Freie Universität Berlin

Fachbereich Mathematik und Informatik
Institut für Mathematik

Bachelorarbeit

Inverse Ungleichungen für die
Stromlinien-Diffusions-Finite-Elemente-Methoden

Name: Vincent Dallmer
Matrikelnummer: 5097352
Betreuer: Prof. Dr. Volker John
Zweitgutachter: Dr. Alfonso Caiazzo

Berlin, den 11. Januar 2022

Inhaltsverzeichnis

Einleitung

Die vorliegende Arbeit befasst sich mit Abschätzungen für inverse Ungleichungen der Form:

$$\|\Delta v\|_{0,T}^2 \leq \lambda_T \|\nabla v\|_{0,T}^2 \qquad \forall v \in \mathcal{P}, \tag{1}$$

die unter anderem für die Analysis von stabilisierten Finite-Elemente-Methoden herangezogen werden. Dabei ist $T \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, ein Gebiet und $\mathcal{P}$ ein endlichdimensionaler Funktionenraum. λ_T hängt von der Wahl von T ab und soll so bestimmt werden, dass die Abschätzung für ein $u \in \mathcal{P}$ scharf ist. Die Arbeit ist in fünf Abschnitte eingeteilt.

Im ersten Abschnitt werden die wichtigsten theoretischen Grundlagen präsentiert, um das Problem in einem mathematischen Anwendungs-Kontext einzuordnen. Am Beispiel der Konvektions-Diffusions-Gleichung stellen wir zunächst die Galerkin-Methode vor - ein klassisches Verfahren zur Approximation von partiellen Differentialgleichungen. Anschließend wird die Stromlinien-Diffusions-Finite-Elemente-Methode (SDFEM) vorgestellt. Diese ist ein stabileres Verfahren und baut auf der Galerkin-Methode auf. Es wird herausgearbeitet, wo die Konstante λ_T zur Analysis der Methode verwendet wird.

Im zweiten Abschnitt wird das Problem mathematisch charakterisiert. Es wird gezeigt, dass (1) äquivalent zu einem verallgemeinerten Eigenwertproblem ist. Damit ist einerseits die Voraussetzung für die Existenz einer optimalen Abschätzung gegeben und andererseits liefert die äquivalente Formulierung ein Problem, für das sowohl explizite, als auch numerische Lösungsmethoden bekannt sind. Weiterhin wird untersucht, wie sich die Abschätzung unter affinen Transformationen verhält.

Im dritten Abschnitt werden die Konstanten λ_T für einige Sonderfälle explizit berechnet. Sofern es eine Menge von Gebieten gibt, sodass für jedes Element T ein eindeutig bestimmtes kleinstes λ_T existiert, können wir λ_T als Funktion auf dieser Menge interpretieren. In [HH92] werden diese Funktionen für bestimmte $\mathcal{P}$ und T explizit angegeben. Wir werden diese Ergebnisse nachvollziehen - teilweise mit alternativen Ansätzen - und wir werden die sich daraus ergebenden Fragestellungen untersuchen.

Seien r_T der Radius der größten in T enthaltenen Kugel und d_T der Durchmesser von T. Der Versuch, die Funktionen λ_T in die Konvergenztheorie für Finite-Elemente zu integrieren, motiviert die Definition einer Element-, oder Gitterzellengröße h_T, sodass Konstanten $C_0, C_1 \in \mathbb{R}$ existieren, mit:

$$C_0 \; r_T \leq h_T \leq C_1 \; d_T \tag{2a}$$

$$\text{und } \lambda_T = \frac{C_{inv}}{h_T^2}. \tag{2b}$$

Diese Definition führt auf die Ungleichung:

$$h_T^2 \|\Delta v\|_{0,T}^2 \leq C_{inv} \|\nabla v\|_{0,T}^2 \qquad \forall v \in \mathcal{P}. \tag{3}$$

Ziel ist es, die Elementgröße h_T so zu definieren, dass C_{inv} für eine Familie von Gebieten konstant bleibt. Insbesondere darf C_{inv} nicht von d_T oder r_T abhängen. In dem Artikel von Harrari und Hughes wird nicht explizit auf die Bedingung (2a) eingegangen. Wir werden in Abschnitt 1.4.2 kurz auf ihre Notwendigkeit eingehen und im dritten Teil zeigen, dass die von Harrari und Hughes definierten Größen (2a) erfüllen.

Die letzten beiden Abschnitte behandeln einige Fragestellungen, die nicht in [HH92] behandelt worden sind. Wir untersuchen, ob die dort gefundenen Funktionen auch für allgemeinere Elementgebiete scharfe Abschätzungen liefern und vergleichen (1) mit einer anderen inversen Ungleichung aus der Finite-Elemente-Analyse.

Es wird vorausgesetzt, dass der Leser mit den folgenden Konzepten vertraut ist:

- Begriffe aus der Vektoranalysis, wie Skalar- und Vektorfeld, Gradient, Divergenz und der Laplace-Operator, inklusive dem Gaußschen-Integralsatz.
- Das Lebesgue-Integral, insbesondere die L^2-Räume auf Teilmengen des $\mathbb{R}^n$, zusammen mit den wichtigsten Ungleichungen (z.B. Cauchy-Schwarz Ungleichung, Youngsche Ungleichung, etc.).
- Grundlegende Konzepte der Funktionalanalys. Dazu gehören z.B. Definitionen der Norm, des Skalarprodukts und der Orthogonalität, sowie die Charakterisierung der Bestapproximation in Hilberträumen, der Rieszsche-Darstellungssatz und der Projektionssatz.

Zur Notation

- In dieser Arbeit werden verschiedene Normen und Skalarprodukte verwendet. Wir legen fest, dass mit $(\cdot,\cdot)$ das L^2 Skalarprodukt gemeint ist und mit $\|\cdot\|_0$ die dadurch induzierte Norm. Wenn eine andere Norm (bzw. Skalarprodukt) verwendet wird, wird das durch ein tiefgestelltes Zeichen rechts unten verdeutlicht. Wenn nicht unmittelbar aus dem Kontext klar ist, auf welches Gebiet sich die Norm bezieht, wird es durch ein tiefgestelltes Zeichen rechts davon notiert. Wir schreiben zum Beispiel $\|\cdot\|_{0,T}$ für die L^2-Norm auf der Menge T.

  Wenn kein tiefgestelltes Zeichen neben der Norm steht, handelt es sich um die euklidische Norm: $\|\underline{v}\| = \sqrt{\sum_{j=1}^n v_j^2}$ für $\underline{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$.

- Für eine Menge $T \subset \mathbb{R}^d$ schreiben wir $d_T = \sup_{\boldsymbol{x},\boldsymbol{y}\in T} d(\boldsymbol{x},\boldsymbol{y})$, wobei $d(\cdot,\cdot)$ der Distanz in der euklidischen Norm entspricht. Den Radius der größten in T enthaltenen Kugel nennen wir r_T.

- Vektoren des $\mathbb{R}^d$ werden fett gedruckt. Wir schreiben $\mathbb{R}^d \ni \boldsymbol{x} = (x_1, \ldots, x_d)^T$. Funktionen mit Wertevorrat im $\mathbb{R}^d$ werden ebenfalls fett gedruckt.

  - Der $\mathbb{R}^d$ mit $d \in \{1, 2, 3\}$ bezeichnet in dieser Arbeit immer die Obermenge des Definitionsbereichs der Funktionen in $\mathcal{P}$. Wenn wir Vektoren des $\mathbb{R}^n$ in einem anderen Kontext betrachten, kennzeichnen wir sie mit einem Strich unter dem Symbol: $\underline{v} \in \mathbb{R}^n$.

- Sei $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$, $|\boldsymbol{\alpha}| := \sum_{j=1}^d \alpha_j$. Für die gemischte partielle Ableitung einer Funktion v auf $\mathbb{R}^d$ der Ordnung $|\boldsymbol{\alpha}|$ nach $\boldsymbol{\alpha}$: $\dfrac{\partial^{|\boldsymbol{\alpha}|} v(\boldsymbol{x})}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$, schreiben wir $D^{\boldsymbol{\alpha}} v$.

  Für die partielle Ableitungen der Ordnung $k \in \mathbb{N}$ bezüglich x_j schreiben wir $\dfrac{\partial^k v(\boldsymbol{x})}{\partial x_j^k}$.

1 Anwendung: Konstruktion der SDFEM

Zunächst wird die stationäre, skalare Konvektions-Diffusions-Gleichung vorgestellt. Die Approximation der Lösung dieser Gleichung ist Gegenstand der hier betrachteten numerischen Methode.

Da die Gleichung im Allgemeinen keine Lösung im klassischen Sinne besitzt, werden schwache Lösungen vorgestellt. Unter bestimmten Voraussetzungen kann für diese Lösungen Existenz und Eindeutigkeit in geeigneten Funktionenräumen garantiert werden. Dafür werden einige Ergebnisse aus der Funktionalanalysis benötigt, die hier, ohne Beweis, mit Verweis auf die Literatur bereitgestellt werden.

Um die schwache Lösung zu approximieren, werden Finite-Elemente-Methoden angewandt. Wir orientieren uns hauptsächlich an der Darstellung in [BS08].

Der Abschnitt endet mit der Vorstellung der SDFEM.

1.1 Die Konvektions-Diffusions-Gleichung

Viele Probleme aus der Physik, wie zum Beispiel der Transport eines Stoffes in einer fließenden Flüssigkeit, lassen sich mittels Konvektions-Diffusions-Gleichungen beschreiben. In [Hir06, S. 35] wird diese Gleichung als Rückgrat der mathematischen Modellierung von Phänomenen der Fluiddynamik bezeichnet.

Auch ökonomische Modelle, wie die Black-Scholes-Gleichung, sind Konvektions-Diffusions-Gleichungen (vgl. [t H17, Kapitel 2]).

In dieser Arbeit befassen wir uns mit einer Modellgleichung, bei der ausschließlich Ableitungen nach Ortsvariablen und keine nach der Zeit auftreten.

1.1.1 Der stationäre skalare Fall

Sei $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$ ein beschränktes Gebiet, $\varepsilon \in \mathbb{R}$, $\boldsymbol{b} : \Omega \to \mathbb{R}^d$ und $c, f : \Omega \to \mathbb{R}$. Den Rand von Ω bezeichnen wir mit Γ. Eine Konvektions-Diffusions-Gleichung mit homogenen Randbedingungen ist gegeben durch:

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = f \tag{4a}$$

$$u|_\Gamma = 0. \tag{4b}$$

Aus der Linearität von ∇, Δ und $(\boldsymbol{b}, \cdot)$ folgt, dass die linke Seite von (4a) linear bzgl. u ist. Wenn u in einem Vektorraum V existiert und f in W, können wir einen linearen Operator

$$\begin{aligned} A_{st} : V &\to W \\ v &\mapsto -\varepsilon \Delta v + \mathbf{b} \cdot \nabla v + cv \end{aligned} \tag{5}$$

definieren, sodass $A_{st} u = f$. (4a) heißt lineare Differentialgleichung 2. Ordnung. Dabei entspricht die Ordnung dem Grad der höchsten Ableitung.

1.2 Schwache Lösungen

Im Allgemeinen wird (4) keine Lösung im klassischen Sinne besitzen. Wenn die Voraussetzungen an die Lösung jedoch abgeschwächt werden, existieren eindeutige Lösungen.

Dazu betrachten wir zunächst die Eigenschaften einer hypothetischen Lösung im klassischen Sinne.

Eine Lösung u_{st} von (4) nennen wir eine starke Lösung. Sei $f \in L^1_{loc}(\Omega)$ und $v \in C_0^\infty(\Omega)$ eine unendlich oft differenzierbare Funktion mit kompaktem Träger auf Ω.

Aus dem dem Gaußschen-Integralsatz folgt:

$$\begin{aligned}(\Delta u, v) &= \int_\Gamma (\nabla u_{st}(\boldsymbol{s}) v(\boldsymbol{s})) \cdot n(\boldsymbol{s}) d\boldsymbol{s} - \int_\Omega \nabla u_{st}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) d\boldsymbol{x} \\ &= -\int_\Omega \nabla u_{st}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) d\boldsymbol{x}.\end{aligned}$$

Mit $A_{st} u_{st} = f$ gilt nun auch:

$$\begin{aligned}(f, v) &= (A_{st} u_{st}, v) \\ &= -\varepsilon(\Delta u_{st}, v) + (\boldsymbol{b} \cdot \nabla u_{st}, v) + (c u_{st}, v) \\ &= \varepsilon(\nabla u_{st}, \nabla v) + (\boldsymbol{b} \cdot \nabla u_{st}, v) + (c u_{st}, v).\end{aligned}$$

Diese Aussage lässt sich nicht umkehren. Eine Lösung von (4) muss zweimal differenzierbar sein - eine Funktion u, die die Gleichung aus der letzten Zeile erfüllt, nur einmal. In diesem Fall existiert Δu nicht. Es kann jedoch eine verallgemeinerte Form der Ableitungen existieren:

Definition 1.1 (Schwache Ableitung)**.** *Sei $v \in L^1_{loc}(\Omega)$, $\boldsymbol{\alpha} \in \mathbb{R}^d$. Die Funktion $g \in L^1_{loc}(\Omega)$ heißt schwache Ableitung von v nach $\boldsymbol{\alpha}$, wenn für alle unendlich oft differenzierbaren Funktionen mit kompaktem Träger auf Ω gilt, dass:*

$$(v, w) = (-1)^{|\boldsymbol{\alpha}|} (g, D^{\boldsymbol{\alpha}} w) \qquad \forall w \in C_0^\infty(\Omega).$$

Wenn die Ableitung $D^{\boldsymbol{\alpha}} v$ existiert, ist sie identisch mit der schwachen Ableitung. Daher schreiben wir die schwache Ableitung auch einfach als $D^{\boldsymbol{\alpha}} v$. Im Folgenden werden auch die in den Operatoren ∇ und Δ vorkommenden partiellen Ableitungen als schwache Ableitungen interpretiert.

Das führt auf die folgende Definition:

Definition 1.2 (Schwache Lösung der Konvektions-Diffusions-Gleichung)**.** *Seien $u, f \in L^1_{loc}(\Omega)$. Wenn für u die schwachen Ableitungen erster Ordnung in allen Variablen $x_1, \ldots, x_d$ existieren und:*

$$\begin{aligned}\varepsilon(\nabla u, \nabla v) + (\boldsymbol{b} \cdot \nabla u, v) + (cu, v) &= (f, v) \\ u|_\Gamma &\equiv 0,\end{aligned} \tag{6}$$

für alle $v \in C_0^\infty(\Omega)$, nennen wir u eine schwache Lösung von (4)*.*

Wir gehen im Folgenden davon aus, dass die schwache Lösung in einem Vektorraum V enthalten ist, sodass $V \supset C_0^\infty(\Omega)$ gilt.

Wir suchen also nach $u \in V$, sodass für:

$$\begin{aligned} a : V \times V &\to \mathbb{R} \\ u, v &\mapsto \varepsilon(\nabla u, \nabla v) + (\boldsymbol{b} \cdot \nabla u, v) + (cu, v) \\ \text{und } F : V &\to \mathbb{R} \\ v &\mapsto (f, v) \end{aligned}$$

gilt:

$$a(u, v) = F(v) \qquad \forall v \in V.$$

1.2.1 Existenz und Eindeutigkeit von schwachen Lösungen

Wenn $a(\cdot, \cdot)$ ein Skalarprodukt und $(V, a(\cdot, \cdot))$ vollständig ist, folgt aus dem Rieszschen Darstellungssatz, dass (6) eine eindeutige Lösung hat.

Für die Formulierung von schwachen Lösungen partieller Differentialgleichungen müssen die Voraussetzungen des Satzes abgeschwächt werden, sodass sie für allgemeinere Bilinearformen gelten. Insbesondere ist die oben definierte Bilinearform $a(\cdot, \cdot)$ nicht symmetrisch, wenn $\boldsymbol{b} \not\equiv 0$.

Für nicht-symmetrische Bilinearformen liefert der Satz von Lax Milgram die Existenz und Eindeutigkeit einer Lösung:

Der Satz von Lax Milgram ([BS08, (2.7.7)])**.** *Sei V ein Hilbertraum, $a(\cdot, \cdot) : V \times V \to$ bilinear und $F \in V'$. Es gebe Konstanten C_k und C_s, sodass für alle $v, w \in V$ gilt:*

1. $|a(v, w)| \leq C_s \|v\|_V \|w\|_V$ *(Stetigkeit),*
2. $a(v, v) \geq C_k \|v\|_V^2$ *(Koerzivität),*

dann existiert ein eindeutiges $u \in V$, sodass

$$a(u, v) = F(v) \qquad \forall v \in V.$$

1.2.2 Der geeignete Lösungsraum

Die Suche nach Funktionenräumen, die die Voraussetzuungen des Satzes von Lax-Milgram erfüllen führt auf eine Klasse von Hilberträumen, bestehend aus Funktionen, die die Differenzierbarkeitseigenschaften von schwachen Lösungen besitzen - die Sobolev-Räume.

Für Funktionen aus $C_0^\infty \subset L^2$ ist ein Skalarprodukt, dass die schwachen Ableitungen bis zur Ordnung $k \in \mathbb{N}$ involviert durch

$$(u, v)_T := \sum_{|\boldsymbol{\alpha}| \leq k} (D^{\boldsymbol{\alpha}} u, D^{\boldsymbol{\alpha}} v) \tag{7}$$

gegeben. Symmetrie und Bilinearität folgen aus den Eigenschaften des L^2-Skalarprodukts und die postive Definitheit folgt, da $\|\cdot\|_0$ eine Norm in L^2 ist:

$$(u, u)_k = \|u\|_0^2 + \sum_{1 \leq |\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} u\|_0^2 \geq \|u\|_0^2.$$

Folglich induziert $(\cdot, \cdot)_k$ eine Norm, die wir $\|\cdot\|_k$ nennen. Gesucht sind Funktionenräume, deren Funktionen die Randbedingung (4b) erfüllen. Die Funktionen

aus $C_0^\infty(\Omega)$ erfüllen die Randbedingung, sie sind mit $(\cdot,\cdot)_k$ aber kein Hilbertraum. Um zu garantieren, dass wir einen Hilbertraum erhalten, betrachten wir die Vervollständigung von $C_0^\infty(\Omega)$ bzgl. $\|\cdot\|_k$.

Definition 1.3. *Sei $k \in \mathbb{N}$, dann ist:*

$$H_0^k(\Omega) := \left\{ v \in L^2(\Omega) : \ \exists \{v_j\}_{j\in\mathbb{N}} \subset C_0^\infty(\Omega), \textit{ sodass } \lim_{j\to\infty} \|v - v_j\|_k = 0 \right\}.$$

Für $k, l \in \mathbb{N}$ und $k \leq l$ gilt:

$$H_0^k(\Omega) \supset H_0^l(\Omega).$$

In [Alt12, A6.10] wird für ein Lipschitzgebiet Ω und $k = 1$ gezeigt, dass

$$v|_\Gamma \equiv 0 \qquad\qquad \forall v \in H_0^1(\Omega)$$

gilt. Die Randbedingungen werden also automatisch durch die Wahl des Funktionenraums erfüllt. Es ist üblich, die folgende Seminorm auf $H_0^k(\Omega)$ zu definieren:

$$|v|_k := \left(\sum_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}} v\|_0^2 \right)^{1/2}. \tag{8}$$

Damit können wir den Term $\|\nabla \cdot\|_0$ als $|\cdot|_1$ schreiben. Es gilt: $\|\cdot\|_k^2 = \|\cdot\|_0^2 + \sum_{j=1}^{k} |\cdot|_j^2$.

Definition 1.4. *Der Dualraum $\left(H_0^1(\Omega)\right)'$, der stetigen, linearen Funktionale auf $H_0^1(\Omega)$ wird mit $H^{-1}(\Omega)$ bezeichnet.*

1.2.3 Beschränktheit und Koerzivität der Bilinearform

Für den ersten Term der in (6) definierten Bilinearform $a(\cdot,\cdot)$ folgt aus der Cauchy-Schwarz Ungleichung und $\|\cdot\|_1^2 = \|\cdot\|_0^2 + |\cdot|_1^2$:

$$|\varepsilon(\nabla v, \nabla w)| \leq \varepsilon \|\nabla v\|_0 \|\nabla w\|_0 = \varepsilon |v|_1 |w|_1 \leq \varepsilon \|v\|_1 \|w\|_1.$$

Falls $\boldsymbol{b}$ und c beschränkt sind, folgt für den zweiten und dritten Term:

$$\begin{aligned} |(\boldsymbol{b}\cdot\nabla v, w)| &\leq \|\boldsymbol{b}\|_{L^\infty(\Omega)} |v|_1 \|w\|_0 \leq \|\boldsymbol{b}\|_{L^\infty(\Omega)} \|v\|_1 \|w\|_1 \\ |(cv, w)| &\leq \|c\|_{L^\infty(\Omega)} \|v\|_0 \|w\|_0 \leq \|c\|_{L^\infty(\Omega)} \|v\|_1 \|w\|_1. \end{aligned}$$

Insgesamt gilt mit der Dreiecksungleichung:

$$|a(v,w)| \leq (\varepsilon + \|\boldsymbol{b}\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)}) \|v\|_1 \|w\|_1.$$

Folglich ist $a(\cdot,\cdot)$ in der $\|\cdot\|_1$-Norm beschränkt und somit stetig.

Um die Koerzivität von $a(\cdot,\cdot)$ zu zeigen wird die Poincaré-Ungleichung benötigt:

Die Poincaré Ungleichung ([BS08, (5.3.5)])**.** *Sei Ω eine Vereinigung endlich vieler Sterngebiete, deren Zentrum eine Kugel enthält, dann existiert eine positive Konstante $C_P \leq \infty$, sodass für alle $v \in H_0^1(\Omega)$ gilt:*

$$\|v\|_0 \leq C_P |v|_1. \tag{9}$$

Damit folgt sofort die Koerzivität des Diffusionsterms, da:

$$\|v\|_1 = \left(\|v\|_0^2 + |v|_1^2\right)^{1/2} \leq \left(\frac{C_p + 1}{\varepsilon}\right)^{1/2} \sqrt{\varepsilon}|v|_1$$

und folglich:

$$\varepsilon(\nabla v, \nabla v) \geq \frac{\varepsilon}{C_p + 1}\|v\|_1^2.$$

Für den Konvektionsterm gilt:

$$\begin{aligned}
(\boldsymbol{b} \cdot \nabla v, v) &= \int_\Omega \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x} \\
&= \frac{1}{2} \int_\Omega \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla v^2(\boldsymbol{x}) d\boldsymbol{x} \\
&= \frac{1}{2} \int_\Gamma (\boldsymbol{b} v^2)(\boldsymbol{s}) \cdot n(\boldsymbol{s}) d\boldsymbol{s} - \frac{1}{2} \int_\Omega \nabla \cdot \boldsymbol{b}(\boldsymbol{x}) v^2(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\frac{1}{2} \int_\Omega \nabla \cdot \boldsymbol{b}(\boldsymbol{x}) v^2(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

Das Integral über den Rand verschwindet aufgrund der Randbedingung (4b). Wenn $c(\boldsymbol{x}) - \frac{1}{2}\nabla \cdot \boldsymbol{b}(\boldsymbol{x}) \geq 0$ für alle $\boldsymbol{x} \in \Omega$, gilt insgesamt:

$$\begin{aligned}
a(v, v) &\geq \frac{\varepsilon}{C_p + 1}\|v\|_1^2 + \left(\left(-\frac{1}{2}\nabla \cdot \boldsymbol{b} + c\right) v, v\right) \\
&\geq \frac{\varepsilon}{C_p + 1}\|v\|_1^2
\end{aligned}$$

für alle $v \in H_0^1(\Omega)$. Damit ist die Existenz und Eindeutigkeit einer Lösung von (6) gezeigt. Für die Analyse von numerischen Lösungen wird häufig noch eine andere Norm auf $H_0^1(\Omega)$ definiert.

1.2.4 Die Energienorm

Die Poincaré-Ungleichung impliziert, dass die symmetrische Bilinearform:

$$\begin{aligned}
(\cdot, \cdot)_E : H_0^1(\Omega) \times H_0^1(\Omega) &\to \mathbb{R} \\
v, w &\mapsto (\nabla v, \nabla w)
\end{aligned} \tag{10}$$

positiv definit ist und folglich ein Skalarprodukt auf $H_0^1(\Omega)$ definiert. Die induzierte Norm ist äquivalent zur H^1-Norm:

$$\frac{1}{(1 + C_P)^{1/2}}\|v\|_1 \leq \|v\|_E \leq \|v\|_1.$$

Diese Norm heißt Energienorm und sie erhält wegen der Äquivalenz auch die Existenz und Eindeutigkeit der Lösung von (6).

1.3 Approximation der schwachen Lösung

Aus der Numerik ist das folgende Vorgehen für die Approximation von stetigen Funktionalen auf einem Hilbertraum in endlich-dimensionalen Unterräumen bekannt:

Sei $V_h \subset V$ ein n-dimensionaler Unterraum des Hilbertraums $(V, (\cdot,\cdot)_V)$, mit Basis $\{\varphi_i\}_{i=1}^n$ und $(u,\cdot)_V := F \in V'$ für ein $u \in V$.

Für die Lösung $(\beta_1, \ldots, \beta_n)^T$ des linearen Gleichungssystems:

$$\sum_{i=1}^{n} \beta_i (\varphi_j, \varphi_i)_V = F(\varphi_j), \qquad j = 1, ..., n,$$

ist $u_h := \sum_{i=1}^{n} \beta_i \varphi_i$ die beste Approximation von u in V_h bezüglich der durch $(\cdot,\cdot)_V$ induzierten Norm.

Bemerkung 1.1. *Insbesondere gilt, falls in* (4a) *die Funktionen* $\boldsymbol{b}$ *und* c *gleich Null sind, dass wir* $(V, (\cdot,\cdot)_V) := (H_0^1(\Omega), (\cdot,\cdot)_E)$ *wählen können und so die Bestapproximation in* V_h *bezüglich* $\|\cdot\|_E$ *für die schwache Lösung berechnen. Diese Methode heißt Ritz-Methode und wird in [BS08, 2.5] besprochen.*

1.3.1 Die Galerkin-Methode

Im Allgemeinen wird das Problem nicht symmetrisch sein. Die Galerkin-Methode ersetzt nun das Skalarprodukt in der Ritz-Methode durch die Bilinearform $a(\cdot,\cdot)$. Damit ergibt sich das folgende Problem:

Seien $V_h \subset V := (H_0^1(\Omega), (\cdot,\cdot)_E)$, $N = dim(V_h)$, $F \in H^{-1}(\Omega)$ und $\{\varphi_i\}_{i=1}^N$ eine Basis von V_h. Sei weiterhin u die schwache Lösung der Konvektions-Diffusions-Gleichung (4). Die Lösung der Galerkin-Methode ist eine Funktion $u_h \in V_h$, sodass

$$a(u_h, v) = F(v) \qquad \forall v \in V_h. \tag{11}$$

Die Berechnung der Lösung erfolgt genauso wie bei der Ritz-Methode und die Existenz der Lösung folgt aus dem Satz von Lax-Milgram, wie schon die Existenz der schwachen Lösung.

Erwartungsgemäß wird die damit berechnete Näherungslösung nicht die beste Approximation an u liefern. Es kann jedoch gezeigt werden, dass die Güte der Approximation nur um einen konstanten Faktor abweicht.

Der Satz von Céa ([BS08, (2.8.1)])**.** *Seien* C_s, C_k *definiert wie oben,* u_h *die Lösung von* (11) *und* u *die Lösung von* (6)*, dann gilt:*

$$\|u - u_h\|_V \leq \frac{C_s}{C_k} \|u - v\|_V \qquad \forall v \in V_h.$$

Die Galerkin-Methode ist **konsistent**: Wenn die schwache Lösung u in V_h ist, ist sie die Lösung der Galerkin-Approximation: u_h. Das folgt aus der Konsistenz der Formulierung der schwachen Lösung und impliziert die sogenannte
Galerkin-Orthogonalität:

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h.$$

Nun müssen noch die geeigneten endlich-dimensionalen Vektorräume definiert werden.

1.3.2 Finite Elemente

Zur Approximation von Funktionen auf $\Omega \subset \mathbb{R}$ werden häufig stückweise polynomiale Funktionen verwendet. Die Finite-Elemente-Methode verallgemeinert diese Idee für Funktionen auf $\mathbb{R}^d$. Wir beschränken uns hier der Einfachheit halber auf $\mathbb{R}^2$.

Definition 1.5 (Finites Element [BS08, (3.1.1)]). *Sei $T \subset \mathbb{R}^d$ eine abgeschlossene und beschränkte Menge mit stückweise glattem Rand, $\mathcal{P}$ ein endlich-dimensionaler Funktionenraum auf T und $\mathcal{N} = \{N_1, \dots, N_n\}$ eine Basis von $\mathcal{P}'$, dann heißt $(T, \mathcal{P}, \mathcal{N})$ ein finites Element. Dabei wird hier angenommen, dass $\mathcal{N} \subset H^{-1}(\Omega)$.*

Wichtige Spezialfälle sind Polygone im $\mathbb{R}^2$, insbesondere Dreiecke und Rechtecke. Für dreieckige Elemente nehmen wir an, dass $\mathcal{P} = P_k(\Omega)$ polynomiale Funktionen vom Grad $k \in \mathbb{N}$ sind. Für rechteckige Elemente ist der Funktionenraum:

$$Q_k(\Omega) = \left\{ \sum_{j=1}^{k} c_j p_j(x_1) q_j(x_2) : \ p_j, q_j \in P_j(\mathbb{R}) \right\}$$

gebräuchlich.

Wir nehmen der Einfachheit halber an, dass Ω sich in solche einfache Formen zerlegen lässt. Genauer sagen wir, dass eine Menge von Elementgebieten $\mathcal{T} = \{T_j |\ j = 1, \dots, m\}$ eine **Zerlegung** von Ω ist, wenn:

1. $\mathring{T}_i \cap \mathring{T}_j = \emptyset$ für $i \neq j$ und
2. $\overline{\Omega} = \bigcup_{j=1}^{m} T_j$.

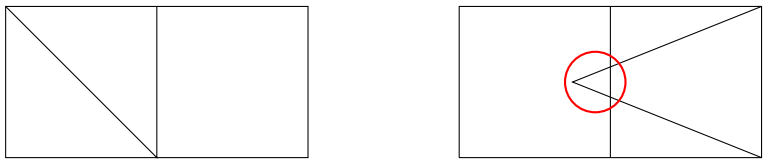

Abbildung 1: Erlaubte und verbotene Zerlegungen

1.3.3 Lagrange Elemente

Bei dieser Klasse von Elementen besteht $\mathcal{N}$ aus Auswertungsfunktionalen an paarweise verschiedenen Knotenpunkten in T. Es müssen also $n := \dim(\mathcal{P})$ Punkte $\{\boldsymbol{z}^1, \dots, \boldsymbol{z}^n\}$ gefunden werden, sodass für $i = 1, \dots, n$ die Funktionen

$$\begin{aligned} N_i : \mathcal{P} &\to \mathbb{R} \\ v &\mapsto v(\boldsymbol{z}^i) \end{aligned}$$

eine Basis von $\mathcal{P}$ bilden. Eine solche Wahl von Punkten ist für $P_k(T)$ auf Dreiecken und für $Q_k(K)$ für alle $k \in \mathbb{N}$ möglich (vgl [BS08, (3.2.7)]). Dabei ist für $k = 0$ ein beliebiger Punkt in $\mathring{T}$ Knotenpunkt. Für $k > 0$ sind die Ecken Knotenpunkte, sowie jeweils $k - 1$ weitere Punkte auf jeder Kante. Die restlichen Punkte befinden sich im Inneren von T und werden induktiv so gewählt, dass das sie die Knotenpunkte für ein in T liegendes Element mit $\tilde{\mathcal{P}} = \begin{cases} P_{k-3}(\tilde{T}) & \text{für Dreicke} \\ Q_{k-2}(\tilde{T}) & \text{für Rechtecke} \end{cases}$ bilden.

Nun fordern wir noch, dass für alle Elemente $(T_j, \mathcal{P}_j, \mathcal{N}_j)$:

1. die T_j drei- bzw. rechteckig sind, mit $\mathcal{P}_j = P_k(T_j)$ bzw. $Q_k(T_j)$,
2. die Schnittmenge zweier Elementgebiete entweder eine gemeinsame Ecke, oder eine gemeinsame Kante, also die Verbindungsgerade zwischen zwei gemeinsamen Ecken ist und

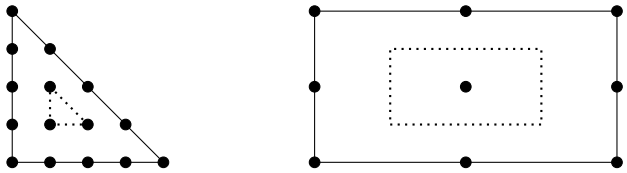

Abbildung 2: Dreieck mit $\mathcal{P} = P_4(T)$ und Rechteck mit $\mathcal{P} = Q_2(T)$

3. Stützstellen: $0 = \xi_0 < \xi_1 < \cdots < \xi_k = 1$ existieren, die symmetrisch bezüglich $\xi = \frac{1}{2}$ sind, sodass für jede Kante $\overline{\boldsymbol{z}^1, \boldsymbol{z}^2}$ die $k+1$ Knotenpunkte auf $\overline{\boldsymbol{z}^1, \boldsymbol{z}^2}$ durch $\boldsymbol{z}^i = \boldsymbol{z}^1 + \xi_i(\boldsymbol{z}^2 - \boldsymbol{z}^1)$ gegeben sind.

Wenn alle Elemente dreieckig sind, sagen wir zur 2. Voraussetzung auch, dass die Zerlegung eine **Triangulierung** ist. Die 3. Voraussetzung garantiert uns, dass die Knotenpunkte der Schnittmenge von zwei benachbarten Elementen zusammenfallen. Zwei benachbarte Knotenpunkte werden so zu einem Knotenpunkt der Basis von V_h'.

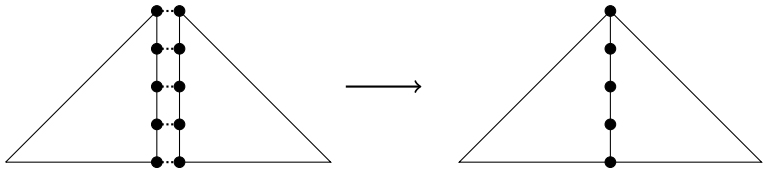

Abbildung 3: Zusammensetzung zweier Elemente mit zusammenfallenden Knotenpunkten

Bemerkung 1.2. *Das setzt voraus, dass zwei benachbarte Elemente in den zusammenfallenden Punkten dieselben Werte haben.*

Sofern V_h nicht aus stückweise konstanten Funktionen besteht, impliziert diese Voraussetzung, dass die Funktionen in V_h stetig sind. Denn die Einschränkung einer Funktion aus $P_k(T)$, oder $Q_k(T)$ auf eine Kante stellt ein Polynom vom Grad k in einer Variablen dar, das nach dem Fundamentalsatz der Algebra durch die $k+1$ vorgegebenen Werte an den Knotenpunkten eindeutig bestimmt ist.

Wenn die Knotenpunkte auf dem Rand von Ω gleich Null gesetzt werden, ist

$$V_h \subset H_0^1(\Omega)$$

(vgl. [BS08, (3.3.17)]). In diesem Fall sprechen wir von einem **konformen** Finite-Elemente Raum. Wir können noch eine Art Standardbasis definieren:

Definition 1.6 ([BS08, (3.1.2)])**.** *Die zu $\mathcal{N}$ duale Basis heißt **nodale Basis** von $\mathcal{P}$. Das sind die Basisvektoren $\{\phi_1, \ldots, \phi_n\} \subset \mathcal{P}$, für die gilt:*

$$N_i(\phi_j) = \delta_{i,j}.$$

Beispiel: Das lineare Lagrange-Dreieck. *Seien*

$$\begin{aligned}\boldsymbol{z}^1 &= (0,0)^T \\ \boldsymbol{z}^2 &= (1,0)^T \\ \boldsymbol{z}^3 &= (0,1)^T,\end{aligned}$$

dann ist die nodale Basis zu $\mathcal{N} = \{N_i(v) = v(\boldsymbol{z}^1)\}$*:*

$$\begin{aligned}\phi_1(\boldsymbol{x}) &= 1 - x_1 - x_2\\ \phi_2(\boldsymbol{x}) &= x_1\\ \phi_3(\boldsymbol{x}) &= x_2.\end{aligned}$$

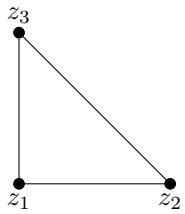

Abbildung 4: Das lineare Lagrange-Dreieck

Definition 1.7. *Sei* $\mathcal{N}_g$ *eine Basis von* V_h'*, dann ist die zu* $\mathcal{N}_g$ *duale Basis die* ***globale Basis*** *von* V_h*.*

Wenn die nodalen Basen der Elemente bekannt sind, ist die Konstruktion der globalen Basis unkompliziert. Sei $\mathcal{N}_g = \{N_1, \dots, N_N\}$, dann nutzen wir die Isomorphie $V_h \cong \mathbb{R}^N$, die durch

$$\begin{aligned}\Phi : V_h &\to \mathbb{R}^N\\ u &\mapsto \begin{pmatrix} N_1(u) \\ \vdots \\ N_N(u) \end{pmatrix}\end{aligned}$$

begründet wird. Dann entspricht die globale Basis gerade den Einheitsvektoren des $\mathbb{R}^N$ und für e_j existiert ein Knotenpunkt $z_j \in \overline{\Omega}$ zu dem jeweils eine Funktion der nodalen Basen der angrenzenden Elemente gehören. Wir weiten diese lokalen Basisfunktionen auf den Ts durch Multiplikation mit der Indikatorfunktion χ_T auf Ω aus und summieren sie auf, um den gesuchten Basisvektor zu erhalten.

Mit einer weiteren Voraussetzung an die Zerlegung müssen wir sogar nur die nodale Basis eines Referenzelements kennen.

Definition 1.8 ([BS08, (3.4.1)])**.** *Sei* $(\hat{T}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ *ein finites Element,*
$F(\boldsymbol{x}) = A\boldsymbol{x} + b$ *eine affine Abbildung mit invertierbarer Matrix* A*. Wir sagen, dass das finite Element* $(T, \mathcal{P}, \mathcal{N})$ ***affin äquivalent*** *zu* $(\hat{T}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ *ist, wenn:*

1. $F(\hat{T}) = T$*,*
2. $\hat{\mathcal{P}} = \{\hat{u} = u \circ F |\ u \in \mathcal{P}\}$ *und*
3. $\mathcal{N} = \Big\{N :\ N(u) = \hat{N}(u \circ F),\ \hat{N} \in \hat{\mathcal{N}},\ u \in \mathcal{P}\Big\}$*.*

Eine Familie von Elementen, die äquivalent zu einem Referenzelement sind, nennen wir ***affine Familie****.*

Es gibt noch weitere Methoden, um Finite-Elemente zu konstruieren. Zum Beispiel können Funktionale als Auswertungsfunktional einer partiellen Ableitung der Funktion definiert werden, um Kontrolle über die Ableitungen der Näherungslösung zu erhalten. Es lassen sich ebenfalls $C^1(\Omega)$-Funktionen definieren und es ist möglich, statt Polynomen andere Funktionenklassen zu wählen.

Uns reicht jedoch hier die Erkenntnis, dass endlichdimensionale Funktionenräume $V_h \subset H_0^1(\Omega)$ existieren, sodass für $v \in V_h, T \in \mathcal{T}$ gilt, dass $v|_T \in P_k(T)$, bzw. $Q_k(T)$.

1.3.4 Approximationseigenschaften von Sobolevfunktionen

Analog zur Approximation mit stückweise polynomialen Funktionen in $\mathbb{R}$ kann durch sukzessive Wahl von immer feineren Zerlegungen eine Folge von Näherungslösungen gefunden werden, die gegen die schwache Lösung der partiellen Differentialgleichung konvergiert. Dafür wird eine zusätzliche Voraussetzung an die Zerlegungen gestellt:

Definition 1.9 ([BS08, 4.4.13])**.** *Sei r_T der Radius der größten Kugel B_T, sodass $T \in \mathcal{T}$ ein Sterngebiet ist und B_T Teilmenge des Zentrums von T ist. Die Familie heißt* ***regulär****, wenn ein $\rho > 0$ existiert, sodass:*

$$r_T \geq \rho d_T \qquad \forall\, T \in \mathcal{T}.$$

Bemerkung 1.3. *Reguläre Zerlegungen sind leicht zu konstruieren. Wenn wir beispielswedie Mittelpunkte der Kanten eines Dreiecks T miteinander verbinden, erhalten wir vier Dreiecke $T_1, \ldots, T_4$ mit $d_{T_i}/r_{T_i} = d_T/r_T$ für alle $i \in \{1, 2, 3, 4\}$.*

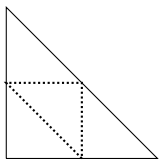

Abbildung 5: Verbindung der Kanten-Mittelpunkte

Wenn für eine Funktion u alle schwachen Ableitungen der Ordnung kleinergleich $k+1$ existieren und quadratintegrierbar sind, kann gezeigt werden, dass ein Interpolationsoperator $\mathcal{I}_h$ existiert, sodass für $\mathbb{N} \ni m \leq k$ gilt:

$$|u - \mathcal{I}_h u|_{m,T} \leq C_{\mathcal{I}_h} d_T^{k+1-m} |u|_{k+1,T}. \tag{12}$$

Eine Darstellung der Approximationstheorie in Sobolevräumen findet sich zum Beispiel in Kapitel 4 von [BS08].

Diese Abschätzung, kombiniert mit dem Satz von Céa, liefert eine obere Begrenzung für den lokalen Fehler der Approximation. Sie spielt daher eine zentrale Rolle in der Finite-Elemente-Analyse. Die Größen d_T können nach oben gegen d_{max}, den maximalen Durchmesser aller Ts abgeschätzt werden, um eine globale Fehlerabschätzung zu erhalten.

1.3.5 Probleme der Galerkin-Methode

Die Dirichlet-Randbedingungen sind aus numerischer Sicht problematisch, da diese sogenannte Grenzschichten (boundary layer) verursachen können. Ein einfaches Beispiel findet sich in [ST08, 1.1- S. 11]. An diesen Stellen gilt:

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} \lim_{\varepsilon \to 0} u(\boldsymbol{x}) \neq \lim_{\varepsilon \to 0} \lim_{\boldsymbol{x} \to \boldsymbol{x}_0} u(\boldsymbol{x}).$$

Wir sagen, dass das Problem **singulär gestört** ist. Die Störung bewirkt, dass die numerische Lösung in der Nähe der Grenzschicht oszilliert.

Die Dirichlet-Randbedingung gibt einen festen Wert für die Punkte in Γ vor. Dieser weicht von den Werten in Ω ab, die durch die Differentialgleichung in der Nähe des Randes vorgegeben werden. Eine stetige Lösung erfordert dann eine steile

Änderung der Werte. Das heißt insbesondere, dass die Ableitungen ersten Grades sehr groß werden.

Zur stabilen Diskretisierung müssen sogenannte stabilisierte Finite-Elemente-Methoden verwendet werden.

1.4 Die Stromlinien-Diffusions-Finite-Elemente-Methode

Die SDFEM ist auch unter dem Namen streamline upwind Petrov-Galerkin Methode (SUPG) bekannt. Die Monographie von Hans G. Roos, Martin Stynes und Lutz Tobiska [ST08] bietet eine umfassende Darstellung von stabilisierten Methoden zur Lösung von singulär gestörten Differentialgleichungen. Dort werden beide Namen erläutert, jedoch wird der Name SDFEM für die Methode verwendet, weshalb wir ihn hier ebenfalls verwenden.

Wir definieren zunächst eine Norm, die dem Umstand Rechnung trägt, dass die Ableitungen ersten Grades der Lösung lokal sehr groß werden.

Sei $V_h \subset H_0^1(\Omega)$, $\mathcal{T}$ eine Zerlegung von Ω, sodass für $v \in V_h$, $T \in \mathcal{T}$ gilt: $v|_T \in C^2(T)$. Weiterhin seien δ_T für $T \in \mathcal{T}$ positive Konstanten und es gelte:

$$c - \frac{1}{2}\nabla \cdot \boldsymbol{b} \geq \omega > 0 \tag{13}$$

Definition 1.10. *Die **Stromlinien-Diffusions-Norm** auf V_h ist gegeben durch:*

$$\|v\|_{SD} = \left(\varepsilon |v|_1^2 + \omega \|v\|_0^2 + \sum_{T \in \mathcal{T}} \delta_T \|\boldsymbol{b} \cdot \nabla v\|_{0,T}^2 \right)^{1/2}. \tag{14}$$

Bemerkung 1.4. *Die Voraussetzung* (13) *ist stärker als die Voraussetzung, die wir für die Existenz der schwachen Lösung gefordert haben.*

Es ist leicht zu sehen, dass $\|\cdot\|_{SD}$ stärker als $\|\cdot\|_E$ ist, denn es gilt:

$$\begin{aligned}
\|v\|_E &= \frac{1}{\varepsilon^{1/2}} \left(\varepsilon |v|_1^2\right)^{1/2} \\
&\leq \frac{1}{\varepsilon^{1/2}} \left(\varepsilon |v|_1^2 + \omega \|v\|_0^2 + \sum_{T \in \mathcal{T}} \delta_T \|\boldsymbol{b} \cdot \nabla v\|_{0,T}^2 \right)^{1/2} \\
&= C \|v\|_{SD}.
\end{aligned}$$

Konvergenz in der $\|\cdot\|_{SD}$-Norm impliziert also auch Konvergenz in der $\|\cdot\|_E$-Norm.

Definition 1.11. *Seien $a(\cdot,\cdot)$ und A_{st} definiert wie oben. Die **SDFEM**-Bilinearform ist definiert als:*

$$\begin{aligned}
a_h : V_h \times V_h &\to \mathbb{R} \\
v, w &\mapsto a(v,w) + \sum_{T \in \mathcal{T}} \delta_T (A_{st} v, \boldsymbol{b} \cdot \nabla w)_T.
\end{aligned}$$

Weiterhin definieren wir:

$$\begin{aligned}
F_h : V_H &\to \mathbb{R} \\
v &\mapsto (f,v) + \sum_{T \in \mathcal{T}} \delta_T (f, \boldsymbol{b} \cdot \nabla v)_T.
\end{aligned}$$

*Die **Lösung der SDFEM** ist $u_h \in V_h$, sodass:*

$$a_h(u_h, v) = F_h(v) \qquad \forall v \in V_h$$

Bemerkung 1.5. *Die SDFEM ist konsistent, denn wenn u_{st} eine Lösung von* (6) *ist, gilt:*

$$\begin{aligned} a(u_{st}, v) &= (f, v), \\ A_{st}u &= f \\ \Rightarrow a_h(u_{st}, v) &= (f, v) + \sum_{T\in\mathcal{T}} \delta_T (f, \boldsymbol{b}\cdot\nabla v)_{0,T} = F_h(v) = a_h(u_h, v). \end{aligned}$$

Weiterhin ist F_h stetig bzgl. der $\|\cdot\|_{SD}$-Norm, sofern f stetig bzgl. der $\|\cdot\|_0$-Norm ist:

$$\begin{aligned} F_h(v) &\leq \|f\|_0\|v\|_0 + \sum_{T\in\mathcal{T}} \delta_T \|f\|_{0,T}\|\boldsymbol{b}\cdot\nabla v\|_{0,T} \\ &\leq \|f\|_0 \left(\|v\|_0 + \sum_{T\in\mathcal{T}} \delta_T \|\boldsymbol{b}\cdot\nabla v\|_{0,T} \right) \\ &\leq \frac{\|f\|_0}{\min\{1,\omega\}} \left(\omega\|v\|_0 + \sum_{T\in\mathcal{T}} \delta_T \|\boldsymbol{b}\cdot\nabla v\|_{0,T} \right) \leq C\|v\|_{SD}. \end{aligned}$$

1.4.1 Existenz und Eindeutigkeit der Lösung

Die Existenz und Eindeutigkeit der Lösung von (15) folgt wieder aus dem Satz von Lax-Milgram. Dazu muss gezeigt werden, dass $a_h(\cdot,\cdot)$ beschränkt und koerziv ist.

Koerzivität Sei $\lambda_T = C_{inv}h_T^{-2}$ die in der Einleitung definierte Konstante, sodass $\lambda_T^{-1}\|\Delta v\|_{0,T} \leq |v|_{1,T}$. Wir fordern, dass:

$$\delta_T \leq \frac{1}{2}\min\left\{ \frac{1}{\varepsilon\lambda_T}, \frac{\omega}{\|c\|^2_{L^\infty(T)}} \right\}. \tag{15}$$

Mit (13) und $a(v,v) = \varepsilon|v|_1^2 + \left(\left(-\frac{1}{2}\nabla\cdot\boldsymbol{b} + c\right)v, v\right)$ folgt:

$$\begin{aligned} a_h(v,v) &\geq \varepsilon|v|_1^2 + \omega\|v\|_0^2 + \sum_{T\in\mathcal{T}} \delta_T\|\boldsymbol{b}\cdot\nabla v\|_0^2 + \delta_T(-\varepsilon\Delta v + cv, \boldsymbol{b}\cdot\nabla v)_{0,T} \\ &\geq \|v\|_{SD} - \left| \sum_{T\in\mathcal{T}} \delta_T(\varepsilon\Delta v - cv, \boldsymbol{b}\cdot\nabla v)_{0,T} \right|. \end{aligned}$$

Mit (1.4.1), der Youngschen- und der Cauchy-Schwarz-Ungleichung folgt nun:

$$\begin{aligned} a_h(v,v) &\geq \|v\|_{SD} - \sum_{T\in\mathcal{T}} \left| \frac{\delta_T}{2} \left(\|\varepsilon\Delta v\|^2_{0,T} - \|cv\|^2_{0,T} \right) \|\boldsymbol{b}\cdot\nabla v\|^2_{0,T} \right| \\ &\geq \|v\|_{SD} - \frac{1}{2}\left(\sum_{T\in\mathcal{T}} \delta_T\varepsilon^2\lambda_T|v|^2_{1,T} + \delta_T\|c\|^2_{\infty,T}\|v\|^2_{0,T} + \delta_T\|\boldsymbol{b}\cdot\nabla v\|^2_{0,T} \right) \\ &\geq \frac{1}{2}\|v\|_{SD}. \end{aligned}$$

Beschränktheit Mit der Beschränktheit von $a(\cdot,\cdot)$ bezüglich der Energienorm und Bemerkung 1.4 gilt:

$$|a_h(v,w)| \leq C\|v\|_{SD}\|w\|_{SD} + \left| \sum_{T\in\mathcal{T}} \delta_T(A_{st}v, \boldsymbol{b}\cdot w)_{0,T} \right|.$$

Wir müssen also nur noch zeigen, dass der letzte Term auf der linken Seite durch die SD-Norm abgeschätzt werden kann:

$$\begin{aligned}
&\left|\sum_{T\in\mathcal{T}} \delta_T (A_{st} v, \boldsymbol{b}\cdot w)_{0,T}\right| \\
\leq &\sum_{T\in\mathcal{T}} \left(\sqrt{\delta_T}\varepsilon \|\Delta v\|_{0,T} + \sqrt{\delta_T}\|c\|_{L^\infty(T)}\|v\|_{0,T} + \delta_T \|\boldsymbol{b}\cdot\nabla v\|_{0,T}\right) \|\boldsymbol{b}\cdot\nabla w\|_{0,T} \\
\leq &\sum_{T\in\mathcal{T}} \left(\frac{\sqrt{\varepsilon}}{\sqrt{2}}|v|_{1,T} + \frac{\sqrt{\omega}}{\sqrt{2}}\|v\|_{0,T} + \sqrt{\delta_T}\|\boldsymbol{b}\cdot\nabla v\|_{0,T}\right) \|\boldsymbol{b}\cdot\nabla w\|_{0,T} \\
\leq &\left(\frac{\varepsilon}{2}|v|_{1,T}^2 + \frac{\omega}{2}\|v\|_{0,T}^2 + \sum_{T\in\mathcal{T}} \|\boldsymbol{b}\cdot\nabla v\|_{0,T}^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|\boldsymbol{b}\cdot\nabla w\|_{0,T}^2\right)^{1/2} \\
\leq &\|v\|_{SD}\|w\|_{SD}.
\end{aligned}$$

Somit ist $a_h(\cdot,\cdot)$ auch beschränkt und es existiert eine eindeutige Lösung von (15).

1.4.2 Fehlerabschätzungen

Sei $u \in H_0^1(\Omega)$ die schwache Lösung der partiellen Differentialgleichung und es gelte die Abschätzung (12). Dann gilt:

$$\|u - u_h\|_{SD} \leq \|u - \mathcal{I}_h u\|_{SD} + \|\mathcal{I}_h u - u_h\|_{SD}$$

Für den zweiten Term erhalten wir wegen der Koerzivität von a_h und der Galerkin-Orthogonalität von u und u_h:

$$\frac{1}{2}\|\mathcal{I}_h u - u_h\|_{SD} \leq a_h(\mathcal{I}_h u - u, \mathcal{I}_h u - u_h).$$

Wir betrachten ausschließlich den Term $\left|\sum_{T\in\mathcal{T}} \delta_T (A_{st}(\mathcal{I}_h u - u), \boldsymbol{b}\cdot\nabla(\mathcal{I}_h u - u_h))_T\right|$ und verweisen auf [ST08, Kapitel 3.2].

$$\begin{aligned}
&\left|\sum_{T\in\mathcal{T}} \delta_T \left(-\varepsilon\Delta(\mathcal{I}_h u - u) + \boldsymbol{b}\cdot\nabla(\mathcal{I}_h u - u) + c(\mathcal{I}_h u - u), \boldsymbol{b}\cdot\nabla(\mathcal{I}_h - u_h)\right)\right| \\
&\qquad \overset{\text{(2b)}}{\leq} C \sum_{T\in\mathcal{T}} (d_T^{k+1}(\varepsilon h_T^{-2} + 1) + d_T^k)|u|_{k+1,T}\|\mathcal{I}_h u - u_h\|_{SD} \qquad (16)
\end{aligned}$$

Bei einer Verfeinerung des Gitters variieren die Größen h_T und d_T. Wir wollen den Term (16) so umschreiben, dass nur eine der beiden Größen darin vorkommt.

Sei $\mathcal{T}$ eine reguläre Zerlegung und h_T eine Größe, für die (2a) gilt: $C_0 r_T \leq h_T \leq C_1 d_T$, dann gilt

$$d^{k+1}|u|_{k+1,T} \leq C h_T^{l+1-m}|u|_{k+1,T}, \qquad (17)$$

mit $C = (C_0^{-1}\rho)^{l+1-m}$. Somit gilt:

$$\begin{aligned}
&\left|\sum_{T\in\mathcal{T}} \delta_T (A_{st}(\mathcal{I}_h u - u), \boldsymbol{b}\cdot\nabla(\mathcal{I}_h u - u_h))_T\right| \\
&\qquad \leq C \sum_{T\in\mathcal{T}} \left((\delta_T\varepsilon)^{1/2}\varepsilon^{1/2} h_T^{k-1} + \delta_T^{1/2}\left(h_T^k + h_T^{k+1}\right)\right)|u|_{k+1,T}\|\mathcal{I}_h u - u_h\|_{SD}
\end{aligned}$$

Aus (1.4.1) folgt: $(\varepsilon\delta_T)^{1/2} \leq C_{inv}^{-1/2} h_T$ und somit können wir den oben stehenden Term nach oben gegen:

$$C \sum_{T\in\mathcal{T}} \left(\varepsilon^{1/2} + \delta_T^{1/2}(1+h_T)\right) h_T^k |u|_{k+1,T} \|\mathcal{I}_h u - u_h\|_{SD}$$

abschätzen. Für die anderen Terme können ähnliche Abschätzungen gemacht werden. Mit $h := \max_{T\in\mathcal{T}} h_T$ erhalten wir die globale Abschätzung:

$$\|\mathcal{I}_h u - u_h\|_{SD} \leq C(\varepsilon^{1/2} + h^{1/2}) h^k |u|_{k+1}.$$

Mit einer zusätzlichen Voraussetzung an die SD-Paramter kann die gleiche Abschätzung für $\|\mathcal{I}_h u - u\|_{SD}$ gemacht werden.

In der Quelle wird $h_T = d_T$ verwendet. Wir können h_T bzw. h auch wieder nach oben gegen d_T bzw. d_{max} abschätzen. Somit sind die beiden Größen austauschbar und die Konvergenzsätze für die SDFEM gelten ebenfalls für allgemeine h_T, unter der Bedingung, dass (2a) gilt.

2 Mathematische Beschreibung des Problems

Wenn für alle $v \in \mathcal{P}$ gilt $\nabla v = 0$, so ist nichts zu zeigen, da dann $\lambda_T = 0$. Andernfalls ist leicht zu sehen, dass (1) äquivalent zu

$$\lambda_T = \sup_{\substack{v \in \mathcal{P} \\ \nabla v \neq 0}} \frac{\|\Delta v\|^2}{\|\nabla v\|^2} \tag{18}$$

ist. (18) heißt verallgemeinerter Rayleigh-Quotient.

2.1 Formulierung als verallgemeinertes Eigenwertproblem

Wir wollen zeigen, dass (18) äquivalent zum folgenden Problem ist:
Finde $w \in \mathcal{P}$, mit $\nabla w \neq 0$ und den maximalen Eigenwert $\lambda_{\max}$, so dass

$$(\Delta w, \Delta v) - \lambda_{\max}(\nabla w, \nabla v) = 0 \qquad \forall v \in \mathcal{P}. \tag{19}$$

Wir schränken zunächst den Definitionsbereich des Problems ein. Der folgende Hilfssatz zeigt, dass diese Einschränkung das Ergebnis nicht beeinflusst.

Lemma 2.1. *Sei $\mathcal{P} = P_0(T) \oplus \tilde{\mathcal{P}}$ und $w = w_0 + w_1$ eine Lösung von* (18) *oder* (19) *mit $w_0 \in P_0(T)$ und $w_1 \in \tilde{\mathcal{P}}$, dann gelten folgenden Aussagen:*

1. *w_1 ist auch eine Lösung und beide Probleme können äquivalent auf $\tilde{\mathcal{P}}$ eingeschränkt werden.*
2. *$(\nabla\cdot, \nabla\cdot)$ ist positiv definit auf $\tilde{\mathcal{P}}$.*

Beweis. **1.** Sei $w = w_0 + w_1$ eine Lösung von (18), dann gilt:

$$\lambda_T = \sup_{\substack{v \in \mathcal{P} \\ \nabla v \neq 0}} \frac{\|\Delta v\|^2}{\|\nabla v\|^2} = \frac{\|\nabla \cdot \nabla(w_0 + w_1)\|^2}{\|\nabla(w_0 + w_1)\|^2} = \frac{\|\Delta w_1\|^2}{\|\nabla w_1\|^2}.$$

Wenn w eine Lösung von (19) ist und $v \in \mathcal{P}$, gilt:

$$\begin{aligned} 0 &= (\nabla \cdot \nabla(w_0 + w_1), \Delta v) - \lambda_T(\nabla(w_0 + w_1), \nabla v) \\ &= (\Delta w_1, \Delta v) - \lambda_T(\nabla w_1, \nabla v). \end{aligned}$$

2. Sei $v \in \mathcal{P}$. Aus der positiven Definitheit des Skalarprodukts folgt:

$$\begin{aligned} (\nabla v, \nabla v) = 0 &\Rightarrow \nabla v = 0 \\ &\Rightarrow \frac{\partial v}{\partial x_i} = 0 \qquad \forall i = 1, ..., d \\ &\Rightarrow v \in P_0 \cap \mathcal{P} \end{aligned}$$

aus dem Mittelwertsatz der Differentialrechnung. Folglich ist $(\nabla\cdot, \nabla\cdot)$ positiv definit auf $\tilde{\mathcal{P}}$. □

Bemerkung 2.1. *Im Allgemeinen wird $\mathcal{P} \cap P_0(T)$ nicht leer sein. Für*

$$\mathcal{P} = P_l(T) = \operatorname{span}\left\{ \boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} x_j^{\alpha_j} : \ \boldsymbol{x} \in T, \ \boldsymbol{\alpha} \in \mathbb{N}^d, \ |\boldsymbol{\alpha}| \leq l \right\},$$

den Raum der Polynome vom Grad kleiner-gleich l gilt mit:

$$\tilde{\mathcal{P}}_1 := \operatorname{span}\left\{ \boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} x_j^{\alpha_j} : \ \boldsymbol{x} \in T, \ \boldsymbol{\alpha} \in \mathbb{N}^d, \ 1 \leq |\boldsymbol{\alpha}| \leq l \right\}$$

$$\textit{und } \tilde{\mathcal{P}}_2 := \operatorname{span}\left\{ \boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} (x_j + 1)^{\alpha_j} : \ \boldsymbol{x} \in T, \ \boldsymbol{\alpha} \in \mathbb{N}^d, \ 1 \leq |\boldsymbol{\alpha}| \leq l \right\},$$

dass $\mathcal{P} = P_0(T) \oplus \tilde{\mathcal{P}}_1 = P_0(T) \oplus \tilde{\mathcal{P}}_2$. Es wird sich herausstellen, dass für unterschiedliche Berechnungen verschiedene Zerlegungen sinnvoll sind. Lemma 2.1 garantiert jedoch, dass eine Lösung in $\tilde{\mathcal{P}}_j$, $j = 1, 2$ auch eine Lösung in $\mathcal{P}$ ist.
Ein Ansatz wird sein, die Anzahl der Variablen zu reduzieren, sodass das Problem in einem Raum mit kleinerer Dimension zu lösen ist. Lemma 2.1 ist ein Spezialfall davon.

Bei dem Beweis verwenden wir die Cholesky-Zerlegung einer positiv definiten und symmetrischen Matrix:

Cholesky-Zerlegung ([Bjö15, 1.3.2])**.** *Für eine positive definite Matrix $A \in \mathbb{R}^{n \times n}$ existiert eine invertierbare untere Dreiecksmatrix L, sodass:*

$$A = LL^T.$$

Nun können wir die Äquivalenz der oben definierten Probleme beweisen.

Satz 2.1. (18) *ist äquivalent zu* (19)*.*

Beweis. Sei $\varphi_1, \ldots, \varphi_n$ eine Basis von $\tilde{\mathcal{P}}$. Wir definieren die symmetrisch, positiv semidefiniten Matrizen

$$K_\Delta = \left[(\Delta\varphi_j, \Delta\varphi_i) \right]_{i,j=1}^{n},$$
$$K_\nabla = \left[(\nabla\varphi_j, \nabla\varphi_i) \right]_{i,j=1}^{n}.$$

Nach Lemma 2.1 ist K_∇ sogar positiv definit.

Für $v = \sum_{i=1}^{n} v_i \varphi_i \in \tilde{\mathcal{P}}$ sei $\underline{v} = (v_1, ..., v_n)^T \in \mathbb{R}^n$. Wir formulieren ein zu (19) äquivalentes verallgemeinertes Eigenwertproblem:

Finde $w \in \tilde{\mathcal{P}}$, sodass $\underline{w}$ der Eigenvektor zum maximalen Eigenwert $\lambda_{\max}$ von K_Δ, K_∇ ist:

$$\underline{v}^T K_\Delta \underline{w} - \lambda_{\max} \underline{v}^T K_\nabla \underline{w} = 0 \qquad \forall\ \underline{v} \in \mathbb{R}^n. \tag{20}$$

Seien $K_\nabla = LL^T$ die Cholesky-Zerlegung von K_∇ und $A := L^{-1} K_\Delta L^{-T}$. Für $\underline{v}, \underline{w} \in \mathbb{R}^n$ definieren wir $L^T \underline{v} := \underline{x}$ und $L^T \underline{w} := \underline{y}$. Für einen verallgemeinerten Eigenwert λ zu K_Δ, K_∇ mit zugehörigem Eigenvektor $\underline{w} \in \mathbb{R}^n$ gilt:

$$\begin{aligned} 0 = \underline{v}^T K_\Delta \underline{w} - \lambda \underline{v}^T K_\nabla \underline{w} &= \underline{v}^T K_\Delta \underline{w} - \lambda \underline{v}^T L L^T \underline{w} && \forall \underline{v} \in \mathbb{R}^n \\ &= \underline{x}^T A \underline{y} - \lambda \underline{x}^T \underline{y} && \forall \underline{x} \in \mathbb{R}^n. \end{aligned}$$

Folglich ist jeder verallgemeinerte Eigenwert von K_Δ, K_∇ auch ein Eigenwert von A und umgekehrt. Insbesondere ist $\lambda_{\max}$ der maximale Eigenwert von A und (20) ist äquivalent zu:

Finde $w \in \tilde{\mathcal{P}}$, sodass $L^{-T}\underline{w} = \underline{y}$ der Eigenvektor zum maximalen Eigenwert $\lambda_{\max}$ von $A = L^{-1}K_\Delta L^{-T}$ ist:

$$A\underline{y} = \lambda_{\max}\underline{y}. \tag{21}$$

Wir zeigen (18) $\Longleftrightarrow$ (21). Zunächst schätzen wir λ_T nach oben gegen $\lambda_{\max}$ ab.

$\boxed{\lambda_T \leq \lambda_{\max}}$

Aus der Invertierbarkeit der Matrix L^T folgt insbesondere, dass sie bijektiv ist. Folglich gilt:

$$\lambda_T = \sup_{\substack{\underline{v}\in\mathbb{R}^n \\ \underline{v}\neq 0}} \frac{\underline{v}^T K_\Delta \underline{v}}{\underline{v}^T K_\nabla \underline{v}} = \sup_{\substack{\underline{v}\in\mathbb{R}^n \\ \underline{v}\neq 0}} \frac{(\underline{v}^T L)(L^{-1}K_\Delta L^{-T})(L^T\underline{v})}{(\underline{v}^T L)(L^T\underline{v})} = \sup_{\substack{\underline{x}\in\mathbb{R}^n \\ \underline{x}\neq 0}} \frac{\underline{x}^T A\underline{x}}{\underline{x}^T\underline{x}}.$$

Die Matrix A ist symmetrisch, denn

$$A^T = \left(L^{-1}K_\Delta L^{-T}\right)^T = (K_\nabla L^{-T})^T (L^{-1})^T = L^{-1}K_\Delta^T L^{-T} = A.$$

Sie ist auch positiv semidefinit, denn $\underline{y}^T A\underline{y} = \underline{w}^T K_\Delta \underline{w}$ und K_Δ ist positiv semidefinit.

Seien $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte von A mit zugehörigen normierten Eigenvektoren $p_1, \ldots, p_n$. Für symmetrische, positiv semidefinite Matrizen ist bekannt, dass die Eigenwerte reell und die zugehörigen Eigenwerte paarweise orthogonal sind und eine Basis des $\mathbb{R}^n$ bilden.
Es existiert also eine eindeutige Darstellung $\underline{x} = \sum_{j=1}^n x_j p_j$ für alle $\underline{x} \in \mathbb{R}^n$. Sei $\underline{x} \in \mathbb{R}^n$ beliebig gewählt, dann gilt:

$$\frac{\underline{x}^T A\underline{x}}{\|\underline{x}\|^2} \underset{(i)}{\leq} \frac{\|\underline{x}\|\|A\underline{x}\|}{\|\underline{x}\|^2} \underset{(ii)}{=} \frac{\|\underline{x}\|\sqrt{\sum_{j=1}^n \lambda_j^2 x_j^2}}{\|\underline{x}\|^2} \underset{(iii)}{\leq} \frac{\|\underline{x}\|\sqrt{\sum_{j=1}^n \lambda_{\max}^2 x_j^2}}{\|\underline{x}\|^2} = \lambda_{\max}. \tag{22}$$

Dabei haben wir bei (i) die Cauchy Schwarz Ungleichung verwendet, bei (ii), dass die p_j paarweise orthogonal sind und bei (iii), dass $\lambda_{\max} \geq \lambda_j$ für $j = 1, \ldots, n$. Da $\underline{x}$ beliebig gewählt war, gilt $\lambda_T \leq \lambda_{\max}$.

$\boxed{(21) \Longrightarrow (18)}$

Sei w eine Lösung von (21) $\Longrightarrow$ für $\underline{y} = L^{-T}\underline{w}$ gilt $A\underline{y} = \lambda_{\max}\underline{y}$. In diesem Fall ist (22) scharf

$$\Longrightarrow \qquad \lambda_{\max} = \frac{\underline{y}^T A\underline{y}}{\|\underline{y}\|^2} = \sup_{\substack{\underline{x}\in\mathbb{R}^n \\ \underline{x}\neq 0}} \frac{\underline{x}^T A\underline{x}}{\|\underline{x}\|^2} = \lambda_T$$

$$\Longrightarrow \qquad \frac{\|\Delta w\|^2}{\|\nabla w\|^2} = \sup_{\substack{v\in\mathcal{P} \\ \nabla v\neq 0}} \frac{\|\Delta v\|^2}{\|\nabla v\|^2}.$$

$\boxed{(18) \Longrightarrow (21)}$

Seien $\underline{x} := p_1$ und $\underline{v} = (v_1, \ldots, v_n)^T = L^{-T}\underline{x}$. Wenn wir $\underline{x}$ in (22) einsetzen, ist die Abschätzung scharf. Somit ist die Lösungsmenge von (18) nicht leer und es gilt $\lambda_{\max} = \lambda_T$.

Für eine Lösung $w \in \tilde{\mathcal{P}}$ von (18), mit $\underline{y} := L^{-T}\underline{w}$ gilt

$$\frac{\underline{y}^T A \underline{y}}{\|\underline{y}\|^2} = \lambda_{\max}.$$

Bei der Cauchy-Schwarz-Ungleichung gilt Gleichheit nur dann, wenn die beiden Vektoren linear abhängig sind. Deshalb impliziert (i) in (22), dass $\underline{y}$ ein Eigenvektor von A ist. Offensichtlich impliziert (iii), dass der zugehörige Eigenwert $\lambda_{\max}$ ist und somit gilt $A\underline{y} = \lambda_{\max}\underline{y}$.

□

Mit Satz 2.1 ist eine äquivalentes Problem gefunden, für das explizite Lösungsmethoden bekannt sind. In Abschnitt 3 werden wir die Formulierung (20) verwenden. Die eingebaute MATLAB funktion `eig` [MAT21] kann mit zwei Argumenten $A = K_\Delta,\ B = K_\nabla$ aufgerufen werden und produziert einen Spaltenvektor, der die verallgemeinerten Eigenwerte von A, B enthält. Wenn A, B symmetrisch sind und B positiv definit, wird das Problem per Voreinstellung mittels einer Cholesky-Zerlegung auf (21) reduziert.

2.2 Affine Familien von Elementen

Für affin äquivalente Elemente kann (18) so umformuliert werden, dass nur Integrale über dem Referenzelement vorkommen.

Satz 2.2. *Sei $T \subseteq \Omega$, und $\hat{T}$ ein Referenzelement, sodass eine affine Funktion $F_T\hat{\boldsymbol{x}} = t + M_T\hat{\boldsymbol{x}}$, $t \in \mathbb{R}^d, M_T \in GL(d)$ mit $T = F_T(\hat{T})$ existiert. Seien weiterhin p ein Polynom auf T und $\hat{p} := p \circ F_T$. Wir bezeichnen mit $H_{\hat{p}}(\hat{\boldsymbol{x}}) = \left[\partial_{\hat{x}_i,\hat{x}_j}\hat{p}(\hat{\boldsymbol{x}})\right]_{i,j=1}^d$ die symmetrische Hessematrix von $\hat{p}$ in $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_d)^T$ und mit $\nabla_{\hat{T}}$ den Gradienten bezüglich der $\hat{x}_i$, dann gilt:*

$$\frac{\|\Delta p\|_{0,T}^2}{\|\nabla p\|_{0,T}^2} = \frac{\|\mathrm{spur}(M_T^{-1}M_T^{-T}H_{\hat{p}})\|_{0,\hat{T}}^2}{\|M_T^{-T}\nabla_{\hat{T}}\hat{p}\|_{0,\hat{T}}^2}.$$

Beweis. Es ist $\hat{\boldsymbol{x}} = F_T^{-1}(\boldsymbol{x}) = M_T^{-1}(\boldsymbol{x} - t)$ für alle $\boldsymbol{x} \in T$. Durch Anwendung der Kettenregel, sehen wir, dass für die partiellen Ableitungen gilt:

$$\begin{aligned}\frac{\partial p(\boldsymbol{x})}{\partial x_i} &= \nabla_{\hat{T}}\hat{p}(\hat{\boldsymbol{x}})^T(M_T^{-1})_i \\ &= (M_T^{-T}\nabla_{\hat{T}}\hat{p}(\hat{\boldsymbol{x}}))_i.\end{aligned}$$

M_T^{-T} ist konstant und daher ergibt sich durch nochmalige Anwendung der Kettenregel und des Satz' von Schwarz:

$$\begin{aligned}\frac{\partial^2 p(\boldsymbol{x})}{\partial^2 x_i} &= (M_T^{-T}H_{\hat{p}}(\hat{\boldsymbol{x}}))^T(M_T^{-1})_i \\ &= (M_T^{-T}H_{\hat{p}}(\hat{\boldsymbol{x}})M_T^{-1})_{i,i}.\end{aligned}$$

Nun können wir den Transformationssatz anwenden und erhalten:

$$\frac{\|\Delta p\|_{0,T}^2}{\|\nabla p\|_{0,T}^2} = \frac{\int_T \operatorname{spur}(M_T^{-T} H_{\hat{p}}(F_T^{-1}(\boldsymbol{x})) M_T^{-1})^2 d\boldsymbol{x}}{\int_T \|M_T^{-T} \nabla_{\hat{T}} \hat{p}(F_T^{-1}(\boldsymbol{x}))\|^2 d\boldsymbol{x}}$$

$$= \frac{|\det(M_T^{-1})|}{|\det(M_T^{-1})|} \frac{\int_{\hat{T}} \operatorname{spur}(M_T^{-T} H_{\hat{p}}(\hat{\boldsymbol{x}}) M_T^{-1})^2 d\hat{\boldsymbol{x}}}{\int_{\hat{T}} \|M_T^{-T} \nabla_{\hat{T}} \hat{p}(\hat{\boldsymbol{x}})\|^2 d\hat{\boldsymbol{x}}}$$

$$= \frac{\|\operatorname{spur}((M_T^{-T} H_{\hat{p}}) M_T^{-1})\|_{0,\hat{T}}^2}{\|M_T^{-T} \nabla_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2}.$$

□

Da die Spurabbildung invariant gegenüber zyklischen Vertauschungen der Faktoren ist, folgt die Behauptung.

Folgerungen aus Satz 2

1. Wir definieren $\Delta_{\hat{T}}$ analog zu $\nabla_{\hat{T}}$. Wenn M_T eine orthogonale Matrix ist, sodass $M_T^{-1} = M_T^T$, folgt: $\Delta p(\boldsymbol{x}) = \operatorname{spur}(H_{\hat{p}}(\hat{\boldsymbol{x}})) = \Delta_{\hat{T}} \hat{p}(\hat{\boldsymbol{x}})$

$$\Rightarrow \|\Delta p\|_{0,T}^2 = \|\Delta_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2.$$

2. In diesem Fall hat auch die L^2-Norm von ∇p in T eine einfache Form:

$$\|\nabla p\|_{0,T}^2 = \int_{\hat{T}} \nabla_{\hat{T}} \hat{p}(\hat{\boldsymbol{x}})^T M_T^{-1} M_T^{-T} \nabla_{\hat{T}} \hat{p}(\hat{\boldsymbol{x}}) d\hat{\boldsymbol{x}} = \|\nabla_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2.$$

3. Aus den ersten beiden Punkten folgt, dass λ_T invariant gegenüber Kongruenzabbildungen ist. Seien $\hat{T}$ ein Referenzelement, $t \in \mathbb{R}^d$, $A \in \mathbb{R}^{d,d}$, sodass $M_T^{-1} = M_T^T$ und $\Omega \supset T = \{M_T \hat{\boldsymbol{x}} + t : \ \hat{\boldsymbol{x}} \in \hat{T}\}$, dann gilt:

$$\lambda_T = \sup_{p \in \mathcal{P}} \frac{\|\Delta p\|_{0,T}^2}{\|\nabla p\|_{0,T}^2} = \sup_{\hat{p} \in \hat{\mathcal{P}}} \frac{\|\Delta_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2}{\|\nabla_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2} = \lambda_{\hat{T}}.$$

4. Für ähnliche Elemente, also die Bilder von Kongruenzabbildungen verknüpft mit einer Skalierung $S_T(\hat{\boldsymbol{x}}) = s_T \hat{\boldsymbol{x}}$, $s_T > 0$ ist leicht einzusehen, dass

$$\lambda_T = \sup_{\hat{p} \in \hat{\mathcal{P}}} \frac{s_T^{-4} \|\Delta_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2}{s_T^{-2} \|\nabla_{\hat{T}} \hat{p}\|_{0,\hat{T}}^2} = s_T^{-2} \lambda_{\hat{T}}$$

   Für ein Rechteck erhalten wir zum Beispiel vier halb so große, ähnliche Figuren, indem wir die Mittelpunkte gegenüberliegender Kanten miteinander verbinden. Bei Dreiecken, indem wir alle Mittelpunkte der Kanten miteinander verbinden.

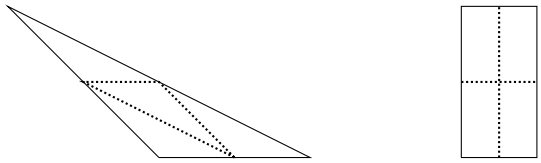

Abbildung 6: Zerlegung in Ähnliche Figuren

5. Die Einträge der Matrizen K_Δ und K_∇ aus Satz 2.1 können somit als

$$(K_\Delta)_{i,j} = \left(\mathrm{spur}(M_T^{-1} M_T^{-T} H_{\hat{\varphi}_j}), \mathrm{spur}(M_T^{-1} M_T^{-T} H_{\hat{\varphi}_i})\right)$$
$$(K_\nabla)_{i,j} = \left(M_T^{-T} \nabla_{\hat{T}} \hat{\varphi}_j, M_T^{-T} \nabla_{\hat{T}} \hat{\varphi}_i\right)$$

geschrieben werden.

3 Explizite Lösung des Problems

In [HH92] werden für einige Sonderfälle explizite Lösungen für das Problem gefunden. Wir werden einige von diesen Fällen nachvollziehen und die verwendeten Techniken vorstellen. Im Gegensatz zum Vorgehen in der Quelle werden die algebraischen Manipulationen, die zur Herleitung der Lösung notwendig sind in Matrixschreibweise dargestellt. Das heißt, wir werden die Matrizen K_∇ und K_Δ explizit aufschreiben. Trotz umfangreicherer Notation ist so die Lösung leichter auf einen Blick zu sehen.

3.1 Der eindimensionale Fall

Für $T = (a, b) \subset \mathbb{R}$ und $\tilde{v} \in \tilde{\mathcal{P}} = \{v \in P_k(T) : \ v(0) = 0\}$ gilt:

$$\nabla\tilde{v} = \tilde{v}' \in P_{k-1}(T) \qquad\qquad \Delta\tilde{v} = \tilde{v}'' \in P_{k-2}(T).$$

Für $k \geq 1$, $\mathcal{P} = P_k(T)$ und $v = \tilde{v}'$ reduziert sich das Problem somit auf:

$$\|v'\|_{0,T}^2 \leq \lambda_T \|v\|_{0,T}^2 \qquad\qquad \forall v \in P_{k-1}(T).$$

Sei $v = \sum_{i=1}^{k-1} v_i\varphi_i$. Wir wählen $\hat{T} = (-1, 1)$ und definieren die positive semidefiniten Matrizen:

$$K_\nabla = \left[\int_{-1}^{1} \varphi_j'(\hat{\boldsymbol{x}})\varphi_i'(\hat{\boldsymbol{x}})d\hat{\boldsymbol{x}}\right]_{i,j=0}^{k-1}, \qquad K_I = \left[\int_{-1}^{1} \varphi_j(\hat{\boldsymbol{x}})\varphi_i(\hat{\boldsymbol{x}})d\hat{\boldsymbol{x}}\right]_{i,j=0}^{k-1}.$$

Wegen der positiven Definitheit des L^2-Skalarproduktes ist K_I positiv definit, also können wir Satz 2.1 analog anwenden:

$$\lambda_{\hat{T}} = \sup_{v \in P_{k-1}(\hat{T})} \frac{|v|_{1,\hat{T}}}{\|v\|_{0,\hat{T}}} = \max_{\underline{w} \in \mathbb{R}^d} \left\{\lambda : \ \underline{v}K_\nabla\underline{w} = \lambda\underline{v}K_I\underline{w} \quad \forall\underline{v} \in \mathbb{R}^k\right\}.$$

Wir setzen $h_T = d_T = b - a$. Wegen Satz 2.2 können wir oBdA. annehmen, dass $T = \left(\frac{-h_T}{2}, \frac{h_T}{2}\right)$ ist, sodass $F_T(\hat{\boldsymbol{x}}) = \frac{h_T}{2}\hat{\boldsymbol{x}}$. Diese Definition erfüllt die Voraussetzung (2a) und mit $h_{\hat{T}} = 2$ gilt:

$$C_{inv} = h_{\hat{T}}^2\lambda_{\hat{T}} = 4 \max_{\underline{w} \in \mathbb{R}^d} \{\lambda : \ K_\nabla\underline{w} = \lambda K_I\underline{w}\}.$$

Die Ungleichung gilt dann für alle reellen, beschränkten Intervalle, wobei C_{inv} konstant bleibt und $h_T = d_T$ leicht zu berechnen ist. In [ÖRW10] und [CZ13] wird der Ansatz gemacht, orthogonale Polynome zu verwenden. In diesem Fall ist $K_I = Id$ und das Problem ist darauf reduziert K_∇ zu berechnen und den maximalen Eigenwert zu finden.

3.1.1 Legendre Polynome

Seien p_i, p_j die Legendrepolynome vom Grad i und $j \in \mathbb{N}$, dann gilt:

$$\int_{-1}^{1} p_j(\hat{\boldsymbol{x}})p_i(\hat{\boldsymbol{x}})d\hat{\boldsymbol{x}} = \frac{2}{2n+1}\delta_{i,j}$$

Insbesondere sind die Polynome

$$\left\{\varphi_i : \left(\frac{2n+1}{2}\right)^{1/2} p_i, \quad p_i \text{ ist das Legendrepolynom vom Grad i}\right\}$$

orthonormal. Für $i = 1, \ldots, 4$ sind sie explizit gegeben durch

$$\varphi_1(\hat{\boldsymbol{x}}) = \left(\frac{3}{2}\right)^{1/2} \hat{\boldsymbol{x}} \qquad \varphi_2(\hat{\boldsymbol{x}}) = \left(\frac{5}{8}\right)^{1/2} (3\hat{\boldsymbol{x}}^2 - 1)$$
$$\varphi_3(\hat{\boldsymbol{x}}) = \left(\frac{7}{8}\right)^{1/2} (5\hat{\boldsymbol{x}}^3 - 3\hat{\boldsymbol{x}}) \qquad \varphi_4(\hat{\boldsymbol{x}}) = \left(\frac{9}{128}\right)^{1/2} (35\hat{\boldsymbol{x}}^4 - 30\hat{\boldsymbol{x}}^2 + 3).$$

Für $k = 5$ hat $K_\nabla = K_\nabla(5)$ die Form:

$$\begin{pmatrix} 3 & 0 & \sqrt{21} & 0 \\ 0 & 15 & 0 & 9\sqrt{5} \\ \sqrt{21} & 0 & 42 & 0 \\ 0 & 9\sqrt{5} & 0 & 90 \end{pmatrix}.$$

Da für $2 \leq k \leq 5$ die zugehörige Matrix $K_\nabla(k)$ nur die Einschränkung auf die $k \times k$ Blockmatrix beginnend in der oberen linken Ecke ist, können wir $C_{inv} = C_{inv}(k)$ für $k = 2, 3$ sofort ablesen:

$$C_{inv}(2) = 12, \qquad C_{inv}(3) = 60. \tag{23}$$

Um nun $C_{inv}(4)$ zu erhalten stellen wir zunächst fest, dass e_2 ein Eigenvektor von $K_\nabla(4)$ zum Eigenwert 15 ist. Da die Eigenvektoren von $K_\nabla(4)$ eine orthogonale Basis von $\mathbb{R}^3$ bilden sind die übrigen Eigenvektoren aus dem orthogonalen Komplement der linearen Hülle von e_2 - das ist die lineare Hülle von e_1 und e_3. Alternativ vertauschen wir den zweiten und den dritten Basisvektor und erhalten:

$$K_\nabla(5) = \begin{pmatrix} 3 & \sqrt{21} & 0 & 0 \\ \sqrt{21} & 42 & 0 & 0 \\ 0 & 0 & 15 & 9\sqrt{5} \\ 0 & 0 & 9\sqrt{5} & 90 \end{pmatrix}$$

$$M := \begin{pmatrix} 3 & \sqrt{21} \\ \sqrt{21} & 42 \end{pmatrix}$$

zu berechnen. Dazu betrachten wir das charakteristische Polynom:

$$\begin{aligned} det(M - tId) &= (3-t)(42-t) - 21 \qquad t \in \mathbb{R} \\ &= t^2 - 45t + 105. \end{aligned}$$

Die Nullstellen sind

$$\frac{45 + \sqrt{1605}}{2} \approx 42.5312, \quad \frac{45 - \sqrt{1605}}{2} \approx 2.4688$$

und somit

$$0 < \underbrace{90 - 2\sqrt{1605}}_{\approx 9.8751} < 60 < \underbrace{90 + 2\sqrt{1605}}_{\approx 170.1249} = C_{inv}(4)$$

Die Eigenwerte von $K_\nabla(5)$ lassen sich so ebenfalls ablesen. Es sind nur noch die Nullstellen des charakteristischen Polynoms von

$$\begin{pmatrix} 15 & 9\sqrt{5} \\ 9\sqrt{5} & 90 \end{pmatrix}$$

zu berechnen:

$$(15-t)(90-t) - 405 = t^2 - 105t + 945.$$

Die Nullstellen sind:

$$\frac{105 - 3\sqrt{805}}{2} \approx 9.9412, \quad \frac{105 + 3\sqrt{805}}{2} \approx 95.0588$$

und folglich

$$0 < \underbrace{90 - 2\sqrt{1605}}_{\approx 9.8751} < \underbrace{210 - 6\sqrt{805}}_{\approx 39.7649} < \underbrace{90 + 2\sqrt{1605}}_{\approx 170.1249} < \underbrace{210 + 6\sqrt{805}}_{\approx 380.2351} = C_{inv}(5).$$

3.2 Zwei Dimensionen

Im zweidimensionalen Fall muss die geometrische Form des Elements beachtet werden. Es gibt einige Techniken, um die Berechnungen zu vereinfachen.

3.2.1 Reduzierung der Anzahl der Variablen

Der folgende Ansatz wird in [HH92](39) benutzt um die Anzahl der unabhängigen Variablen zu reduzieren.

Sei $\mathcal{P} = P_l(T)$. Für Funktionen $v_1, v_2 \in \mathcal{P}$ mit $\Delta v_1 = \Delta v_2$ genügt es, nur diejenige zu beachten, für die die Norm des Gradienten kleiner ist. Für ein Polynom $p(\boldsymbol{x}) = \sum a_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$ ist eine solche Menge gegeben durch alle Polynome mit gleichen Koeffizienten für $|\boldsymbol{\alpha}| \geq 2$, da $P_1(T) \in ker(\Delta)$ und $(\Delta\cdot, \Delta\cdot)$ bilinear. Wir variieren also die Koeffizienten für $|\boldsymbol{\alpha}| = 1$ bei fest gewählten Koeffizienten für $|\boldsymbol{\alpha}| \geq 2$.

Satz 3.1. *Für $p \in P_l(T)$ sei M_p die Menge der Polynome mit identischen Koeffizienten für $|\boldsymbol{\alpha}| \geq 2$ und $p_0 \in M_p$, sodass $\nabla p_0 \perp P_0(T)^d$, dann ist $\|\nabla p_0\|^2 \leq \|\nabla p\|^2$ für alle $p \in M_p$. Dieses Polynom existiert für alle p und ist unter der zusätzlichen Voraussetzung $p(0) = 0$ eindeutig bestimmt.*

Beweis. Wir nummerieren zunächst die Basis, sodass $p(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \boldsymbol{x}^{\boldsymbol{\alpha}^j}$. Dabei sind die ersten $d+1$ Monome gegeben durch: $\boldsymbol{x}^{\boldsymbol{\alpha}^1} = 1$ und $\boldsymbol{x}^{\boldsymbol{\alpha}^{j+1}} = x_j$, sodass die Koeffizienten a_j für $j > d+1$ fest gewählt sind und die $a_1, ..., a_d$ in Abhängigkeit davon gewählt werden müssen.

Minimaleigenschaft

$$\begin{aligned} \|\nabla p\|^2 &= (\nabla(p_0 + r), \nabla(p_0 + r)) && r = p - p_0 \\ &= \|\nabla p_0\|^2 + 2(\nabla p_0, \nabla r) + \|\nabla r\|^2 \\ &\geq \|\nabla p_0\|^2, \end{aligned}$$

denn $r \in P_1(T) \Rightarrow \nabla r \in P_0(T)^d$.

Existenz und Eindeutigkeit Die Bedingung $(\nabla p_0, \nabla r) = 0$ für alle $r \in P_1(T)$ ist genau dann erfüllt, wenn

$$\int_T \frac{\partial p_0(\boldsymbol{x})}{\partial x_i}\, d\boldsymbol{x} = 0, \quad i = 1, \ldots, d.$$

Verwendet man den Ansatz für p_0, so erhält man

$$0 = \int_T \sum_{j=1}^{N} a_j \frac{\partial \boldsymbol{x}^{\boldsymbol{\alpha}^j}}{\partial x_i}\, d\boldsymbol{x} = \int_T \left(a_{i+1} + \sum_{j=d+2}^{N} a_j \frac{\partial \boldsymbol{x}^{\boldsymbol{\alpha}^j}}{\partial x_i} \right) d\boldsymbol{x}, \quad i = 1, \ldots, d.$$

Das ist äquivalent mit

$$a_{i+1} = -\frac{1}{|T|} \int_T \sum_{j=d+2}^{N} a_j \frac{\partial \boldsymbol{x}^{\boldsymbol{\alpha}^j}}{\partial x_i}\, d\boldsymbol{x}, \quad i = 1, \ldots, d,$$

wobei $|T|$ das Volumen von T ist. Somit sind die Koeffizienten $a_2, \ldots, a_{d+1}$ eindeutig aus den Koeffizienten $a_{d+2}, \ldots, a_N$ bestimmt. Mit $a_1 = 0$ ist p_0 eindeutig bestimmt. □

Bemerkung 3.1. *Die weiteren Schritte des direkten Ansatzes aus [HH92] bestehen darin, die Norm des Gradienten mit arithmetischen Operationen so zu minimieren, dass die Norm des Laplace-Operators schließlich als Faktor auftaucht. Satz 2.1 zeigt, dass dieser Ansatz zum richtigen Ergebnis führt, und dass der andere Term des Produkts konstant ist. Damit ist das verallgemeinerte Eigenwertproblem auf ein einfaches Eigenwertproblem reduziert, das es noch zu lösen gilt - was für die gewählten Elemente und Funktionenräume jedoch ohne zusätzlichen Aufwand möglich ist.*

3.2.2 Dreiecke

Der folgende Ansatz ist konzeptionell dem aus Satz 3.1 entlehnt und baut darauf auf. Dabei wird eine Basis des Kerns von $\Delta|_{\mathcal{P}}$ zu einer Basis von $\mathcal{P}$ erweitert. Anschließend werden die Koeffizienten der Kernelemente so gewählt, dass die übrigen in Frage kommenden Lösungen orthogonal zu $ker\Delta \cap \mathcal{P}$ bzgl. $(\cdot,\cdot)_E$ sind. Nach Satz 2.1 erfüllt jede Lösung von (18) diese Voraussetzung.

P_2 über dem 2-Einheitssimplex Sei $\hat{T} = \{\hat{\boldsymbol{x}} \in \mathbb{R}^2_{\geq 0} : \sum_{i=1}^{2} \hat{x}_i \leq 1\}$ der Einheitssimplex in $\mathbb{R}^2$. Um die Berechnung zu vereinfachen werden

$$\hat{x}_t = \hat{x}_1 - \frac{1}{3},\ \hat{y}_t = \hat{x}_2 - \frac{1}{3}$$

definiert, sodass der geometrische Schwerpunkt des verschobenen Dreiecks über dem Nullpunkt liegt. Damit verschwinden die Integrale von $a\hat{x}_t + b\hat{y}_t$ für alle $a, b \in \mathbb{R}$. Es ist

$$ker(\Delta|_{P_2(\hat{T})}) = \text{span}(\{1, \hat{x}_t, \hat{y}_t, \hat{x}_t\hat{y}_t, \hat{x}_t^2 - \hat{y}_t^2\}).$$

Wir definieren $\mathcal{P}$ als die lineare Hülle von $\{\hat{x}_t, \hat{y}_t, \hat{x}_t\hat{y}_t, \hat{x}_t^2 - \hat{y}_t^2, \hat{y}_t^2\}$ ist. Eine Lösung in $\mathcal{P}$ ist nach Lemma 2.1 auch eine Lösung in $P_2(\hat{T})$. Für die oben definierte Basis hat K_Δ die einfache Form:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

Für eine Funktion $p_0(\hat{\boldsymbol{x}}) = a_1\hat{x}_t + a_2\hat{y}_t + a_3\hat{x}_t\hat{y}_t + a_4(\hat{x}_t^2 - \hat{y}_t^2) + a_5\hat{y}_t^2$, die eine Lösung von (18) ist gilt:

$$\frac{1}{36}\begin{pmatrix} 18 & 0 & 0 & 0 & 0 \\ 0 & 18 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 8 & -4 \\ 0 & 0 & -1 & -4 & 4 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \frac{2}{\lambda_{\hat{T}}}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ a_5 \end{pmatrix}.$$

Wir betrachten nur die ersten vier Gleichungen und erhalten:

$$\begin{pmatrix} 18 & 0 & 0 & 0 \\ 0 & 18 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = a_5\begin{pmatrix} 0 \\ 0 \\ 1 \\ 4 \end{pmatrix}.$$

Dieses System ist für alle $a_5 \neq 0$ lösbar. Wir wählen $a_5 := 2$ und erhalten $a_3 = a_4 = 1$

$$\begin{aligned} \Longrightarrow p_0 &= \hat{x}_t\hat{y}_t + 2\hat{y}_t^2 + \hat{x}_t^2 - \hat{y}_t^2 \\ &= \frac{1}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \hat{x}_1 - \hat{x}_2 + \hat{x}_1\hat{x}_2 + \hat{x}_1^2 + \hat{x}_2^2 \end{aligned}$$

und schließlich

$$\begin{aligned} \frac{1}{\lambda_{\hat{T}}} &= \frac{-1-4+8}{144} \\ &= \frac{1}{48} \\ \lambda_{\hat{T}} &= 48. \end{aligned}$$

Das Problem kann in diesem Fall also auf die Lösung eines linearen Gleichungssystems reduziert werden.

Affine Abbildungen Jedes Dreieck T im $\mathbb{R}^2$ ist das Bild einer affinen Abbildung $F_T(\hat{\boldsymbol{x}}) = M_T\hat{\boldsymbol{x}} + t$ mit dem zwei-dimensionalen Einheitssimplex $\hat{T}$ als Referenzelement. Satz 2.2 zeigt, dass es genügt nur Dreiecke der mit den folgenden Eckpunkten zu betrachten:

$$\left\{0, \begin{pmatrix} X_1 \\ 0 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix} : \quad X_1, Y_2 \in \mathbb{R}_+,\ X_2 \in \mathbb{R}\right\}.$$

Die Matrizen M_T und M_T^{-1} sind gegeben durch:

$$M_T = \begin{pmatrix} X_1 & X_2 \\ 0 & Y_2 \end{pmatrix} \qquad X_1, Y_2 \in \mathbb{R}_+,\ X_2 \in \mathbb{R}$$
$$M_T^{-1} = Det(M_T)^{-1} \begin{pmatrix} Y_2 & -X_2 \\ 0 & X_1 \end{pmatrix}$$
$$M_T^{-1} M_T^{-T} = Det(M_T)^{-2} \begin{pmatrix} X_2^2 + Y_2^2 & -X_1 X_2 \\ -X_1 X_2 & X_1^2 \end{pmatrix}.$$

Aus Satz 2.2 folgt:

$$\nabla p(\boldsymbol{x}) = M_T^{-T} \nabla_{\hat{T}} \hat{p}(\hat{\boldsymbol{x}}) = Det(M_T)^{-1} \begin{pmatrix} Y_2 \dfrac{\partial \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} - X_2 \dfrac{\partial \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_2} \\ X_1 \dfrac{\partial \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_2} \end{pmatrix} \tag{24}$$

$$\begin{aligned} \Delta p(\boldsymbol{x}) &= \operatorname{spur}(M_T^{-1} M_T^{-T} H_{\hat{p}}(\hat{\boldsymbol{x}})) \\ &= det(M_T)^{-2} \left((X_2^2 + Y_2^2) \frac{\partial^2 \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1^2} + X_1^2 \frac{\partial^2 \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_2^2} - 2 X_1 X_2 \frac{\partial^2 \hat{p}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1 \partial \hat{x}_2} \right). \end{aligned} \tag{25}$$

Sei $\boldsymbol{x} = (x_1, x_2)^T = M_T \hat{\boldsymbol{x}}$. Wir wollen wieder Koordinaten definieren, sodass der geometrische Schwerpunkt c von T über dem Nullpunkt liegt. Der geometrisch Schwerpunkt ist durch $1/3(X_1 + X_2, Y_2)^T = (c_x, c_y)^T$ gegeben. Wir definieren die affinen Koordinaten:

$$\begin{aligned} x_a &:= x_1 - c_x \\ &= \hat{x}_t X_1 + \hat{y}_t X_2 \\ y_a &:= x_2 - c_y \\ &= \hat{y}_t Y_2, \end{aligned}$$

mit $\hat{x}_t, \hat{y}_t$ wie oben.

Sei $\tilde{\mathcal{P}}$ die lineare Hülle von

$$\{x_a, y_a, x_a y_a, x_a^2 - y_a^2, y_a^2\}.$$

K_Δ hat dann die Form:

$$2 Det(M_T) \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

und K_∇

$$\frac{Det(M_T)}{36} \begin{pmatrix} 36 & 0 & 0 & 0 & 0 \\ 0 & 36 & 0 & 0 & 0 \\ 0 & 0 & 2(X_1^2 + X_2^2 + Y_2^2 - X_1 X2) & 0 & Y_2(2X_2 - X_2 1) \\ 0 & 0 & 0 & 4(X_1^2 + X_2^2 + Y_2^2 - X_1 X2) & -4Y_2^2 \\ 0 & 0 & Y_2(2X_2 - X_2 1) & -4Y_2^2 & 4Y_2^2 \end{pmatrix}.$$

Also ist:

$$a_1 = 0,\ a_2 = 0,\ a_3 = (-1)\frac{a_5 Y_2(2X_2 - X_1)}{X_1^2 + X_2^2 + Y_2^2 - X_1X2},\ a_4 = \frac{a_5 Y_2^2}{X_1^2 + X_2^2 + Y_2^2 - X_1X2}$$

Mit $a_5 := X_1^2 + X_2^2 + Y_2^2 - X_1X2$ bleibt:

$$\begin{aligned}
\frac{a_5}{\lambda_T} &= \frac{1}{72}(-Y_2^2(2X_2 - X_1)^2 - 4Y_2^4 + 4Y_2^2(X_1^2 + X_2^2 + Y_2^2 - X_1X_2)) \\
&= \frac{1}{24}(Y_2^2 X_1^2) \\
\lambda_T &= 24\frac{(X_1^2 + X_2^2 - X_1X_2 + Y_2^2)}{Det(M_T)^2}.
\end{aligned}$$

Sei $X_0 = Y_0 = Y_1 = 0$ und $A = \frac{Det(M_T)}{2}$ der Flächeninhalt von T. Wir legen fest, dass C_{inv} der Größe $\lambda_{\hat{T}}$ auf dem Einheitssimplex $\hat{T}$ entsprechen soll und setzen $C_{inv} = 48$. Die Gitterzellengröße h_T ist:

$$h_T^2 = \frac{48}{\lambda_T} = \frac{8A^2}{(X_1^2 + X_2^2 - X_1X_2 + Y_2^2)}. \tag{26}$$

Um die gleiche Darstellung, wie in [HH92](80) zu erhalten, betrachten wir zunächst die Abstände der Ecken zum geometrischen Schwerpunkt von T:

$$\begin{aligned}
\begin{pmatrix} X_0 - c_x \\ Y_0 - c_y \end{pmatrix} &= \frac{1}{3}\begin{pmatrix} -X_1 - X_2 \\ -Y_2 \end{pmatrix}, \\
\begin{pmatrix} X_1 - c_x \\ Y_1 - c_y \end{pmatrix} &= \frac{1}{3}\begin{pmatrix} 2X_1 - X_2 \\ -Y_2 \end{pmatrix}, \\
\begin{pmatrix} X_2 - c_x \\ Y_2 - c_y \end{pmatrix} &= \frac{1}{3}\begin{pmatrix} -X_1 + 2X_2 \\ 2Y_2 \end{pmatrix}.
\end{aligned}$$

Jetzt können wir den Nenner des letzten Terms aus (26) folgendermaßen umschreiben:

$$\begin{aligned}
X_1^2 + X_2^2 - X_1X_2 &= \frac{1}{6}((4X_1^2 + X_2^2 - 4X_1X_2) \\
&\quad + (X_1^2 + 4X_2^2 - 4X_1X_2) + (X_1^2 + X_2^2 + 2X_1X_2)) \\
&= \frac{1}{6}\left((2X_1 - X_2)^2 + (X_1 - 2X_2)^2 + (X_1 + X_2)^2\right) \\
&= \frac{3}{2}\sum_i (X_i - c_x)^2 \\
Y_2^2 &= \frac{1}{6}(Y_2^2 + 4Y_2^2 + Y_2^2) \\
&= \frac{3}{2}\sum_i (Y_i - c_x)^2 \\
\Longrightarrow h_T &= \frac{4A}{\sqrt{3 \cdot \sum_i (X_i - c_x)^2 + (Y_i - c_y)^2}}.
\end{aligned} \tag{27}$$

Nun ist noch zu zeigen, dass h_T Bedingung 2a erfüllt.

Die Seitenlängen von T sind:

$$a = X_1 \qquad b = \sqrt{X_2^2 + Y_2^2} \qquad c = \sqrt{(X_1 - X_2)^2 + Y_2^2}.$$

Wir formen wieder den Nenner von (26) um:

$$\begin{aligned} X_1^2 + X_2^2 - X_1X_2 + Y_2^2 &= \frac{(X_1 - X_2)^2 + X_1^2 + X_2^2 + 2Y_2^2}{2} \\ &= \frac{a^2 + b^2 + c^2}{2}. \end{aligned}$$

Damit erhalten wir

$$h_T^2 = \frac{16A^2}{a^2 + b^2 + c^2} \geq \left(\frac{4A}{a + b + c}\right)^2 = 4\left(\frac{2A}{a + b + c}\right)^2.$$

Der Radius des Inkreises ist $\frac{2A}{a+b+c} = r_T$ (vgl. [KK07, S.179]).

Sei oBdA. $a = \max\{a, b, c\} = d_T$. Wir schreiben $h_a a = 2A \Rightarrow A^2 = \frac{a^2 h_a^2}{4}$. Es gilt $h_a \leq \min\{b, c\} \Rightarrow h_a \leq a$ und damit:

$$h_T^2 = \frac{16A^2}{a^2 + b^2 + c^2} \leq \frac{4a^2 h_a^2}{3a^2} \leq \frac{4}{3} d_T$$

Somit gilt:

$$2r_T \leq h_T \leq \frac{2}{\sqrt{3}} d_T. \tag{28}$$

Damit ist gezeigt, dass h_T tatsächlich eine Größe der Gitterzelle definiert.

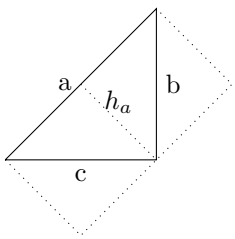

Abbildung 7: Zur Veranschaulichung $2A = ah_a$ und $h_a \leq \min\{b, c\}$

3.2.3 Rechtecke

Wir werden darauf verzichten erst die Konstante für Quadrate auszurechnen, da sie sich als Spezialfall für allgemeine Rechtecke ergibt. Wir nehmen zunächst an, dass

$$T = \left\{\boldsymbol{x} \in \mathbb{R}^2 : \quad -h_x \leq x_1 \leq h_x,\ -h_y \leq x_2 \leq h_y,\ h_x, h_y \in \mathbb{R} \setminus \{0\}\right\}.$$

Mit Satz 2.2 können wir dann die Aussagen auf beliebige Rechtecke verallgemeinern. Wir versuchen analog zur Vorgehensweise oben die Basis von $\mathcal{P}$ so zu wählen, dass alle harmonischen Polynome zuerst auftauchen. Sei $\mathcal{P}$ die lineare Hülle von

$$\left\{x_1,\ x_2,\ x_1x_2,\ x_1^2 - x_2^2,\ x_1x_2^2,\ x_1^2x_2,\ x_2^2,\ x_1^2x_2^2\right\}.$$

Damit erhalten wir

$$K_\Delta = h_x h_y \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{h_x^2}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{h_y^2}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & \frac{h_x^2+h_y^2}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{h_x^2+h_y^2}{3} & \frac{9h_x^4+9h_y^4+10h_x^2h_y^2}{180} \end{pmatrix},$$

$$K_\nabla = h_x h_y \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & \frac{h_y^2}{12} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \frac{h_x^2}{12} & 0 & 0 \\ 0 & 0 & \frac{h_x^2+h_y^2}{12} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{h_x^2+h_y^2}{3} & 0 & 0 & -\frac{h_y^2}{3} & 0 \\ \hline \frac{h_y^2}{12} & 0 & 0 & 0 & h_y^2\frac{9h_x^2+20h_y^2}{720} & 0 & 0 & 0 \\ 0 & \frac{h_x^2}{12} & 0 & 0 & 0 & h_x^2\frac{20h_x^2+9h_y^2}{720} & 0 & 0 \\ 0 & 0 & 0 & -\frac{h_y^2}{3} & 0 & 0 & \frac{h_y^2}{3} & \frac{h_x^2h_y^2}{36} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{h_x^2h_y^2}{36} & h_x^2h_y^2\frac{h_x^2+h_y^2}{240} \end{array}\right)$$

$$:= h_x h_y \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right).$$

Sei $w \in \mathcal{P}$ eine Lösung und $\underline{w} \in \mathbb{R}^8$ definiert wie oben. Wegen Satz 2.1 gilt:

$$\begin{pmatrix} A & B \end{pmatrix} \underline{w} = 0$$

$$\Longrightarrow \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{h_x^2+h_y^2}{12} & 0 \\ 0 & 0 & 0 & \frac{h_x^2+h_y^2}{3} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} -\frac{h_y^2}{12} & 0 & 0 & 0 \\ 0 & -\frac{h_x^2}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{h_y^2}{3} & 0 \end{pmatrix} \begin{pmatrix} w_5 \\ w_6 \\ w_7 \\ w_8 \end{pmatrix}$$

$$\Longrightarrow \qquad \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} -\frac{h_y^2}{12} w_5 \\ -\frac{h_x^2}{12} w_6 \\ 0 \\ \frac{h_y^2}{h_x^2+h_y^2} w_7 \end{pmatrix}.$$

Damit können wir die Anzahl der Variablen auf 4 verringern und definieren $\mathcal{P}$ als die lineare Hülle von

$$\left\{ x_1 x_2^2 - \frac{h_y^2 x_1}{12},\ x_1^2 x_2 - \frac{h_x^2 x_2}{12},\ x_2^2 + \frac{h_y^2(x_1^2 - x_2^2)}{h_x^2+h_y^2},\ x_1^2 x_2^2 \right\}.$$

Nun müssen wir nur noch das folgende verallgemeinerte Eigenwertproblem lösen: finde $\underline{\tilde{w}} \in \mathbb{R}^4$, sodass

$$\begin{pmatrix} \frac{h_x^2}{3} & 0 & 0 & 0 \\ 0 & \frac{h_y^2}{3} & 0 & 0 \\ 0 & 0 & 4 & \frac{h_x^2+h_y^2}{3} \\ 0 & 0 & \frac{h_x^2+h_y^2}{3} & \frac{9h_x^4+9h_y^4+10h_x^2h_y^2}{180} \end{pmatrix} \underline{\tilde{w}}$$

$$= \lambda \begin{pmatrix} \frac{h_y^2(5h_x^2+h_y^2)}{180} & 0 & 0 & 0 \\ 0 & \frac{h_x^2(h_x^2+5h_y^2)}{180} & 0 & 0 \\ 0 & 0 & \frac{h_x^2h_y^2}{3(h_x^2+h_y^2)} & \frac{h_x^2h_y^2}{36} \\ 0 & 0 & \frac{h_x^2h_y^2}{36} & \frac{h_x^2h_y^2(h_x^2+h_y^2)}{240} \end{pmatrix} \underline{\tilde{w}}.$$

Zwei Eigenwerte lassen sich sofort ablesen: $\frac{60h_x^2}{h_y^2(5h_x^2+h_y^2)}$ und $\frac{60h_y^2}{h_x^2(h_x^2+5h_y^2)}$. Um die anderen beiden Eigenwerte zu finden berechnen wir die Nullstellen in $\lambda \in \mathbb{R}$,

für:

$$\det\begin{pmatrix} 4-\lambda\dfrac{h_x^2h_y^2}{3(h_x^2+h_y^2)} & \dfrac{h_x^2+h_y^2}{3}-\lambda\dfrac{h_x^2h_y^2}{36} \\ \dfrac{h_x^2+h_y^2}{3}-\lambda\dfrac{h_x^2h_y^2}{36} & \dfrac{9h_x^4+9h_y^4+10h_x^2h_y^2}{180}-\lambda\dfrac{h_x^2h_y^2(h_x^2+h_y^2)}{240} \end{pmatrix}$$

$$= \frac{(h_x^6h_y^4+h_x^4h_y^6)\lambda^2-24(h_x^6h_y^2+h_x^4h_y^4+h_x^2h_y^6)\lambda+144(h_x^6+h_x^4h_y^2+h_x^2h_y^4+h_y^6)}{1620(h_x^2+h_y^2)}.$$

Wir teilen durch $\dfrac{(h_x^6h_y^4+h_x^4h_y^6)}{1620(h_x^2+h_y^2)}$ und erhalten die quadratische Gleichung:

$$0=\lambda^2-24\frac{h_x^4+h_x^2h_y^2+h_y^4}{h_x^2h_y^2(h_x^2+h_y^2)}\lambda+144\frac{h_x^4+h_y^4}{h_x^4h_y^4},$$

mit den Nullstellen:

$$12\frac{h_x^4+h_y^4}{h_x^2h_y^2(h_x^2+h_y^2)}\leq 12\frac{h_x^2+h_y^2}{h_x^2h_y^2}.$$

Für die anderen Eigenwerte machen wir die Abschätzungen:

$$\begin{aligned}\frac{60h_x^2}{h_y^2(5h_x^2+h_y^2)}&\leq 12\frac{h_x^2}{h_x^2h_y^2}\leq 12\frac{h_x^2+h_y^2}{h_x^2h_y^2},\\ \frac{60h_y^2}{h_x^2(h_x^2+5h_y^2)}&\leq 12\frac{h_y^2}{h_x^2h_y^2}\leq 12\frac{h_x^2+h_y^2}{h_x^2h_y^2}.\end{aligned}$$

Somit erhalten wir schließlich

$$\lambda_T=12\frac{h_x^2+h_y^2}{h_x^2h_y^2}. \tag{29}$$

Nun definieren wir wie in [HH92, (94)]:

$$h_T:=h_xh_y\sqrt{\frac{2}{h_x^2+h_y^2}}\Longrightarrow C_{inv}=24. \tag{30}$$

Es ist leicht zu zeigen, dass die Definition von h_T Bedingung 2a erfüllt:

$$\begin{aligned}2r_T=\min\{h_x,h_y\}&=h_xh_y\sqrt{\frac{2}{2\max\{h_x,h_y\}^2}}\\ &\leq h_T\leq\frac{h_xh_y}{\sqrt{h_xh_y}}=\sqrt{h_xh_y}\leq\sqrt{\frac{h_x^2+h_y^2}{2}}\leq\frac{d_T}{\sqrt{2}}.\end{aligned}$$

Diese Abschätzungen sind scharf für Quadrate.

3.3 Beobachtungen zur Gitterzellengröße

Abhängigkeit von r_T Sowohl für Drei-, als auch für Rechtecke lässt sich h_T auch nach oben gegen r_T abschätzen. Für Dreiecke gilt mit:

$$\begin{aligned} a^2+b^2+c^2 &= \frac{1}{3}\left(a^2+b^2+c^2+\underbrace{(a^2+b^2)}_{\geq 2ab}+\underbrace{(a^2+c^2)}_{\geq 2ac}+\underbrace{(b^2+c^2)}_{\geq 2bc}\right) \\ &\geq \frac{(a+b+c)^2}{3}, \end{aligned}$$

dass

$$\begin{aligned} h_T &= \frac{4A}{\sqrt{a^2+b^2+c^2}} \\ &\leq \sqrt{12}\ \frac{2A}{a+b+c} \\ &= \sqrt{12}\ r_T. \end{aligned}$$

Für Rechtecke gilt:

$$\begin{aligned} h_T &= h_x h_y \sqrt{\frac{2}{h_x^2+h_y^2}} \\ &\leq h_x h_y \sqrt{\frac{2}{\max\{h_x,h_y\}^2}} \\ &= 2\sqrt{2}\ \frac{\min\{h_x,h_y\}}{2} = \sqrt{8}\ r_T. \end{aligned}$$

Es ist also

$$h_T = O(r_T), \tag{31}$$

unabhängig von der Konstante ρ, aus Definition 1.9.

Zusammenhang der verschiedenen Charakterisierungen Für Dreiecke haben wir verschiedene Charakterisierungen gefunden, die entweder die Quadrate der Seitenlängen, oder die quadratischen Abstände der Eckpunkte zum geometrischen Schwerpunkt involvieren. Diese Darstellungen können wir auch für Rechtecke formulieren.

Für ein Rechteck T mit den Eckpunkten $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$ für $i = 1,\dots 4$, definieren wir A als den Flächeninhalt und $c = (c_x, c_y)$ als den geometrischen Schwerpunkt. Mit $h_x/2 = |X_i - c_x|$ und $h_y/2 = |Y_i - c_y|$ für alle i erhalten wir:

$$\begin{aligned} h_T &= \frac{\sqrt{2}A}{\sqrt{\frac{1}{4}\left(\sum_i (2(X_i-c_x))^2 + (2(Y_i-c_y))^2\right)}} \\ &= \frac{\sqrt{2}A}{\sqrt{\left(\sum_i (X_i-c_x)^2 + (Y_i-c_y)^2\right)}}. \end{aligned} \tag{32}$$

Die Seitenlängen des Rechtecks sind $a = c = h_x$ und $b = d = h_y$ und somit können wir schreiben:

$$h_T = \frac{2A}{\sqrt{a^2+b^2+c^2+d^2}}. \tag{33}$$

4 Ein numerisches Experiment

Wir wollen überprüfen, ob sich die in [HH92] gefundenen Definition von h_T für biquadratische Polynome auf Rechtecken auf Parallelogramme übertragen lässt. Seien $C_{inv} := 24$,

$$\begin{aligned}
\hat{T} &= \{\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2)^T : \ -1/2 \leq x_j \leq 1/2, \ j = 1, 2\}, \\
M_T &= \begin{pmatrix} b & b_1 \\ 0 & h_b \end{pmatrix}, \\
T &= \left\{ \boldsymbol{x} \in \mathbb{R}^2 : \ M_T \hat{\boldsymbol{x}} + t, \ t \in \mathbb{R}^2, \ \hat{\boldsymbol{x}} \in \hat{T} \right\} \\
\text{und } \mathcal{P} &= \text{span}\left\{ \hat{x}_2, \hat{x}_2^2, \hat{x}_1, \hat{x}_1\hat{x}_2, \hat{x}_1\hat{x}_2^2, \hat{x}_1^2, \hat{x}_1^2\hat{x}_2, \hat{x}_1^2\hat{x}_2^2 \right\}.
\end{aligned}$$

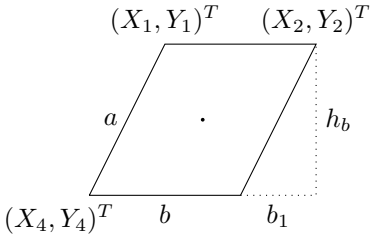

Abbildung 8: Parallelogramm

Wir definieren die Gitterzellengröße h_T als:

$$h_T^2 := \frac{2h_a^2 h_b^2}{h_a^2 + h_b^2}, \tag{34}$$

mit $h_b = A/b$ und $h_a = A/a$. Die Seitenlänge a ist gegeben durch $a = \sqrt{h_b^2 + b_1^2}$ und der Flächeninhalt ist $A = bh_b$. Somit erhalten wir:

$$h_T^2 = \frac{2A^4/(a^2b^2)}{A^2/a^2 + A^2/b^2} = \underbrace{\frac{2A^2}{a^2 + b^2}}_{:=(*)} = \frac{2b^2h_b^2}{h_b^2 + b_1^2 + b^2}$$

$$\Longrightarrow \frac{24}{h_T^2} = \frac{12(h_b^2 + b_1^2 + b^2)}{b^2h_b^2}$$

Diese Definition ist konsistent zu den Charakterisierungen der Gitterzellengröße von Rechtecken, die in Abschnitt 3.3 erläutert wurden:

Der Term $(*)$ impliziert, dass die Definition konsistent mit der Charakterisierung (33) ist, die die Seitenlängen von T involviert.

Für die quadratischen Abstände der Eckpunkte zum Mittelpunkt gilt:

$$\begin{aligned}
\sum_i (X_i - c_x)^2 + (Y_i - c_y)^2 &= 2(b/2 + b_1/2)^2 + 2(b/2 - b_1/2)^2 + h_b^2 \\
&= b^2 + b_1^2 + h_b^2.
\end{aligned}$$

Also ist die Definition auch konsistent zu Chahrakterisierung (32).

Wir vermuten

$$\lambda_T = \sup_{v \in \mathcal{P}} \frac{\|\nabla v\|_{0,T}^2}{\|\Delta v\|_{0,T}^2} = \frac{C_{inv}}{h_T^2}.$$

Wir plotten für $0.0001 \leq b_1 \leq 150$, und $0.01 \leq b \leq 150$, $h_b = 1$ und numerisch ermitteltem λ_T

$$\lambda_T - \frac{C_{inv}}{h_T^2} = \lambda_T - \frac{12(b^2 + b_1^2 + h_b^2)}{b^2 h_b^2}.$$

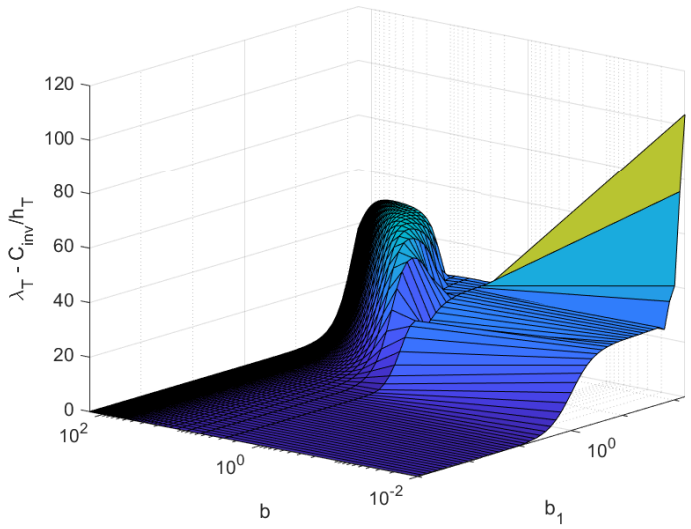

Abbildung 9: $\lambda_T - \frac{24}{h_T^2}$

Die Matrizen K_Δ und K_∇ wurden mit der Symbolic Math Toolbox von MATLAB als symbolischer Term erzeugt und dann mittels des Befehls `matlabFunction` in eine MATLAB Funktion umgewandelt. Die λ_T wurden mit der eingebauten Funktion `eig` berechnet.

Folgerung Die Definition der Gitterzellengröße h_T aus [HH92] lässt sich nicht auf Parallelogramme verallgemeinern. Für $b_1 \approx 0$ ist der Fehler erwartungsgemäß klein. Bei größerer Verzerrung b_1, wird auch der Fehler größer. Man beachte, dass hier nicht der absolute Betrag des Fehlers dargestellt ist, sondern die numerisch ermittelte Näherung des exakten Wertes minus den Wert des vereinfachten Modells. Insbesondere ist in dem betrachteten Bereich

$$\lambda_T \geq \frac{C_{inv}}{h_T^2}$$

und somit kann $\frac{C_{inv}}{h_T^2}$ nicht als Näherung für λ_T verwendet werden.

Bemerkung 4.1. *Es ist nicht ausgeschlossen, dass es eine andere, als die oben gewählte Definition der Gitterzellengröße gibt, mit der sich die Ergebnisse von Harari und Hughes auf Parallelogramme übertragen lassen.*

5 Eine andere inverse Ungleichung

Im Fall $d = 1$ kann das Problem, wie in Abschnitt 3.1 auf die Ungleichung

$$\|\nabla v\|_{0,T} \leq \lambda_T \|v\|_{0,T} \qquad \forall v \in \mathcal{P} \tag{35}$$

reduziert werden. Wir untersuchen, ob das auch im zweidimensionalen Fall möglich ist.

In [ÖRW10] wird die folgende Abschätzung für allgemeine Dreiecke bewiesen:

Sei T ein allgemeines Dreieck mit Umfang U und Flächeninhalt A. Weiterhin seien $\hat{T}$ der Einheitssimplex,

$$C_k := \sup_{\hat{v} \in P_k(\hat{T})} \frac{\left\| \frac{\partial \hat{v}}{\partial \hat{x}_1} \right\|_{0,\hat{T}}^2}{\|\hat{v}\|_{0,\hat{T}}^2},$$

dann gilt:

$$\|\nabla v\|_{0,T} \leq \sqrt{C_k} \frac{U}{A} \|v\|_{0,T}, \tag{36}$$

mit

$$C_1 = 3 \qquad C_2 = 15 \qquad C_3 = \frac{45 + \sqrt{1605}}{2} \approx 42.5312.$$

In dem Artikel sind Werte für $k = 1, \ldots, 10$ angegeben.

Wir zeigen, dass eine ähnliche Aussage auch für Parallelogramme gilt.

Sei:

$$\tilde{C}_k := \sup_{\hat{v} \in P_k(\hat{T})} \frac{\left\| \frac{\partial \hat{v}}{\partial \hat{x}_1} \right\|_{0,\hat{T}}^2}{\|\hat{v}\|_{0,\hat{T}}^2} = \sup_{\hat{v} \in P_k(\hat{T})} \frac{\left\| \frac{\partial \hat{v}}{\partial \hat{x}_2} \right\|_{0,\hat{T}}^2}{\|\hat{v}\|_{0,\hat{T}}^2}.$$

Die Gleichheit gilt aus Symmetriegründen.
Seien T ein Parallelogram mit Referenzelement $\hat{T}$ und

$$M_T = \begin{pmatrix} b & b_1 \\ 0 & h_b \end{pmatrix} \qquad \Longrightarrow M_T^{-T} = \frac{1}{A} \begin{pmatrix} h_b & 0 \\ -b_1 & b \end{pmatrix}$$

definiert, wie in Abschnitt 4. Mit Satz 2.2 und $U = 2\left(b + \sqrt{h_b^2 + b_1^2}\right)$ folgt:

$$\begin{aligned} \|\nabla v\|_{0,T} &= \left|\det\left(M_T^{-T}\right)\right|^{1/2} \|M_T^{-T} \nabla_{\hat{T}} \hat{v}\|_{0,\hat{T}} \\ &= \frac{\left|\det\left(M_T^{-T}\right)\right|^{1/2}}{A} \left\| \begin{pmatrix} h_b \\ -b_1 \end{pmatrix} \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} + \begin{pmatrix} 0 \\ b \end{pmatrix} \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} \right\|_{0,\hat{T}} \end{aligned}$$

Anwendung der Dreiecksungleichung liefert

$$
\begin{aligned}
&\leq \frac{\left|\det\left(M_T^{-T}\right)\right|^{1/2}}{A} \left\| \begin{pmatrix} h_b \\ -b_1 \end{pmatrix} \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} \right\|_{0,\hat{T}} + \left\| \begin{pmatrix} 0 \\ b \end{pmatrix} \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} \right\|_{0,\hat{T}} \\
&= \frac{\left|\det\left(M_T^{-T}\right)\right|^{1/2}}{A} \left(\left\| \begin{pmatrix} h_b \\ -b_1 \end{pmatrix} \right\| \left\| \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} \right\|_{0,\hat{T}} + \left\| \begin{pmatrix} 0 \\ b \end{pmatrix} \right\| \left\| \frac{\partial \hat{v}(\hat{\boldsymbol{x}})}{\partial \hat{x}_1} \right\|_{0,\hat{T}} \right) \\
&\leq \sqrt{\tilde{C}_k} \frac{\left|\det\left(M_T^{-T}\right)\right|^{1/2}}{A} \left(\left\| \begin{pmatrix} h_b \\ -b_1 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 \\ b \end{pmatrix} \right\| \right) \|\hat{v}\|_{0,\hat{T}} \\
&= \frac{\sqrt{\tilde{C}_k}}{A} \left(\left\| \begin{pmatrix} h_b \\ -b_1 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 \\ b \end{pmatrix} \right\| \right) \|v\|_{0,T} \\
&= \frac{\sqrt{h_b^2 + b_1^2} + b}{A} \sqrt{\tilde{C}_k} \|v\|_{0,T} \\
&= \frac{U}{2A} \sqrt{\tilde{C}_k} \|v\|_{0,T}.
\end{aligned}
$$

Insgesamt erhalten wir:

$$
\|\nabla v\|_{0,T}^2 \leq \frac{U^2}{4A^2} \tilde{C}_k \|v\|_{0,T}^2 \qquad \forall v \in \mathcal{P}. \tag{37}
$$

Wir sehen, dass sich die Aussage (36) aus [ÖRW10] mit dem Vorfaktor 1/2 (bzw. 1/4 für das Quadrat) auf Parallelogramme übertragen lässt.

Wir können nun (37) nutzen, um λ_T für Parallelogramme abzuschätzen:

$$
\begin{aligned}
\Delta v(\boldsymbol{x})^2 &= \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} + \frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2} \right)^2 \\
&= \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2} \right)^2 + \underbrace{2 \frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} \frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2}}_{\leq \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2} \right)^2} \\
&\leq 2 \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 v(\boldsymbol{x})}{\partial x_2^2} \right)^2 \\
&\leq 2 \left\| \nabla \frac{\partial v(\boldsymbol{x})}{\partial x_1} \right\|^2 + 2 \left\| \nabla \frac{\partial v(\boldsymbol{x})}{\partial x_2} \right\|^2 \\
\implies \|\Delta v\|_{0,T}^2 &\leq 2 \left\| \nabla \frac{\partial v}{\partial x_1} \right\|_{0,T}^2 + 2 \left\| \nabla \frac{\partial v}{\partial x_2} \right\|_{0,T}^2.
\end{aligned}
$$

Aus (37) folgt nun

$$
\|\Delta v\|_{0,T}^2 \leq \tilde{C}_k \frac{U^2}{2A^2} \left(\left\| \frac{\partial v}{\partial x_1} \right\|_{0,T}^2 + \left\| \frac{\partial v}{\partial x_2} \right\|_{0,T}^2 \right) = \tilde{C}_k \frac{U^2}{2A^2} \|\nabla v\|_{0,T}^2. \tag{38}
$$

In [KNR16] werden asymptotische Abschätzungen für $\sup_{\hat{v} \in P_k(\hat{T})} \frac{\left\| \frac{\partial \hat{v}}{\partial \hat{x}_1} \right\|_{0,\hat{T}}^2}{\|\hat{v}\|_{0,\hat{T}}^2}$ für beliebige k auf dem Referenzelement $\hat{T} = (-1,1)^2$ gegeben. Wir geben hier nur einige numerisch ermittelten Werte für $\tilde{C}_k$, $k = 1,2,3$ an.

$$
\tilde{C}_1 = 12 \qquad \tilde{C}_2 = 60 \qquad \tilde{C}_3 \approx 170.1249. \tag{39}
$$

Für Dreiecke gilt mit derselben Begründung:

$$\|\Delta v\|_{0,T}^2 \leq C_k \frac{2U^2}{A^2} \|\nabla v\|_{0,T}^2. \tag{40}$$

Bemerkung 5.1. *Wir hatten bereits in Abschnitt 3.2.2 gezeigt, dass für die in [HH92] gefundene Definition von* h_T

$$h_T = O(r_T) = O\left(\frac{A}{U}\right)$$

gilt. Folglich gilt für $k = 2$ *auch*

$$\lambda_T = \frac{C_{inv}}{h_T^2} = O\left(C_k \frac{U^2}{A^2}\right).$$

Dadurch ist gezeigt, dass der Ansatz $\lambda_T = \frac{C_{inv}}{h_T^2}$ *für allgemeinere Fälle sinvoll ist.*

6 Fazit

Wir haben (1) als verallgemeinertes Eigenwertproblem umformuliert, das Matrizen mit symbolischen Einträgen involviert. Damit können optimale Abschätzungen für λ_T gefunden werden, die mit den Ergebnissen in [HH92] übereinstimmen. Weiterhin wurde in Abschnitt 4 gezeigt, dass sich die optimalen Ergebnisse für Rechtecke nicht ohne weiteres auf Parallelogramme übertragen lassen.

In Abschnitt 3 wurde gezeigt, dass die Definitionen von h_T aus [HH92] tatsächlich jeweils eine Größe der Gitterzelle definieren und dass sie sich in die Konvergenztheorie der Finite-Elemente Analyse integrieren lassen. Die Ungleichungen (38) und (40) zeigen außerdem, dass ein Ansatz wie in (2b) in allgemeineren Fällen gerechtfertigt ist.

Im eindimensionalen Fall kann (1) auf das Problem (35) reduziert werden. Im zweidimensionalen Fall kann (35) herangezogen werden, um gröbere Abschätzungen für (1) zu finden. Die Ergebnisse für (35) für allgemeine Dreiecke aus [ÖRW10] konnten wir auf Parallelogramme übertragen.

Quellen

[HH92] Isaac Harari and Thomas J.R. Hughes. "What are C and h?: Inequalities for the analysis and design of finite element methods". In: *Computer Methods in Applied Mechanics and Engineering* 97.2 (1992), pp. 157–192.

[Hir06] C. Hirsch. *Numerical Computation of Internal and External Flows*. Butterworth-Heinemann Limited, 2006. ISBN: 9780750665957.

[KK07] Max Koecher and Aloys Krieg. "Das Dreieck und seine Kreise". In: *Ebene Geometrie*. Springer Berlin Heidelberg, 2007, pp. 143–196. ISBN: 978-3-540-49328-0. URL: https://doi.org/10.1007/978-3-540-49328-0_5.

[BS08] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. New York, NY: Springer New York, 2008.

[ST08] Roos H.G.and M. Stynes and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion-Reaction and Flow Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. DOI: 10.1007/978-3-540-34467-4_1.

[ÖRW10] Sevtap Özısık, Beatrice Riviere, and Tim Warburton. "On the constants in inverse inequalities in L2". In: *CAAM Technical Reports* (2010). URL: https://hdl.handle.net/1911/102161.

[Alt12] Hans Wilhelm Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[CZ13] Shaochun Chen and Jikun Zhao. "Estimations of Constants in Inverse Inequalities for Finite Element Functions". In: *Journal of Computational Mathematics* 31.5 (2013), pp. 522–531.

[Bjö15] Åke Björck. "Direct Methods for Linear Systems". In: *Numerical Methods in Matrix Computations*. Cham: Springer International Publishing, 2015, pp. 1–209. DOI: 10.1007/978-3-319-05089-8_1.

[KNR16] Christoph Koutschan, Martin Neumüller, and Cristian-Silviu Radu. "Inverse inequality estimates with symbolic computation". In: *Advances in Applied Mathematics* 80 (2016), pp. 1–23. ISSN: 0196-8858. DOI: https://doi.org/10.1016/j.aam.2016.04.005.

[t H17] Karel in 't Hout. "Partial Differential Equations". In: *Numerical Partial Differential Equations in Finance Explained: An Introduction to Computational Finance*. London: Palgrave Macmillan UK, 2017, pp. 9–14. DOI: 10.1057/978-1-137-43569-9_2.

[Kno21] Florian Knorn. *M-code LaTeX Package version 2.7.0.0 Retrieved December 31, 2021.* MATLAB Central File Exchange, 2021. URL: https://www.mathworks.com/matlabcentral/fileexchange/8015-m-code-latex-package.

[MAT21] MATLAB. *version: 9.11 (R2021b)*. Natick, Massachusetts: The MathWorks Inc., 2021. URL: https://de.mathworks.com/help/matlab/ref/eig.html.

Anhang

MATLAB-Skripte Das Layout der hier abgebildeten Skripte wurde mit dem mcode-package [Kno21] erstellt.

MATLAB-Code zu Abbildung 9 in Absatz 4:

```
syms x y
syms x_t y_t b h_b b_1

x_t = b*x + b_1*y;
y_t = h_b*y;

inv_M_T = 1/(h_b*b)*[h_b, -b_1; 0, b];

base  = [y_t, y_t^2, x_t, x_t*y_t, x_t*y_t^2, x_t^2, x_t^2*y_t, ...
    x_t^2*y_t^2];

base_nabla = transpose(inv_M_T)*cat(1,diff(base, x), diff(base,y));
base_Delta = 1/(b^2*h_b^2)*((b_1^2 + h_b^2)*diff(base, x,2) + ...
    b^2*diff(base, y,2) - 2*b*b_1*diff(base, x,y));

K_nabla(b, h_b, b_1) = Int_T(transpose(base_nabla)*base_nabla);
K_nabla = matlabFunction(K_nabla);

K_Delta(b, h_b, b_1) = Int_T(transpose(base_Delta)*base_Delta);
K_Delta = matlabFunction(K_Delta);

X = 10.^(-3:0.1:2.2);
Y = linspace(0.01,150,length(X));
Z = zeros(length(X), length(X));


val_h_b = 1;

for j = 1:length(Y)
    val_b = Y(j);
    for k = 1:length(X)
        val_b_1 = X(k);
        approx = 12*(val_b^2 + val_h_b^2 + ...
            val_b_1^2)/(val_b^2*val_h_b^2);
        ex = max(eig(K_Delta(val_b, val_h_b, ...
            val_b_1),K_nabla(val_b, val_h_b, val_b_1)));
        Z(j,k) = ex - approx;
    end
end


surf(X,Y,Z);
xlabel('b_1')
ylabel('b')
set(gca, 'xscale', 'log')
set(gca, 'yscale', 'log')
zlabel('\lambda_T - C_{inv}/h_T')


function Int_T = Int_T(f)
    syms x y
    Int_T = int(int(f,y,-1/2,1/2),x,-1/2,1/2);
end
```

Matlab-Code zur Berechnung der Konstanten (39) in Absatz 5:

```
syms x y

base_1 = [1, y, x, x*y];
base_2 = [1, y, y^2, x, x*y, x*y^2, x^2, x^2*y, x^2*y^2];
base_3 = [1, y, y^2, y^3, x, x*y, x*y^2, x*y^3, x^2, x^2*y, ...
    x^2*y^2, x^2*y^3, x^3, x^3*y, x^3*y^2, x^3*y^3];


lam_1 = lambda_T(base_1);
lam_2 = lambda_T(base_2);
lam_3 = lambda_T(base_3);

function lambda_T = lambda_T(base)
    syms x y
    base_dx = diff(base, x);

    K_Id = Int_T(transpose(base)*base);
    K_Id = double(K_Id);

    K_dx = Int_T(transpose(base_dx)*base_dx);
    K_dx = double(K_dx);

    lambda_T = max(eig(K_dx, K_Id));
end


function Int_T = Int_T(f)
    syms x y
    Int_T = int(int(f,x,-1/2,1/2),y,-1/2,1/2);
end
```