P_k C u) \quad (2.15)$$

Die Variable  $n_{cells}$  ist somit die Anzahl der Teilgebiete  $\Omega_j$ . Die Matrix  $A_k$  ist der elementspezifische Operator A. Die Matrix  $P_k$  ist genau diese Transformation von der wir gerade sprachen, nämlich die Transformation von den lokalen Freiheitsgraden in die globalen Freiheitsgrade. Es gilt nur noch zu klären, welche Aufgabe die Matrix  $C$  in unserer Gleichung besitzt. Diese Frage klären wir im Rahmen des nächsten Unterkapitels zum *Diskontinuierlichen Galerkin-Verfahren*.

Vorher definieren wir uns die Strukturen, die diese Arbeit untersucht. Das ist einerseits die Masse Matrix mit folgender Bilinearform

$$a_M(u, v) = \int_{\Omega} u v \, dx \quad (2.16)$$

Eine einfache Struktur, die dazu dient sich an die Thematik ranzutasten und ein Gefühl dafür zu bekommen. Desweiteren wollen wir die Bilinearform des bereits erwähnten elliptischen Problems untersuchen mit

$$a_L(u, v) = \int_{\Omega} \Delta v \Delta v \, dx \quad (2.17)$$

Außerdem brauchen wir etwas Hintergrundwissen, wie wir die Integrale in (3.1) und in (2.17) berechnen können. In der Praxis erfolgt die Berechnung von Integralen über sogenannte *Quadraturformeln*.

#### 2.1.4 Quadratur

Allgemein beschäftigt uns das Integrationsproblem

$$I(f) = \int_a^b f(x) \, dx \quad (2.18)$$

mit  $a, b \in \mathbb{R}$ ,  $a < b$  und  $f \in C[a, b]$ . Wir werden von interpolatorischen Quadraturformeln gebrauch machen. Intuitiv machen wir eine Polynominterpolation für die zu integrierende Funktion  $f$  und summieren über Stützstellen.

Es seien  $x_i \in \mathbb{R}$ ,  $a \leq x_0 < \dots < x_n \leq b$  Stützstellen mit Gewichten  $\alpha_i \in \mathbb{R}$ . Dann können wir das Integral (2.18) approximieren durch

$$I(f) = \int_a^b f(x) \, dx \approx \sum_{i=0}^n \alpha_i f(x_i) = I^{(n)}(f) \quad (2.19)$$

Doch wie genau ist diese Approximation? Interpolatorische Quadraturformeln  $I^{(n)}(\cdot)$  zu  $n + 1$  Stützstellen sind mindestens von Ordnung  $n + 1$ . Das heißt sie integrieren alle Polynome von maximalen Grad  $n$  exakt.

Einer Quadraturformel wollen wir besondere Beachtung schenken, da diese für unsere Anwendung besondere praktikabilität gezeigt hat.

**Definition 2.10.** (*Gauss Lobatto*)

Es sei  $x_i$  die  $(i - 1)$  te Nullstelle des Legendre Polynoms  $P'_{n-1}(x)$  und  $n$  die Anzahl der Stützstellen, bzw.  $n - 1$  der Grad unseres Legendre Polynoms. Dann können wir

das Integral auf dem Gebiet  $[-1, 1]$  der Funktion  $f$  approximieren durch

$$\int_{-1}^1 f(x) dx \approx \frac{2}{n(n-1)}[f(1) + f(-1)] + \sum_{i=2}^{n-1} w_i f(x_i).$$

mit Gewichten

$$w_i = \frac{2}{n(n-1)[P_{n-1}(x_i)]^2}, \quad x_i \neq \pm 1$$

Gauss Lobatto liefert uns eine exakte Approximation von Polynomen bis zu Grad  $2n - 3$ . Eine ausführliche Ausarbeitung der Thematik finden Sie in [Ran05, 79].

### 2.1.5 Diskontinuierliche Galerkin-Methode

Die diskontinuierliche Galerkin-Methode wurde 1973 erstmals von Reed und Hill eingeführt für hyperbolische Gleichung erster Ordnung. Es gab eine Reihe von Untersuchungen seither bezüglich hyperbolische Probleme erster Ordnung als auch für die Diskretisierung instationärer Probleme. Unabhängig davon wurde die diskontinuierliche Galerkin-Methode für elliptische Gleichungen vorgeschlagen. Man wollte mehr Flexibilität bezüglich der Stetigkeitsvoraussetzungen an unsere lokalen Funktionen und mehr Freiraum bei der Gitter Generierung, insbesondere bei adaptiven h-p-Methoden. Die Art von dGFEM-Code erleichtert Parallelisierbarkeit was in Zeiten von GPU Programmierung eine große Effizienzsteigerung erlaubt. Die Idee von dGFEM ist maßgeblich Strafterme einzuführen, welche Unstetigkeit zwar erlauben aber in ihrem Ausmaß einschränkt. Diese Freiheit erlaubt uns aber zum Beispiel lokal Polynome höherer Ordnung zu benutzen und Singularitäten damit gekonnt zu beseitigen, ohne von Vorne zu beginnen zu müssen. Zu mal pro Element ein hängender Knoten erlaubt ist. Das heißt eine Verfeinerung des Gitters ist lokal möglich, ohne direkt Probleme zu bekommen [CG05, 292].

$$v = A(u) = \sum_{k=1}^{n_{cells}} C^T P_k^T A_k (P_k C u) \quad (2.20)$$

Die Matrix C kommt aus der Diskontinuierlichen Galerkin Methode und kümmert sich um *hängende Knoten* und sorgt dafür, das wir keine Probleme mit der Stetigkeit der Lösung bekommen.

Nun können wir diese Gleichung besser fassen und haben eine Idee davon, was sie aussagt. In dieser Arbeit werden wir uns nur mit dem Term

$$v_k = A_k u \quad (2.21)$$

beschäftigen, für  $A_k$  Elementsteifigkeitsmatrix der Masse Matrix und der Laplace Bilinearform für Referenzzellen im zweidimensionalen. Die Verallgemeinerung auf die dritte Dimension ist mit wendig Mehraufwand verbunden und wird an gegebener Stelle erläutert. Außerdem wir machen uns keine Gedanken um  $P_k$  noch um  $C$ .

Das Ziel des nächsten Kapitels ist die Untersuchung der Transformation von (2.21) zu

$$A_k^+ v_k = u \tag{2.22}$$

Vorher sollten wir uns jedoch mit der Tensor Dekomposition beschäftigen und dafür eine theoretische Grundlage schaffen.

## 2.2 Tensor Dekomposition

### 2.2.1 Einführung in die Tensor Architektur

In Kapitel 3 werden wir sehen, dass es möglich ist die Elementsteifigkeitsmatrix der Masse Matrix und der Laplace Bilinearform als einen Tensor umzudefinieren. Nachdem wir dies gemacht haben, werden wir die Strukturen dieses Tensors untersuchen, mit Hinblick einen einfachen Weg zu finden, die Pseudoinverse zu bestimmen. Was genau ein Tensor ist, was wir unter der Pseudoinversen eines Tensors verstehen und wie wir einen Tensor analysieren, wird in diesem Unterkapitel beantwortet.

**Definition 2.11.** (*Tensor*)

Ein Tensor ist eine multidimensionale Matrix  $\mathfrak{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ . Die Ordnung ist die Anzahl der Dimensionen, in diesem Fall  $N$ .

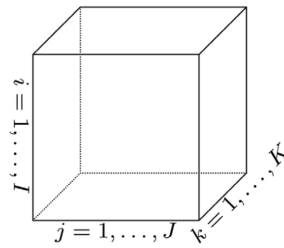


Abbildung 2: Tensor dritter Ordnung  $\mathfrak{X} \in \mathbb{R}^{I \times J \times K}$  [TK09, 456]

Vorsicht zu haben, gilt es bei den Begriffen Ordnung und Rang bei Tensoren. Es sollte vermieden werden diese Begriffe zu verwechseln. Außerdem sollte man nicht den Begriff des Rangs einer Matrix, mit dem Begriff des Ranges eines Tensors verwechseln. Die Definition des Rangs eines Tensors werden wir hier auslassen, da dies ein überaus schwieriger Begriff ist und es vermieden werden kann in dieser Arbeit damit zu hantieren. Eine ausführliche Erklärung des Begriffs finden Sie in [TK09, 464].



Um unsere Tensoren zu klassifizieren und charakterisieren, brauchen wir Eigenschaftsbegriffe. Es sei  $\mathfrak{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  ein Tensor.

**Definition 2.12.** (*Symmetrie*)

- a) Den Tensor  $\mathfrak{X}$  nennt man *kubisch genau* dann wenn  $I_i = I_j$  für alle  $i, j$ .
- b) Einen kubischen Tensor nennt man *supersymmetrisch genau* dann wenn die Elemente des Tensors konstant bleiben unter jeglicher Permutation der Indizes.
- c) Einen Tensor nennt man *stückweise symmetrisch*, wenn die Elemente konstant bleiben unter der Permutation von mindestens 2 Indizes.

**Definition 2.13.** (*Diagonal*)

Den Tensor  $\mathfrak{X}$  nennt man *diagonal*, wenn  $x_{i_1, \dots, i_N} \neq 0$  genau dann wenn  $i_1 = \dots = i_N$ .

**Definition 2.14.** (*Faser*)

Eine Faser ist das multidimensionale Analog zu Matrixspalten und Matrixzeilen. Wir definieren eine Faser, indem wir jeden Index abgesehen von einem festhalten.

Einen Tensor kann man entfalten. Dies impliziert eine Neuordnung der Tensorelemente in eine Matrix. Wir betrachten nur die sogenannte *mode-n Entfaltung*, da dies die einzig relevante Form der Entfaltung für uns ist.

**Bemerkung 2.15.** (*Entfaltung*)

Eine mode-n Entfaltung des Tensors  $\mathfrak{X}$  wird mit  $\mathbf{X}_{(n)}$  geschrieben und ordnet die mode-n Fasern in die Spalten der Ergebnismatrix. Formal ist es eine Abbildung des Indize  $N$ -tupels  $(i_1, \dots, i_N)$  auf Matrixindizes  $(i_n, j)$

$$j = 1 + \sum_{\substack{k=1 \\ k \neq n}}^N (i_k - 1) J_k \text{ mit } J_k = \prod_{\substack{m=1 \\ m \neq n}}^{k-1} I_m \quad (2.23)$$

Nun fehlt uns noch eine Tensor Multiplikation um mit der Dekomposition von Tensoren anzufangen.

**Definition 2.16.** (*n-mode Produkt*)

Das n-mode Produkt des Tensors  $\mathfrak{X}$  mit einer Matrix  $\mathbf{U} \in \mathbb{R}^{J \times I_n}$  wird als  $\mathfrak{X} \times_n \mathbf{U}$  notiert. Die Ergebnismatrix hat die Größe  $I_1 \times \dots \times I_{n-1} \times J \times I_{n+1} \times \dots \times I_N$

$$(\mathfrak{X} \times_n \mathbf{U})_{i_1 \dots i_{n-1} j i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 \dots i_N} u_{j i_n} \quad (2.24)$$

**Bemerkung 2.17.** *Jedes mode- $n$  Produkt kann mit Hilfe von entfalteten Tensoren äquivalent ausgedrückt werden.*

$$\mathbf{Y} = \mathbf{X} \times_n \mathbf{U} \iff \mathbf{Y}_{(n)} = \mathbf{U} \mathbf{X}_{(n)} \quad (2.25)$$

### 2.2.2 Singulärwertzerlegung höherer Ordnung

Die *Singulärwertzerlegung höherer Ordnung* bzw. *Higher Order Singular Value Decomposition* (HOSVD) oder auch bekannt unter der Tucker Dekomposition ist eine uminterpretierte multidimensionale Hauptkomponentenanalyse. Man versucht durch Hauptachsentransformationen die Korrelation zwischen den verschiedenen Komponenten durch Überführung in eine neue Basis zu minimieren. Die HOSVD zerlegt den Tensor in einen sogenannten *Core Tensor* multipliziert mit einer Matrix in jeder Ordnung.

Allgemein ist die HOSVD des Tensors  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  gegeben durch

$$\mathbf{X} = \mathbf{G} \times_1 A^{(1)} \dots \times_N A^{(N)} \quad (2.26)$$

Man kann äquivalent die HOSVD auch mit entfalteten Tensoren wie folgt angeben

$$\mathbf{X}_{(n)} = A^{(n)} \mathbf{G}_{(n)} (A^{(1)} \otimes \dots \otimes A^{(n+1)} \otimes A^{(n-1)} \otimes \dots \otimes A^{(1)})^T \quad (2.27)$$

**Beispiel 2.18.** (*HOSVD Tensor dritter Ordnung*)

Es sei  $\mathbf{X} \in \mathbb{R}^{I \times J \times K}$ . Dann kann man den Tensor  $\mathbf{X}$  zerlegen in

$$\mathbf{X} \approx \mathbf{G} \times_1 A \times_2 B \times_3 C \quad (2.28)$$

wobei  $A \in \mathbb{R}^{I \times P}$ ,  $B \in \mathbb{R}^{J \times Q}$  und  $C \in \mathbb{R}^{K \times R}$  die Faktormatrizen sind, welche orthogonal sind.  $\mathbf{G}$  bezeichnet den Core Tensor und zeigt wie hoch die Korrelation zwischen den verschiedenen Komponenten ist.

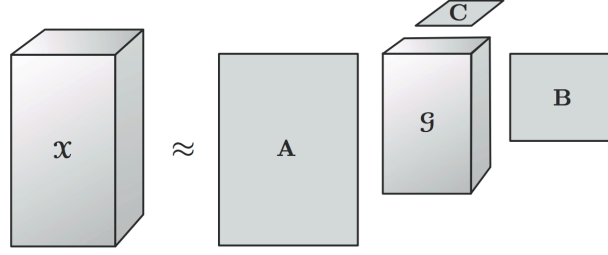


Abbildung 3: HOSVD eines Tensors dritter Ordnung [TK09, 475]

**Bemerkung 2.19.** (Berechnung der HOSVD)

Die Berechnung der HOSVD von  $\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$  geht wie folgt

1. Berechne die mode- $k$  Entfaltungen  $\mathbf{X}_{(k)}$  für alle  $k$
2. Berechne die Singulärwertzerlegung  $\mathbf{X}_{(k)} = U_k \Sigma_k V_k^T$  und speichere  $U_k$
3. Der Kerntensor  $\mathbf{G}$  ergibt sich aus der Projektion des Tensors auf die Tensorbasis geformt von den Faktormatrizen  $\{U_k\}_{k=1}^N$  also  $\mathbf{G} = \mathbf{X} \times_{n=1}^N U_n^T$

Die HOSVD existiert für alle Tensoren. Wie sieht es aber mit der Eindeutigkeit aus? Die HOSVD ist keine eindeutige Zerlegung. Dies führen wir an einem Tensor dritter Ordnung an.

**Beispiel 2.20.** (Eindeutigkeit der HOSVD)

Es seien  $U \in \mathbb{R}^{P \times P}$ ,  $V \in \mathbb{R}^{Q \times Q}$  und  $W \in \mathbb{R}^{R \times R}$ . Es gilt

$$\mathbf{X} = \mathbf{G} \times_1 A \times_2 B \times_3 C = (\mathbf{G} \times_1 U \times_2 V \times_3 W) \times_1 AU^{-1} \times_2 BV^{-1} \times_3 CW^{-1} \quad (2.29)$$

In anderen Worten: Wir können den Core Tensor  $\mathbf{G}$  modifizieren, ohne die Gleichung zu ändern, solange wir das Inverse der Modifizierung auf den zugehörigen Faktormatrizen multiplizieren.

Mit dieser Kenntniss können wir nun zum Beispiel versuchen, so viele Elemente des Kerntensor wie möglich auf Null zu bekommen oder so klein wie möglich zu machen, damit wir bei der späteren Herleitung der Pseudoinversen weniger Probleme bekommen.

Wir brauchen noch einige Eigenschaften des Kronecker Produkts, die wir uns später für die Berechnung der Pseudoinversen zu Nutze machen wollen.

**Lemma 2.21.** (*Invertieren des Kronecker Produkts*)

Es seien  $A \in \mathbb{R}^{i \times i}$  und  $B \in \mathbb{R}^{j \times j}$  invertierbar, so ist auch  $(A \otimes B)$  invertierbar. Mit der Inversen

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

Für die Moore Penrose Pseudoinversen gilt analog

$$(A \otimes B)^+ = A^+ \otimes B^+$$

**Lemma 2.22.** (*Matrixprodukt und Kronecker Produkt*)

Es seien  $A, B, C, D$  Matrizen, deren Matrizenprodukte  $AC$  und  $BD$  definiert sind, dann gilt

$$AC \otimes BD = (A \otimes B)(C \otimes D).$$

Dieses Ergebnisse sind entscheidend für die Herleitung der Pseudoinversen. Die theoretische Grundlage ist nun geschaffen. Es ist Zeit sich dem Herz dieser Arbeit zu widmen, nämlich der Tensorstruktur der Elementarsteifigkeitsmatrizen und die Herleitung der Pseudoinversen.

## 3 Tensorstruktur und Pseudoinverse der Zellmatrizen

### 3.1 Summenfaktorisierung

In [Tea] wird die effektive Berechnung der Masse Matrix mit einem Vektor vorgestellt. Diese Methodik sollten wir uns kurz vor Augen führen und daraufaufbauend um einige eigene Gedanken erweitern um diese Methodik für unsere Anwendung zugänglich zu machen. Das Ziel dieses Unterkapitels ist die Tensorstruktur für die Massematrix und der Laplace Bilinearform herzuleiten und für die Berechnung der Pseudoinversen zu nutzen.

#### 3.1.1 Berechnung der Elementmassenmatrix

Es sei  $T$  die Referenzzelle für Rechtecke und  $\varphi_i(\mathbf{x})$  Basisfunktion des diskreten Raumes  $V_n$  mit  $\mathbf{x} = (x, y)$ .

$$M_{ik} = \int_T \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \quad (3.1)$$

Die Basisfunktionen haben eine Tensorstruktur, die wie folgt aussieht

$$\varphi_i(\mathbf{x}) = \varphi_{i_1+(N+1)i_2}(x, y) = \varphi_{i_1}(x)\varphi_{i_2}(y) \quad (3.2)$$

Wir werden eine lexikographische Ordnung der Freiheitsgrade benutzen. Die Reichweite von  $i_1$  und  $i_2$  reicht von 0 bis  $N$ . Das heißt der Index  $i$  geht von 0 bis  $N_p = (N + 1)^2 - 1$ . In Abbildung (4) sehen Sie ein Beispiel für  $N = 3$ .

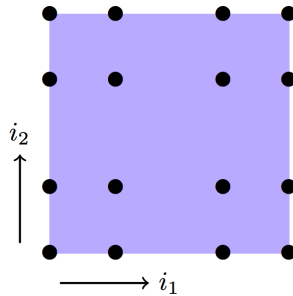


Abbildung 4: [Tea, 3]

Wir wollen für die Berechnung der Integrale in (3.1) die Gauss Quadratur benut-

zen. Es seien  $\mathbf{x}_q = (x_{q1}, x_{q2})$  die Stützstellen und  $\mathbf{w}_q = w_{q1}w_{q2}$  die Gewichte. Dann können wir (3.1) approximieren wie folgt

$$\begin{aligned}
M_{ij} &= \int_T \varphi_i(\mathbf{x}) \varphi_j(\mathbf{x}) d\mathbf{x} \\
&\approx \sum_{q=1}^Q \mathbf{w}_q \varphi_i(\mathbf{x}_q) \varphi_j(\mathbf{x}_q) \\
&= \sum_{q_1=1}^{Q_{1D}} \sum_{q_2=1}^{Q_{1D}} \varphi_{i_1}(x_{q1}) \varphi_{i_2}(x_{q2}) \varphi_{j_1}(x_{q1}) \varphi_{j_2}(x_{q2}) w_{q1} w_{q2}
\end{aligned} \tag{3.3}$$

Wir wählen die Anzahl der Quadraturpunkte  $Q_{1D}$  per Dimension so, dass wir exakt integrieren. Wir wählen  $Q_{1D}$  gleich der Anzahl der Basisfunktionen  $N + 1$ , da wir mit der Gauss Quadratur mit  $N + 1$  Stützstellen bis  $2N + 1$  exakt integrieren und der höchste Grad bei uns  $2N$  ist.

Wir definieren uns  $\mathcal{N}_{iq} = \varphi_i(x_q)$  und weiterhin die Matrix  $\mathcal{W}_{ii} = \mathbf{w}_i$ , die in der Diagonalen die Quadraturgewichte hat und sonst Nullen. Damit können wir nun die Massematrix schreiben als

$$M = \mathcal{N} \mathcal{W} \mathcal{N}^T \tag{3.4}$$

Dies ist die Form mit in  $\mathcal{N}$  an Stützstellen evaluierte zweidimensionale Basisfunktionen. Wir können dies aber weiter aufspalten in eindimensionale Basisfunktionen und dadurch eine Effizienzsteigerung erzielen bei der Berechnung des Matrix-Vektor Produkts mit der Elementmassenmatrix.

### 3.1.2 Berechnung der Elementmassenmatrix-Vektor Produkt

Das Ziel dieses Unterkapitels ist die effiziente Berechnung des Matrix-Vektor Produkts  $Mu = v$  mit  $M$  als Massematrix.

Wir fangen damit an, die Matrix  $\mathcal{N}$  in ein Tensorprodukt aufzuspalten.

$$\mathcal{N} = \mathcal{N}^{1D} \otimes \mathcal{N}^{1D} \tag{3.5}$$

Die Matrix  $\mathcal{N}^{1D}$  ist nun äquivalent definiert wie  $\mathcal{N}$  bloß mit eindimensionalen Basisfunktionen. Äquivalent ausgeschrieben sieht die Gleichung (3.5) wie folgt aus

$$\begin{bmatrix} \varphi_1^{2D}(\mathbf{x}_1) & \dots & \varphi_N^{2D}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1^{2D}(\mathbf{x}_Q) & \dots & \varphi_N^{2D}(\mathbf{x}_Q) \end{bmatrix} = \begin{bmatrix} \varphi_1^{1D}(x_1) & \dots & \varphi_n^{1D}(x_1) \\ \vdots & \ddots & \vdots \\ \varphi_1^{1D}(x_{Q_{1D}}) & \dots & \varphi_n^{1D}(x_{Q_{1D}}) \end{bmatrix} \otimes \begin{bmatrix} \varphi_1^{1D}(x_1) & \dots & \varphi_n^{1D}(x_1) \\ \vdots & \ddots & \vdots \\ \varphi_1^{1D}(x_{Q_{1D}}) & \dots & \varphi_n^{1D}(x_{Q_{1D}}) \end{bmatrix}$$

Wir nutzen absofort  $Q_{1D} = N$  und  $Q = (N + 1)^2$ , da wie wir oben bereits argumentiert haben, damit exakt integrieren können. Wir erinnern uns  $v = Mu = \mathcal{N}\mathcal{W}\mathcal{N}^T$ . Da  $\mathcal{W}$  eine Diagonalmatrix ist, ist es naheliegend diese Multiplikation bereits durchzuführen. Definiere  $\mathcal{W}_N = \mathcal{N}\mathcal{W}$  und die Spaltung von  $\mathcal{W}_N$  sieht wie folgt aus

$$\mathcal{W}_N = \mathcal{W}_N^{1D} \otimes \mathcal{W}_N^{1D} \quad (3.6)$$

Genauer

$$\begin{bmatrix} \mathcal{N}_{11}\mathbf{w}_1 & \dots & \mathcal{N}_{1N}\mathbf{w}_N \\ \vdots & \ddots & \vdots \\ \mathcal{N}_{N1}\mathbf{w}_1 & \dots & \mathcal{N}_{NN}\mathbf{w}_N \end{bmatrix} = \begin{bmatrix} \mathcal{N}_{11}^{1D}w_1 & \dots & \mathcal{N}_{1N}^{1D}w_N \\ \vdots & \ddots & \vdots \\ \mathcal{N}_{N1}^{1D}w_1 & \dots & \mathcal{N}_{NN}^{1D}w_N \end{bmatrix} \otimes \begin{bmatrix} \mathcal{N}_{11}^{1D}w_1 & \dots & \mathcal{N}_{1N}^{1D}w_N \\ \vdots & \ddots & \vdots \\ \mathcal{N}_{N1}^{1D}w_1 & \dots & \mathcal{N}_{NN}^{1D}w_N \end{bmatrix}$$

Damit können wir folgende Umformulierung bereits vornehmen

$$v = Mu = \mathcal{N}\mathcal{W}\mathcal{N}^T = \mathcal{W}_N\mathcal{N}^T = [(\mathcal{W}_N^{1D} \otimes \mathcal{W}_N^{1D})(\mathcal{N}_{1D} \otimes \mathcal{N}_{1D})^T]u \quad (3.7)$$

Mit Lemma (2.22) folgt

$$v = [(\mathcal{N}_{1D} \otimes \mathcal{N}_{1D})\mathcal{W}(\mathcal{N}_{1D} \otimes \mathcal{N}_{1D})^T]u = [(\mathcal{N}_{1D}\mathcal{N}_{1D}) \otimes (\mathcal{N}_{1D}\mathcal{N}_{1D})^T]u \quad (3.8)$$

Angenommen wir hätten eine Lösung  $u \in V$  der Form

$$u(x, y) = \sum_{h=1}^{n^2} \psi_h(x, y) u_h, \quad u_h \in \mathbb{R} \quad \forall h = 1, \dots, n^2 =: N \quad (3.9)$$

mit  $\psi_h$  ist eine globale Ansatzfunktion. Diese Ansatzfunktionen können als das Tensorprodukt von zwei ein-dimensionalen Polynomen ausgedrückt werden. Die Überführung ist definiert mit

$$h = j(n-1) + i, \quad i < n \leftrightarrow (i, j) \quad (3.10)$$

was uns eine alternative Representation von  $u$  gibt

$$u(x, y) = \sum_{i=1}^n \sum_{j=1}^n \varphi_i(x) \varphi_j(y) u_{ij}, \quad i, j = 1, \dots, n \quad (3.11)$$

Seien  $q = (q_1, \dots, q_n)^T$  Quadraturpunkte einer 1-dimensionalen Quadraturegel mit zugehörigen Gewichten  $w = (w_1, \dots, w_n)^T$ . Wir können die 2-dimensionale Regel erhalten indem wir wieder das Tensorprodukt ausnutzen und erhalten  $(\mathbf{q}_1, \dots, \mathbf{q}_N)$  und Gewichten  $(\mathbf{w}_1, \dots, \mathbf{w}_N)$  mit  $\mathbf{q}_h = (q_i, q_j)$  and  $\mathbf{w}_h = w_i w_j$ .

Erstmals wollen wir  $u$  in jedem Quadraturpunkt evaluieren. Das können wir mit einem Matrix Vektor Produkt ausdrücken:

$$\begin{pmatrix} \psi_1(\mathbf{q}_1) & \dots & \psi_N(\mathbf{q}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\mathbf{q}_N) & \dots & \psi_N(\mathbf{q}_N) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} u(\mathbf{q}_1) \\ \vdots \\ u(\mathbf{q}_N) \end{pmatrix} \quad (3.12)$$

Die Matrix kann als Tensorprodukt zweier, in diesem Fall sogar identischer, Matrizen:

$$\mathcal{N}^T \otimes \mathcal{N}^T = \begin{pmatrix} \varphi_1(q_1) & \dots & \varphi_n(q_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(q_n) & \dots & \varphi_n(q_n) \end{pmatrix} \otimes \begin{pmatrix} \varphi_1(q_1) & \dots & \varphi_n(q_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(q_n) & \dots & \varphi_n(q_n) \end{pmatrix} \quad (3.13)$$

Sei  $(\mathcal{U}_{ij}) \in \mathbb{R}^{n \times n}$  die Matrix mit den Koeffizienten  $u_{ij}$  als Elemente. Durch die Nutzung der Summenfaktorisierung können wir diese Matrizen multiplizieren anstatt



der Formel in (4). Wir erhalten:

$$\mathcal{N}^T \mathcal{U} \mathcal{N} = \bar{\mathcal{U}}, \quad \bar{\mathcal{U}}_{ij} = u(q_i, q_j) \quad (3.14)$$

Die Quadraturgewichte und die Summe über die Elemente kann in einer effiziente Weise vollzogen werden, wenn wir die Tensorprodukt Struktur ausnutzen. Die Multiplikation  $\bar{\mathcal{U}}$  mit den Quadraturgewichten  $w$  von der rechten Seite wird uns die gewichteten Spalten aufsummieren, multipliziert mit  $w^T$  wird den Rest erledigen. Die ganze Operation sieht wie folgt aus:

$$\int \int u(x, y) dx dy \approx w^T \mathcal{N}^T \mathcal{U} \mathcal{N} w \quad (3.15)$$

Wir haben nun  $u$  integriert. Jedoch können wir nun mit Ansatzfunktionen testen ohne viel zu verändern. Wir schauen uns  $f(x, y) = g(x)h(y)$  an. Also ist  $f$  separabel und somit müssen wir nur den letzten Schritt ändern. Anstatt nur den Gewichten zu multiplizieren, werden wir auch mit den Ansatzfunktionen, an den bestimmten Quadraturpunkten evaluiert, multiplizieren:

$$\bar{w}^x = (w_1 g(q_1), \dots, w_n g(q_n))^T \quad \bar{w}^y = (w_1 h(q_1), \dots, w_n h(q_n))^T$$

$$\int \int u(x, y) f(x, y) dx dy \approx (\bar{w}^x)^T \mathcal{N}^T \mathcal{U} \mathcal{N} \bar{w}^y \quad (3.16)$$

Wenn wir eine Menge von Ansatzfunktionen haben, welche hergeleitet sind von dem Tensorprodukt von identischen 1-dimensionalen Ansatzfunktionen, können wir die Matrizen multiplizieren anstatt die Vektoren im letzten Schritt, was nun die Operation vollständig macht. Wir definieren:

$$W = \begin{pmatrix} w_1 \varphi_1(q_1) & \dots & w_n \varphi_1(q_n) \\ \vdots & \ddots & \vdots \\ w_1 \varphi_n(q_1) & \dots & w_n \varphi_n(q_n) \end{pmatrix}$$

Da  $\mathcal{W} \mathcal{N}^T$  auf Beiden Seiten auftaucht, können wir diese Operation mit 3 Matrixmultiplikationen ausdrücken. Finale Form der Gleichung:

$$\mathcal{V} = \mathcal{W} \mathcal{N}^T \mathcal{U} (\mathcal{W} \mathcal{N}^T)^T \quad (3.17)$$

Die Formel leicht geändert kann benutzt werden um andere Bilinearformen auszudrücken:

$$(\nabla u, \nabla v).$$

$$(\nabla u, \nabla v) = (\partial_x u, \partial_x v) + (\partial_y u, \partial_y v) \quad (3.18)$$

$$= \sum_{i,j} u_{ij}(\varphi'_i(x)\varphi_j(y), \phi'(x)\phi(y)) \quad (3.19)$$

$$+ \sum_{i,j} u_{ij}(\varphi_i(x)\varphi'_j(y), \phi(x)\phi'(y)) \quad (3.20)$$

Nun werden wir 2 Matrizen  $\mathcal{W}'$  und  $\mathcal{N}'$  vorstellen. Diese Matrizen sind ähnlich zu den oben, aber anstatt die Ansatzfunktionen zu evaluieren, evaluieren sie die Ableitung der Ansatzfunktionen. Die resultierende Form ergibt sich aus:

$$\mathcal{V} = \mathcal{W}'\mathcal{N}'^T\mathcal{U}(\mathcal{W}\mathcal{N}^T)^T + \mathcal{W}\mathcal{N}^T\mathcal{U}(\mathcal{W}'\mathcal{N}'^T)^T$$

Sei  $(-\Delta u, v) = -(\partial_{xx}u, v) + -(\partial_{yy}u, v)$ .

Wir führen  $\mathcal{N}''$  ein, was die Werte der 2.Ableitung der Ansatzfunktionen in den Quadraturgewichten speichert. Bemerke, dass die Matrix  $\mathcal{W}$  nicht modifiziert wurde. Als Resultat erhalten wir

$$\mathcal{V} = -\mathcal{W}\mathcal{N}''^T\mathcal{U}(\mathcal{W}\mathcal{N}^T)^T - \mathcal{W}\mathcal{N}^T\mathcal{U}(\mathcal{W}\mathcal{N}''^T)^T$$

Zusammengefasst erhalten wir für die Masse Matrix die Form

$$M * u = (\mathcal{W}\mathcal{N}^T) \otimes (\mathcal{W}\mathcal{N}^T) * u \quad (3.21)$$

und für die Laplace Bilinearform

$$\mathcal{V} * u = -((\mathcal{W}\mathcal{N}''^T) \otimes (\mathcal{W}\mathcal{N}^T) - (\mathcal{W}\mathcal{N}^T) \otimes (\mathcal{W}\mathcal{N}''^T)) * u \quad (3.22)$$

Die Berechnung der Pseudoinversen passiert mit diesem Ansatz quasi on the fly, da wir hier gleich als Ergebnis nicht die Pseudoinverse bekommen, sondern das Matrix Vektor Produkt mit der Pseudoinversen als Matrix.

Für die effiziente Berechnung der Inversen brauchen wir folgendes Ergebnis:

**Bemerkung 3.1.**

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \quad (3.23)$$

Nun erhalten wir fast direkt unser gewünschtes Ergebnis

$$M^{-1} * u = ((\mathcal{W}\mathcal{N}^T) \otimes (\mathcal{W}\mathcal{N}^T))^{-1} * u = ((\mathcal{W}\mathcal{N}^T)^{-1} \otimes (\mathcal{W}\mathcal{N}^T)^{-1}) * u \quad (3.24)$$

$\mathcal{W}\mathcal{N}^T \in \mathbb{R}^{n \times n}$  somit also quadratisch. Das heißt um das Pseudoinverse-Vektor Produkt zu berechnen müssen wir nur eine SVD von einer  $n \times n$  Matrix berechnen. Falls uns das zu teuer wird können wir auch eine truncated SVD benutzen. In welchem Komplexitätsbereich wir uns befinden, wird im nächsten Subkapitel angeführt.

Für die Laplace Bilinearform ähnlich:

$$\mathcal{V}^{-1} * u = -((\mathcal{W}\mathcal{N}''^T) \otimes (\mathcal{W}\mathcal{N}^T) - (\mathcal{W}\mathcal{N}^T) \otimes (\mathcal{W}\mathcal{N}''^T))^{-1} * u = -((\mathcal{W}\mathcal{N}''^T)^{-1} \otimes (\mathcal{W}\mathcal{N}^T)^{-1} - (\mathcal{W}\mathcal{N}^T)^{-1} \otimes (\mathcal{W}\mathcal{N}''^T)^{-1}) * u \quad (3.25)$$

Praktisch müssen wir also nur die Inversen von  $(\mathcal{W}\mathcal{N}''^T)$  und von  $(\mathcal{W}\mathcal{N}^T)$  berechnen.

Nun war dies der 2 Dimensionale Fall. Die Erweiterung auf den 3 Dimensionalen Fall ist analog und kann im “Efficient Evaluation ..... Teachlet” angeschaut werden.

Zu der Implementierungsstrategie kann man in Kapitel 4 mehr erfahren.

## 3.2 Singulärwertzerlegung höherer Ordnung

Wir wollen nun mit Hilfe von der Theorie zur Singulärwertzerlegung höherer Ordnung eine Theorie entwickeln, wie wir die Pseudoinverse zur Masse Matrix und zur globalen Steifigkeitsmatrix der Laplace Bilinearform effizient berechnen können.

Allgemein wissen wir:

$$\mathcal{A} = \mathcal{S} \times_{n=1}^N U^{(n)} \quad (3.26)$$

D.h. wir können einen Tensor  $\mathcal{A}$  in einen Kerntensor  $\mathcal{S}$  und zugehörige Faktormatrizen  $U$  zerlegen. Wie bekommen wir nun die Pseudoinverse zu  $\mathcal{A}$ ? Was bedeutet in dem Kontext eines Tensors überhaupt Pseudoinverse?

Wir kennen noch die Eigenschaften der Moore Penrose Pseudoinverse für Matrizen:

**Lemma 3.2.** *Moore Penrose Pseudoinverse (Wikipedia)*

- $AA^+A = A$
- $A^+AA^+ = A^+$  ( $A^+$  ist eine schwache Inverse der multiplikativen Halbgruppe)
- $(AA^+)^* = AA^+$  ( $AA^+$  ist hermitisch)
- $(A^+A)^* = A^+A$  ( $A^+A$  ist hermitisch)

Jetzt gilt es diese Eigenschaften für Tensoren zu übertragen. Da wir erstmal keine intuitive Tensor Tensor Multiplikation haben, gilt es diese zu definieren. Diese

Tensor-Tensor Multiplikation macht dann nur für unsere Anwendung einen Sinn und ist sonst zweckfrei.

**Definition 3.3.** *Tensor-Tensor Multiplikation  $ttp(\cdot, \cdot)$*

Nun ist kommt auch ein weiterer Trick den wir nutzen können. Trick kann nur mit Vorsicht genossen werden. Die Gleichheit ist mit einem Rechner nicht zu erzielen, daher wird die Bemerkung abgeschwächt und für Tensoren. Wir erhalten:

**Lemma 3.4.** *Moore Penrose Pseudoinverse für Tensoren*

- $ttp(A, ttp(A^+, A)) - A < \epsilon$
- $ttp(A^+, ttp(A, A^+)) - (A^+ < \epsilon$  ( $A^+$  ist eine schwache Inverse der multiplikativen Halbgruppe)
- $(ttp(A, A^+))^* - ttp(A, A^+) < \epsilon$  ( $AA^+$  ist hermitisch)
- $(ttp(A^+, A))^* - ttp(A^+, A) < \epsilon$  ( $A^+A$  ist hermitisch)

Die Wahl des Epsilons ist hier entscheidend. Man könnte Maschinengenauigkeit wählen, doch ist für unser Zweck vielleicht zu Hoch gezielt. Letztlich wollen wir mit unserer Pseudoinversen einen Präkonditionierer bauen. Wenn wir durch die Wahl eines etwas größeren Epsilons erheblichen Aufwand sparen, sollten wir dies in Erwägung ziehen. Nun wissen wir, wie wir einen Tensor als Pseudoinverse klassifizieren können. Doch wie bekommen wir die Pseudoinverse?

Die Pseudoinverse setzt sich zusammen aus:

$$\mathcal{A}^\dagger = \mathcal{S}^\dagger \times_{n=1}^N U^{(n)T} \quad (3.27)$$

Da die Faktormatrizen orthogonal sind, reicht es also einfach die Transponierte zu nehmen. Das Invertieren des Kerntensors erweist sich nun aber als problematisch. Hier ist es nützlich die Struktur des Kerntensors zu kennen. Der Kerntensor ist leider in den meisten Fälle vollbesetzt. Doch genaueres Hinsehen zeigt 2 Arten von Zahlen. Ziemlich große Zahlen von größer als 1 und ziemlich kleine Zahlen von kleiner als  $10^{-10}$ . Die kleinen Zahlen sind in diesem Fall unbrauchbar und beinhalten wenig Informationen. Doch das Auslöschen vieler kleiner Zahlen nimmt uns in der Summe vielleicht relevante Informationen. Wir können also kleine Zahlen einfach ausradieren und erhalten plötzlich einen super-diagonalen Tensor. Die Invertierung des Tensors beschränkt sich darauf einfach jedes Diagonalelement zu Invertieren.

**Als nächstes wollen wir klären wie wir überhaupt von Massen Matrix Form und Laplace Bilinearform zu einer Tensorform kommen.** Die Masse Matrix  $M$  besteht aus Einträgen  $M_{ij} = (\phi_i, \phi_j)$ , wobei  $(\cdot, \cdot)$  die zugehörige Bilinearform ist und  $\phi_k$  eine 2 dimensionale Ansatzfunktion. Dies kann man vereinfachen mit  $M_{ij} = (\phi_i, \phi_j) = (\phi_{i_1}\phi_{i_2}, \phi_{j_1}\phi_{j_2}) = M^*_{i_1i_2j_1j_2}$ , wobei wir nun 1 dimensionale Ansatzfunktionen haben.  $M^*$  ist ein Tensor 4.Ordnung, den wir nun mit Hilfe der Singulärwertzerlegung höherer Ordnung zerlegen können. Für die Laplace Bilinearform ist es ähnlich, denn wir haben in schwacher Form  $L_{ij} = (\nabla\phi_i, \nabla\phi_j) = (\nabla\phi_{i_1}\nabla\phi_{i_2}, \nabla\phi_{j_1}\nabla\phi_{j_2}) = L^*_{i_1i_2j_1j_2}$

Wir wissen nun wie wir unsere Tensoren berechnen können und wissen auch wie die Pseudoinverse sich gewinnen lässt mittels der Singulärwertzerlegung höherer Ordnung. Der nächste Punkt ist die effiziente Berechnung der Pseudoinversen.

## 4 Effiziente Implementierung

In Kapitel 3 haben wir uns zwei Möglichkeiten angeschaut, das Matrix-Vektor Produkt zu berechnen mit der Pseudoinversen. Beide Alternativen bargen eine Tensorprodukt Struktur, in Form des Matrix-Kronecker Produkt.

Das heißt um dies effizient zu implementieren sollten wir uns Gedanken dadrüber machen wie wir diese Struktur ausnutzen können.

Im Exa-dg-teachlet wird eine Strategie vorgestellt ein Matrix-vektor Produkt mit Kronecker Matrizen also  $z = (\mathcal{B} \otimes \mathcal{A})y$  berechnet werden kann.

Sei  $\mathcal{A} \in \mathbb{R}^{m \times n}$  und  $\mathcal{B} \in \mathbb{R}^{p \times q}$ . Das Kronecker Produkt dieser Matrizen kann man schreiben als,

$$\mathcal{B} \times \mathcal{A} = \begin{pmatrix} b_{11}\mathcal{A} & \dots & b_{1q}\mathcal{A} \\ \vdots & \ddots & \vdots \\ b_{p1}\mathcal{A} & \dots & b_{pq}\mathcal{A} \end{pmatrix}$$

Nun wir sehen also die sich wiederholende Struktur von  $\mathcal{A}$ . Genau diese wollen wir uns nun zu nutze machen. Nehmen wir an  $y$  sei geordnet in der Indexierung.

$$y = (y_1, y_2, \dots, y_n, \dots, \dots, y_{(q-1)n+1}, y_{(q-1)n+2}, \dots, y_{qn})^T$$

Wir denken uns nun die Faktoren  $b_{ij}$  die mit  $\mathcal{A}$  multipliziert werden erstmal weg. Definiere  $y^{(1)} = (y_1, y_2, \dots, y_n)^T$ .

$$w^{(1)} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \dots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathcal{A}y^{(1)}$$

Auf ähnliche Weise können wir uns  $y^{(2)} = (y_{n+1}, \dots, y_{2n})^T$  definieren und dann

$$w^{(2)} = \mathcal{A}y^{(2)}$$

Wir führen dies so weiter und erhalten

$$w = ((w^{(1)})^T, \dots, (w^{(q)})^T) \in \mathbb{R}^{mq}$$

Nun müssen wir die Informationen der Matrix  $B$  noch mit reinbringen. Dazu berechnen wir wie im Teachlet auch vorgeschlagen  $z_i$  mit

$$z_i = \sum_{i=1}^q b_{1i} w_m^{(i)}$$

Mit Hilfe diesen Algorithmus haben wir die Komplexität von  $2m^4$  auf  $4m^3$  reduziert. Nun wollen wir dies erweitern auf  $z = (\mathcal{C} \otimes \mathcal{B} \otimes \mathcal{A})v$ .

$$z := \begin{pmatrix} c_{11}b_{11}A & \dots & c_{11}b_{1n}A & \dots & \dots & c_{1n}b_{11}A & \dots & c_{1n}b_{1n}A \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ c_{11}b_{n1}A & \dots & c_{11}b_{nn}A & \dots & \dots & c_{1n}b_{n1}A & \dots & c_{1n}b_{nn}A \\ c_{21}b_{n1}A & \dots & c_{21}b_{nn}A & \dots & \dots & c_{2n}b_{n1}A & \dots & c_{2n}b_{nn}A \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ c_{n1}b_{n1}A & \dots & c_{n1}b_{nn}A & \dots & \dots & c_{nn}b_{n1}A & \dots & c_{nn}b_{nn}A \end{pmatrix} * v \quad (4.1)$$

Wir sehen hier sich zwei wiederholende Strukturen die wir ausnutzen können um bei einem Matrix-Vektor Produkt operationen zu sparen.

$$z = \begin{pmatrix} c_{11}\textcolor{red}{b}_{11}\textcolor{red}{A} & \dots & c_{11}\textcolor{red}{b}_{1n}\textcolor{red}{A} & \dots & \dots & c_{1n}\textcolor{red}{b}_{11}\textcolor{red}{A} & \dots & c_{1n}\textcolor{red}{b}_{1n}\textcolor{red}{A} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ c_{11}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{11}\textcolor{red}{b}_{nn}\textcolor{red}{A} & \dots & \dots & c_{1n}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{1n}\textcolor{red}{b}_{nn}\textcolor{red}{A} \\ c_{21}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{21}\textcolor{red}{b}_{nn}\textcolor{red}{A} & \dots & \dots & c_{2n}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{2n}\textcolor{red}{b}_{nn}\textcolor{red}{A} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ c_{n1}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{n1}\textcolor{red}{b}_{nn}\textcolor{red}{A} & \dots & \dots & c_{nn}\textcolor{red}{b}_{n1}\textcolor{red}{A} & \dots & c_{nn}\textcolor{red}{b}_{nn}\textcolor{red}{A} \end{pmatrix} * v \quad (4.2)$$

Nun wollen wir uns dies zu nutze machen. Wir schauen uns erstmal die einzelnen Einträge von  $z$  an und bekommen. Vorher definieren wir unser  $v$  um zu einem Tensor  $\mathcal{V} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ . Der erste Index repräsentiert in welcher Spalteneintrag von  $C$  wir uns befinden, der zweite in welchem Spalteneintrag von  $B$  und der Dritte in welchem Spalteneintrag von  $A$ .

$$z_1 = \mathcal{V}(1, 1, 1)c_{11}\textcolor{red}{b}_{11}\textcolor{red}{a}_{11} + \dots + \mathcal{V}(1, 1, n)c_{11}\textcolor{red}{b}_{11}\textcolor{red}{a}_{1n} + \dots + \mathcal{V}(1, n, 1)c_{11}\textcolor{red}{b}_{1n}\textcolor{red}{a}_{11} \\ + \dots + \mathcal{V}(n, 1, 1)c_{1n}\textcolor{red}{b}_{11}\textcolor{red}{a}_{11} + \dots + \mathcal{V}(n, n, n)c_{1n}\textcolor{red}{b}_{1n}\textcolor{red}{a}_{1n} \quad (4.3)$$

Definiere  $w_1(i, j) := \mathcal{V}(i, j, 1)a_{11} + \dots + \mathcal{V}(i, j, n)a_{1n}$ . Dann erhalten wir:

$$z_1 = w_1(1, 1)c_{11}b_{11} + \dots + w_1(1, n)c_{11}b_{1n} + \dots + w_1(n, 1)c_{1n}b_{11} + \dots + w_1(n, n)c_{1n}b_{1n}$$

Damit haben wir uns die sich wiederholende Struktur von der Matrix  $\mathbf{A}$  zu nutze gemacht. Im nächsten Schritt machen wir uns die sich wiederholende Struktur von  $b_{ij}$  zu nutze. Definiere hierfür  $\mathfrak{w}_{1,k}(i) := w_k(i, 1)b_{11} + \dots + w_k(i, n)b_{1n}$ . Damit erhalten wir:

$$z_1 = \mathfrak{w}_{1,1}(1)c_{11} + \dots + \mathfrak{w}_{1,1}(n)c_{1n} \quad (4.4)$$

Wir wollen nun  $z$  genau so umformen wie wir das auch für  $v$  gemacht haben. Damit erhalten wir für allgemeines  $z_i$  folgende Formel:

$$\mathcal{Z}(i, j, k) = \mathfrak{w}_{j,k}(1)c_{i1} + \dots + \mathfrak{w}_{j,k}(n)c_{in} \quad (4.5)$$

Wobei  $j$  und  $k$  den Zeilen jeweils in den Matrizen  $B$  und  $C$  entsprechen.

Der komplette Algorithmus würde nun wie folgt aussehen:

```

for k=1 < n do
  for i= 1 < n do
    for j= 1 < n do
       $w_k(i, j) = \mathcal{V}(i, j, 1)a_{k1} + \dots + \mathcal{V}(i, j, n)a_{kn}$ 
    end for
  end for
end for
for k=1 < n do
  for i= 1 < n do
    for j= 1 < n do
       $\mathfrak{w}_{i,j}(k) := w_k(i, 1)b_{11} + \dots + w_k(i, n)b_{1n}$ 
    end for
  end for
end for
for k=1 < n do
  for i= 1 < n do
    for j= 1 < n do
       $\mathcal{Z}(i, j, k) = \mathfrak{w}_{j,k}(1)c_{i1} + \dots + \mathfrak{w}_{j,k}(n)c_{in}$ 
    end for
  end for
end for

```



Wenn wir annehmen, dass die Matrizen  $A, B, C \in \mathbb{R}^{n \times n}$ . Dann haben wir bei 4.2 eine Matrix-Vektor Multiplikation von einer Matrix der Größe  $n^3 \times n^3$ . Dementsprechend hätten wir  $n^6$  Multiplikationen und  $n^6$  Additionen. Die Komplexität des vorgeschlagenen Algorithmuses reduziert sich auf  $3n^3$  Multiplikationen und genau so viele Additionen. Ein enorme Reduktion, vor allem für großes  $n$ .

**Wie können wir uns nun diese Algorithmen zu nutze machen für die in Kapitel 3 besprochenen Strategien?**

### Summenfaktorisierung

Rekapituliere die Formeln für die Masse Matrix und für die Laplace Bilinearform  
**Masse Matrix**

$$M^{-1} * u = ((\mathcal{W}\mathcal{N}^T) \otimes (\mathcal{W}\mathcal{N}^T))^{-1} * u = ((\mathcal{W}\mathcal{N}^T)^{-1} \otimes (\mathcal{W}\mathcal{N}^T)^{-1}) * u$$

### Laplace Bilinearform

$$\mathcal{V}^{-1} * u = -((\mathcal{W}\mathcal{N}''^T)^{-1} \otimes (\mathcal{W}\mathcal{N}^T)^{-1} - (\mathcal{W}\mathcal{N}^T)^{-1} \otimes (\mathcal{W}\mathcal{N}''^T)^{-1}) * u$$

Um dies effizient zu berechnen nutzen wir einfach die Formeln aus dem Techart und erhalten eine Komplexität von: Matrizenmultiplikation ist kubisch also  $n^3$ . Das Kroneckerprodukt von 2 Matrizen ist wie wir bereits wissen  $4n^3$ . Nun müssen wir uns um die Invertierung und deren Komplexität kümmern. Nach einer normalen Singulärwertzerlegung bekommen wir  $O(\min(mn^2, m^2n))$ . Nun müssen wir wieder die Inverse daraus herleiten. Oder wir können den Gauß Algorithmus oder mit der Neumann Reihe arbeiten. Gauß Algorithmus gibt uns eine Komplexität von  $O(n^3)$ . Neumann Reihe macht keinen Sinn, da die Komplexität mindestens genau so hoch ist wie bei Gauß. Insgesamt haben wir eine Komplexität mit Gauß:  $O(n^3) + O(n^3) + O(4n^3) = O(6n^3) \in O(n^3)$ . Das heißt wir sind immer noch kubisch, was ein großer Vorteil ist. Die Singulärwertzerlegung bringt uns erst einen Vorteil, wenn wir eine trunkierte Zerlegung wählen. Die mögliche Ersparnis ist aber nicht so hoch. **BEGRÜNDUNG.**

**Singulärwertzerlegung höherer Ordnung** Nun haben wir uns mit Hilfe der Tucker Dekomposition eine Herleitung für die Pseudoinverse erarbeitet. Nun geht es um die effiziente Berechnung dieser Formel. Dazu wollen wir uns die Summenfaktorisierung zu nutze machen, wie sie auch in [Tea, 9-11] vorgeschlagen wird. Sei  $\mathcal{A} \in \mathbb{R}^{I_1 \times \dots \times I_n}$ . Die Formel für die Pseudoinverse lautet:

$$\mathcal{A}^\dagger = \mathcal{S}^\dagger \times_{n=1}^N U^{(n)T} \quad (4.6)$$

Wobei  $\mathcal{S} \in \mathbb{R}^{I_1 \times \dots \times I_n}$  und  $U^{(n)} \in \mathbb{R}^{J_n \times I_n}$ . Man kann 4.6 nach [TK09, 462] äquivalent umformen zu

$$\begin{aligned} \mathcal{A}_{(n)}^\dagger &= U^{(n)T} \mathcal{S}_{(n)}^\dagger (U^{(N)T} \otimes \dots \otimes U^{(n+1)T} \otimes U^{(n-1)T} \otimes \dots \otimes U^{(1)T})^T \\ \iff \mathcal{A}_{(n)}^\dagger &= U^{(n)T} \mathcal{S}_{(n)}^\dagger (U^{(N)} \otimes \dots \otimes U^{(n+1)} \otimes U^{(n-1)} \otimes \dots \otimes U^{(1)}) \end{aligned} \quad (4.7)$$

Nun betrachten wir uns das Matrix-Vektor Produkt und überlegen uns wie wir uns die Strukturen dort zu nutze machen.

$$\mathcal{A}_{(n)}^\dagger v = U^{(n)T} \mathcal{S}_{(n)}^\dagger (U^{(N)} \otimes \dots \otimes U^{(n+1)} \otimes U^{(n-1)} \otimes \dots \otimes U^{(1)})v \quad (4.8)$$

Wir schauen uns die Struktur mal für den Fall, dass  $\mathcal{A} \in \mathbb{R}^{I_1 \times \dots \times I_4}$ . Das heißt 4.8 reduziert sich auf:

$$\mathcal{A}_{(n)}^\dagger v = U^{(n)T} \mathcal{S}_{(n)}^\dagger (U^{(N_1)} \otimes U^{(N_2)} \otimes U^{(N_3)})v \quad (4.9)$$

mit  $N_i \neq n$ .

Wir erkennen auf der Rechten Seite können wir unsere Methodik, die wir entwickelt haben, um dieses doppelte Kronecker Produkt mit dem Vektor effizient zu berechnen. Wir können dies mit einer Komplexität von  $O(6n^3)$  machen. Nun geht es darum den Kerntensor zu invertieren. Wie wir wissen müssen wir dazu erstmal kleine Zahlen, die wir Rauschen nennen, auf 0 setzen und die Diagonalelemente invertieren. Dazu müssen wir, da der Kerntensor vollbesetzt ist, durch alle Elemente des Tensors durchiterieren! Das heißt wir haben eine Komplexität von  $O(n^4)$ , da unser Tensor Ordnung 4 hat. Nun folgt eine letzte Matrix-Matrix Multiplikation, aber von Matrizen der Größe  $n^2 \times n^2$ . Das heißt eine Komplexität von  $O(n^6)$  was uns die ganze Komplexität mit raketengeschwindigkeit zerbombt. Doch dank der Struktur des Kerntensors reduziert sich die Komplexität auf  $O(n^4)$ . Insgesamt haben wir eine Komplexität von  $O(6n^3) + 2O(n^4) = O(6n^3 + 2n^4)$ . Dies ist aber nicht annähernd so gut wie Option 1. Wie können wir unsere Effizienz weiter steigern? Wir können uns an die trunkierte HOSVD ranmachen oder eine andere Alternative finden. Nun anstatt den Kerntensor zu modifizieren, können wir bei der Matrix-Matrix Multiplikation einfach nur die Diagonalelemente für die Multiplikation in Erwägung ziehen. Damit sparen wir uns schon  $O(n^4)$  Operationen. Was wir tun könnten ist den Kerntensor in

das Kronecker Produkt reinmultiplizieren bevor wir dieses ausrechnen. Selbst wenn wir das hinkriegen würden wir es nicht hinbekommen die Komplexität von Option 1 zu übertreffen. Außerdem ist das Problem, dass wir nicht mit der Pseudoinverse eine Matrix-Vektor Multiplikation durchführen, sondern mit dem entfalteten Tensor, der die Pseudoinverse darstellt. Doch wie dieser Tensor entfaltet wird, ist für die korrekte Berechnung von enormer Wichtigkeit. D.h. selbst wenn wir es schaffen dies effizient zu berechnen, ist das Ergebnis das was wir wollen? Die Antwort ist: Es kommt drauf an. Worauf? Wie wir unser  $v$  aufsetzen. Unser  $v$  könnte so aufgebaut werden, dass dies in Übereinstimmung mit den Elementen des entfalteten Tensors übereinstimmt. Dadurch würden wir genau das erreichen was wir erreichen wollten. Doch wie gesagt, macht uns die Komplexität einen gewaltigen Schnitt durch die Rechnung.

Eine letzte Alternative zeigt wäre eben die trunkierte HOSVD. Wir hauen eine Dimension raus und die korrespondierenden Singulärwerte raus. Am Meisten macht es Sinn natürlich die kleinsten Singulärwerte rauszuhauen.

Wir erhalten:

$$\mathbf{A}_{(n)}^\dagger v = U^{(n)T} \mathbf{S}_{(n)}^\dagger (U^{(N_1)} \otimes U^{(N_2)}) v \quad (4.10)$$

Die Komplexität dies auszurechnen beträgt für das Kronecker-Matrix Vektor Produkt  $O(4n^3)$ . Nun folgt ein weiterer Trick. Wir nutzen folgende Bemerkung:

**Bemerkung 4.1.** Sei  $A$  eine Diagonalmatrix mit  $A \in \mathbb{R}^{n^3 \times n^3}$  und  $y \in \mathbb{R}^{n^3}$ . Dann gilt:

$$A(\mathcal{B} \otimes \mathcal{C})y = (\mathcal{B} \otimes \mathcal{C}) \begin{pmatrix} a_{11}y_1 \\ \vdots \\ a_{n^3 n^3}y_n \end{pmatrix}$$

Um diese Transformation zu berechnen brauchen wir  $O(n^3)$  Operationen. Insgesamt erhalten wir  $O(5n^3)$ . Nun müssen wir noch das letzte Matrix-Matrix Produkt berechnen, was uns wieder alles kaputt macht.

Wenn wir noch eine Dimension rausstreichen enden wir bei:

$$\mathbf{A}_{(n)}^\dagger v = U^{(n)T} \mathbf{S}_{(n)}^\dagger (U^{(N_1)}) v \quad (4.11)$$

## 5 Resultate

## Literatur

- [CG05] Hans-Görg Roos Christian Großmann. *Numerische Behandlung partieller Differentialgleichungen*. Teubner, 2005.
- [Joh08] Claes Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Dover Publications, 2008.
- [MK12] Katharina Kormann Martin Kronbichler. *A generic interface for parallel cell-based finite element operator application*. Elsevier, 2012.
- [Ran05] Prof. Dr. Rannacher. *Einführung in die Numerische Mathematik Vorlesungsskriptum*. 2005.
- [Tea] *Efficient evaluation of weak forms in discontinuous Galerkin methods*.
- [TK09] Brett Bader Tamara Kolda. *Tensor Decompositions and Applications*. SIAM, 2009.