TECHNISCHE
UNIVERSITÄT
WIEN
Vienna | Austria

Bachelorarbeit

Stabilität der L^2-Orthogonalprojektion im Sobolev-Raum H^1_D und in lokal gewichteten L^2-Räumen

Ausgeführt am Institut für
Analysis und Scientific Computing
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Dipl.Math. Dr.techn. Dirk Praetorius
Dipl.-Ing. Dr.techn. Michael Karkulik

durch
Carl-Martin Pfeiler
Pernerstorfergasse 42
1100 Wien

Wien, 28. September 2015

Inhaltsverzeichnis

1 Einleitung

1.1 Zur adaptiven FEM

Die grundlegende Vorgangsweise bei adaptiver FEM ist es, zunächst die diskrete Lösung auf dem aktuellen Gitter $\mathcal{T}$ zu berechnen. Dann wird auf jedem Element ein a posteriori Fehlerschätzer berechnet, und jene Elemente mit den größten Fehlerschätzern werden zur Verfeinerung markiert. Man erhält eine Menge $\mathcal{M} \subset \mathcal{T}$ markierter Elemente. Zuletzt wird nun ein neues Gitter $\mathcal{T}'$ durch Verfeinerung einiger Elemente erzeugt. Im Allgemeinen werden zusätzlich zu den markierten Elementen $\mathcal{M}$ noch weitere Elemente verfeinert. Dies wird als Vervollständigung der Partition bezeichnet und geschieht, um strukturelle Eigenschaften der Partition zu erhalten, die für die Fehlerschätzung in der nächsten Iteration notwendig sind.

1.2 Die L^2-Orthogonalprojektion

Für eine Partition $\mathcal{T}$ eines Gebietes $\Omega \subset \mathbb{R}^d$ definiere den Raum der global stetigen und stückweise affinen Funktionen

$$\mathcal{S}^1_D(\mathcal{T}) := \{V \in C(\Omega) \mid V|_{\Gamma_D} = 0, \forall T \in \mathcal{T} : V|_T \text{ ist affin}\}\,, \tag{1}$$

wobei $\Gamma_D \subset \partial\Omega$ den Dirichlet-Rand von Ω bezeichnet, und $\Gamma := \partial\Omega \setminus \overline{\Gamma}_D$. Wir betrachten die L^2-Orthogonalprojektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$, welche eindeutig definiert ist durch

$$\int_\Omega (v - \Pi_D(\mathcal{T})v)V\,dx = 0 \quad \text{für alle } v \in L^2(\Omega) \text{ und } V \in \mathcal{S}^1_D(\mathcal{T})\,. \tag{2}$$

Wir untersuchen H^1_D-Stabilität von $\Pi_D(\mathcal{T})$. Genauer gesagt interessieren wir uns dafür, unter welchen Voraussetzungen

$$\|\Pi_D(\mathcal{T})v\|_{H^1(\Omega)} \le C_1\|v\|_{H^1(\Omega)} \quad \text{für alle } v \in H^1_D(\Omega) \tag{3}$$

mit einer Konstanten $C_1 > 0$ gilt.
Aus der Funktionalanalysis ist bekannt, dass $\Pi_D(\mathcal{T})$ eine stetige Projektion bezüglich der L^2-Norm mit Operatornorm 1 ist. Es reicht demnach zu überprüfen, ob

$$\|\nabla\Pi_D(\mathcal{T})v\|_{L^2(\Omega)} \le C_2\|\nabla v\|_{L^2(\Omega)} \quad \text{für alle } v \in H^1_D(\Omega) \tag{4}$$

mit einer Konstanten $C_2 > 0$ gilt.
Bevor wir uns der Überprüfung dieser Bedingung widmen, wollen wir eine Netzverfeinerungsstrategie in beliebiger Raumdimension d präsentieren.

2 Newest Vertex Bisection in $\mathbb{R}^D$

In 2D ist ***Newest Vertex Bisection (NVB)*** eine populäre Netzverfeinerungsstrategie, welche die Grundlage für adaptive ***Finite Elemente Methoden (FEM)*** bildet. In diesem Kapitel wird nach der Vorlage von [Ste08] eine Netzverfeinerungsstrategie für beliebige Raumdimension $d \ge 2$ vorgestellt. Da diese Strategie für $d = 2$ mit NVB in 2D zusammenfällt, werden wir sie ebenfalls mit NVB bezeichnen.

2.1 Definitionen

Sei $2 \leq d \leq D$. Ein d-***Simplex*** T in $\mathbb{R}^D$ ist die konvexe Hülle $T = \operatorname{conv}\{x_0, ..., x_d\}$ von $d+1$ ***Knoten*** $x_0, ..., x_d \in \mathbb{R}^D$, welche nicht auf einer $(d-1)$-dimensionalen Hyperebene liegen.

Definition 2.1. Sei $T = \operatorname{conv}\{x_0, \ldots, x_d\}$ ein d-Simplex in $\mathbb{R}^D$. Mit

$$\mathcal{N}(T) := \{x_0, \ldots, x_d\}$$

bezeichnen wir die Menge der Knoten von T. Sei $\{y_0, \ldots, y_n\} \subset \mathcal{N}(T)$, $0 \leq n < d$. Wir nennen die konvexe Hülle $\operatorname{conv}\{y_0, \ldots, y_n\}$ eine n-***dimensionale Hyperfläche*** von T. Falls $n = 1$, bezeichnen wir diese eindimensionale Hyperfläche E als ***Kante*** von T. Falls $n = d-1$ bezeichnen wir diese $(d-1)$-dimensionale Hyperfläche F als ***Seite*** von T. Wir definieren die Mengen:

$$\begin{aligned} \mathcal{E}(T) &:= \{E \subset T \mid E \text{ ist eine Kante von } T\} \\ \mathcal{F}(T) &:= \{F \subset T \mid F \text{ ist eine Seite von } T\}\ . \end{aligned}$$

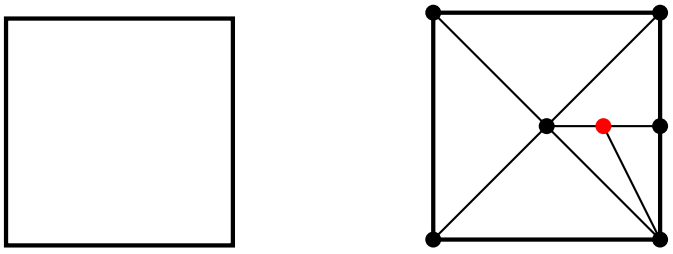
Eine endliche Menge $\mathcal{T}$ von d-Simplizes in $\mathbb{R}^D$ wird ***Partition*** einer d-dimensionalen Mannigfaltigkeit $\Omega \subset \mathbb{R}^D$ genannt, falls $\overline{\Omega} = \bigcup \{T \mid T \in \mathcal{T}\}$ und der Durchschnitt zweier verschiedener Simplizes $T, T' \in \mathcal{T}$ stets d-dimensionales Maß 0 hat. Normalerweise wird eine Partition $\mathcal{T}$ ***konform*** genannt, falls der Durchschnitt zweier verschiedener Simplizes $T, T' \in \mathcal{T}$ entweder leer, oder eine gemeinsame Hyperfläche von T und T' ist. Im Fall, dass Ω auf beiden Seiten eines $(d-1)$-dimensionalen Teils seines Randes liegt, ist diese Bedingung jedoch unnötig einschränkend, siehe Abbildungen 1 und 2 für eine Illustration. Anstatt dessen nennen wir $\mathcal{T}$ konform, wenn Folgendes gilt:

(C1) Für alle $T \in \mathcal{T}$ ist $\overline{\Gamma_D \cap T}$ die Vereinigung von Hyperflächen von T. Für alle $T \in \mathcal{T}$ ist $\overline{\Gamma \cap T}$ die Vereinigung von Hyperflächen von T.

(C2) Seien $T, T' \in \mathcal{T}$. Jedes $x \in T \cap T'$, für welches für jede offene Kugel B mit $x \in B$ alle $y \in T \cap B \cap \Omega$, $y' \in T' \cap B \cap \Omega$ durch einen Weg durch $B \cap \Omega$ verbunden sind, liegt auf einer gemeinsamen Hyperfläche von T und T'.

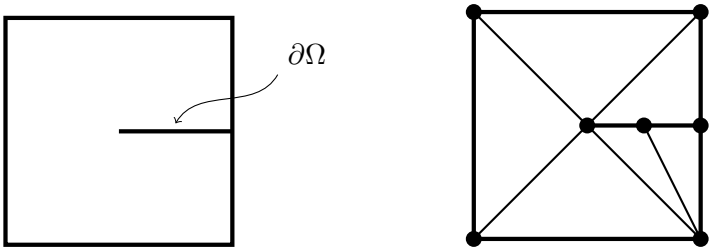
Diese Definition ist etwas unhandlich. Der folgende Satz vereinfacht die Beantwortung der Frage nach der Konformität einer Partition. Verschiedene $T, T' \in \mathcal{T}$ die eine gemeinsame Seite $F = T \cap T'$ haben und für die $T \cap T' \cap \Omega$ nichtleer ist, werden ***Nachbarn*** genannt.

Satz 2.2 ([Ste08, Theorem 3.2.])**.** *Eine Partition erfüllt **(C2)** genau dann, wenn zwei verschiedene $T, T' \in \mathcal{T}$, für die $T \cap T' \cap \Omega$ einen inneren Punkt einer Seite von T (oder T') enthält, Nachbarn sind.* □



(a) Gebiet Ω (b) Partition $\mathcal{T}$ auf Ω

Abbildung 1: $\mathcal{T}$ ist keine konforme Partition auf $\Omega = [0,1] \times [0,1]$.



(a) Gebiet Ω (b) Partition $\mathcal{T}$ auf Ω

Abbildung 2: $\mathcal{T}$ ist eine konforme Partition auf $\Omega = [0,1] \times [0,1] \setminus [0.5,1] \times \{0.5\}$.

Eine konforme Partition $\mathcal{T}$ nennen wir ein ***Gitter***. Von nun an wollen wir nur noch konforme Partitionen $\mathcal{T}$ betrachten und verwenden die Notation

$$\begin{aligned} \mathcal{N} &:= \mathcal{N}(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} \mathcal{N}(T)\,, \\ \mathcal{E} &:= \mathcal{E}(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} \mathcal{E}(T)\,, \\ \mathcal{F} &:= \mathcal{F}(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} \mathcal{F}(T)\,. \end{aligned}$$

2.2 Bisektion eines Simplex

Wir identifizieren einen Simplex $T \in \mathcal{T}$ mit der Menge seiner Eckpunkte $\{x_0, ..., x_d\}$, und unterscheiden zwischen $d(d+1)!$ ausgezeichneten Simplizes, gegeben durch alle möglichen geordneten Sequenzen $(x_0, ..., x_d)_\gamma$ und Typen $\gamma \in \{0, ..., d-1\}$. Zu einem gegebenen ***ausgezeichneten Simplex*** $T = (x_0, ..., x_d)_\gamma$ definieren wir die ausgezeichneten ***Söhne***

$$T' = (x_0, \frac{x_0 + x_d}{2}, x_1, \ldots, x_\gamma, x_{\gamma+1}, \ldots, x_{d-1})_{(\gamma+1) \bmod d} \tag{5}$$

und

$$T'' = (x_d, \frac{x_0 + x_d}{2}, x_1, \ldots, x_\gamma, x_{d-1}, \ldots, x_{\gamma+1})_{(\gamma+1) \bmod d}\,, \tag{6}$$

wobei die Sequenz $(x_{d-1}, ..., x_{\gamma+1})$ rückwärts gezählt wird. $(x_{\gamma+1}, ..., x_{d-1})$ und $(x_1, ..., x_\gamma)$ werden für $\gamma = d-1$ respektive $\gamma = 0$ als leere Folgen verstanden.
Die Söhne entstehen also durch Bisektion der Kante $\overline{x_0 x_d}$, welche wir die ***Verfeinerungskante*** von T nennen und mit $e(T)$ bezeichnen. Für eine Illustration siehe Abbildungen 3 und 4 für $d = 2$ und $d = 3$. Zu einem ausgezeichneten Simplex $T = (x_0, ..., x_d)_\gamma$ definieren wir außerdem den ausgezeichneten Simplex

$$T_R := (x_d, x_1, ..., x_\gamma, x_{d-1}, ..., x_{\gamma+1}, x_0)_\gamma\,, \tag{7}$$

welcher zu T in dem Sinne äquivalent ist, dass durch Bisektion dieselben ausgezeichneten Söhne T', T'' entstehen. Wieder wird hier die Sequenz $(x_{d-1}, ..., x_{\gamma+1})$ rückwärts gezählt. Also unterscheiden wir tatsächlich zwischen $d(d+1)!/2$ ausgezeichneten Simplizes. Wir nennen zwei Simplizes T und T' ***gespiegelte Nachbarn***, wenn die Folge der geordneten Knoten von $T' = (x'_0, \ldots, x'_d)_{\gamma'}$ mit $T = (x_0, ..., x_d)_\gamma$ oder T_R an allen außer einer Position übereinstimmt. Bei gespiegelten Nachbarn T, T' gilt offensichtlich $T \cap T' \in \mathcal{F}$.

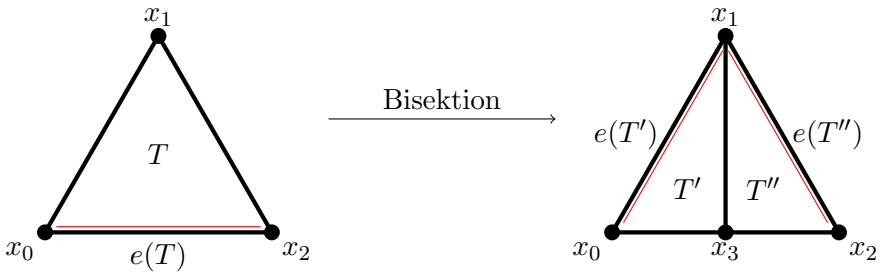


Abbildung 3: $T \subset \mathbb{R}^2$ wird zu $T' = (x_0, x_3, x_1)_1$ und $T'' = (x_2, x_3, x_1)_1$ verfeinert .

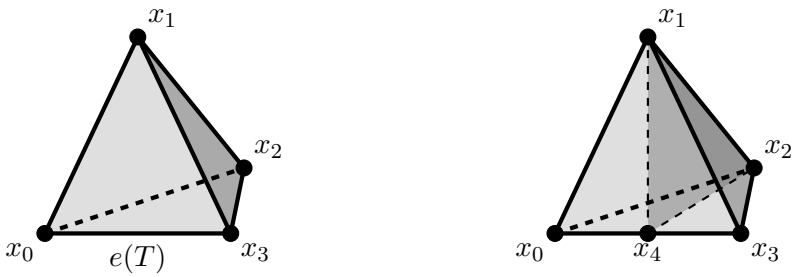


Abbildung 4: $T \subset \mathbb{R}^3$ wird zu $T' = (x_0, x_4, x_1, x_2)_1$ und $T'' = (x_3, x_4, x_2, x_1)_1$ verfeinert.

2.3 Das Anfangsgitter $\mathcal{T}_0$

Bisektion von markierten Elementen $\mathcal{M} \subset \mathcal{T}$ führt im Allgemeinen nicht auf konforme Gitter. Um die Konformität der Partition nach Verfeinern der markierten Elemente wiederherzustellen, sind weitere Verfeinerungen nötig. Damit der rekursive Algorithmus aus Abschnitt 2.3.1 terminiert, sind gewisse Anforderungen an das Anfangsgitter $\mathcal{T}_0$ notwendig. Wir nennen das Anfangsgitter $\mathcal{T}_0$ ***zulässig***, falls

- $\mathcal{T}_0$ konform ist,
- für alle $T, T' \in \mathcal{T}_0$, mit $T \cap T' = F \in \mathcal{F}_0$ gilt entweder

(Z1) $e(T) = e(T') \subset F$ und T, T' sind gespiegelte Nachbarn, oder

(Z2) es existieren Söhne $t \subset T$ und $t' \subset T'$ mit $F = t \cap t'$, die gespiegelte Nachbarn sind.

In den Abbildungen 5–7 sind alle Möglichkeiten, wie zwei benachbarte Simplizes in 2D zueinander stehen können, skizziert. Die Verfeinerungskante $e(\cdot)$ ist an der Innenseite des jeweiligen Elements in rot doppelt gestrichen.

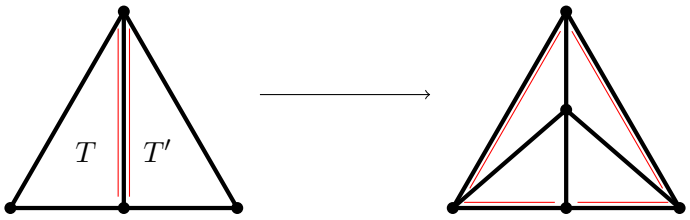


Abbildung 5: T und T' erfüllen ***(Z1)***. Verfeinerung führt zu keinen hängenden Knoten.

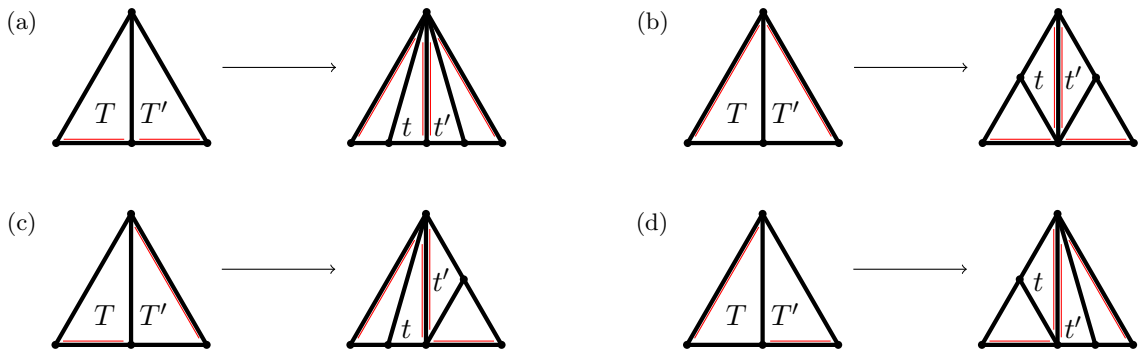


Abbildung 6: Es gibt vier Möglichkeiten (a)–(d), sodass T und T' Bedingung ***(Z2)*** erfüllen. Die Söhne t und t' von T und T' mit derselben Seite $F = t \cap t' = T \cap T'$ erfüllen ***(Z1)***.

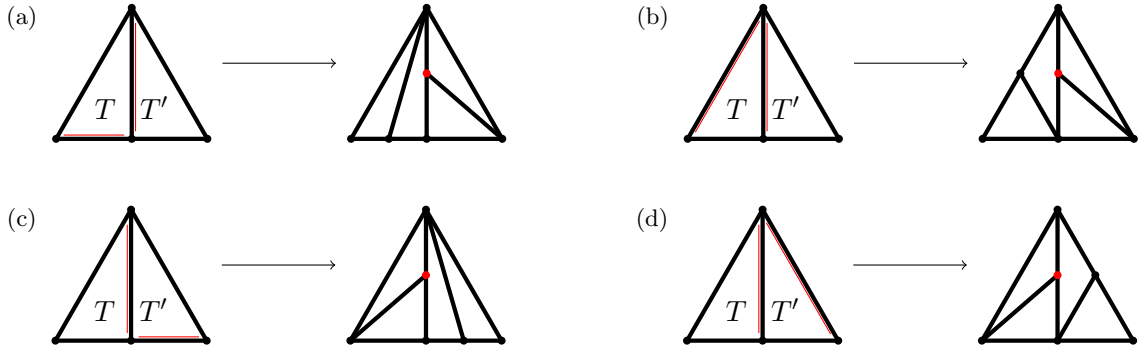


Abbildung 7: T und T' erfüllen hier weder ***(Z1)*** noch ***(Z2)***. Die Wahl der Verfeinerungungskanten in (a)–(d) führt auf hängende Knoten. T und T' die so zueinander stehen, können nicht Elemente eines zulässigen Anfangsgitters $\mathcal{T}_0$ sein.

Nun stellt sich die Frage, ob ein gegebenes Gitter $\mathcal{T}_{-1}$, stets als zulässiges Anfangsgitter dienen kann. In [BDD04, Chapter 2] wurde gezeigt, dass dies für $d = 2$ stets möglich ist. Für allgemeines $d \geq 3$ ist es bislang unklar. Zumindest kann aber jedes Gitter $\mathcal{T}_{-1}$ durch passende Verfeinerung und entsprechende Nummerierung der Knoten zu einem zulässigen Anfangsgitter $\mathcal{T}_0$ verfeinert werden:

2.3.1 Algorithmus zur Sicherstellung der Zulässigkeit von $\mathcal{T}_0$

Gegeben sei ein Gitter $\mathcal{T}_{-1}$ von d-Simplizes. Wir wollen $\mathcal{T}_{-1}$ nun zu einem zulässigen Anfangsgitter $\mathcal{T}_0$ verfeinern. Der folgende Algorithmus stammt aus [Ste08, Appendix A].
Zunächst konstruieren wir eine konforme Verfeinerung eines d-Simplex in $\frac{1}{2}(d+1)!$ Subsimplizes gemeinsam mit einer globalen Nummerierung der Knoten und einer Markierung der Kanten in dieser Verfeinerung, sodass die folgenden Eigenschaften erfüllt sind:

- Ein Knoten an einer markierten Kante hat keinen Index.
- Die anderen Knoten sind mit $1, ..., d-1$ durchnummeriert.
- Jeder Subsimplex hat Knoten mit Indizes $1, ..., d-1$ und zwei Knoten an einer markierten Kante.
- Die Unterteilung, Nummerierung und Kantenmarkierung sind symmetrisch in den baryzentrischen Koordinaten des ursprünglichen Simplex.

Für $d = 2$ unterteilen wir ein Dreieck in drei Subdreiecke indem wir den Schwerpunkt mit den Eckpunkten verbinden. Der Schwerpunkt erhält die Nummer 1 und die Kanten des ursprünglichen Dreiecks werden markiert. Offensichtlich sind obige Eigenschaften erfüllt.

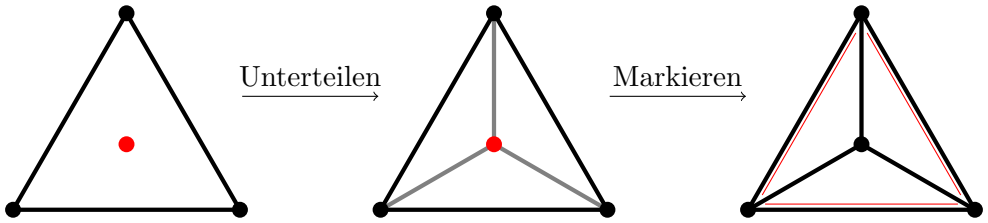


Abbildung 8: Das Dreieck wird durch Verbinden der Eckpunkte mit dem Schwerpunkt in drei Dreiecke unterteilt. Die Kanten des ursprünglichen Dreiecks werden markiert.

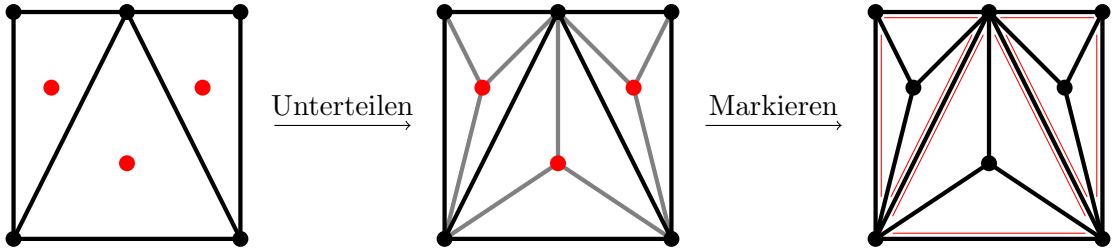


Abbildung 9: Die Unterteilung wird für alle Simplizes in $\mathcal{T}_{-1}$ durchgeführt. Sobald die Kanten markiert wurden, erfüllen alle Paare von Nachbarn ***(Z1)*** oder ***(Z2)***.

Für $d \geq 3$, angenommen wir haben bereits eine den obigen Eigenschaften entsprechende Unterteilung, Nummerierung sowie Markierung der $(d-1)$-dimensionalen Hyperflächen, gehen wir wie folgt vor: Für jeden d-Simplex erhalten wir $(d+1)$ Subsimplizes durch Verbinden der Knoten mit dessen Schwerpunkt und nummerieren diesen mit $d-1$. Jeder Subsimplex hat nun eine Seite gemeinsam mit dem ursprünglichen Simplex. Verwende die Unterteilung dieser Seite, um diese in insgesamt $\frac{1}{2}d!$ durchnummerierte und markierte $(d-1)$-Simplizes zu unterteilen. Verbinde nun jeden dieser $\frac{1}{2}d!$ $(d-1)$-Simplizes mit dem Schwerpunkt des d-Simplex um eine Unterteilung in $\frac{1}{2}(d+1)!$ d-Simplizes zu erhalten.
Also ausgehend von einer gegebenen konformen Partition von d-Simplizes unterteilen wir jeden Simplex in $\frac{1}{2}(d+1)!$ Subsimplizes wie oben beschrieben. Nach Konstruktion ist diese verfeinerte Partition, welche als $\mathcal{T}_0$ dienen wird, konform. Um die Simplizes in $\mathcal{T}_0$ auszuzeichnen, bleibt noch ein Typ zu definieren, welcher $(d-1)$ sein wird, sowie die Knoten in jedem Simplex lokal zu ordnen. Dazu vererben wir einfach jedem Simplex die Nummerierung des entsprechenden Supersimplex aus dem er entstanden ist, wobei zusätzlich die zwei Knoten an der markierten Kante in beliebiger Reihenfolge mit 0 und d nummeriert werden. Nachbarn innerhalb desselben Supersimplex sind offensichtlich gespiegelte Nachbarn, da die Knoten auf der gemeinsamen Hyperfläche gleich nummeriert sind modulo Vertauschung von 0 und d. Selbiges gilt nun aber auch aufgrund der Symmetrie der Nummerierung in den baryzentrischen Koordinaten für zwei Nachbarn, die aus verschiedenen Supersimplizes entstanden sind. Insgesamt folgt, dass $\mathcal{T}_0$ zulässig ist.

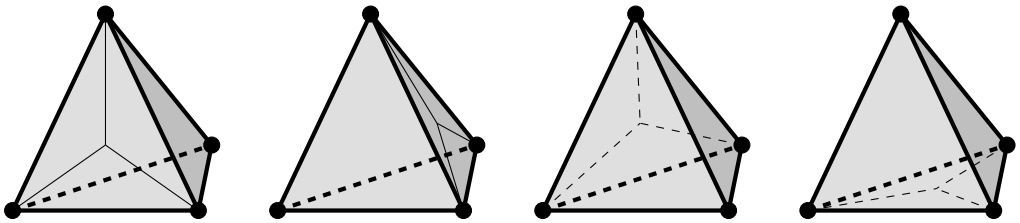


Abbildung 10: Die Seiten eines Tetraeders in $\mathbb{R}^3$ sind Dreiecke. Die Knoten eines Dreiecks verbinden wir mit dessen Schwerpunkt. Der Schwerpunkt eines Dreiecks wird mit dem lokalen Index 1 nummeriert und die ihm gegenüberliegende Kante wird markiert. Jede der vier Seiten des Tetraeders wird in drei Dreiecke unterteilt, also erhalten wir eine Unterteilung der Seiten des Tetraeders insgesamt 12 Dreiecke.

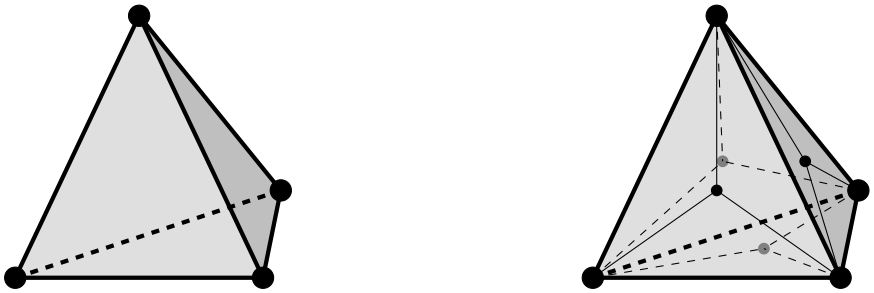


Abbildung 11: Wir haben eine Unterteilung der Seiten mit den gewünschten Eigenschaften. Verbinden wir nun noch jeden Knoten mit dem Schwerpunkt des Tetraeders, so erhalten wir die gewünschte Unterteilung. Für jedes entstandene Tetraeder wird der Schwerpunkt des ursprünglichen Tetraeders mit dem lokalen Index 2 nummeriert. Jedes der 12 entstandenen Tetraeder hat genau eine Kante gemeinsam mit dem ursprünglichen Tetraeder. Diese Kante ist die Verfeinerungskante.

2.4 Verfeinern eines Gitters mittels NVB

Sei $\mathcal{M} \subset \mathcal{T}$ die Menge der zur Verfeinerung markierten Elemente aus $\mathcal{T}$. Wenn wir nur die markierten Elemente verfeinern, wird die entstehende Partition $\mathcal{T}'$ im Allgemeinen nicht konform sein. Um eine konforme Partition zu erhalten, sind im Allgemeinen zusätzliche Bisektionen erforderlich. Zunächst geben wir einen Algorithmus an, der aus einem gegebenem Gitter $\mathcal{T}$ und gegebenem $T \in \mathcal{T}$ ein Gitter $\mathcal{T}'$ erzeugt, in dem zumindest $T \in \mathcal{T}$ verfeinert wurde.
Wir nennen zwei Nachbarn $T, T' \in \mathcal{T}$ ***kompatibel teilbar***, wenn sie dieselbe Verfeinerungskante besitzen, d.h. $e(T) = e(T')$. Für eine Partition $\mathcal{T}$ und $T \in \mathcal{T}$ definieren wir

$$N(\mathcal{T}, T) := \left\{ T' \in \mathcal{T} \;\middle|\; T' \cap T \in \mathcal{F}, e(T) \subset T' \right\} \tag{8}$$

die Menge der Nachbarn von T, welche $e(T)$ enthalten. Die nun vorgestellte Routine $\mathcal{T}' = \texttt{refine}(\mathcal{T}, T)$ erzeugt unter gewissen Voraussetzungen, siehe dazu Satz 2.3, eine konforme Verfeinerung $\mathcal{T}'$ des Gitters $\mathcal{T}$, in der $T \in \mathcal{T}$ verfeinert wurde:

$\mathcal{T}' := \texttt{refine}(\mathcal{T}, T)$ {
    $K := \emptyset$
    $F := \{T\}$

    `while` $\#F > 0$

```
        F_new := ∅
        forall T' ∈ F do
            forall T'' ∈ N(T, T') mit T'' ∉ F ∪ K do
                if T'' kompatibel teilbar mit T' then
                    F_new := F_new ∪ {T''}
                else
                    T := refine(T, T'')
                    Füge zu F_new den Sohn von T'' hinzu, welcher ein Nachbar von T' ist
                endif
            endfor
        endfor
        K := K ∪ F
        F := F_new
    endwhile

    Erzeuge T' aus T durch gleichzeitiges Verfeinern aller T' ∈ K
    return T'
}
```

Satz 2.3 ([Ste08, Theorem 5.1.]). *Sei $\mathcal{T}_0$ ein zulässiges Anfangsgitter und es existiere ein $n \in \mathbb{N}_0$, eine Folge von Gittern $\mathcal{T}_1, ..., \mathcal{T}_n$ und eine Folge von Elementen $T_j \in \mathcal{T}_j$ für $j = 0, ..., n-1$ mit $\mathcal{T}_{j+1} = \texttt{refine}(\mathcal{T}_j, T_j)$ für alle $j = 0, ..., n-1$ und $\mathcal{T} = \mathcal{T}_n$. Dann gilt:*
$\mathcal{T}' := \texttt{refine}(\mathcal{T}, T)$ terminiert, und $\mathcal{T}'$ ist die gröbste (im Sinne von Definition 2.6 und Satz 2.7) konforme Verfeinerung von $\mathcal{T}$, in der T verfeinert wurde. □

Der folgende Algorithmus $\mathcal{T}' = \texttt{refine}(\mathcal{T}, \mathcal{M})$ erzeugt unter den Voraussetzungen von Satz 2.3 aus einem Gitter $\mathcal{T}$ und einer Menge markierter Elemente $\mathcal{M} \subset \mathcal{T}$ eine konforme Verfeinerung $\mathcal{T}'$ des Gitters $\mathcal{T}$, in der alle Elemente $T \in \mathcal{M}$ verfeinert wurden:

```
T' := refine(T, M) {
    T' := T
    for  T ∈ M do
        if  T ∈ T' then
            T' := refine(T', T)
        endif
    endfor

    return  T'
}
```

Anmerkung 2.4 ([Ste08, Section 5]). Falls $\mathcal{T}$ die Voraussetzungen von Satz 2.3 erfüllt, terminiert der Algorithmus und $\mathcal{T}' := \texttt{refine}(\mathcal{T}, \mathcal{M})$ ist die gröbste (im Sinne von Definition 2.6 und Satz 2.7) konforme Verfeinerung von $\mathcal{T}$, in der alle Elemente $T \in \mathcal{M}$ verfeinert wurden.

Definition 2.5. Wir schreiben $\mathcal{T}' \in \texttt{refine}(\mathcal{T})$ genau dann, wenn ein $n \in \mathbb{N}_0$, eine Folge $\mathcal{T}_0, ..., \mathcal{T}_n$ von Gittern und eine Folge $\mathcal{M}_0, ..., \mathcal{M}_{n-1}$ von markierten Elementen mit $\mathcal{M}_j \subset \mathcal{T}_j$ für $j = 0, ..., n-1$ existieren, sodass $\mathcal{T} = \mathcal{T}_0$, $\mathcal{T}' = \mathcal{T}_n$ und $\mathcal{T}_{j+1} = \texttt{refine}(\mathcal{T}_j, \mathcal{M}_j)$ für alle $j = 0, ..., n-1$ gilt.

Definition 2.6. Für $\mathcal{T}, \mathcal{T}' \in \texttt{refine}(\mathcal{T}_0)$ nennen wir $\mathcal{T}'$ ***feiner*** als $\mathcal{T}$ (bezüglich $\mathcal{T}_0$), falls $\mathcal{T}' \in \texttt{refine}(\mathcal{T})$. In diesem Fall nennen wir auch $\mathcal{T}$ ***gröber*** als $\mathcal{T}'$ (bezüglich $\mathcal{T}_0$).

Satz 2.7. *Sei $\mathcal{T}_0$ ein zulässiges Anfangsgitter. Dann gilt:*

(i). *Feiner zu sein ist eine Halbordnung auf* $\texttt{refine}(\mathcal{T}_0)$.

(ii). *$\mathcal{T}_0$ ist das gröbste Element in* $\texttt{refine}(\mathcal{T}_0)$*, d.h. jedes $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ ist feiner als $\mathcal{T}_0$.*

(iii). *Für $\mathcal{T}', \mathcal{T}'' \in \texttt{refine}(\mathcal{T}_0)$ ist $\mathcal{T}''$ genau dann feiner als $\mathcal{T}'$, wenn für jedes $T'' \in \mathcal{T}''$ ein $T' \in \mathcal{T}'$ existiert mit $T'' \subset T'$.*

Beweis. (i). und (ii). folgen sofort aus der Definition von $\texttt{refine}(\mathcal{T}_0)$.
(iii). Sei zunächst $\mathcal{T}''$ feiner als $\mathcal{T}'$, d.h. $\mathcal{T}'' \in \texttt{refine}(\mathcal{T}')$. Da durch $\texttt{refine}(\cdot)$ Elemente nur verfeinert werden, existiert nach unserer Bisektionsvorschrift (5)–(6) für jedes $T'' \in \mathcal{T}''$ ein kanonisches Element $T' \in \mathcal{T}'$ mit $T'' \subset T'$.
Wenn hingegen für jedes $T'' \in \mathcal{T}''$ ein $T' \in \mathcal{T}'$ existiert mit $T'' \subset T'$, bleibt zu zeigen, dass wir ein $n' \in \mathbb{N}_0$, Gitter $\mathcal{T}'_0, ..., \mathcal{T}'_{n'}$ und Mengen markierter Elemente $\mathcal{M}'_0, ..., \mathcal{M}'_{n'-1}$, $\mathcal{M}'_j \subset \mathcal{T}'_j$ für $j = 0, ..., n'-1$ wählen können, sodass $\mathcal{T}' = \mathcal{T}'_0$, $\mathcal{T}'' = \mathcal{T}'_{n'}$ und $\mathcal{T}'_{j+1} = \texttt{refine}(\mathcal{T}'_j, \mathcal{M}'_j)$ für alle $j = 0, ..., n'-1$ gilt. Wir wählen $n' := \#\mathcal{T}'' - \#\mathcal{T}'$ und

- $\mathcal{T}'_0 := \mathcal{T}'$,
- $\mathcal{M}'_j := \left\{ T \in \mathcal{T}'_j \mid T \notin \mathcal{T}'' \right\}$,
- $\mathcal{T}'_{j+1} := \texttt{refine}(\mathcal{T}'_j, \mathcal{M}'_j)$,

für $j = 0, ..., n'-1$. Es bleibt zu zeigen, dass $\mathcal{T}'_{n'} = \mathcal{T}''$:
Die Wahl der Mengen $\mathcal{M}'_j$ garantiert, dass für alle $j = 0, ..., n'$ und jedes $T'' \in \mathcal{T}''$ ein $T'_j \in \mathcal{T}'_j$ existiert, mit $T'' \subset T'_j$. Falls $\mathcal{M}'_j = \emptyset$ für ein $j \in \{0, ..., n'-1\}$, folgt $\mathcal{M}_k = \emptyset$ für alle $k \in \{j, ..., n'-1\}$ und damit $\mathcal{T}'' = \mathcal{T}'_j = \mathcal{T}'_{j+1} = ... = \mathcal{T}'_{n'}$.
Falls $\mathcal{M}'_j \neq \emptyset$ für alle $j = 0, ..., n'-1$, dann gilt $\#\mathcal{T}'_{j+1} > \#\mathcal{T}'_j$ für alle $j = 0, ..., n'-1$ und damit $\#\mathcal{T}'_{n'} = \#\mathcal{T}''$, da $n' := \#\mathcal{T}'' - \#\mathcal{T}'$ gewählt wurde. Insgesamt folgt $\mathcal{T}'_{n'} = \mathcal{T}''$. □

2.5 Durch NVB erzeugte Partitionen

$\mathcal{T}_0$ sei von nun an stets ein zulässiges Anfangsgitter und $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$. Wir definieren eine ***Levelfunktion*** $\texttt{level} : \mathcal{T} \to \mathbb{N}_0$, die jedem Element $T \in \mathcal{T}$ eine natürliche Zahl zuordnet, sein sogenanntes Level: Für alle $T \in \mathcal{T}_0$ definiere $\texttt{level}(T) := 0$. Wenn T zu T', T'' verfeinert wird, definiere $\texttt{level}(T') := \texttt{level}(T'') := \texttt{level}(T) + 1$. Wir nennen einen ausgezeichneten Simplex, welcher durch ℓ rekursive Bisektionen aus T entstanden ist, einen ***Level ℓ Nachfahren*** von T.

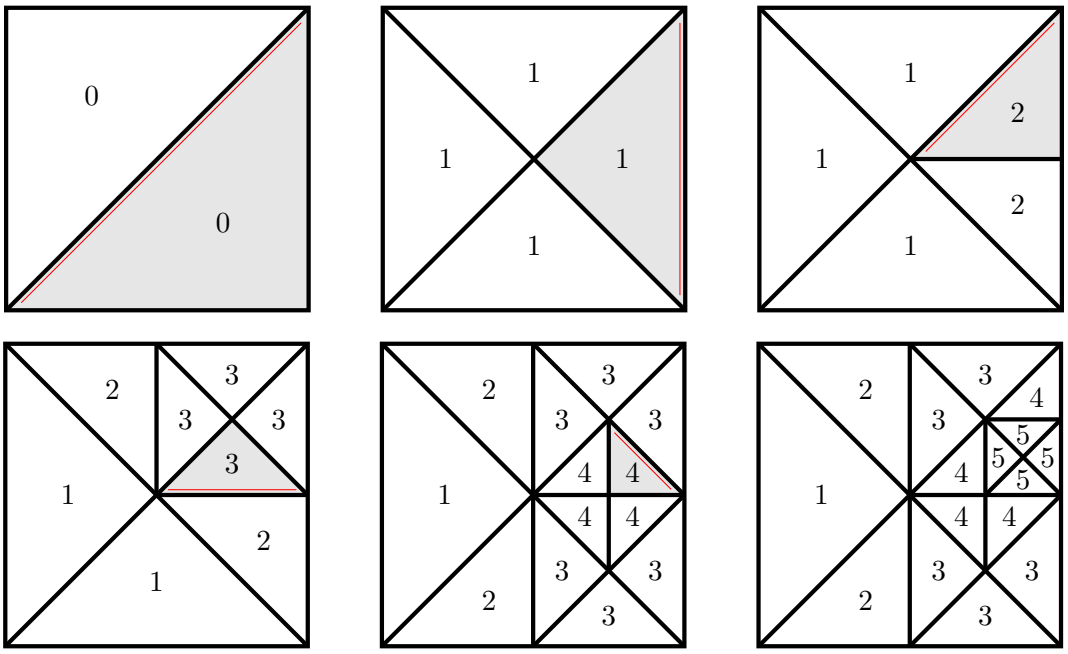


Abbildung 12: Die Entwicklung der Level anhand eines einfachen Beispiels in 2D: Wir starten von einem Ausgangsgitter mit 2 Elementen, dort ist der Level jedes Elements gleich 0. Das grau ausgefüllte Element ist jeweils zur Verfeinerung markiert, seine Verfeinerungskante ist in rot doppelt gestrichen. Das nächste Gitter ist immer das gröbste Gitter in dem das markierte Element verfeinert wurde. Die Zahl innerhalb eines Dreiecks ist sein aktuelles Level.

Weiters nennen wir die Verfeinerung $\mathcal{T}_\ell$ von $\mathcal{T}_0$, bei der alle $T \in \mathcal{T}_\ell$ den gleichen Level ℓ haben, die ***uniforme Level ℓ Verfeinerung*** von $\mathcal{T}_0$. Für uniforme Verfeinerungen $\mathcal{T}_\ell \in \texttt{refine}(\mathcal{T}_0)$ gilt folgender Satz:

Satz 2.8 ([Ste08, Theorem 4.3.])**.** *Jede uniforme Verfeinerung $\mathcal{T}_\ell$ von $\mathcal{T}_0$ ist konform.* □

Der folgende Satz hält fest, dass der Levelunterschied zweier benachbarter Elemente stets beschränkt ist. Ihre Level unterscheiden sich sogar maximal um 1:

Satz 2.9. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$, und $T, T' \in \mathcal{T}$ seien Nachbarn in $\mathcal{T}$. Dann gilt*

$$|\texttt{level}(T) - \texttt{level}(T')| \leq 1\,. \tag{9}$$

Beweis. Seien T, T' Nachbarn mit $\texttt{level}(T) = \texttt{level}(T') - p$ und $p \geq 2$. Dann existiert ein Level p Nachfahre S von T, der einen Punkt von $T' \cap \Omega$ innerhalb einer Seite enthält. Nach Satz 2.8 ist S ein Nachbar von T'. Somit hat S eine gemeinsame Hyperfläche mit T' und folglich auch mit T. Da ein Level $p \geq 2$ Nachfahre von T weniger als d Knoten gemeinsam mit T hat, kommen wir auf einen Widerspruch. Also kann sich das Level von Nachbarn höchstens um 1 unterscheiden. □

Satz 2.10 ([Ste08, Theorem 5.1.])**.** *Wenn $T' \in \mathcal{T}' \setminus \mathcal{T}$ neu durch den Aufruf von $\mathcal{T}' := \texttt{refine}(\mathcal{T}, T)$ erzeugt wurde, dann gilt $\texttt{level}(T') \leq \texttt{level}(T) + 1$.* □

Als Konsequenz dieses Satzes folgt unmittelbar folgendes Korollar:

Korollar 2.11. *Sei $\mathcal{T}' := \texttt{refine}(\mathcal{T}, T)$. Dann gilt:*

(i). *Die Level 1 Söhne T_1, T_2 von T wurden durch den Aufruf nicht weiter verfeinert und es gilt $T_1, T_2 \in \mathcal{T}'$.*

(ii). *Für $T' \in \mathcal{T}$ mit $\texttt{level}(T') > \texttt{level}(T)$ gilt $T' \in \mathcal{T}'$.*

Als Nächstes wollen wir untersuchen, wie groß der Levelunterschied zweier Elemente sein kann, die eine gemeinsame Kante haben. Dafür beweisen wir zunächst folgendes Lemma:

Lemma 2.12. *Sei $T \in \mathcal{T}$ und T' ein Level $2d-2$ Nachfahre von T. Dann haben T und T' keine gemeinsame Kante in $\mathcal{T}$. Das heißt, dass für alle Kanten $E \in \mathcal{E}$ gilt $E \not\subset T \cap T'$.*

Beweis. **1. Fall:** Betrachte zunächst einen Simplex T vom Typ $\gamma \in \{0,1\}$

$$T = (x_0, x_1, \dots, x_d)_\gamma \,.$$

Nach der Bisektionsvorschrift (5)–(6) haben die 2^{d-1} Level $d-1$ Söhne von T die Gestalt

$$T_i = (x_{i_0}, y_{i_1}, y_{i_2}, \dots, y_{i_{d-1}}, x_{i_d})_{(\gamma+d-1) \bmod d} \qquad i = 1, \dots, 2^{d-1} \,,$$

mit $0 \le i_0, i_d \le d$, $i_0 \ne i_d$ und $y_{i_j} \notin \{x_0, \dots, x_d\}$ für alle $j = 1, \dots, d-1$. Ein Sohn von T_i hat nur noch einen Knoten, entweder x_{i_0} oder x_{i_d}, mit T gemeinsam. Folglich gilt für einen Simplex T des Typs 0 oder 1, dass ein Level d Sohn von T keine Kante $E \in \mathcal{E}(T)$ enthält.
2. Fall: Betrachte nun T von beliebigem Typ $2 \le \gamma \le d-1$. Die Level $d-\gamma$ Söhne von T haben Typ 0, und auf diese trifft der 1. Fall zu. Somit gilt, dass jeder Level $d + d - \gamma \le 2d-2$ Nachfahre von T nur noch höchstens einen Knoten mit T, und folglich keine Kante mit T gemeinsam hat. □

Mit Hilfe dieses Lemmas können wir nun folgenden Satz zeigen:

Satz 2.13. *Seien $T', T'' \in \mathcal{T}$ mit $E \subset T' \cap T''$ für ein $E \in \mathcal{E}$. Dann gilt*

$$|\texttt{level}(T'') - \texttt{level}(T')| \le 2d-3 \,. \tag{10}$$

Beweis. Sei ohne Einschränkung der Allgemeinheit $\texttt{level}(T'') = \texttt{level}(T') + p$ mit $p \in \mathbb{N}_0$, und halte die Kante $E \in \mathcal{E}$ mit $E \subset T' \cap T''$ fest. Nun wollen wir durch (gegebenenfalls wiederholtes) Aufrufen von $\texttt{refine}(\cdot)$ die gröbste Verfeinerung $\mathcal{T}'$ erzeugen, die alle Level p Söhne von T' enthält. Nach Korollar 2.11(ii) gilt $T'' \in \mathcal{T}'$, und damit $E \in \mathcal{E}(\mathcal{T}')$. Mit Lemma 2.12 folgt nun $p < 2d-2$. □

Nun wollen wir noch zeigen, dass die Anzahl der Elemente in einem Gitter $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$, die einen Knoten $z_0 \in \mathcal{N}_0$ gemeinsam haben, beschränkt ist:

Definition 2.14. Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$. Definiere für $z \in \mathcal{N}$ die Menge der Elemente im Knotenpatch von z:

$$\omega_{\mathcal{T}}(z) := \omega(z) := \{T \in \mathcal{T} \mid z \in T\} \,. \tag{11}$$

Satz 2.15. (i). *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und $z \in \mathcal{N}(\mathcal{T})$. Dann gilt für alle $\mathcal{T}' \in \texttt{refine}(\mathcal{T})$*

$$\#\omega_{\mathcal{T}'}(z) \le S_d \#\omega_{\mathcal{T}}(z) \,, \tag{12}$$

mit

$$S_d := \max_{\substack{j=0,\dots,d \\ \gamma=0,\dots,d-1}} S_d(j,\gamma) \,,$$

$$S_d(j,\gamma) := \begin{cases} 1 \,, & j \in \{0,d\} \\ S_d(j+1, (\gamma+1) \bmod d) + S_d(j+1, (\gamma+1) \bmod d) \,, & 1 \le j \le \gamma < d \\ S_d(j+1, (\gamma+1) \bmod d) + S_d(d-(j-\gamma-1), (\gamma+1) \bmod d) \,, & 0 \le \gamma < j < d \end{cases}$$

Es gilt $S_d < \infty$ und S_d hängt nur von d ab.
(ii). *Es existiert eine Konstante $C_3 > 0$, die nur von d und $\mathcal{T}_0$ abhängt, sodass für alle $\mathcal{T} \in$ $\texttt{refine}(\mathcal{T}_0)$*

$$\#\omega_{\mathcal{T}}(z) \leq C_3 := S_d \max_{z_0 \in \mathcal{N}_0} \#\omega_{\mathcal{T}_0}(z_0)\,, \tag{13}$$

für alle $z \in \mathcal{N}_0$.

Beweis. (i). Betrachte einen Simplex

$$T_0 = (x_0, \ldots, x_d)_\gamma \in \mathcal{T}\,.$$

Die Lösung der Rekursion $S_d(j, \gamma)$ ist die maximale Anzahl an Elementen $T \in \mathcal{T}'$ mit $x_j \in T \subset T_0$, die in einer beliebigen Verfeinerung $\mathcal{T}' \in \texttt{refine}(\mathcal{T})$ vorkommen kann. Das folgt sofort aus den Bisektionsvorschriften (5)–(6). Wir müssen uns nur überlegen, ob diese Rekursion terminiert: Der Typ γ wird bei jeder Bisektion um 1 erhöht, oder falls er $d-1$ war, auf 0 gesetzt. Im Beweis von Satz 2.12 haben wir gesehen, dass ein Level d Sohn T'' eines Simplex T' des Typs 0, nur noch einen Knoten mit T' gemeinsam hat, und dieser Knoten steht an 0-ter Stelle von T''. Es folgt, dass die Rekursion terminiert, denn $S_d(0, 0) = 1$.
Multiplizieren wir nun für $z \in \mathcal{N}(\mathcal{T})$ die Anzahl der Elemente im Knotenpatch $\omega_{\mathcal{T}}(z)$ mit der maximalen Anzahl S_d an Elementen, die innerhalb eines Elements dieses Patches an einem Knoten entstehen können, so folgt die Behauptung.
(ii). Nach (i) gilt

$$\#\omega_{\mathcal{T}}(z_0) \leq S_d \#\omega_{\mathcal{T}_0}(z_0)\,,$$

für alle $z_0 \in \mathcal{N}_0$. Mit $C_3 := S_d \max_{z_0 \in \mathcal{N}_0} \#\omega_{\mathcal{T}_0}(z_0)$ gilt die Aussage mit einer Konstanten C_3, die nur von $\mathcal{T}_0$ und d abhängt. □

2.6 γ-Formregularität

In diesem Kapitel zeigen wir, dass die Winkel in allen möglichen durch NVB aus einem zulässigen Anfangsgitter $\mathcal{T}_0$ entstandenen Partitionen gleichmäßig durch eine Konstante, welche nur von $\mathcal{T}_0$ und der Raumdimension d abhängt, nach unten beschränkt sind. Dabei folgen wir der Präsentation in [Tra97].
Ein Gitter $\mathcal{T}$ von d-Simplizes nennen wir γ-***formregulär***, wenn eine Konstante $\gamma \in \mathbb{R}$ existiert, sodass

$$\max_{T \in \mathcal{T}} \frac{\texttt{diam}(T)^d}{|T|} \leq \gamma < \infty. \tag{14}$$

Zur einfacheren Handhabung definieren wir einen Referenzsimplex, den sogenannten Kuhn-Simplex:

Definition 2.16. Seien $\{e_1, ..., e_d\}$ die kanonischen Einheitsvektoren des $\mathbb{R}^d$. Ein ***Kuhn-Simplex*** ist ein Simplex des Typs 0 mit den Knoten $\{x_0^\pi, ..., x_d^\pi\}$, wobei $x_i^\pi = \sum_{j=1}^{i} e_{\pi(j)}$ und π eine Permutation von $\{1, ..., d\}$ bezeichnet. Wir merken an, dass die Menge der $d!$ Kuhn Simplizes eine Partition des d-dimensionalen Einheitswürfels ist.

Satz 2.17 ([Tra97, Section 5])**.** *Es gibt nur endliche viele Winkel, die durch beliebig viele (möglicherweise unendliche viele) Verfeinerungen eines Kuhn Simplex entstehen können.* □

Satz 2.18. *Sei $\mathcal{T}_0$ ein zulässiges Anfangsgitter. Dann gibt es nur endlich viele Winkel, die in allen möglichen $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ auftreten können.*

Beweis. Sei $T = (x_0, ..., x_d)_0$ ein beliebiger ausgezeichneter d-Simplex des Typs 0 und $K_0 = (0, k_1, ..., k_d)_0$ der Kuhn Simplex mit $k_i = \sum_{j=1}^{i} e_j$. Dann existiert eine eindeutige affine Abbildung $F_T : \mathbb{R}^d \to \mathbb{R}^d$ so, dass $F(x_i) = k_i$, $0 \leq i \leq d$. Unsere Bisektionsvorschrift (5)–(6) zeigt, dass die Level ℓ Nachfahren von T genau die Bilder der Level ℓ Nachfahren von K_0 unter der Abbildung F_T^{-1} sind. Mit Satz 2.17 sehen wir, dass der kleinste Winkel in den potenziellen Nachfolgern von T nur abhängig von dem kleinsten Winkel in T von Null fernbleibt. Dasselbe gilt für einen ausgezeichneten Simplex T des Typs $g \in \{0, ..., d-1\}$, da seine 2^{d-g} Level $d-g$ Nachfolger Typ 0 haben. Die Tatsache, dass $\mathcal{T}_0$ nur endlich viele Elemente hat, schließt den Beweis ab. □

Satz 2.19. *Sei $\mathcal{T}_0$ ein zulässiges Anfangsgitter. Dann existieren Konstanten C_4, C_5 die nur von $\mathcal{T}_0$ abhängen, sodass für alle Verfeinerungen $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und alle $T \in \mathcal{T}$*

$$C_4 \texttt{diam}(T) \leq |T|^{1/d} \leq C_5 \texttt{diam}(T)$$

gilt.

Beweis. Nach Satz 2.18 existiert ein Winkel $\alpha > 0$, der nur von $\mathcal{T}_0$ abhängt, sodass für alle $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ alle dort auftretenden Winkel durch α nach unten beschränkt sind. Seien $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und $T \in \mathcal{T}$ beliebig. Definiere $\widetilde{T} \subset \mathbb{R}^d$ als $\widetilde{T} := \texttt{diam}(T)^{-1} T$. Mit dem Transformationssatz gilt nun

$$|T| = \texttt{diam}(T)^d |\widetilde{T}| \,. \tag{15}$$

Es gilt $\texttt{diam}(\widetilde{T}) = 1$ und die Winkel in $\widetilde{T}$ sind die gleichen wie in T, insbesondere sind damit auch alle in $\widetilde{T}$ auftretenden Winkel durch $\alpha > 0$ nach unten beschränkt. Das impliziert, dass alle Kanten $E \in \mathcal{E}(\widetilde{T})$ durch eine Konstante $\widetilde{C} > 0$, die nur von $\mathcal{T}_0$ abhängt, nach unten beschränkt sind. Also gilt $|\widetilde{T}| \simeq 1$, wobei die versteckte Konstante nur von $\mathcal{T}_0$ abhängt. In (15) eingesetzt, folgt dann

$$|T|^{1/d} \simeq \texttt{diam}(T)$$

und damit die Behauptung. □

3 H^1_D-Stabilität

Die Beweisideen in diesem Abschnitt orientieren sich an [KPP13a],[KPP13b], wobei angemerkt sei, dass die Beweise von [KPP13a, Theorem 6] beziehungsweise [KPP13b, Proposition 13] fehlerhaft sind. Der dort definierte Knoten-Abstand $\delta(\cdot, \cdot)$ erfüllt nicht die im Beweis geforderte Eigenschaft $|\delta(z_k, T') - \delta(z_j, T')| \leq 1$, beziehungsweise $|\delta(z_k, z_{k'}) - \delta(z_j, z_{k'})| \leq 1$ falls $\delta(z_j, z_k) \leq 1$.

Definition 3.1. Zu einem zulässigen Anfangsgitter $\mathcal{T}_0$ definiere

$$\begin{aligned} C_6 &:= \min_{T_0 \in \mathcal{T}_0} |T_0| \,, \\ C_7 &:= \max_{T_0 \in \mathcal{T}_0} |T_0| \,. \end{aligned}$$

3.1 Technik

Für jeden Knoten $z_\ell \in \mathcal{N}(\mathcal{T}) = \{z_1, ..., z_N\}$ bezeichne $\varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$ die zu z_ℓ gehörige Hutfunktion, d.h. $\varphi_\ell(z_k) = \delta_{\ell k}$ mit dem Kroneckersymbol $\delta_{\ell k}$. Ferner sei $N = \#\mathcal{N}$ die Anzahl der Knoten in $\mathcal{T}$. Wir definieren die Funktion

$$[\cdot, \cdot] : \mathcal{T} \times \{1, ..., d+1\} \to \{1, ..., N\}, \tag{16}$$

welche den bezüglich eines Simplex $T \in \mathcal{T}$ lokalen Index eines Knotens auf den globalen Index des Knotens abbilden soll, d.h. es gilt $T = \operatorname{conv}\left\{z_{[T,1]}, ..., z_{[T,d+1]}\right\}$ für alle $T \in \mathcal{T}$.
Für jedes Element $T \in \mathcal{T}$, definiere die lokale Massenmatrix

$$\mathbf{M}_T \in \mathbb{R}^{(d+1)\times(d+1)}_{\text{sym}}, \quad \text{durch } (\mathbf{M}_T)_{jk} = \int_T \varphi_{[T,j]}\varphi_{[T,k]}\, d\mathbf{x}. \tag{17}$$

Für jeden Knoten $z_\ell \in \mathcal{N}(\mathcal{T})$ sei $h_\ell > 0$ ein positiver Skalar, der sich wie die lokale Größe des Elements verhält, also der $h_\ell \simeq \texttt{diam}(T)$ erfüllt. Wir werden diese Skalare später passend wählen. Diese Knotenwerte ermöglichen die Definition einer geglätteten Netzweitenfunktion $h = \sum_{\ell=1}^N h_\ell \varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$. Die folgende diagonale Skalierungsmatrix gibt an, wie sehr sich $h|_T$ von $\texttt{diam}(T)$ unterscheidet. Wir definieren

$$\mathbf{\Lambda}_T \in \mathbb{R}^{(d+1)\times(d+1)}_{\text{diag}}, \quad (\mathbf{\Lambda}_T)_{jk} = \frac{\texttt{diam}(T)}{h_{[T,k]}}\delta_{jk}. \tag{18}$$

In Lemma 3.4 werden wir sehen, dass $\mathbf{M}_T$ nur von $|T|$ abhängt. Zunächst zeigen wir folgendes Lemma:

Lemma 3.2. *Sei $1 \le \ell \le d$. Dann gilt*

$$\int_0^{1-\sum_{i=1}^{\ell-1} x_i} x_\ell^m \Big(1 - \sum_{i=1}^{\ell} x_i\Big)^n dx_\ell = \frac{m!\cdot n!}{(m+n+1)!}\Big(1 - \sum_{i=1}^{\ell-1} x_i\Big)^{m+n+1} \tag{19}$$

für alle $m, n \in \mathbb{N}$.

Beweis. Durch m-maliges partielles Integrieren, wobei die Randterme immer 0 sind, erhält man

$$\begin{aligned}\int_0^{1-\sum_{i=1}^{\ell-1} x_i} x_\ell^m \Big(1 - \sum_{i=1}^{\ell} x_i\Big)^n dx_\ell &= \frac{m!\cdot n!}{(n+m)!}\int_0^{1-\sum_{i=1}^{\ell-1} x_i} \Big(1 - \sum_{i=1}^{\ell} x_i\Big)^{n+m} dx_\ell \\ &= \frac{m!\cdot n!}{(m+n+1)!}\Big(1 - \sum_{i=1}^{\ell-1} x_i\Big)^{n+m+1}.\end{aligned}$$

□

Lemma 3.3. *Es existiert ein Referenzsimplex $\widehat{T} \subset \mathbb{R}^d$, sodass die zugehörige Massenmatrix*

$$\widehat{\mathbf{M}}_{jk} := \int_{\widehat{T}} \widehat{\varphi}_k\widehat{\varphi}_j\, d\mathbf{x} = 1 + \delta_{jk}$$

erfüllt.

Beweis. Betrachte zunächst den Simplex $\widetilde{T} := \operatorname{conv}\{\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_d\} \subset \mathbb{R}^d$, wobei $\mathbf{e}_j \in \mathbb{R}^d$ den j-ten Einheitsvektor bezeichnet. Die zugehörige Massenmatrix $\widetilde{\mathbf{M}}$ ist definiert durch

$$\widetilde{\mathbf{M}}_{jk} := \int_{\widetilde{T}} \widetilde{\varphi}_k \widetilde{\varphi}_j \, d\mathbf{x}$$

mit den nodalen Basisfunktionen $\widetilde{\varphi}_j \in \mathcal{P}^1(\widetilde{T})$, $j = 1, ..., d+1$. Es gilt

- $\widetilde{\varphi}_j(\mathbf{x}) = x_j$ für $j = 1, ..., d$
- $\widetilde{\varphi}_{d+1}(\mathbf{x}) = 1 - \sum_{j=1}^{d} x_j \,.$

Mit dem Satz von Fubini gilt nun

$$\begin{aligned}\widetilde{\mathbf{M}}_{jk} &= \int_{\widetilde{T}} \widetilde{\varphi}_k \widetilde{\varphi}_j \, d\mathbf{x} \\ &= \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-\sum_{i=1}^{d-1} x_i} \widetilde{\varphi}_k \widetilde{\varphi}_j \, dx_d \ldots \, dx_3 \, dx_2 \, dx_1 \,.\end{aligned}$$

Löst man nun alle Integrale nacheinander von innen nach außen mit der Gleichheit (19) auf, erhält man schließlich

$$\widetilde{\mathbf{M}}_{jk} = \frac{1}{(d+2)!}(1 + \delta_{jk}) \,. \tag{20}$$

Mit dem Transformationssatz folgt nun für den Referenzsimplex

$$\widehat{T} := (d+2)!^{1/d} \cdot \widetilde{T} := (d+2)!^{1/d} \operatorname{conv}\{\mathbf{0}, \mathbf{e}_1, ..., \mathbf{e}_d\} \,, \tag{21}$$

dass die zugehörige Massenmatrix $\widehat{\mathbf{M}}$ die Eigenschaft $\widehat{\mathbf{M}}_{jk} = 1 + \delta_{jk}$ erfüllt. □

Lemma 3.4. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und $\widehat{T} \subset \mathbb{R}^d$ der Referenzsimplex aus Lemma 3.3 mit zugehöriger Massenmatrix $\widehat{\mathbf{M}}$. Dann gilt für die lokalen Massenmatrizen aus* (17)

$$\mathbf{M}_T = C_8 |T| \widehat{\mathbf{M}} \tag{22}$$

für alle $T \in \mathcal{T}$. Die Konstante $C_8 > 0$ hängt nur von d ab.

Beweis. Sei $T \in \mathcal{T}$ beliebig und $\Phi_T : \widehat{T} \to T$ eine affine Transformation, die $\widehat{T}$ auf T abbildet. Dann gilt mit dem Transformationssatz für die Einträge der lokalen Massenmatrix $\mathbf{M}_T$ von T

$$\begin{aligned}(\mathbf{M}_T)_{jk} &= \int_T \varphi_k \varphi_j \, d\mathbf{x} \\ &= \int_{\widehat{T}} \widehat{\varphi}_k \widehat{\varphi}_j \, d\mathbf{x} \cdot |\det D\Phi_T| \\ &= \frac{|T|}{|\widehat{T}|} \widehat{\mathbf{M}}_{jk} \,.\end{aligned}$$

Mit $C_8 := |\widehat{T}|^{-1} > 0$ folgt die Behauptung. □

Nun definieren wir noch einen Interpolationsoperator, den wir später in dem Beweis von Proposition 3.9 benötigen werden. Dabei verallgemeinern wir die Präsentation in [Pra15] für allgemeines $d \geq 2$:

Definition 3.5. Wähle für $z \in \mathcal{N}$ eine Seite $F_z \in \mathcal{F}$ mit

- $z \in F_z$,
- $F_z \subset \overline{\Gamma}_D$, falls $z \in \overline{\Gamma}_D$,
- $F_z \subset \Gamma$, falls $z \in \Gamma$.

Aus dem Satz von Riesz folgt sofort folgendes Lemma:

Lemma 3.6. *Für $z \in \mathcal{N}$ existiert eine eindeutige duale Funktion $\psi_z \in \mathcal{P}^1(F_z)$ so, dass*

$$\int_{F_z} \psi_z \varphi_{z'} \, ds = \delta_{zz'} \tag{23}$$

für alle $z' \in \mathcal{N}$.

Definition 3.7. Definiere den Scott-Zhang-Projektor $\mathcal{J}_D(\mathcal{T}) : H^1_D(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ durch

$$\mathcal{J}_D(\mathcal{T})v := \sum_{z \in \mathcal{N}} \left(\int_{F_z} \psi_z v \, ds \right) \varphi_z \tag{24}$$

mit ψ_z wie in (23).

Lemma 3.8 ([Pra15, Theorem 4.7.])**.** *Die Scott-Zhang-Projektion $\mathcal{J}_D(\mathcal{T}) : H^1_D(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ aus Definition 3.7 erfüllt die folgenden Eigenschaften:*

(i). *$\mathcal{J}_D(\mathcal{T})$ ist eine Projektion auf $\mathcal{S}^1_D(\mathcal{T})$, d.h.*

$$\mathcal{J}_D(\mathcal{T})v = v \,, \tag{25}$$

für alle $v \in \mathcal{S}^1_D(\mathcal{T})$.

(ii). *$\mathcal{J}_D(\mathcal{T})$ ist lokal H^1_D-stabil, d.h.*

$$\|\nabla \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)} \,, \tag{26}$$

für alle $v \in H^1_D(\Omega)$ und alle $T \in \mathcal{T}$, wobei $\omega_T := \bigcup_{T \in \omega(T)}$ und $\omega(T) := \{T' \in \mathcal{T} \mid T \cap T' \cap \Omega \neq \emptyset\}$.

(iii). *$\mathcal{J}_D(\mathcal{T})$ erfüllt eine Approximationseigenschaft erster Ordnung, d.h.*

$$\|v - \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \mathtt{diam}(T) \|\nabla v\|_{L^2(\omega_T)} \,, \tag{27}$$

für alle $v \in H^1_D(\Omega)$ und alle $T \in \mathcal{T}$.

Die versteckten Konstanten in (26)–(27) hängen nur von der γ-Formregularität von $\mathcal{T}$ ab. □

Anhand dieser Definitionen lautet unser Stabilitätskriterium wie folgt:

Proposition 3.9. *Sei $\mathcal{T}$ ein γ-formreguläres Gitter. Für alle Elemente $T \in \mathcal{T}$ gelte*

$$C_9^{-1} \mathbf{x}^\top \mathbf{\Lambda}_T^2 \mathbf{M}_T \mathbf{\Lambda}_T^2 \mathbf{x} \leq \mathbf{x}^\top \mathbf{M}_T \mathbf{x} \leq C_{10} \mathbf{x}^\top \mathbf{\Lambda}_T^2 \mathbf{M}_T \mathbf{x} \quad \text{für alle } \mathbf{x} \in \mathbb{R}^{d+1}, \tag{28}$$

mit den Konstanten $C_9, C_{10} > 0$. Dann gilt H^1_D-Stabilität (3) der L^2-orthogonalen Projektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\Omega)$, und die Konstante $C_2 > 0$ hängt nur von der γ-Formregularität von $\mathcal{T}$ und den Konstanten $C_9, C_{10} > 0$ ab.

Beweis. Sei $\mathcal{J}_D(\mathcal{T}) : H^1_D(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ der Scott-Zhang Projektor aus Definition 3.7, d.h. nach Lemma 3.8 gilt

$$\|\nabla \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)} \text{ sowie } \|v - \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \mathtt{diam}(T)\|\nabla v\|_{L^2(\omega_T)} \tag{29}$$

für alle $v \in H^1_D(\Omega)$ und alle $T \in \mathcal{T}$. Die versteckten Konstanten hängen nur von der γ-Formregularität ab. Sei $q := \Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v = \sum_{\ell=1}^N q_\ell \varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$. Wir definieren $p := \sum_{\ell=1}^N q_\ell h_\ell^{-2} \varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$. Mit den lokalen Koeffizientenvektoren $\mathbf{x} = (q_{[T,1]}, \dots, q_{[T,d+1]})^\top \in \mathbb{R}^{d+1}$ und $\mathbf{y} = (q_{[T,1]} h_{[T,1]}^{-2}, \dots, q_{[T,d+1]} h_{[T,d+1]}^{-2})^\top = \mathtt{diam}(T)^{-2} \mathbf{\Lambda}_T^2 \mathbf{x} \in \mathbb{R}^{d+1}$ erhalten wir durch die untere Abschätzung in (28)

$$\begin{aligned} \|p\|^2_{L^2(T)} = \mathbf{y}^\top \mathbf{M}_T \mathbf{y} = \mathtt{diam}(T)^{-4} \mathbf{x}^\top \mathbf{\Lambda}_T^2 \mathbf{M}_T \mathbf{\Lambda}_T^2 \mathbf{x} &\leq C_9 \mathtt{diam}(T)^{-4} \mathbf{x}^\top \mathbf{M}_T \mathbf{x} \\ &= C_9 \mathtt{diam}(T)^{-4} \|q\|^2_{L^2(T)} \,. \end{aligned}$$

Weiters liefert die obere Abschätzung in (28)

$$\|q\|^2_{L^2(T)} = \mathbf{x}^\top \mathbf{M}_T \mathbf{x} \leq C_{10} \mathbf{x}^\top \mathbf{\Lambda}_T^2 \mathbf{M}_T \mathbf{x} = C_{10} \mathtt{diam}(T)^2 \int_T pq \, dx \,.$$

Summieren wir nun die letzten zwei Ungleichungen über alle Elemente $T \in \mathcal{T}$, so sehen wir durch Ausnutzen der Symmetrie von Orthogonalprojektionen und $\Pi_D(\mathcal{T})p = p$ für $p \in \mathcal{S}^1_D(\mathcal{T})$, dass

$$\begin{aligned} \sum_{T \in \mathcal{T}} \mathtt{diam}(T)^{-2} \|q\|^2_{L^2(T)} &\lesssim \int_\Omega pq \, dx = \int_\Omega p(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v) \, dx = \int_\Omega p(v - \mathcal{J}_D(\mathcal{T})v) \, dx \\ &\leq \left(\sum_{T \in \mathcal{T}} \mathtt{diam}(T)^2 \|p\|^2_{L^2(T)} \right)^{1/2} \left(\sum_{T \in \mathcal{T}} \mathtt{diam}(T)^{-2} \|v - \mathcal{J}_D(\mathcal{T})v\|^2_{L^2(T)} \right)^{1/2} \\ &\overset{(27)}{\lesssim} \left(\sum_{T \in \mathcal{T}} \mathtt{diam}(T)^{-2} \|q\|^2_{L^2(T)} \right)^{1/2} \|\nabla v\|_{L^2(\Omega)} \,. \end{aligned}$$

Dies zeigt

$$\left(\sum_{T \in \mathcal{T}} \mathtt{diam}(T)^{-2} \|q\|^2_{L^2(T)} \right)^{1/2} \lesssim \|\nabla v\|_{L^2(\Omega)} \,. \tag{30}$$

Nun verwenden wir eine inverse Abschätzung $\|\nabla q\|_{L^2(T)} \lesssim \mathtt{diam}(T)^{-1} \|q\|_{L^2(T)}$ für alle $T \in \mathcal{T}$ und lokale H^1_D-Stabilität von $\mathcal{J}_D(\mathcal{T})$, d.h. $\|\nabla \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)}$ für alle $T \in \mathcal{T}$. Dies zeigt

$$\begin{aligned} \|\nabla \Pi_D(\mathcal{T})v\|_{L^2(\Omega)} &\leq \|\nabla(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v)\|_{L^2(\Omega)} + \|\nabla \mathcal{J}_D(\mathcal{T})v\|_{L^2(\Omega)} \\ &= \left(\sum_{T \in \mathcal{T}} \|\nabla(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v)\|^2_{L^2(T)} \right)^{1/2} + \left(\sum_{T \in \mathcal{T}} \|\nabla \mathcal{J}_D(\mathcal{T})v\|^2_{L^2(T)} \right)^{1/2} \\ &\overset{(26)}{\lesssim} \left(\sum_{T \in \mathcal{T}} \mathtt{diam}(T)^{-2} \|q\|^2_{L^2(T)} \right)^{1/2} + \left(\sum_{T \in \mathcal{T}} \|\nabla v\|^2_{L^2(\omega_T)} \right)^{1/2} \\ &\overset{(30)}{\lesssim} \|\nabla v\|_{L^2(\Omega)} \end{aligned}$$

und schließt den Beweis. □

3.2 Scharfe Abschätzung des minimalen Eigenwerts

Später in Proposition 3.20 werden wir untersuchen unter welchen Bedingungen die Voraussetzungen (28) in Proposition 3.9 erfüllt werden können. Um möglichst scharfe Bedingungen zu formulieren, die (28) implizieren, ist es essenziell, scharfe untere Schranken für die Eigenwerte der Matrix $\mathbf{A}$ aus (31) zu finden.
Ziel dieses Kapitels ist es nun, die Eigenwerte der Matrix $\mathbf{A}$ aus (31) nach unten scharf abzuschätzen. Genauer gesagt, werden wir Schritt für Schritt folgenden Satz beweisen:

Satz 3.10. *Seien $d \in \mathbb{N}$ und $1 < M \in \mathbb{R}$. Seien $\alpha_1, ..., \alpha_{d+1} \in \mathbb{R} \setminus \{0\}$ mit $M^{-1} \leq \alpha_i^2/\alpha_j^2 \leq M$ für $i, j = 1, ..., d+1$. Die Matrix $\mathbf{A} \in \mathbb{R}^{(d+1)\times(d+1)}$ habe folgende Gestalt:*

$$\mathbf{A}_{ij} = \begin{cases} 4 & \text{für } i = j\,, \\ \frac{\alpha_i}{\alpha_j} + \frac{\alpha_j}{\alpha_i} & \text{für } i \neq j. \end{cases} \tag{31}$$

Sei λ ein Eigenwert von $\mathbf{A}$, dann gilt

$$\lambda \geq 3 + d - \sqrt{\frac{d(d+2)}{4}(M + M^{-1}) + \frac{(d+1)^2+1}{2}} \qquad \text{für } d \text{ gerade,}$$

$$\lambda \geq 3 + d - \sqrt{\frac{(d+1)^2}{4}(M + M^{-1}) + \frac{(d+1)^2}{2}} \qquad \text{für } d \text{ ungerade.}$$

Zunächst benötigen wir folgende Definitionen:

Definition 3.11. Definiere die Funktion

$$\begin{aligned} F : \mathbb{R}^d &\to \mathbb{R} \\ \mathbf{x} &\mapsto F(\mathbf{x}) := \sum_{i=1}^{d}\sum_{j=i}^{d}\left(\prod_{\ell=i}^{j} x_\ell + \prod_{\ell=i}^{j} x_\ell^{-1}\right), \end{aligned}$$

sowie für $1 < M \in \mathbb{R}$ die Menge

$$B_M := \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; M^{-1} \leq \prod_{\ell=i}^{j} x_\ell \leq M, 1 \leq i \leq j \leq d \right\}.$$

Für einen Vektor $\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ und Indizes $1 \leq i \leq j \leq d$ definieren wir

$$x_{ij} := \prod_{\ell=i}^{j} x_\ell, \tag{32}$$

und die verallgemeinerten Intervalle

$$\begin{aligned} \mathcal{I}_{k,\delta}(\mathbf{x}) &:= \{x_1\} \times \cdots \times \{x_{k-1}\} \times (x_k - \delta, x_k + \delta) \times \{x_{k+1}\} \times \cdots \times \{x_d\} \subset \mathbb{R}^d \\ \mathcal{J}_{k,\delta}(\mathbf{x}) &:= ((1+\delta)^{-1}x_k, (1+\delta)x_k) \subset \mathbb{R} \end{aligned}$$

für $\mathbf{x} \in \mathbb{R}^d$, $1 \leq k \leq d$ und $\delta > 0$. Mit f werden wir die Einschränkung von F auf B_M bezeichnen, also hat f mit der Notation aus (32) die Form

$$\begin{aligned} f : B_M &\to \mathbb{R} \\ \mathbf{x} &\mapsto f(\mathbf{x}) := \sum_{i=1}^{d}\sum_{j=i}^{d}\left(x_{ij} + x_{ij}^{-1}\right). \end{aligned}$$

Schließlich definieren wir noch die Menge

$$\widetilde{B}_M := B_M \cap \left\{M^{-1}, 1, M\right\}^d$$

für $1 < M \in \mathbb{R}$.

Satz 3.12. *Sei $1 < M \in \mathbb{R}$. Dann existiert ein $\mathbf{x} \in B_M$ mit $f(\mathbf{x}) = \max_{\mathbf{y} \in B_M} f(\mathbf{y}) \in \mathbb{R}$.*

Beweis. Da $1 < M \in \mathbb{R}$, gilt $0 \notin [M^{-1}, M] \neq \emptyset$. Damit ist $f : B_M \to \mathbb{R}$ eine stetige Abbildung auf einer kompakten Menge und nimmt ihr Maximum an. □

Zunächst wollen wir das Maximum von f auf B_M explizit bestimmen. Den ersten Schritt dazu macht der folgende Satz, welcher die überabzählbar vielen Kandidaten f zu maximieren auf endlich viele reduziert:

Satz 3.13. *Sei $f(\mathbf{x}^*) = \max_{\mathbf{x} \in B_M} f(\mathbf{x})$ für ein $\mathbf{x}^* = (x_1^*, ..., x_d^*) \in B_M$. Dann gilt $\mathbf{x}^* \in \widetilde{B}_M$.*

Beweis. Auf jedes $\mathbf{x}^* \in B_M$ trifft einer der folgenden Fälle zu:

(a) $\mathbf{x}^* \in \left\{M^{-1}, 1, M\right\}^d$.

(b) Es existieren ein $k \in \{1, ..., d\}$ und ein $\delta > 0$ so, dass $\mathcal{I}_{k,\delta}(\mathbf{x}^*) \subset B_M$.

(c) Es trifft weder (a) noch (b) zu, d.h. $\mathbf{x}^* \notin \left\{M^{-1}, 1, M\right\}^d$ und für alle $i \in \{1, ..., d\}$ und alle $\delta > 0$ gilt $\mathcal{I}_{i,\delta}(\mathbf{x}^*) \not\subset B_M$.

Wir werden zeigen, dass es für alle $\mathbf{x}^* \in B_M$, die (b) oder (c) erfüllen, ein $\mathbf{y}^* \in B_M$ gibt, welches $f(\mathbf{y}^*) > f(\mathbf{x}^*)$ erfüllt.

1. Fall: $\mathbf{x}^*$ erfülle (b). Wir wählen das entsprechende k und δ und definieren die Funktion

$$\begin{aligned} g : (x_k^* - \delta, x_k^* + \delta) &\to \mathbb{R} \\ y &\mapsto f(x_1^*, ..., x_{k-1}^*, y, x_{k+1}^*, ..., x_d^*). \end{aligned}$$

Aufgrund der Wahl von δ und k ist g wohldefiniert. Wir berechnen die zweite Ableitung von g:

$$g''(y) = \sum_{i=1}^{k} \sum_{j=k}^{d} 2y^{-3} \prod_{\substack{\ell=i \\ \ell \neq k}}^{j} x_\ell^{*^{-1}} > 0$$

und sehen, dass g strikt konvex ist. Also existiert ein $y^* \in (x_k^* - \delta, x_k^* + \delta)$ mit $g(y^*) > g(x_k^*)$ und es gilt $f(x_1^*, ..., x_{k-1}^*, y^*, x_{k+1}^*, ..., x_d^*) > f(\mathbf{x}^*)$. Insgesamt folgt, dass f in einem Punkt $\mathbf{x}^*$, auf den Fall (b) zutrifft, sein Maximum nicht annehmen kann.

2. Fall: Falls $\mathbf{x}^*$ nun (c) erfüllt, ist das Ganze etwas trickreicher. Zunächst wählen wir zwei Indizes $1 \leq i \leq j \leq d$ so, dass die folgenden drei Eigenschaften erfüllt sind:

(E1) Es existiert ein Index $k \in \{i, \ldots, j\}$ mit $x_k^* \notin \left\{M^{-1}, 1, M\right\}$.

(E2) $x_{ij}^* \in \left\{M^{-1}, M\right\}$.

(E3) (i, j) ist minimal, d.h. es existiert kein Indexpaar $(m, n) \neq (i, j)$ mit $i \leq m < n \leq j$, das (E1)–(E2) erfüllt.

Da auf $\mathbf{x}^*$ nicht Fall (a) zutrifft, kann (E1) erfüllt werden.
Falls (E2) nicht erfüllt werden kann, also für alle i, j, k, die (E1) erfüllen, $x_{ij}^* \in (M^{-1}, M)$ gilt, dann gibt es $k \in \{1, ..., d\}$ und $\delta > 0$ so, dass nicht (c), sondern Fall (b) auf $\mathbf{x}^*$ zutrifft: Um das einzusehen, wähle zum Beispiel k fest wie in (E1) und

$$\delta := \min_{i \leq k \leq j} \left\{ \frac{x_k^*}{x_{ij}^*}(M - x_{ij}^*), \frac{x_k^*}{x_{ij}^*}(x_{ij}^* - M^{-1}) \right\} > 0. \tag{33}$$

Mit dieser Wahl von k und δ gilt $\mathcal{I}_{k,\delta}(\mathbf{x}^*) \subset B_M$. Dafür genügt es zu zeigen, dass für alle $\mathbf{y} = (y_1, \ldots, y_d) \in \mathcal{I}_{k,\delta}(\mathbf{x}^*)$ auch $\mathbf{y} \in B_M$ gilt, d.h., dass

$$M^{-1} \leq y_{ij} \leq M \quad \text{für alle } i,j \in \{1,\ldots,d\} \text{ mit } i \leq j. \tag{34}$$

Dies ist für $i \leq j < k$ und $k < i \leq j$ offensichtlich, denn dann gilt $y_{ij} = x^*_{ij}$. Es bleibt der Fall $i \leq k \leq j$ in (34) zu zeigen. Beachte $\mathbf{y} \in \mathcal{I}_{k,\delta}(\mathbf{x}^*)$ impliziert $x^*_k - \delta < y_k < x^*_k + \delta$. Ausnutzen der Definition von δ zeigt

$$\begin{aligned}
M^{-1} &= x^*_{ij} - x^*_{ij} + M^{-1} = x^*_{ij} - \frac{x^*_k}{x^*_{ij}}(x^*_{ij} - M^{-1}) \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell \\
&\overset{(33)}{\leq} x^*_{ij} - \delta \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell = (x^*_k - \delta) \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell \\
&< y_k \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell = \prod_{\ell=i}^{j} y_\ell \\
&= y_{ij} \\
&= \prod_{\ell=i}^{j} y_\ell = y_k \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell \\
&< (x^*_k + \delta) \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell = x^*_{ij} + \delta \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell \\
&\overset{(33)}{\leq} x^*_{ij} + \frac{x^*_k}{x^*_{ij}}(M - x_{ij}) \prod_{\substack{\ell=i \\ \ell\neq k}}^{j} x^*_\ell = x^*_{ij} + M - x^*_{ij} = M
\end{aligned}$$

und damit die Behauptung (34).
Es seien also i,j,k so gewählt, dass sie (E1) und (E2) erfüllen. Falls i,j (E3) nicht erfüllen, existieren Indizes m,n mit $i \leq m < n \leq j$, $(m,n) \neq (i,j)$ die (E1)–(E2) erfüllen. Wir definieren $i := m$, $j := n$. Das machen wir so lange bis i,j (E3) erfüllen.

Wir können also i,j,k so wählen, dass (E1)–(E3) erfüllt sind. Fixiere beliebige solche Indizes i,j,k. Aus (E1) und (E2) folgt außerdem $j - i \geq 1$, denn $x_{ii*} = x^*_i$ nach Definition. Aufgrund von (E2) existiert außer k noch ein Index $p \neq k$, $i \leq p \leq j$ mit $x^*_p \notin \left\{M^{-1}, 1, M\right\}$. Wähle so einen Index p beliebig aber fest, und sei ohne Einschränkung der Allgemeinheit $p > k$. Wir definieren

$$\begin{aligned}
\delta_k &:= \min_{\substack{i\leq m\leq k\leq n\leq j \\ (m,n)\neq(i,j)}} \left\{Mx^*_{mn} - 1, M/x^*_{mn} - 1\right\}, \\
\delta_p &:= \min_{\substack{i\leq m\leq p\leq n\leq j \\ (m,n)\neq(i,j)}} \left\{Mx^*_{mn} - 1, M/x^*_{mn} - 1\right\}, \\
\delta &:= \min\left\{\delta_k, \delta_p\right\}.
\end{aligned}$$

Aufgrund von (E3) gilt $\delta > 0$. Wir definieren die Funktion

$$\begin{aligned}
g : \mathcal{J}_{k,\delta}(\mathbf{x}^*) &\to \mathbb{R} \\
y &\mapsto f(x^*_1, ..., x^*_{k-1}, y, x^*_{k+1}, ..., x^*_{p-1}, \frac{x^*_k x^*_p}{y}, x^*_{p+1}, ..., x^*_d).
\end{aligned}$$

Wir wollen wieder zeigen, dass g strikt konvex ist, jedoch müssen wir uns zuerst überlegen, dass g wohldefiniert ist. Es gilt also zu zeigen, dass

$$(y_1^*, ..., y_d^*) = \mathbf{y}^* = \mathbf{y}^*(y) := (x_1^*, ..., x_{k-1}^*, y, x_{k+1}^*, ..., x_{p-1}^*, \frac{x_k^* x_p^*}{y}, x_{p+1}^*, ..., x_d^*) \tag{35}$$

für alle $y \in \mathcal{J}_{k,\delta}(\mathbf{x}^*)$ ein Element von B_M ist. Es gilt genau dann $\mathbf{y}^* \in B_M$, wenn für alle $1 \leq m \leq n \leq d$ gilt, dass $M^{-1} \leq y_{mn}^* \leq M$. Nach Wahl von k, p gilt $i \leq k < p \leq j$. Für die Lage von $m \leq n$ bezüglich i, k, p, j sind 15 Fälle möglich:

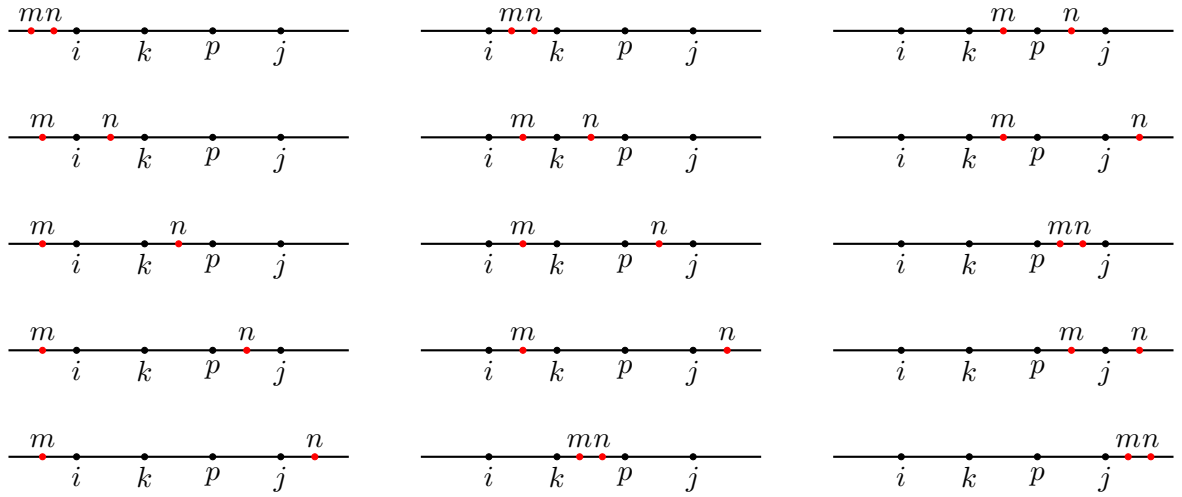


Abbildung 13: Die 15 möglichen Fälle, wie $m \leq n$ bezüglich i, k, p, j liegen können.

Für jeden dieser 15 Fälle gilt genau eine der folgenden Eigenschaften(i)–(viii):

(i). $m \leq n < k$

(ii). $m < i \leq k \leq n < p$

(iii). $i \leq m \leq k \leq n < p$

(iv). $m \leq k < p \leq n$

(v). $k < m \leq n < p$

(vi). $k < m \leq p \leq n \leq j$

(vii). $k < m \leq p \leq j < n$

(viii). $p < m \leq n$

Für (i), (v) und (viii) gilt $y_{mn}^* = x_{mn}^* \in [M^{-1}, M]$, da sich $\mathbf{y}^*$ von $\mathbf{x}^*$ nur an der k-ten und p-ten Stelle unterscheiden kann und $\mathbf{x}^* \in B_M$. Für (iv) gilt ebenfalls $y_{mn}^* = x_{mn}^*$, da $y_k^* y_p^* = x_k^* x_p^*$.
Nun zu (iii): Wir haben δ so gewählt, dass

$$Mx_{mn}^* - 1 \geq \delta,$$

was äquivalent zu

$$(1+\delta)^{-1} x_{mn}^* \geq M^{-1}$$

ist. Da $y_k^* = y > (1+\delta)^{-1} x_k^*$ gilt auch $y_{mn}^* > (1+\delta)^{-1} x_{mn}^* \geq M^{-1}$. Eine analoge Rechnung zeigt $y_{mn}^* < (1+\delta) x_{mn}^* \leq M$, was (iii) abschließt.
(vi) lässt sich analog zu (iii) behandeln: Wir haben δ so gewählt, dass

$$Mx_{mn}^* - 1 \geq \delta,$$

was äquivalent zu

$$(1+\delta)^{-1}x^*_{mn} \geq M^{-1}$$

ist. Da $y^*_p = x^*_k x^*_p / y > (1+\delta)^{-1}x^*_p$ gilt auch $y^*_{mn} > (1+\delta)^{-1}x^*_{mn} \geq M^{-1}$. Eine analoge Rechnung zeigt $y^*_{mn} < (1+\delta)x^*_{mn} \leq M$, was (vi) abschließt.
Für (ii) brauchen wir Ergebnisse einiger anderer, schon behandelter Fälle. Es gelte also $m < i \leq k \leq n < p$. Aus (iii) und (iv) wissen wir bereits, dass

$$M^{-1} \leq y^*_{in} \leq M \quad \text{sowie} \quad y^*_{ij} = x^*_{ij}.$$

Sei ohne Einschränkung der Allgemeinheit $y^*_{ij} = x^*_{ij} = M$ in (E2), der Fall $y^*_{ij} = x^*_{ij} = M^{-1}$ ließe sich analog behandeln. Damit gilt sogar $1 \leq y^*_{in} \leq M$, da ansonsten $y^*_{n+1,j} > M$ aufgrund von $y^*_{ij} = y^*_{in}y^*_{n+1,j} = M$ gelten würde, was aber im Widerspruch zu (vi) steht. Ferner gilt $M^{-1} \leq y^*_{m,i-1} \leq 1$, da ansonsten $x^*_{mj} = y^*_{mj} = y^*_{m,i-1}y^*_{ij} > M$ folgen würde, was ein Widerspruch zu $\mathbf{x}^* \in B_M$ ist. Also gilt insgesamt

$$y^*_{mn} = y^*_{m,i-1}y^*_{in} \in [M^{-1}, M].$$

(vii) lässt sich analog zu (ii) behandeln: Auch für (vii) brauchen wir Ergebnisse einiger anderer, schon behandelter Fälle. Es gelte also $k < m \leq p \leq j < n$. Aus (vi) und (iv) wissen wir bereits, dass

$$M^{-1} \leq y^*_{mj} \leq M \quad \text{sowie} \quad y^*_{ij} = x^*_{ij}.$$

Sei ohne Einschränkung der Allgemeinheit $y^*_{ij} = x^*_{ij} = M$ in (E2), der Fall $y^*_{ij} = x^*_{ij} = M^{-1}$ ließe sich analog behandeln. Damit gilt sogar $1 \leq y^*_{mj} \leq M$, da ansonsten $y^*_{i,m-1} > M$ aufgrund von $y^*_{ij} = y^*_{i,m-1}y^*_{m,j} = M$ gelten würde, was aber im Widerspruch zu (iii) steht. Ferner gilt $M^{-1} \leq y^*_{j+1,n} \leq 1$, da ansonsten $x^*_{in} = y^*_{in} = y^*_{ij}y^*_{j+1,n} > M$ folgen würde, was ein Widerspruch zu $\mathbf{x}^* \in B_M$ ist. Also gilt insgesamt

$$y^*_{mn} = y^*_{mj}y^*_{j+1,n} \in [M^{-1}, M].$$

Wir haben für (i)–(viii), also für alle $1 \leq m \leq n \leq d$ gezeigt, dass $M^{-1} \leq y^*_{mn} \leq M$. Insbesondere ist g wohldefiniert. Es bleibt noch die strikte Konvexität von g zu zeigen. Wir verwenden die Darstellung (35) und wiederholen die Definition von g:

$$\begin{aligned}
g(y) &= f(\mathbf{y}^*) \\
&= \sum_{1\leq i\leq j\leq d} (y^*_{ij} + y^{*^{-1}}_{ij}) \\
&= \sum_{\substack{1\leq i\leq j\leq d \\ k<i\leq p\leq j}} (y^*_{ij} + y^{*^{-1}}_{ij}) + \sum_{\substack{1\leq i\leq j\leq d \\ i\leq k\leq j<p}} (y^*_{ij} + y^{*^{-1}}_{ij}) + \sum_{\substack{1\leq i\leq j\leq d \\ i\leq j<k}} (y^*_{ij} + y^{*^{-1}}_{ij}) \\
&\quad + \sum_{\substack{1\leq i\leq j\leq d \\ i\leq k<p\leq j}} (y^*_{ij} + y^{*^{-1}}_{ij}) + \sum_{\substack{1\leq i\leq j\leq d \\ k<i\leq j<p}} (y^*_{ij} + y^{*^{-1}}_{ij}) + \sum_{\substack{1\leq i\leq j\leq d \\ p<i\leq j}} (y^*_{ij} + y^{*^{-1}}_{ij}) .
\end{aligned}$$

Wir verwenden $y_k^* = y$, $y_p^* = x_k^* x_p^*/y$ sowie $y_\ell^* = x_\ell^*$ für $\ell \neq k, p$ und erhalten

$$\begin{aligned} g(y) = y^{-1}x_k^* &\left(\sum_{\substack{1\le i\le j\le d\\ k<i\le p\le j}} x_{ij}^* + \sum_{\substack{1\le i\le j\le d\\ i\le k\le j<p}} {x_{ij}^*}^{-1} \right) + y{x_k^*}^{-1} \left(\sum_{\substack{1\le i\le j\le d\\ i\le k\le j<p}} x_{ij}^* + \sum_{\substack{1\le i\le j\le d\\ k<i\le p\le j}} {x_{ij}*}^{-1} \right) \\ &+ \sum_{\substack{1\le i\le j\le d\\ i\le j<k}} (x_{ij}^* + {x_{ij}^*}^{-1}) + \sum_{\substack{1\le i\le j\le d\\ i\le k<p\le j}} (x_{ij}^* + {x_{ij}^*}^{-1}) \\ &+ \sum_{\substack{1\le i\le j\le d\\ k<i\le j<p}} (x_{ij}^* + {x_{ij}^*}^{-1}) + \sum_{\substack{1\le i\le j\le d\\ p<i\le j}} (x_{ij}^* + {x_{ij}^*}^{-1})\,. \end{aligned}$$

Für die zweite Ableitung von g ergibt sich nun

$$g''(y) = 2y^{-3}x_k^* \left(\sum_{\substack{1\le i\le j\le d\\ k<i\le p\le j}} x_{ij}^* + \sum_{\substack{1\le i\le j\le d\\ i\le k\le j<p}} {x_{ij}^*}^{-1} \right) > 0\,.$$

Mit $g''(y) > 0$ für alle $y \in \mathcal{J}_{k,\delta}(\mathbf{x}^*)$ haben wir gezeigt, dass $g : \mathcal{J}_{k,\delta}(\mathbf{x}^*) \to \mathbb{R}$ strikt konvex ist. Also existiert ein $y^* \in \mathcal{J}_{k,\delta}(\mathbf{x}^*)$ mit $g(y^*) > g(x_k^*)$, und es folgt

$$f((x_1^*, ..., x_{k-1}^*, y^*, x_{k+1}^*, ..., x_{p-1}^*, \frac{x_k^* x_p^*}{y^*}, x_{p+1}^*, ..., x_d^*)) = g(y^*) > g(x_k^*) = f(\mathbf{x}^*).$$

Damit wäre gezeigt, dass f auch an keinem Punkt $\mathbf{x}^*$ auf den Fall (c) zutrifft, sein Maximum annehmen kann.

Insgesamt folgt, dass, falls f in $\mathbf{x}^*$ sein Maximum annimmt, Fall (a) auf $\mathbf{x}^*$ zutrifft. Das schließt den Beweis. □

Definition 3.14. Für Vektoren $\mathbf{x} = (x_1, ..., x_d) \in \widetilde{B}_M$, $\mathbf{y} = (y_1, ..., y_n) \in \left\{M^{-1}, 1, M\right\}^n$ definieren wir

$$\begin{aligned} p(\mathbf{x}) &:= \#\left\{(i,j) \in \{1,...,d\} \times \{1,...,d\} \mid i \le j, x_{ij} = 1\right\}, \\ q(\mathbf{x}) &:= \#\left\{(i,j) \in \{1,...,d\} \times \{1,...,d\} \mid i \le j, x_{ij} \neq 1\right\} \\ r(\mathbf{y}) &:= \#\left\{i \in \{1,...,n\} \mid y_i = 1\right\}, \\ s(\mathbf{y}) &:= \#\left\{i \in \{1,...,n\} \mid y_i \neq 1\right\}, \end{aligned}$$

sowie analog zu $q(\mathbf{x})$

$$\widetilde{q}(x_1, ..., x_{d-1}) := \#\left\{(i,j) \in \{1,...,d-1\} \times \{1,...,d-1\} \mid i \le j, x_{ij} \neq 1\right\},$$

wobei # eine Menge auf die Anzahl ihrer Elemente abbilden soll. Offensichtlich gilt $p(\mathbf{x}) + q(\mathbf{x}) = d(d+1)/2$ sowie $r(\mathbf{y}) + s(\mathbf{y}) = n$.

Lemma 3.15. *Sei $\mathbf{x} = (x_1, ..., x_d) \in \widetilde{B}_M$. Dann gilt mit den Notationen aus Definition 3.11 und Definition 3.14 für alle $1 \le i \le j \le d$:*

(i). $x_{ij} \in \left\{M^{-1}, 1, M\right\}$.

(ii). *Aus $x_{ij} = M$ folgt $x_{i\ell}, x_{kj} \in \{1, M\}$ für alle $k \in \{1, ..., j\}, \ell \in \{i, ..., d\}$.*

(iii). *Aus* $x_{ij} = M^{-1}$ *folgt* $x_{i\ell}, x_{kj} \in \left\{M^{-1}, 1\right\}$ *für alle* $k \in \{1, ..., j\}, \ell \in \{i, ..., d\}$.

(iv). $q(\mathbf{x}) = \widetilde{q}(x_1, ..., x_{d-1}) + s(x_{1d}, x_{2d}, ..., x_{dd})$.

(v). $f(\mathbf{x}) = d(d+1) + (M + M^{-1} - 2)q(\mathbf{x})$.

(vi). *Aus* $q(\mathbf{x}) = \max\limits_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} q(\widetilde{\mathbf{x}})$ *folgt auch* $f(\mathbf{x}) = \max\limits_{\mathbf{y} \in B_M} f(\mathbf{y})$.

Beweis. Sei $\mathbf{x} = (x_1, ..., x_d) \in \widetilde{B}_M$ beliebig aber fest.
(i). Es gilt $x_\ell \in \left\{M^{-1}, 1 = M^0, M\right\}$ für alle $1 \leq \ell \leq d$ und damit

$$x_{ij} = \prod_{\ell=i}^{j} x_\ell = \prod_{\ell=i}^{j} M^{k_\ell} = M^{\sum\limits_{\ell=i}^{j} k_\ell} = M^k,$$

wobei $k_\ell \in \{-1, 0, 1\}$ und $k := \sum\limits_{\ell=i}^{j} k_\ell$. Es gilt offensichtlich $k \in \mathbb{Z}$. Da $\mathbf{x} \in B_M$, gilt $x_{ij} \in [M^{-1}, M]$ und damit insgesamt $k \in \{-1, 0, 1\}$.
(ii). Sei $\ell \in \{i, ..., d\}$, nach (i) bleibt nur noch $x_{i\ell} \neq M^{-1}$ zu zeigen. Falls $\ell = j$ gilt nach Voraussetzung $x_{i\ell} = M$. Im Fall $\ell < j$ folgt die Behauptung aus $M = x_{ij} = x_{i\ell} x_{\ell+1,j}$, da $x_{\ell+1,j} \neq M^2$ nach (i). Falls $\ell > j$ folgt die Behauptung aus $x_{i\ell} = x_{ij} x_{j+1,\ell} = M x_{j+1,\ell} \neq M^{-1}$, da $x_{j+1,\ell} \neq M^{-2}$ nach (i). Für $k \in \{1, ..., j\}$ verläuft der Beweis analog.
(iii). Der Beweis verläuft analog zum Beweis von (ii).
(iv). Gemäß Definition 3.14 gilt:

$$\begin{aligned} q(\mathbf{x}) &= \#\left\{(i,j) \in \{1, ..., d\} \times \{1, ..., d\} \mid i \leq j, x_{ij} \neq 1\right\} \\ &= \#\left\{(i,j) \in \{1, ..., d-1\} \times \{1, ..., d-1\} \mid i \leq j, x_{ij} \neq 1\right\} \\ &\quad + \#\left\{i \in \{1, ..., d\} \mid x_{id} \neq 1\right\} \\ &= \widetilde{q}(x_1, ..., x_{d-1}) + s(x_{1d}, x_{2d}, ..., x_{dd}). \end{aligned}$$

(v). Da $\mathbf{x} \in \widetilde{B}_M$, gilt nach Punkt (i), dass $x_{ij} \in \left\{M^{-1}, 1, M\right\}$ für alle $1 \leq i \leq j \leq d$. Weiters erinnern wir daran, dass nach Definition 3.14 die Gleichung $p(\mathbf{x}) + q(\mathbf{x}) = d(d+1)/2$ gilt. Insgesamt folgt

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^{d} \sum_{j=i}^{d} x_{ij} + x_{ij}^{-1} \\ &\overset{\text{(i)}}{=} (1 + 1^{-1}) p(\mathbf{x}) + (M + M^{-1}) q(\mathbf{x}) \\ &= 2(d(d+1)/2 - q(\mathbf{x})) + (M + M^{-1}) q(\mathbf{x}) \\ &= d(d+1) + (M + M^{-1} - 2) q(\mathbf{x}). \end{aligned}$$

(vi). Nach Satz 3.13 nimmt $f : B_M \to \mathbb{R}$ sein Maximum auf $\widetilde{B}_M \subset B_M$ an. Nach Punkt (v) gilt $f(\widetilde{\mathbf{x}}) = d(d+1) + (M + M^{-1} - 2) q(\widetilde{\mathbf{x}})$ für alle $\widetilde{\mathbf{x}} \in \widetilde{B}_M$. Weiters gilt $M + M^{-1} - 2 =$

$(M^{1/2} - M^{-1/2})^2 > 0$. Sei nun $q(\mathbf{x}) = \max_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} q(\widetilde{\mathbf{x}})$, dann folgt:

$$\begin{aligned}
\max_{\mathbf{y} \in B_M} f(\mathbf{y}) &\overset{3.13}{=} \max_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} f(\widetilde{\mathbf{x}}) \\
&\overset{(v)}{=} \max_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} (d(d+1) + \underbrace{(M + M^{-1} - 2)}_{>0} q(\widetilde{\mathbf{x}})) \\
&= d(d+1) + (M + M^{-1} - 2) \max_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} q(\widetilde{\mathbf{x}}) \\
&= d(d+1) + (M + M^{-1} - 2) q(\mathbf{x}) \\
&\overset{(v)}{=} f(\mathbf{x}) \,.
\end{aligned}$$

Also falls $q(\mathbf{x}) = \max_{\widetilde{\mathbf{x}} \in \widetilde{B}_M} q(\widetilde{\mathbf{x}})$ für ein $\mathbf{x} \in \widetilde{B}_M$, gilt auch $f(\mathbf{x}) = \max_{\mathbf{y} \in B_M} f(\mathbf{y})$. □

Mit Hilfe des folgenden Satzes können wir die Suche nach dem Maximum von f abschließen:

Satz 3.16. *Seien $d \in \mathbb{N}$ und $1 < M \in \mathbb{R}$ beliebig. Sei $\mathbf{x} \in \widetilde{B}_M$ mit $r(x_{1d}, x_{2d}, ..., x_{dd}) = k$. Dann gilt $q(\mathbf{x}) = (k+1)(d-k)$.*

Beweis durch Induktion nach d. Für $d = 1$ und $\mathbf{x} = x \in \left\{M^{-1}, 1, M\right\}$ sei $r(x) = k \in \{0, 1\}$. Wir müssen $q(x) = (k+1)(d-k)$ zeigen. Wir unterscheiden zwei Fälle:

(a). Falls $k = 1$ folgt aus der Definition von $r(\cdot)$ direkt $x = 1$ und damit

$$q(x) := \#\left\{(i,j) \in \{1\} \times \{1\} \mid x_{ij} \neq 1\right\} = 0 = (1+1)(1-1) = (k+1)(d-k)\,.$$

(b). Falls hingegen $k = 0$ folgt dieses Mal aus der Definition von $r(\cdot)$, dass $x \in \left\{M^{-1}, M\right\}$ und damit

$$q(x) := \#\left\{(i,j) \in \{1\} \times \{1\} \mid x_{ij} \neq 1\right\} = 1 = (0+1)(1-0) = (k+1)(d-k)\,.$$

Also gilt die Aussage für $d = 1$.

Sei nun $d \in \mathbb{N}_{>1}$ beliebig und die Aussage gelte für $d-1$. Sei $r(x_{1d}, x_{2d}, ..., x_{dd}) = k$, d.h. wir müssen $q(\mathbf{x}) = (k+1)(d-k)$ zeigen. Wir unterscheiden drei Fälle:
1. Fall: Sei zunächst $x_d = x_{dd} = 1$. Es gilt $x_{i,d-1} = x_{i,d-1} x_d = x_{id}$ für alle $1 \leq i \leq d-1$. Damit gilt

$$r(x_{1,d-1}, x_{2,d-1}, ..., x_{d-1,d-1}) = r(x_{1d}, x_{2d}, ..., x_{d-1,d}) = k - 1\,.$$

Laut Induktionsannahme gilt nun

$$\widetilde{q}(x_1, ..., x_{d-1}) = ((k-1)+1)((d-1)-(k-1)) = k(d-k)\,.$$

Wir erinnern uns, dass aus Definition 3.14 direkt $r(x_{1d}, ..., x_{dd}) + s(x_{1d}, ..., x_{dd}) = d$ folgt, und erhalten mit Lemma 3.15(iv)

$$\begin{aligned}
q(\mathbf{x}) &= \widetilde{q}(x_1, ..., x_{d-1}) + s(x_{1d}, ..., x_{dd}) \\
&= k(d-k) + (d - r(x_{1d}, ..., x_{dd})) \\
&= k(d-k) + (d-k) \\
&= (k+1)(d-k)\,.
\end{aligned}$$

2. Fall: Sei nun $x_d = x_{dd} = M$. Nach Lemma 3.15(ii) gilt $x_{id} \in \{1, M\}$ für alle $i = 1, ..., d-1$. Weiters folgt mit der Beziehung $x_{id} = x_{i,d-1}x_d = x_{i,d-1}M$

$$x_{i,d-1} = \begin{cases} 1, & \text{falls } x_{id} = M \\ M^{-1}, & \text{falls } x_{id} = 1 \end{cases} \qquad i = 1, ...d-1\,. \tag{36}$$

Wir erinnern uns, dass aus Definition 3.14 direkt $r(x_{1d}, ..., x_{dd}) + s(x_{1d}, ..., x_{dd}) = d$ folgt, verwenden (36) und erhalten

$$\begin{aligned} r(x_{1,d-1}, x_{2,d-1}, ..., x_{d-1,d-1}) &\overset{(36)}{=} s(x_{1d}, x_{2d}, ...x_{d-1,d}) \\ &= s(x_{1d}, ...x_{dd}) - 1 \\ &= d - r(x_{1d}, ..., x_{dd}) - 1 \\ &= d - k - 1\,. \end{aligned}$$

Laut Induktionsannahme gilt nun

$$\widetilde{q}(x_1, ..., x_{d-1}) = ((d-k-1)+1)((d-1)-(d-k-1)) = k(d-k)\,. \tag{37}$$

Wieder verwenden wir die Beziehung $r(x_{1d}, ..., x_{dd}) + s(x_{1d}, ..., x_{dd}) = d$, sowie Lemma 3.15(iv) und erhalten

$$\begin{aligned} q(\mathbf{x}) &= \widetilde{q}(x_1, ..., x_{d-1}) + s(x_{1d}, ..., x_{dd}) \\ &= k(d-k) + (d - r(x_{1d}, ..., x_{dd})) \\ &= k(d-k) + (d-k) \\ &= (k+1)(d-k)\,. \end{aligned}$$

3. Fall: Der Fall $x_d = x_{dd} = M^{-1}$ kann analog zum 2. Fall behandelt werden. □

Satz 3.17. *Sei $1 < M \in \mathbb{R}$ und $d \in \mathbb{N}$. Dann gilt*

$$\begin{aligned} \max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) &= \frac{d(d+2)}{4} \quad \textit{für } d \textit{ gerade,} \\ \max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) &= \frac{(d+1)^2}{4} \quad \textit{für } d \textit{ ungerade.} \end{aligned}$$

Beweis. Nach der Definition von $r(\cdot)$ gilt $r(y_1, ..., y_d) = r(y_{\pi(1)}, ..., y_{\pi(d)})$ für jede Permutation π von $\{1, ..., d\}$. Dementsprechend wollen wir $\widetilde{B}_M$ in $d+1$ Äquivalenzklassen

$$\mathbf{X}^{(k)} := \left\{ \mathbf{x} \in \widetilde{B}_M \mid r(x_{1d}, x_{2d}, ..., x_{dd}) = k \right\} \qquad k = 0, ..., d$$

zerlegen. Diese Äquivalenzklassen sind trivialerweise paarweise disjunkt, und ihre Vereinigung überdeckt ganz $\widetilde{B}_M$. Um zu zeigen, dass sie auch alle nichtleer sind, wählen wir jeweils einen Repräsentanten: Für $0 \le k \le d$ wählen wir $\mathbf{x}^{(k)}$ als Repräsentant der Äquivalenzklasse $\mathbf{X}^{(k)}$. Für $k = 0, ..., d$ seien die $\mathbf{x}^{(k)} = (x_1^{(k)}, ..., x_d^{(k)})$ wie folgt definiert:

$$x_i^{(k)} := \begin{cases} M & \text{für } i = d-k \\ 1 & \text{sonst} \end{cases} \qquad i = 1, ..., d.$$

Durch Nachrechnen sieht man für $1 \leq i \leq j \leq d$, $k = 0, ..., d$ sofort

$$x_{ij}^{(k)} = \begin{cases} M & \text{für } i \leq d-k \leq j \\ 1 & \text{sonst,} \end{cases} \quad \text{insb.} \quad x_{id}^{(k)} = \begin{cases} M & \text{für } i \leq d-k \\ 1 & \text{sonst.} \end{cases}$$

Also $\mathbf{x}^{(k)} \in \widetilde{B}_M$, $r(\mathbf{x}^{(k)}) = k$ und somit $\mathbf{x}^{(k)} \in \mathbf{X}^{(k)}$. Insbesondere sehen wir auch, dass alle Äquivalenzklassen nichtleer sind. Nach Satz 3.16 gilt, dass alle Elemente derselben Äquivalenzklasse $\mathbf{X}^{(k)}$ dasselbe Bild unter q haben. Ferner gilt $q(\mathbf{x}^{(k)}) = q(\mathbf{x}^{(d-k-1)})$ für $0 \leq k < d$ und q nimmt sein Maximum mit Sicherheit nicht in $\mathbf{x}^{(d)}$ an, da $q(\mathbf{x}^{(d)}) = 0$. Wenn wir nun diese Vorarbeit kombinieren, sehen wir

$$\max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) = \max_{0 \leq k \leq d/2-1} q(\mathbf{x}^{(k)}) \qquad \text{für } d \text{ gerade,} \tag{38}$$

$$\max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) = \max_{0 \leq k \leq (d-1)/2} q(\mathbf{x}^{(k)}) \qquad \text{für } d \text{ ungerade,} \tag{39}$$

da $\bigcup_{k=0}^{d} \mathbf{X}^{(k)} = \widetilde{B}_M$. Aus $\mathbf{x}^{(k)} \in \mathbf{X}^{(k)}$ folgt nach Definition von $\mathbf{X}^{(k)}$ direkt $r(x_{1d}^{(k)}, ..., x_{dd}^{(k)}) = k$. Mit Satz 3.16 folgt nun

$$q(\mathbf{x}^{(k)}) = (k+1)(d-k). \tag{40}$$

Die Funktion $h(k) := -k^2 + (d-1)k + d$ ist auf $[0, (d-1)/2] \subset \mathbb{R}$ monoton steigend und erfüllt $q(\mathbf{x}^{(k)}) = h(k)$. Also gilt mit (38)–(40) insgesamt

$$\max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) = q(\mathbf{x}^{(d/2-1)}) = \frac{d}{2}\frac{d+2}{2} \qquad \text{für } d \text{ gerade,}$$

$$\max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) = q(\mathbf{x}^{((d-1)/2)}) = \frac{d+1}{2}\frac{d+1}{2} \qquad \text{für } d \text{ ungerade.}$$

Das schließt den Beweis. □

Um Satz 3.10 zu beweisen, benötigen wir noch folgendes Resultat aus [BPS02]:

Lemma 3.18 ([BPS02, Proposition 6.1.])**.** *Seien $\alpha_1, ..., \alpha_{d+1} \in \mathbb{R} \setminus \{0\}$, die Matrix $\mathbf{A} \in \mathbb{R}^{(d+1)\times(d+1)}$ wie in* (31) *und definiere*

$$\lambda_\pm := 3 + d \pm \sqrt{\sum_{i=1}^{d+1} \alpha_i^2 \cdot \sum_{i=1}^{d+1} \alpha_i^{-2}}. \tag{41}$$

Dann liegen alle Eigenwerte von $\mathbf{A}$ in der Menge $\{\lambda_+, \lambda_-, 2\}$. □

Nun können wir Satz 3.10 beweisen:

Beweis von Satz 3.10. Nach Lemma 3.18 liegen alle Eigenwerte λ von $\mathbf{A}$ in der Menge $\{\lambda_+, \lambda_-, 2\}$. Da $(M + M^{-1}) = (M^{1/2} - M^{-1/2})^2 + 2 > 2$, gilt die Aussage für $\lambda = 2$, denn

$$\begin{aligned} 2 &= 3 + d - \sqrt{(d+1)^2} \\ &= 3 + d - \sqrt{\frac{d(d+2)}{4} \cdot 2 + \frac{(d+1)^2 + 1}{2}} \\ &\geq 3 + d - \sqrt{\frac{d(d+2)}{4}(M + M^{-1}) + \frac{(d+1)^2 + 1}{2}} \qquad \text{für } d \text{ gerade,} \end{aligned}$$

sowie

$$\begin{aligned} 2 &= 3 + d - \sqrt{\frac{(d+1)^2}{4} \cdot 2 + \frac{(d+1)^2}{2}} \\ &\geq 3 + d - \sqrt{\frac{(d+1)^2}{4}(M + M^{-1}) + \frac{(d+1)^2}{2}} \quad \text{für } d \text{ ungerade.} \end{aligned}$$

Offensichtlich gilt $\lambda_+ \geq \lambda_-$ in (41), also genügt es die Aussage für $\lambda = \lambda_-$ zu zeigen. Es gilt

$$\lambda_- = 3 + d - \sqrt{\sum_{i=1}^{d+1} \alpha_i^2 \cdot \sum_{i=1}^{d+1} \alpha_i^{-2}}. \tag{42}$$

Durch Umformen von (42) erhalten wir

$$\lambda_- = 3 + d - \sqrt{d + 1 + \sum_{i=1}^{d} \sum_{j=i+1}^{d+1} \left(\frac{\alpha_i^2}{\alpha_j^2} + \frac{\alpha_j^2}{\alpha_i^2} \right)}.$$

Wir definieren $a_{ij} := \alpha_i^2/\alpha_j^2$ und sehen

$$\lambda_- = 3 + d - \sqrt{d + 1 + \sum_{i=1}^{d} \sum_{j=i+1}^{d+1} (a_{ij} + a_{ij}^{-1})}.$$

Es gilt $a_{ij} = \prod_{\ell=i}^{j-1} a_{\ell,\ell+1}$, also

$$\lambda_- = 3 + d - \sqrt{d + 1 + \sum_{i=1}^{d} \sum_{j=i+1}^{d+1} \left(\prod_{\ell=i}^{j-1} a_{\ell,\ell+1} + \prod_{\ell=i}^{j-1} a_{\ell,\ell+1}^{-1} \right)}.$$

Und schließlich

$$\lambda_- = 3 + d - \sqrt{d + 1 + F(\mathbf{a})}\,, \tag{43}$$

mit einem Vektor $\mathbf{a} = (a_{12}, a_{23}, ..., a_{d,d+1}) \in \mathbb{R}^d$ und der Funktion

$$\begin{aligned} F : \mathbb{R}^d &\to \mathbb{R} \\ \mathbf{x} &\mapsto F(\mathbf{x}) := \sum_{i=1}^{d} \sum_{j=i}^{d} \left(\prod_{\ell=i}^{j} x_\ell + \prod_{\ell=i}^{j} x_\ell^{-1} \right), \end{aligned}$$

aus Definition 3.11. Wir nutzen die Voraussetzung $M^{-1} \leq \alpha_i^2/\alpha_j^2 \leq M$ für $i, j = 1, ..., d+1$ aus, und erhalten

$$\lambda_- \geq 3 + d - \sqrt{d + 1 + \max_{\mathbf{y} \in B_M} f(\mathbf{y})}\,,$$

mit $f : B_M \to \mathbb{R}$ aus Definition 3.11. Nach Satz 3.13 nimmt f sein Maximum auf $\widetilde{B}_M \subset B_M$ an, also gilt

$$\lambda_- \geq 3 + d - \sqrt{d + 1 + \max_{\mathbf{x} \in \widetilde{B}_M} f(\mathbf{x})}\,.$$

Mit Lemma 3.15(v)–(vi) gilt nun

$$\lambda_- \geq 3 + d - \sqrt{d + 1 + d(d+1) + (M + M^{-1} - 2) \max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x})}\,.$$

Nach Satz 3.17 gilt

$$\begin{aligned} \max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) &= \frac{d(d+2)}{4} \quad \text{für } d \text{ gerade,} \\ \max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x}) &= \frac{(d+1)^2}{4} \quad \text{für } d \text{ ungerade.} \end{aligned}$$

Durch Einsetzen von $\max_{\mathbf{x} \in \widetilde{B}_M} q(\mathbf{x})$ folgt die Behauptung auch für λ_-, und damit gilt die Aussage für alle Eigenwerte λ der Matrix $\mathbf{A}$. □

Später wird sich die Frage stellen, wie groß M gewählt werden kann, sodass λ aus Satz 3.10 stets positiv ist. Diese Frage werden wir im folgenden Satz beantworten. Zur kompakteren Notation definieren wir

$$\mu_1(d) := \begin{cases} \frac{d(d+2)}{4} & \text{für } d \text{ gerade} \\ \frac{(d+1)^2}{4} & \text{für } d \text{ ungerade} \end{cases} \quad \mu_2(d) := \begin{cases} \frac{(d+1)^2+1}{2} & \text{für } d \text{ gerade} \\ \frac{(d+1)^2}{2} & \text{für } d \text{ ungerade} \end{cases} \tag{44}$$

und

$$\beta(d) := \frac{(d+3)^2 - \mu_2(d)}{2\mu_1(d)} \tag{45}$$

$$B(d) := \exp(\operatorname{arcosh} \beta(d))\,. \tag{46}$$

Satz 3.19. *Es gelten die Voraussetzungen von Satz 3.10. Falls zusätzlich* $1 < M < B(d)$ *erfüllt ist, so gilt für alle Eigenwerte* λ *der Matrix* $\mathbf{A}$ *stets* $\lambda > 0$.

Beweis. Beachte zunächst, dass für $1 < a < b$ gilt

$$(b + b^{-1})ab - (a + a^{-1})ab = ab^2 - b + a - a^2 b = (ab - 1)(b - a) > 0$$

und deshalb $b + b^{-1} > a + a^{-1}$ für alle $1 < a < b$. Folglich gilt für $1 < M < B(d)$, dass $(M + M^{-1}) < (B(d) + B(d)^{-1})$. Damit und mit den Definitionen (45)–(46) schätzen wir λ nach unten ab:

$$\begin{aligned} \lambda &\overset{3.10}{\geq} d + 3 - \sqrt{\mu_1(d)(M + M^{-1}) + \mu_2(d)} \\ &> d + 3 - \sqrt{\mu_1(d)(B(d) + B(d)^{-1}) + \mu_2(d)} \\ &= d + 3 - \sqrt{2\mu_1(d)(\exp(\operatorname{arcosh} \beta(d)) + \exp(-\operatorname{arcosh} \beta(d)))/2 + \mu_2(d)}\,. \end{aligned}$$

Durch Ausnutzen der Identität $\cosh(x) = (\exp(x) + \exp(-x))/2$ erhalten wir schließlich

$$\begin{aligned} \lambda &> d + 3 - \sqrt{2\mu_1(d)\cosh(\operatorname{arcosh} \beta(d)) + \mu_2(d)} \\ &= d + 3 - \sqrt{(d+3)^2 - \mu_2(d) + \mu_2(d)} \\ &= 0\,. \end{aligned}$$

Also gilt $\lambda > 0$, was zu zeigen war. □

3.3 Hinreichende Kriterien

Die nächste Proposition stellt Kriterien zur Verfügung, um die Voraussetzungen (28) von Proposition 3.9 zu überprüfen:

Proposition 3.20. *Sei $\mathcal{T}$ ein γ-formreguläres Gitter. Seien für $z_\ell \in \mathcal{T}$ die Skalare $h_\ell > 0$ so gewählt, dass für alle Elemente $T \in \mathcal{T}$ und alle $z_j, z_k \in \mathcal{N}(T)$ stets*

$$\frac{h_j^2}{h_k^2} \leq C_{11} < B(d)\,, \tag{47}$$

sowie

$$C_{12}\mathtt{diam}(T) \leq h_j \leq C_{13}\mathtt{diam}(T) \tag{48}$$

mit Konstanten $C_{11}, C_{12}, C_{13} > 0$ und $B(d)$ aus (46) gilt. Dann sind die Voraussetzungen (28) von Proposition 3.9 erfüllt. Die Konstante C_9 hängt nur von der γ-Formregularität und C_{12}^{-1} ab. C_{10} hängt nur von der γ-Formregularität, C_{13} und $\varepsilon := B(d) - C_{11}$ ab.

Beweis. **Schritt 1:** Zunächst zeigen wir die untere Abschätzung in (28): Mit Lemma 3.3 und Lemma 3.4 sieht man die Äquivalenz

$$|T|\mathbf{y}^\top\mathbf{y} \simeq \mathbf{y}^\top\mathbf{M}_T\mathbf{y} \quad \text{für alle } \mathbf{y} \in \mathbb{R}^{d+1}\,, \tag{49}$$

wobei die versteckten Konstanten nur von der γ-Formregularität abhängen. Diese Äquivalenz und die untere Abschätzung in (48) zeigen

$$\mathbf{x}^\top\boldsymbol{\Lambda}_T^2\mathbf{M}_T\boldsymbol{\Lambda}_T^2\mathbf{x} = (\boldsymbol{\Lambda}_T^2\mathbf{x})^\top\mathbf{M}_T(\boldsymbol{\Lambda}_T^2\mathbf{x}) \overset{(49)}{\simeq} |T|(\boldsymbol{\Lambda}_T^2\mathbf{x})^\top(\boldsymbol{\Lambda}_T^2\mathbf{x}) \overset{(48)}{\lesssim} |T|\mathbf{x}^\top\mathbf{x} \overset{(49)}{\simeq} \mathbf{x}^\top\mathbf{M}_T\mathbf{x}\,.$$

Das zeigt die untere Abschätzung in (28), und $C_9 > 0$ hängt nur von der γ-Formregularität und C_{12} ab.
Schritt 2: Als Nächstes zeigen wir die obere Abschätzung in (28) bis auf eine verbleibende Eigenwertabschätzung: Wie in Lemma 3.3 fixieren wir den Referenzsimplex $\widehat{T} \subset \mathbb{R}^d$ so, dass die zugehörige Massenmatrix $\widehat{\mathbf{M}}_{jk} = 1 + \delta_{jk}$ mit Kroneckers Delta erfüllt. Nach Lemma 3.4 gilt $\mathbf{M}_T = C_8|T|\widehat{\mathbf{M}}$ mit einer Konstanten $C_8 > 0$. Wir definieren die Matrix

$$\mathbf{A}_T := \boldsymbol{\Lambda}_T^2\mathbf{M}_T + \mathbf{M}_T\boldsymbol{\Lambda}_T^2 = C_8|T|(\boldsymbol{\Lambda}_T^2\widehat{\mathbf{M}} + \widehat{\mathbf{M}}\boldsymbol{\Lambda}_T^2) =: C_8|T|\widehat{\mathbf{A}}_T\,.$$

Die symmetrische Matrix $\widehat{\mathbf{B}}_T := \boldsymbol{\Lambda}_T^{-1}\widehat{\mathbf{A}}_T\boldsymbol{\Lambda}_T^{-1}$ erfüllt

$$(\widehat{\mathbf{B}}_T)_{jk} = \left(\boldsymbol{\Lambda}_T\widehat{\mathbf{M}}\boldsymbol{\Lambda}_T^{-1} + \boldsymbol{\Lambda}_T^{-1}\widehat{\mathbf{M}}\boldsymbol{\Lambda}_T\right)_{jk} = \left(\frac{h_j}{h_k} + \frac{h_k}{h_j}\right)\widehat{\mathbf{M}}_{jk} = \left(\frac{h_j}{h_k} + \frac{h_k}{h_j}\right)(1 + \delta_{jk})\,.$$

Falls nun der minimale Eigenwert $\lambda_{\min}$ von $\widehat{\mathbf{B}}_T$ positiv ist, erhalten wir mit $\mathbf{y} = \boldsymbol{\Lambda}_T\mathbf{x}$

$$\mathbf{x}^\top\mathbf{M}_T\mathbf{x} \overset{(49)}{\simeq} |T|\mathbf{x}^\top\mathbf{x} \overset{(48)}{\lesssim} |T|\mathbf{y}^\top\mathbf{y} \lesssim |T|\mathbf{y}^\top\widehat{\mathbf{B}}_T\mathbf{y}\,,$$

wobei die versteckten Konstanten nur von der γ-Formregularität, $C_{13} > 0$ und $\lambda_{\min} > 0$ abhängen. Symmetrie von $\boldsymbol{\Lambda}_T$ und $\mathbf{M}_T$ liefern $\mathbf{x}^\top\mathbf{A}_T\mathbf{x} = 2\mathbf{x}^\top\boldsymbol{\Lambda}_T^2\mathbf{M}_T\mathbf{x}$ und damit

$$|T|\mathbf{y}^\top\widehat{\mathbf{B}}_T\mathbf{y} = |T|\mathbf{x}^\top\widehat{\mathbf{A}}_T\mathbf{x} \simeq \mathbf{x}^\top\mathbf{A}_T\mathbf{x} = 2\mathbf{x}^\top\boldsymbol{\Lambda}_T^2\mathbf{M}_T\mathbf{x}\,.$$

Kombinieren der letzten zwei Rechnungen zeigt die obere Abschätzung in (28) und $C_{10} > 0$ hängt nur von der γ-Formregularität, $C_{13} > 0$ und $\lambda_{\min} > 0$ ab.
Schritt 3: Es bleibt noch zu zeigen, dass der minimale Eigenwert $\lambda_{\min}$ von $\widehat{\mathbf{B}}_T$ positiv ist. $\widehat{\mathbf{B}}_T$ hat die Form (31) aus Satz 3.10. Mit Voraussetzung (47) folgt aus Satz 3.19, dass $\lambda_{\min}$ positiv ist. □

In der nächsten Proposition geben wir ein handliches, hinreichendes Kriterium an, welches die Erfüllbarkeit der Voraussetzungen (47)–(48) von Proposition 3.20 garantiert:

Proposition 3.21. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$. Existiert eine Konstante $L_{\text{patch}} < \frac{d}{2}\log_2 B(d)$ so, dass für alle $z \in \mathcal{N} \setminus \mathcal{N}_0$ und alle $T, T' \in \omega(z)$*

$$|\texttt{level}(T') - \texttt{level}(T)| \leq L_{\text{patch}} \tag{50}$$

gilt, dann können die Voraussetzungen (47)–(48) von Proposition 3.20 stets erfüllt werden, und die L^2-orthogonale Projektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ ist H^1_D-stabil. Die Konstante C_{11} hängt nur von L_{patch} und d ab, C_{12} hängt nur von $C_7, C_4, \#\mathcal{N}_0, C_3, L_{\text{patch}}$ und d ab, C_{13} hängt nur von C_7, C_5 und d ab.

Beweis.

(i). Wir beginnen mit der Definition der Werte h_j. Dazu definieren wir zunächst einen Knoten-Element-Abstand $\delta : \mathcal{N} \times \mathcal{T} \to \mathbb{N}_0$.

$$\begin{aligned}\delta(z,T) := n \quad & \text{für die kleinste Zahl } n \in \mathbb{N}_0 \text{ so, dass Elemente } T_0, \dots, T_n \in \mathcal{T} \\ & \text{existieren mit } z \in T_0,\ T = T_n, \\ & \text{und } T_i \cap T_{i+1} \neq \emptyset \text{ für alle } i = 0, \dots, n-1.\end{aligned}$$

Für $z_j \in \mathcal{N}$ definiere

$$h_j := \min_{T'_j \in \mathcal{T}} 2^{(L_{\text{patch}}\delta(z_j, T'_j) - \texttt{level}(T'_j))/d}. \tag{51}$$

(ii). Wir überprüfen zunächst Voraussetzung (47) von Proposition 3.20. Sei also $T \in \mathcal{T}$ beliebig und $z_j, z_k \in T$. Nach (51) existiert ein Element $T'_k \in \mathcal{T}$ mit $h_k = 2^{(L_{\text{patch}}\delta(z_k, T'_k) - \texttt{level}(T'_k))/d}$. Es gilt

$$\begin{aligned}\frac{h_j^2}{h_k^2} &\leq \frac{2^{2(L_{\text{patch}}\delta(z_j, T'_k) - \texttt{level}(T'_k))/d}}{2^{2(L_{\text{patch}}\delta(z_k, T'_k) - \texttt{level}(T'_k))/d}} \\ &= 2^{2(L_{\text{patch}}(\delta(z_j, T'_k) - \delta(z_k, T'_k)))/d} \\ &\leq 2^{2L_{\text{patch}}/d} < B(d)\,,\end{aligned}$$

da nach Voraussetzung $L_{\text{patch}} < \frac{d}{2}\log_2 B(d)$. Dass $|\delta(z_j, T'_k) - \delta(z_k, T'_k)| \leq 1$, folgt aus der Definition von $\delta(\cdot,\cdot)$ und $z_j, z_k \in T$.

(iii). Als Nächstes zeigen wir die obere Abschätzung in der Voraussetzung (48) von Proposition 3.20. Sei also $T \in \mathcal{T}$ beliebig und $z_j \in T$. Sei $T_0 \in \mathcal{T}_0$ jenes Element im Anfangsgitter $\mathcal{T}_0$ mit $T \subset T_0$ und beachte $\delta(z_j, T) = 0$. Nach der Bisektionsvorschrift (5)–(6) gilt $2^{-\texttt{level}(T)}|T_0| = |T|$. Dann gilt

$$\begin{aligned}h_j &\leq 2^{-\texttt{level}(T)/d} \\ &= |T_0|^{-1/d}(2^{-\texttt{level}(T)}|T_0|)^{1/d} \\ &= |T_0|^{-1/d}|T|^{1/d} \\ &\leq C_{13}\texttt{diam}(T)\end{aligned}$$

mit $C_{13} := C_6^{-1/d}C_5$ und den Konstanten C_6, C_5 aus Definition 3.1 und Satz 2.19.

(iv). Zuletzt bleibt noch die untere Abschätzung in der Voraussetzung (48) von Proposition 3.20 zu zeigen. Sei also wieder $T \in \mathcal{T}$ beliebig und $z_j \in T$. Nach (51) existiert ein Element $T'_j \in \mathcal{T}$ mit $h_j = 2^{(L_{\mathrm{patch}}\delta(z_j,T'_j)-\mathtt{level}(T'_j))/d}$. Es gilt $\mathtt{level}(T) \leq \mathtt{level}(T'_j)$, da ansonsten $2^{-\mathtt{level}(T)} < h_j$ im Widerspruch zur Definition von h_j gelten würde. Sei wieder $T_0 \in \mathcal{T}_0$ jenes Element im Anfangsgitter $\mathcal{T}_0$ mit $T \subset T_0$. Nach der Bisektionsvorschrift (5)–(6) gilt $2^{-\mathtt{level}(T)} = |T|/|T_0|$. Wir rechnen nach

$$\begin{aligned} h_j &= 2^{-\mathtt{level}(T)/d} \cdot 2^{(L_{\mathrm{patch}}\delta(z_j,T'_j)-|\mathtt{level}(T'_j)-\mathtt{level}(T)|)/d} \\ &\geq C_7^{-1/d} C_4 \mathtt{diam}(T) \cdot 2^{(L_{\mathrm{patch}}\delta(z_j,T'_j)-|\mathtt{level}(T'_j)-\mathtt{level}(T)|)/d} \end{aligned}$$

mit den Konstanten C_7, C_4 aus Definition 3.1 und Satz 2.19. Sei $\delta(z_j, T'_j) = n$. Nach Definition von $\delta(\cdot,\cdot)$ existieren Elemente $T_0, \ldots, T_n$ mit $z_j \in T_0$, $T'_j = T_n$ und $T_i \cap T_{i+1} \neq \emptyset$ für alle $i = 0, \ldots, n-1$. Definiere $T_{-1} := T$. Aus der Definition von $\delta(\cdot,\cdot)$ trifft auf einen Knoten $z \in \mathcal{N}$ genau einer der folgenden drei Fälle zu:

- $z \notin T_i$ für alle $i \in \{-1, \ldots, n\}$.
- $z \in T_i$ für genau ein $i \in \{-1, \ldots n\}$.
- $z \in T_i \cap T_{i+1}$ für genau ein $i \in \{-1, \ldots, n-1\}$.

In allen anderen Fällen wäre n nicht minimal. Sei nun m die Anzahl der $i \in \{-1, \ldots, n-1\}$, für die $T_i \cap T_{i+1}$ keinen Punkt aus $\mathcal{N} \setminus \mathcal{N}_0$ enthält. Für solche i gilt nach Satz 2.9 und Satz 2.15(ii) $|\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \leq C_3$. Aufgrund der Minimalität von n gilt $m \leq \#\mathcal{N}_0$. Falls $T_i \cap T_{i+1}$ hingegen einen Punkt aus $\mathcal{N}\setminus\mathcal{N}_0$ enthält, gilt nach Voraussetzung $|\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \leq L_{\mathrm{patch}}$. Nun gilt mit der Dreiecksungleichung

$$\begin{aligned} |\mathtt{level}(T'_j) - \mathtt{level}(T)| &\leq \sum_{i=-1}^{n-1} |\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \\ &\leq L_{\mathrm{patch}}(n+1-m) + C_3 m \\ &\leq L_{\mathrm{patch}}(n+1) + C_3 \#\mathcal{N}_0 \,. \end{aligned}$$

Es folgt

$$\begin{aligned} h_j &\geq C_7^{-1/d} C_4 \mathtt{diam}(T) 2^{(L_{\mathrm{patch}} n - L_{\mathrm{patch}}(n+1) - C_3 \#\mathcal{N}_0)/d} \\ &= C_{12} \mathtt{diam}(T) \end{aligned}$$

mit $C_{12} := C_7^{-1/d} C_4 2^{-(\#\mathcal{N}_0 C_3 + L_{\mathrm{patch}})/d}$.

Alle Voraussetzungen (47)–(48) von Proposition 3.20 können erfüllt werden, und die L^2-Orthogonalprojektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ ist H^1_D-stabil. □

3.4 2D

Wir betrachten den Fall $d = 2$. Sei $\mathcal{T} \in \mathtt{refine}(\mathcal{T}_0)$ eine NVB-Verfeinerung eines zulässigen Anfangsgitters $\mathcal{T}_0$. Wir wollen überprüfen, ob die Voraussetzungen von Proposition 3.21 erfüllt sind:
Es gilt $\frac{2}{2}\log_2 B(2) = 3.307\ldots$. Nach Proposition 3.21 genügt es also zu zeigen, dass für beliebiges $z \in \mathcal{N}\setminus\mathcal{N}_0$ der Levelunterschied zweier Elemente $T, T' \in \mathcal{T}$ mit $z \in T \cap T'$ durch eine Konstante $L_{\mathrm{patch}} < 3.307\ldots$ beschränkt ist. Dies ist tatsächlich der Fall. Wir zeigen zunächst folgendes Lemma:

Lemma 3.22. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und $z \in \mathcal{N} \setminus \mathcal{N}_0$. Dann gilt*

$$\#\omega(z) \leq \begin{cases} 4, & z \in \partial\Omega \\ 8, & z \notin \partial\Omega \end{cases} \tag{52}$$

Beweis. Da es zu jedem Gitter $\mathcal{T}$ ein feineres, uniform verfeinertes Gitter, zum Beispiel $\mathcal{T}_\ell$ mit $\ell = \max\{\texttt{level}(T) \mid T \in \mathcal{T}\}$, gibt, genügt es nur uniforme Gitter zu betrachten. Da $z \in \mathcal{N} \setminus \mathcal{N}_0$, existiert ein maximales n, sodass $z \notin \mathcal{N}_n = \mathcal{N}(\mathcal{T}_n)$, wobei $\mathcal{T}_n$ die uniforme Level n Verfeinerung von $\mathcal{T}_0$ bezeichnet. Aufgrund der Maximalität von n existiert ein Kante $E \in \mathcal{E}_n$ deren Mittelpunkt z ist. Aus $d = 2$ folgt, dass E Teilmenge, und damit Verfeinerungskante von genau einem $T_1 \in \mathcal{T}_n$, falls $z \in \partial\Omega$, beziehungsweise von genau zwei Elementen $T_1, T_2 \in \mathcal{T}_n$, falls $z \notin \partial\Omega$, ist. In $\mathcal{T}_{n+1}$ gilt dann

$$\#\omega_{\mathcal{T}_{n+1}}(z) = 2, \quad \text{falls } z \in \partial\Omega\,,$$

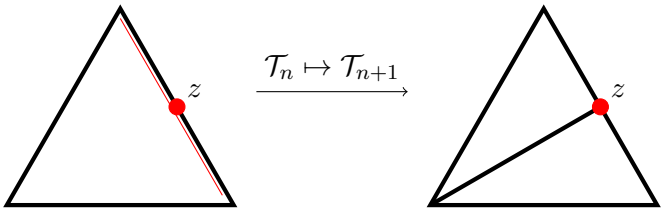


Abbildung 14: Der Knotenpatch lokal um z in $\mathcal{T}_n$ beziehungsweise in $\mathcal{T}_{n+1}$ für $z \in \partial\Omega$.

beziehungsweise

$$\#\omega_{\mathcal{T}_{n+1}}(z) = 4, \quad \text{falls } z \notin \partial\Omega\,.$$

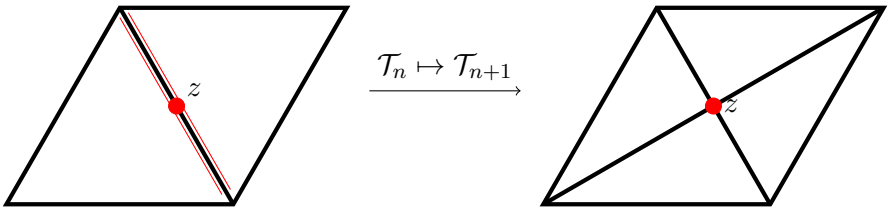


Abbildung 15: Der Knotenpatch lokal um z in $\mathcal{T}_n$ beziehungsweise in $\mathcal{T}_{n+1}$ für $z \notin \partial\Omega$.

Es gilt $S_2 = 2$, und nach Satz 2.15(i) gilt damit in $\mathcal{T}_\ell$

$$\#\omega_{\mathcal{T}_\ell}(z) \leq 2\#\omega_{\mathcal{T}_{n+1}}(z) = \begin{cases} 4, & z \in \partial\Omega \\ 8, & z \notin \partial\Omega \end{cases}$$

und damit die Behauptung. □

Satz 3.23. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$. Sei $z \in \mathcal{N} \setminus \mathcal{N}_0$ und $T, T' \in \mathcal{T}$ mit $z \in T \cap T'$. Dann gilt*

$$|\texttt{level}(T') - \texttt{level}(T)| \leq 3\,. \tag{53}$$

Beweis. Beachte, dass $z \notin \mathcal{N}_0$. Daher gilt mit Lemma 3.22 entweder $z \in \partial\Omega$ und $\#\omega(z) \leq 4$, oder $z \notin \partial\Omega$ und $\#\omega(z) \leq 8$. Nach Satz 2.9 unterscheidet sich das Level von zwei benachbarten Elementen maximal um 1. Also gilt die Aussage jedenfalls, wenn $z \in \partial\Omega$. Ebenfalls aufgrund

von Satz 2.9 ist die Aussage für $z \notin \partial\Omega$ nur dann nicht trivialerweise erfüllt, wenn $\#\omega(z) = 8$, siehe Abbildung 16. Wir führen einen Widerspruchsbeweis.
Die Elemente im Patch $\omega(z)$ seien gegen den Uhrzeigersinn durchnummeriert und es gelte $\texttt{level}(T_1) = \ell$ und $\texttt{level}(T_5) = \ell + 4$. Mit Satz 2.9 gilt $\texttt{level}(T_2) = \texttt{level}(T_8) = \ell + 1$, $\texttt{level}(T_3) = \texttt{level}(T_7) = \ell + 2$ und $\texttt{level}(T_4) = \texttt{level}(T_6) = \ell + 3$.

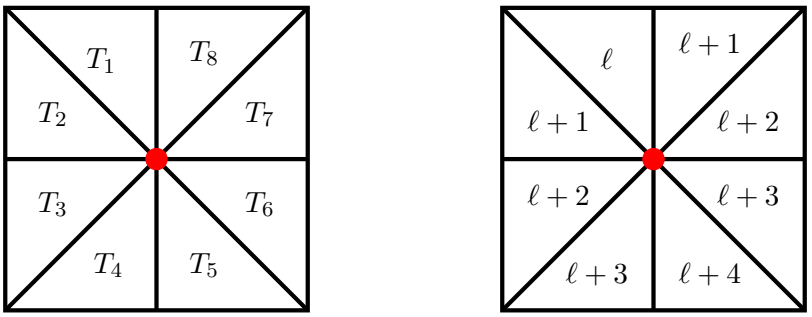
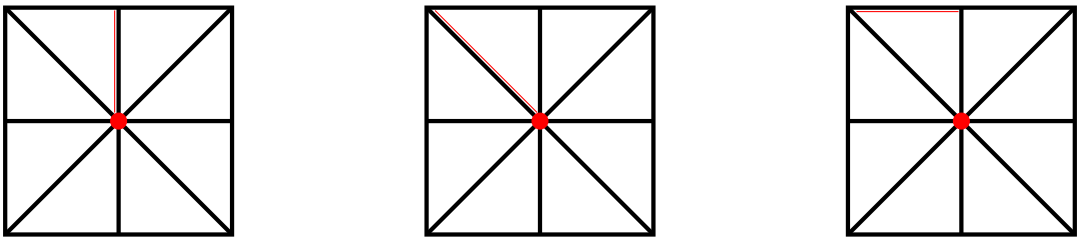


Abbildung 16: Der Patch $\omega(z)$ und die entsprechenden Level.

Da $\texttt{level}(T_2) = \texttt{level}(T_8) > \texttt{level}(T_1)$ folgt aus Korollar 2.11(ii), dass der Aufruf von $\mathcal{T}_\star = \texttt{refine}(\mathcal{T}, T_1)$ weder eine Verfeinerung von T_2, noch von T_8 herbeiführt, also, dass $T_2, T_8 \in \mathcal{T}_\star$. Demnach muss $z \notin e(T_1)$ gelten, siehe Abbildung 17(a)–(b). Dann gilt in $\mathcal{T}_\star$ aber $\#\omega(z) \geq 9$ im Widerspruch zu Lemma 3.22, siehe Abbildung 17(c).



(a) Es folgt $T_8 \notin \mathcal{T}_\star$ im Widerspruch zu Korollar 2.11(ii).

(b) Es folgt $T_2 \notin \mathcal{T}_\star$ im Widerspruch zu Korollar 2.11(ii).

(c) Es folgt $\omega_{\mathcal{T}_\star}(z) \geq 9$ im Widerspruch zu Lemma 3.22.

Abbildung 17: Falls $\omega(z)$ wie in Abbildung 16, kann keine Kante von T_1 die Verfeinerungskante sein.

Insgesamt folgt, dass keine Kante von $T_1 \in \mathcal{T}$ die Verfeinerungskante sein kann, also kommen wir auf einen Widerspruch. □

3.5 Dimension $d \geq 3$

Nun stellt sich die Frage ob wir das Resultat auf allgemeines $d \geq 2$ verallgemeinern können. Nach unserer Analyse genügt es $L_{\text{patch}} \in \mathbb{R}$ zu finden, sodass die Voraussetzungen von Proposition 3.21 erfüllt sind.
In 2D konnten wir für $z \in \mathcal{N} \setminus \mathcal{N}_0$ mit Lemma 3.22 die Anzahl der Elemente im Patch $\omega(z)$ beschränken. Mit Hilfe dieses Lemmas konnte in Satz 3.23 dann $L_{\text{patch}} = 3$ bewiesen werden.
Ein analoges Vorgehen wird für $d \geq 3$ nicht zum Ziel führen. Für $d \geq 3$ existiert kein Analogon zu Lemma 3.22, es gilt sogar folgender Satz:

Satz 3.24. *Für jedes $L \in \mathbb{N}$ existiert ein $\Omega \subset \mathbb{R}^3$ und ein zulässiges Anfangsgitter $\mathcal{T}_0$ auf Ω,*

sodass in der uniformen Level 1 *Verfeinerung* $\mathcal{T}_1 \in \texttt{refine}(\mathcal{T}_0)$ *ein Knoten* $z \in \mathcal{N}_1 \setminus \mathcal{N}_0$ *existiert, mit* $\#\omega(z) > L$.

Beweis. Sei ohne Einschränkung der Allgemeinheit $L \geq 4$ gerade. Definiere die Knoten

$$\begin{aligned} z_0 &:= (0,0,0)^\top \in \mathbb{R}^3\,, \\ z_{L+1} &:= (0,0,1)^\top \in \mathbb{R}^3\,, \\ z_j &:= (\sin(2(j-1)\pi/L), \cos(2(j-1)\pi/L), 0.5)^\top \in \mathbb{R}^3 \quad \text{für } j = 1, ..., L\,. \end{aligned}$$

Definiere die Elemente

$$\begin{aligned} T_L &:= (z_0, z_1, z_L, z_{L+1})_0 \subset \mathbb{R}^3\,, \\ T_j &:= \begin{cases} (z_0, z_j, z_{j+1}, z_{L+1})_0 \subset \mathbb{R}^3, & j \text{ ungerade} \\ (z_0, z_{j+1}, z_j, z_{L+1})_0 \subset \mathbb{R}^3, & j \text{ gerade} \end{cases} \quad \text{für } j = 1, ..., L-1\,. \end{aligned}$$

Dann ist $\mathcal{T}_0 := \{T_j \mid j = 1, ..., L\}$ ein zulässiges Anfangsgitter auf $\Omega := \big(\bigcup_{T\in\mathcal{T}_0} T\big)^\circ$, denn alle $T, T' \in \mathcal{T}_0$, die Nachbarn sind, erfüllen offensichtlich ***(Z1)***:
Da $e(T) = \overline{z_0 z_{L+1}}$ für alle $T \in \mathcal{T}_0$, bleibt nur noch zu zeigen, dass Nachbarn in $\mathcal{T}_0$ auch stets gespiegelte Nachbarn sind. Für $j = 1, ..., L$ sind die Nachbarn von T_j in $\mathcal{T}_0$ genau T_{j-1} und T_{j+1}, wobei $T_{L+1} := T_1$ und $T_0 := T_L$. Die geordnete Sequenz der Knoten von T_j stimmt sowohl mit der von T_{j-1}, als auch mit der von T_{j+1} an allen außer einer Stelle überein. Also erfüllen in $\mathcal{T}_0$ Nachbarn stets ***(Z1)***.
Beachte, dass $e(T) = \overline{z_0 z_{L+1}}$ für alle $T \in \mathcal{T}_0$. Da $\#\mathcal{T}_0 = L$, gilt in der uniformen Level 1 Verfeinerung $\mathcal{T}_1$

$$z := (0,0,0.5)^\top \in \mathcal{N}_1 \quad \text{und} \quad \omega_{\mathcal{T}_1}(z) = 2L > L\,.$$

Dass $L \in \mathbb{N}$ beliebig groß gewählt war, schließt den Beweis. □

Das heißt aber nicht, dass (3) für $d \geq 3$ nicht gilt. Die Gültigkeit von (3) folgt aus unserer Analyse sofort, falls die Existenz von $L_{\text{patch}} \in \mathbb{R}$, welches die Voraussetzungen von Proposition 3.21 erfüllt, bewiesen werden kann.
Wir merken an, dass die Voraussetzungen von Proposition 3.21 hinreichend, aber nicht notwendig sind. Eine andere Möglichkeit wäre es, nur $L_{\text{patch}} < \frac{d}{2}\log_2 B(d)$ zu fordern, und auf die Voraussetzung (50) zu verzichten. Kritisch ist es dann, im Punkt (iv) des Beweises von Proposition 3.21 die Abschätzung

$$|\texttt{level}(T_j') - \texttt{level}(T)| \leq L_{\text{patch}} \cdot n + C\,, \tag{54}$$

mit einer Konstanten $C > 0$, welche nicht von n abhängt, zu zeigen. Falls das jedoch gelingt, wäre (3) gültig.

4 Stabilität in $h_{\mathcal{T}}^{-s}$-gewichteten L^2-Normen

Die Technik aus Kapitel 3 zeigt noch mehr: Mit einigen leichten Modifikationen in den Voraussetzungen und Beweisen der Propositionen 3.9, 3.20 und 3.21 lassen sich Kriterien formulieren, aus denen Stabilität der L^2-Orthogonalprojektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ in $h_{\mathcal{T}}^{-s}$-gewichteten L^2-Normen (56) folgt.

4.1 Vorbereitung

Zunächst definieren wir eine Netzweitenfunktion:

Definition 4.1. Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ ein Gitter auf dem Gebiet $\Omega \subset \mathbb{R}^d$. Definiere die Netzweitenfunktion $h_\mathcal{T} \in L^2(\Omega;\mathbb{R})$ mit

$$h_\mathcal{T}(\mathbf{x}) = \texttt{diam}(T) \tag{55}$$

für $\mathbf{x} \in T$. Die Netzweitenfunktion $h_\mathcal{T} : \Omega \to \mathbb{R}$ ist elementweise konstant und als L^∞-Funktion wohldefiniert.

Sei $s \in \mathbb{R}$. Wir untersuchen Stabilität in $h_\mathcal{T}^{-s}$-gewichteten L^2-Normen, d.h. wir stellen die Frage, unter welchen Voraussetzungen

$$\|h_\mathcal{T}^{-s}\Pi_D(\mathcal{T})v\|_{L^2(\Omega)} \leq C_{14}\|h_\mathcal{T}^{-s}v\|_{L^2(\Omega)} \quad \text{für alle } v \in L^2(\Omega) \tag{56}$$

mit einer Konstanten $C_{14} > 0$ gilt.
Für den Beweis von Proposition 3.9 verwendeten wir den Scott-Zhang-Projektor aus Definition 3.7. Die Spur einer Funktion $v \in L^2(\Omega)$ auf Seiten $F \in \mathcal{F}$ ist aber im Allgemeinen nicht definiert. Deshalb definieren wir ähnlich wie in Definition 3.7 einen Interpolationsoperator für $L^2(\Omega)$-Funktionen:

Definition 4.2. Wähle für $z \in \mathcal{N} \setminus \overline{\Gamma}_D$ ein Element $T_z \in \mathcal{T}$ mit $z \in T_z$.

Aus dem Satz von Riesz folgt sofort folgendes Lemma:

Lemma 4.3. *Für $z \in \mathcal{N} \setminus \overline{\Gamma}_D$ existiert eine eindeutige duale Funktion $\psi_z \in \mathcal{P}^1(T_z)$ so, dass*

$$\int_{T_z} \psi_z \varphi_{z'}\, dx = \delta_{zz'} \tag{57}$$

für alle $z' \in \mathcal{N} \setminus \overline{\Gamma}_D$.

Definition 4.4. Definiere die Scott-Zhang-artige Projektion $\mathcal{J}_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ durch

$$\mathcal{J}_D(\mathcal{T})v := \sum_{z \in \mathcal{N}\setminus\overline{\Gamma}_D} \left(\int_{T_z} \psi_z v\, dx\right) \varphi_z \tag{58}$$

mit ψ_z wie in (57).

Lemma 4.5 ([Pra15, Exercise 46])**.** *Die Scott-Zhang-artige Projektion $\mathcal{J}_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ aus Definition 4.4 erfüllt die folgenden Eigenschaften:*

(i). *$\mathcal{J}_D(\mathcal{T})$ ist eine Projektion auf $\mathcal{S}^1_D(\mathcal{T})$, d.h.*

$$\mathcal{J}_D(\mathcal{T})v = v \tag{59}$$

für alle $v \in \mathcal{S}^1_D(\mathcal{T})$.

(ii). *$\mathcal{J}_D(\mathcal{T})$ ist lokal L^2-stabil, d.h.*

$$\|\mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|v\|_{L^2(\omega_T)} \tag{60}$$

für alle $v \in L^2(\Omega)$ und alle $T \in \mathcal{T}$.

(iii). $\mathcal{J}_D(\mathcal{T})$ *ist lokal* H^1_D*-stabil, d.h.*

$$\|\nabla \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)} \tag{61}$$

für alle $v \in H^1_D(\Omega)$ *und alle* $T \in \mathcal{T}$.

(iv). $\mathcal{J}_D(\mathcal{T})$ *erfüllt eine Approximationseigenschaft erster Ordnung, d.h.*

$$\|v - \mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \mathtt{diam}(T)\|\nabla v\|_{L^2(\omega_T)} \tag{62}$$

für alle $v \in H^1_D(\Omega)$ *und alle* $T \in \mathcal{T}$.

Die versteckten Konstanten in (60)–(62) *hängen nur von der* γ*-Formregularität von* $\mathcal{T}$ *ab.* □

4.2 Adaptierte Technik

Analog zu Proposition 3.9 lautet das Stabilitätskriterium hier wie folgt:

Proposition 4.6. *Sei* $\mathcal{T}$ *ein* γ*-formreguläres Gitter und* $s \in \mathbb{R}$. *Für alle Elemente* $T \in \mathcal{T}$ *gelte*

$$C_{15}^{-1}\mathbf{x}^\top \mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{\Lambda}_T^{2s}\mathbf{x} \le \mathbf{x}^\top \mathbf{M}_T\mathbf{x} \le C_{16}\mathbf{x}^\top \mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{x} \quad \text{für alle } \mathbf{x} \in \mathbb{R}^{d+1}, \tag{63}$$

mit den Konstanten $C_{15}, C_{16} > 0$. *Dann ist die* L^2*-orthogonale Projektion* $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\Omega)$ *stabil bezüglich der* $h_{\mathcal{T}}^{-s}$*-gewichteten* L^2*-Norm* (56)*, und die Konstante* $C_{14} > 0$ *hängt nur von der* γ*-Formregularität von* $\mathcal{T}$ *und den Konstanten* $C_{15}, C_{16} > 0$ *ab.*

Beweis. Sei $\mathcal{J}_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ die Scott-Zhang-artige Projektion aus Definition 4.4, d.h. nach Lemma 4.5 gilt

$$\|\mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|v\|_{L^2(\omega_T)} \tag{64}$$

für alle $v \in L^2(\Omega)$ und alle $T \in \mathcal{T}$. Die versteckte Konstante hängt nur von der γ-Formregularität ab. Sei $q := \Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v = \sum_{\ell=1}^N q_\ell\varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$. Wir definieren $p := \sum_{\ell=1}^N q_\ell h_\ell^{-2s}\varphi_\ell \in \mathcal{S}^1_D(\mathcal{T})$. Mit den lokalen Koeffizientenvektoren $\mathbf{x} = (q_{[T,1]}, \dots, q_{[T,d+1]})^\top \in \mathbb{R}^{d+1}$ und $\mathbf{y} = (q_{[T,1]}h_{[T,1]}^{-2s}, \dots, q_{[T,d+1]}h_{[T,d+1]}^{-2s})^\top = \mathtt{diam}(T)^{-2s}\mathbf{\Lambda}_T^{2s}\mathbf{x} \in \mathbb{R}^{d+1}$ erhalten wir durch die untere Abschätzung in (63)

$$\begin{aligned}\|p\|^2_{L^2(T)} = \mathbf{y}^\top\mathbf{M}_T\mathbf{y} = \mathtt{diam}(T)^{-4s}\mathbf{x}^\top\mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{\Lambda}_T^{2s}\mathbf{x} &\le C_{15}\mathtt{diam}(T)^{-4s}\mathbf{x}^\top\mathbf{M}_T\mathbf{x}\\ &= C_{15}\mathtt{diam}(T)^{-4s}\|q\|^2_{L^2(T)}\,.\end{aligned}$$

Weiters liefert die obere Abschätzung in (63)

$$\|q\|^2_{L^2(T)} = \mathbf{x}^\top\mathbf{M}_T\mathbf{x} \le C_{16}\mathbf{x}^\top\mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{x} = C_{16}\mathtt{diam}(T)^{2s}\int_T pq\,dx\,.$$

Summieren wir nun die letzten zwei Ungleichungen über alle Elemente $T \in \mathcal{T}$, so sehen wir durch Ausnutzen der Symmetrie von Orthogonalprojektionen und $\Pi_D(\mathcal{T})p = p$ für $p \in \mathcal{S}^1_D(\mathcal{T})$, dass

$$\begin{aligned}\sum_{T\in\mathcal{T}}\mathtt{diam}(T)^{-2s}\|q\|^2_{L^2(T)} &\lesssim \int_\Omega pq\,dx = \int_\Omega p(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v)\,dx = \int_\Omega p(v - \mathcal{J}_D(\mathcal{T})v)\,dx\\ &\le \left(\sum_{T\in\mathcal{T}}\mathtt{diam}(T)^{2s}\|p\|^2_{L^2(T)}\right)^{1/2}\left(\sum_{T\in\mathcal{T}}\mathtt{diam}(T)^{-2s}\|v - \mathcal{J}_D(\mathcal{T})v\|^2_{L^2(T)}\right)^{1/2}\\ &\overset{(64)}{\lesssim} \left(\sum_{T\in\mathcal{T}}\mathtt{diam}(T)^{-2s}\|q\|^2_{L^2(T)}\right)^{1/2}\|h_{\mathcal{T}}^{-s}v\|_{L^2(\Omega)}\,.\end{aligned}$$

Dies zeigt

$$\left(\sum_{T\in\mathcal{T}} \mathtt{diam}(T)^{-2s}\|q\|^2_{L^2(T)}\right)^{1/2} \lesssim \|h_{\mathcal{T}}^{-s}v\|_{L^2(\Omega)}\,. \tag{65}$$

Wir erinnern daran, dass $h_{\mathcal{T}}|_T \equiv \mathtt{diam}(T)$ für alle $T\in\mathcal{T}$. Weiters gilt lokale L^2-Stabilität von $\mathcal{J}_D(\mathcal{T})$, d.h. $\|\mathcal{J}_D(\mathcal{T})v\|_{L^2(T)} \lesssim \|v\|_{L^2(\omega_T)}$ für alle $T\in\mathcal{T}$. Dies zeigt

$$\begin{aligned}
\|h_{\mathcal{T}}^{-s}\Pi_D(\mathcal{T})v\|_{L^2(\Omega)} &\leq \|h_{\mathcal{T}}^{-s}(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v)\|_{L^2(\Omega)} + \|h_{\mathcal{T}}^{-s}\mathcal{J}_D(\mathcal{T})v\|_{L^2(\Omega)}\\
&= \left(\sum_{T\in\mathcal{T}} \|h_{\mathcal{T}}^{-s}(\Pi_D(\mathcal{T})v - \mathcal{J}_D(\mathcal{T})v)\|^2_{L^2(T)}\right)^{1/2} + \left(\sum_{T\in\mathcal{T}} \|h_{\mathcal{T}}^{-s}\mathcal{J}_D(\mathcal{T})v\|^2_{L^2(T)}\right)^{1/2}\\
&\overset{(64)}{\lesssim} \left(\sum_{T\in\mathcal{T}} \mathtt{diam}(T)^{-2s}\|q\|^2_{L^2(T)}\right)^{1/2} + \left(\sum_{T\in\mathcal{T}} \mathtt{diam}(T)^{-2s}\|v\|^2_{L^2(\omega_T)}\right)^{1/2}\\
&\overset{(65)}{\lesssim} \|h_{\mathcal{T}}^{-s}v\|_{L^2(\Omega)}
\end{aligned}$$

und schließt den Beweis. □

Das Analogon zu Proposition 3.20 lautet nun

Proposition 4.7. *Sei $\mathcal{T}$ ein γ-formreguläres Gitter und $s\in\mathbb{R}$. Seien für $z_\ell\in\mathcal{T}$ die Skalare $h_\ell>0$ so gewählt, dass für alle Elemente $T\in\mathcal{T}$ und alle $z_j, z_k\in\mathcal{N}(T)$ stets*

$$\frac{h_j^{2s}}{h_k^{2s}} \leq C_{11} < B(d)\,, \tag{66}$$

sowie

$$C_{17}\mathtt{diam}(T)^s \leq h_j^s \leq C_{18}\mathtt{diam}(T)^s \tag{67}$$

mit Konstanten $C_{11}, C_{17}, C_{18} > 0$ und $B(d)$ aus (46) gilt. Dann sind die Voraussetzungen (63) von Proposition 4.6 erfüllt. Die Konstante C_{15} hängt nur von der γ-Formregularität und C_{17}^{-1} ab. C_{16} hängt nur von der γ-Formregularität, C_{18} und $\varepsilon := B(d) - C_{11}$ ab.

Beweis. **Schritt 1:** Zunächst zeigen wir die untere Abschätzung in (63): Mit Lemma 3.3 und Lemma 3.4 sieht man die Äquivalenz

$$|T|\mathbf{y}^\top\mathbf{y} \simeq \mathbf{y}^\top\mathbf{M}_T\mathbf{y} \quad \text{für alle } \mathbf{y}\in\mathbb{R}^{d+1}\,, \tag{68}$$

wobei die versteckten Konstanten nur von der γ-Formregularität abhängen. Diese Äquivalenz und die untere Abschätzung in (67) zeigen

$$\mathbf{x}^\top\boldsymbol{\Lambda}_T^{2s}\mathbf{M}_T\boldsymbol{\Lambda}_T^{2s}\mathbf{x} = (\boldsymbol{\Lambda}_T^{2s}\mathbf{x})^\top\mathbf{M}_T(\boldsymbol{\Lambda}_T^{2s}\mathbf{x}) \overset{(68)}{\simeq} |T|(\boldsymbol{\Lambda}_T^{2s}\mathbf{x})^\top(\boldsymbol{\Lambda}_T^{2s}\mathbf{x}) \overset{(67)}{\lesssim} |T|\mathbf{x}^\top\mathbf{x} \overset{(68)}{\simeq} \mathbf{x}^\top\mathbf{M}_T\mathbf{x}\,.$$

Das zeigt die untere Abschätzung in (63), und $C_{15}>0$ hängt nur von der γ-Formregularität, s und C_{17} ab.
Schritt 2: Als Nächstes zeigen wir die obere Abschätzung in (63) bis auf eine verbleibende Eigenwertabschätzung: Wie in Lemma 3.3 fixieren wir den Referenzsimplex $\widehat{T}\subset\mathbb{R}^d$ so, dass die zugehörige Massenmatrix $\widehat{\mathbf{M}}_{jk} = 1+\delta_{jk}$ mit Kroneckers Delta erfüllt. Nach Lemma 3.4 gilt $\mathbf{M}_T = C_8|T|\widehat{\mathbf{M}}$ mit einer Konstanten $C_8>0$. Wir definieren die Matrix

$$\mathbf{A}_{T,s} := \boldsymbol{\Lambda}_T^{2s}\mathbf{M}_T + \mathbf{M}_T\boldsymbol{\Lambda}_T^{2s} = C_8|T|(\boldsymbol{\Lambda}_T^{2s}\widehat{\mathbf{M}} + \widehat{\mathbf{M}}\boldsymbol{\Lambda}_T^{2s}) =: C_8|T|\widehat{\mathbf{A}}_{T,s}\,.$$

Die symmetrische Matrix $\widehat{\mathbf{B}}_{T,s} := \mathbf{\Lambda}_T^{-s}\widehat{\mathbf{A}}_{T,s}\mathbf{\Lambda}_T^{-s}$ erfüllt

$$(\widehat{\mathbf{B}}_{T,s})_{jk} = \left(\mathbf{\Lambda}_T^s\widehat{\mathbf{M}}\mathbf{\Lambda}_T^{-s} + \mathbf{\Lambda}_T^{-s}\widehat{\mathbf{M}}\mathbf{\Lambda}_T^s\right)_{jk} = \left(\frac{h_j^s}{h_k^s} + \frac{h_k^s}{h_j^s}\right)\widehat{\mathbf{M}}_{jk} = \left(\frac{h_j^s}{h_k^s} + \frac{h_k^s}{h_j^s}\right)(1+\delta_{jk})\,.$$

Falls nun der minimale Eigenwert $\lambda_{\min,s}$ von $\widehat{\mathbf{B}}_{T,s}$ positiv ist, erhalten wir mit $\mathbf{y} = \mathbf{\Lambda}_T^s\mathbf{x}$

$$\mathbf{x}^\top\mathbf{M}_T\mathbf{x} \overset{(68)}{\simeq} |T|\mathbf{x}^\top\mathbf{x} \overset{(67)}{\lesssim} |T|\mathbf{y}^\top\mathbf{y} \lesssim |T|\mathbf{y}^\top\widehat{\mathbf{B}}_{T,s}\mathbf{y}\,,$$

wobei die versteckten Konstanten nur von der γ-Formregularität, $C_{18} > 0$ und $\lambda_{\min,s} > 0$ abhängen. Symmetrie von $\mathbf{\Lambda}_T^s$ und $\mathbf{M}_T$ liefern $\mathbf{x}^\top\mathbf{A}_{T,s}\mathbf{x} = 2\mathbf{x}^\top\mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{x}$ und damit

$$|T|\mathbf{y}^\top\widehat{\mathbf{B}}_{T,s}\mathbf{y} = |T|\mathbf{x}^\top\widehat{\mathbf{A}}_{T,s}\mathbf{x} \simeq \mathbf{x}^\top\mathbf{A}_{T,s}\mathbf{x} = 2\mathbf{x}^\top\mathbf{\Lambda}_T^{2s}\mathbf{M}_T\mathbf{x}\,.$$

Kombinieren der letzten zwei Rechnungen zeigt die obere Abschätzung in (63) und $C_{16} > 0$ hängt nur von der γ-Formregularität, $C_{18} > 0$ und $\lambda_{\min,s} > 0$ ab.
Schritt 3: Es bleibt noch zu zeigen, dass der minimale Eigenwert $\lambda_{\min,s}$ von $\widehat{\mathbf{B}}_{T,s}$ positiv ist. Mit $\alpha_i := h_i^s$ hat $\widehat{\mathbf{B}}_{T,s}$ die Form (31) aus Satz 3.10. Mit Voraussetzung (66) folgt aus Satz 3.19, dass $\lambda_{\min,s}$ positiv ist. □

Das Analogon zu Proposition 3.21 lautet nun

Proposition 4.8. *Sei $\mathcal{T} \in \texttt{refine}(\mathcal{T}_0)$ und $s \in \mathbb{R}$. Existiert eine Konstante $L_{\text{patch}} < \frac{d}{2|s|}\log_2 B(d)$ so, dass für alle $z \in \mathcal{N}\setminus\mathcal{N}_0$ und alle $T, T' \in \omega(z)$*

$$|\texttt{level}(T') - \texttt{level}(T)| \le L_{\text{patch}} \tag{69}$$

gilt, dann können die Voraussetzungen (66)–(67) von Proposition 4.7 stets erfüllt werden, und die L^2-orthogonale Projektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ ist stabil bezüglich der $h_{\mathcal{T}}^{-s}$-gewichteten L^2-Norm (56). Die Konstante C_{11} hängt nur von L_{patch}, s und d ab, C_{17} hängt nur von $C_7, C_4, \#\mathcal{N}_0, C_3, L_{\text{patch}}$, s und d ab, C_{18} hängt nur von C_7, C_5, s und d ab.

Beweis.

(i). Wir beginnen mit der Definition der Werte h_j. Dazu definieren wir zunächst einen Knoten-Element-Abstand $\delta : \mathcal{N}\times\mathcal{T} \to \mathbb{N}_0$.

$$\begin{aligned}\delta(z,T) := n \quad &\text{für die kleinste Zahl } n \in \mathbb{N}_0 \text{ so, dass Elemente } T_0,\dots,T_n \in \mathcal{T}\\ &\text{existieren mit } z \in T_0,\ T = T_n,\\ &\text{und } T_i \cap T_{i+1} \neq \emptyset \text{ für alle } i = 0,\dots,n-1.\end{aligned}$$

Für $z_j \in \mathcal{N}$ definiere

$$h_j := \min_{T_j' \in \mathcal{T}} 2^{(L_{\text{patch}}\delta(z_j,T_j') - \texttt{level}(T_j'))/d}\,. \tag{70}$$

(ii). Wir überprüfen zunächst Voraussetzung (66) von Proposition 4.7. Sei also $T \in \mathcal{T}$ beliebig und $z_j, z_k \in T$. Nach (70) existiert ein Element $T_k' \in \mathcal{T}$ mit $h_k = 2^{(L_{\text{patch}}\delta(z_k,T_k') - \texttt{level}(T_k'))/d}$. Sei zunächst $s \ge 0$, dann gilt

$$\begin{aligned}\frac{h_j^{2s}}{h_k^{2s}} &\le \frac{2^{2s(L_{\text{patch}}\delta(z_j,T_k') - \texttt{level}(T_k'))/d}}{2^{2s(L_{\text{patch}}\delta(z_k,T_k') - \texttt{level}(T_k'))/d}}\\ &= 2^{2s(L_{\text{patch}}(\delta(z_j,T_k') - \delta(z_k,T_k')))/d}\\ &\le 2^{2sL_{\text{patch}}/d} < B(d)\,,\end{aligned}$$

da nach Voraussetzung $L_{\text{patch}} < \frac{d}{2s}\log_2 B(d)$. Dass $|\delta(z_j, T'_k) - \delta(z_k, T'_k)| \leq 1$, folgt aus der Definition von $\delta(\cdot,\cdot)$ und $z_j, z_k \in T$.
Sei nun $s < 0$. Wir haben bereits gezeigt, dass

$$\frac{h_{j'}^{2s'}}{h_{k'}^{2s'}} < B(d) \tag{71}$$

für alle $T \in \mathcal{T}$, $z_{j'}, z_{k'} \in \mathcal{N}(T)$ und alle $s' \geq 0$. Insbesondere gilt (71) mit $z_{j'} := z_k$, $z_{k'} := z_j$ und $s' := -s > 0$. Es gilt

$$\frac{h_j^{2s}}{h_k^{2s}} = \frac{h_k^{-2s}}{h_j^{-2s}} = \frac{h_{j'}^{2s'}}{h_{k'}^{2s'}} \overset{(71)}{<} B(d)$$

und damit die Voraussetzung (66) von Proposition 4.7.

(iii). Als Nächstes zeigen wir für $s \geq 0$ die obere Abschätzung, beziehungsweise für $s < 0$ die untere Abschätzung in der Voraussetzung (67) von Proposition 4.7. Sei also $T \in \mathcal{T}$ beliebig und $z_j \in T$. Sei $T_0 \in \mathcal{T}_0$ jenes Element im Anfangsgitter $\mathcal{T}_0$ mit $T \subseteq T_0$ und beachte $\delta(z_j, T) = 0$. Nach der Bisektionsvorschrift (5)–(6) gilt $2^{-\texttt{level}(T)}|T_0| = |T|$. Sei zunächst $s \geq 0$. Dann gilt

$$\begin{aligned} h_j^s &\leq 2^{-s\cdot\texttt{level}(T)/d} \\ &= |T_0|^{-s/d}(2^{-\texttt{level}(T)}|T_0|)^{s/d} \\ &= |T_0|^{-s/d}|T|^{s/d} \\ &\leq C_{18}\texttt{diam}(T)^s \end{aligned}$$

mit $C_{18} := C_6^{-s/d}C_5^s$ und den Konstanten C_6, C_5 aus Definition 3.1 und Satz 2.19. Das zeigt die obere Abschätzung in (67).
Sei nun $s < 0$. Wir haben eben gezeigt, dass

$$h_j^{s'} \leq C_{18}\texttt{diam}(T)^{s'} \tag{72}$$

für alle $T \in \mathcal{T}$, $z_j \in \mathcal{N}(\mathcal{T})$ und alle $s' \geq 0$. Mit $s' := -s > 0$ gilt insbesondere

$$h_j^{-s} = h_j^{s'} \overset{(72)}{\leq} C_{18}\texttt{diam}(T)^{s'} = C_{18}\texttt{diam}(T)^{-s}\,.$$

Da $(\cdot)^{-1} : \mathbb{R}^+ \to \mathbb{R}^+$ monoton fallend ist, folgt

$$h_j^s \geq C_{18}^{-1}\texttt{diam}(T)^s$$

und damit die untere Abschätzung in (67).

(iv). Zuletzt bleibt noch für $s \geq 0$ die untere Abschätzung, beziehungsweise für $s < 0$ die obere Abschätzung in der Voraussetzung (67) von Proposition 4.7 zu zeigen. Sei also wieder $T \in \mathcal{T}$ beliebig und $z_j \in T$. Nach (70) existiert ein Element $T'_j \in \mathcal{T}$ mit $h_j = 2^{(L_{\text{patch}}\delta(z_j,T'_j) - \texttt{level}(T'_j))/d}$. Es gilt $\texttt{level}(T) \leq \texttt{level}(T'_j)$, da ansonsten $2^{-\texttt{level}(T)} < h_j$ im Widerspruch zur Definition von h_j gelten würde. Sei wieder $T_0 \in \mathcal{T}_0$ jenes Element im Anfangsgitter $\mathcal{T}_0$ mit $T \subset T_0$. Nach der Bisektionsvorschrift (5)–(6) gilt $2^{-\texttt{level}(T)} = |T|/|T_0|$. Sei zunächst wieder $s \geq 0$. Wir rechnen nach

$$\begin{aligned} h_j^s &= 2^{-s\cdot\texttt{level}(T)/d}\cdot 2^{s(L_{\text{patch}}\delta(z_j,T'_j) - |\texttt{level}(T'_j) - \texttt{level}(T)|)/d} \\ &\geq C_7^{-s/d}C_4^s\texttt{diam}(T)^s \cdot 2^{s(L_{\text{patch}}\delta(z_j,T'_j) - |\texttt{level}(T'_j) - \texttt{level}(T)|)/d} \end{aligned}$$

mit den Konstanten C_7, C_4 aus Definition 3.1 und Satz 2.19. Sei $\delta(z_j, T_j') = n$. Nach Definition von $\delta(\cdot,\cdot)$ existieren Elemente $T_0, \dots, T_n$ mit $z_j \in T_0$, $T_j' = T_n$ und $T_i \cap T_{i+1} \neq \emptyset$ für alle $i = 0, \dots, n-1$. Definiere $T_{-1} := T$. Aus der Definition von $\delta(\cdot,\cdot)$ trifft auf einen Knoten $z \in \mathcal{N}$ genau einer der folgenden drei Fälle zu:

- $z \notin T_i$ für alle $i \in \{-1, \dots, n\}$.
- $z \in T_i$ für genau ein $i \in \{-1, \dots n\}$.
- $z \in T_i \cap T_{i+1}$ für genau ein $i \in \{-1, \dots, n-1\}$.

In allen anderen Fällen wäre n nicht minimal. Sei nun m die Anzahl der $i \in \{-1, \dots, n-1\}$, für die $T_i \cap T_{i+1}$ keinen Punkt aus $\mathcal{N} \setminus \mathcal{N}_0$ enthält. Für solche i gilt nach Satz 2.9 und Satz 2.15(ii) $|\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \leq C_3$. Aufgrund der Minimalität von n gilt $m \leq \#\mathcal{N}_0$. Falls $T_i \cap T_{i+1}$ hingegen einen Punkt aus $\mathcal{N} \setminus \mathcal{N}_0$ enthält, gilt nach Voraussetzung $|\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \leq L_{\mathrm{patch}}$. Nun gilt mit der Dreiecksungleichung

$$\begin{aligned} |\mathtt{level}(T_j') - \mathtt{level}(T)| &\leq \sum_{i=-1}^{n-1} |\mathtt{level}(T_{i+1}) - \mathtt{level}(T_i)| \\ &\leq L_{\mathrm{patch}}(n+1-m) + C_3 m \\ &\leq L_{\mathrm{patch}}(n+1) + C_3 \#\mathcal{N}_0 \,. \end{aligned}$$

Es folgt

$$\begin{aligned} h_j^s &\geq C_7^{-s/d} C_4^s \mathtt{diam}(T)^s 2^{s(L_{\mathrm{patch}} n - L_{\mathrm{patch}}(n+1) - C_3 \#\mathcal{N}_0)/d} \\ &= C_{17} \mathtt{diam}(T)^s \end{aligned}$$

mit $C_{17} := C_7^{-s/d} C_4^s 2^{-s(\#\mathcal{N}_0 C_3 + L_{\mathrm{patch}})/d}$ und damit die untere Abschätzung in (67). Sei nun $s < 0$. Wir haben eben gezeigt, dass

$$h_j^{s'} \geq C_{17} \mathtt{diam}(T)^{s'} \tag{73}$$

für alle $T \in \mathcal{T}$, $z_j \in \mathcal{N}(\mathcal{T})$ und alle $s' \geq 0$. Mit $s' := -s > 0$ gilt insbesondere

$$h_j^{-s} = h_j^{s'} \overset{(73)}{\geq} C_{17} \mathtt{diam}(T)^{s'} = C_{17} \mathtt{diam}(T)^{-s} \,.$$

Da $(\cdot)^{-1} : \mathbb{R}^+ \to \mathbb{R}^+$ monoton fallend ist, folgt

$$h_j^s \leq C_{17}^{-1} \mathtt{diam}(T)^s$$

und damit die obere Abschätzung in (67).

Alle Voraussetzungen (66)–(67) von Proposition 4.7 können erfüllt werden, und die L^2-Orthogonalprojektion $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}_D^1(\mathcal{T})$ ist stabil bezüglich der $h_{\mathcal{T}}^{-s}$-gewichteten L^2-Norm. □

4.3 2D

Mit der Vorarbeit aus Kapitel 3.4 können wir (56) für gewisse $s \in \mathbb{R}$ beweisen:

Satz 4.9. *Sei $d = 2$ und $\mathcal{T} \in \mathtt{refine}(\mathcal{T}_0)$. Sei $s \in \mathbb{R}$ mit $|s| < 1.1024....$ Dann ist $\Pi_D(\mathcal{T}) : L^2(\Omega) \to \mathcal{S}_D^1(\mathcal{T})$ stabil in der $h_{\mathcal{T}}^{-s}$-gewichteten L^2-Norm* (56).

Beweis. Nach Satz 3.23 ist (69) in 2D gültig mit $L_{\text{patch}} := 3$. Die Bedingung $L_{\text{patch}} = 3 < \frac{2}{2|s|} \log_2 B(2)$ in Proposition 4.8 gilt für alle $s \in \mathbb{R}$ mit

$$|s| < \log_2 B(2)/3 = 1.1024\ldots .$$

Es können alle Voraussetzungen von Proposition 4.8 erfüllt werden, und es folgt die Behauptung. □

Literatur

[BDD04] Peter Binev, Wolfgang Dahmen, and Ron DeVore. Adaptive finite element methods with convergence rates. *Numerische Mathematik*, 97(2):219–268, 2004.

[BPS02] James H. Bramble, Joseph E. Pasciak, and Olaf Steinbach. On the stability of the L^2 projection in $H^1(\Omega)$. *Math. Comp.*, 71:147–156, 2002.

[KPP13a] Michael Karkulik, David Pavlicek, and Dirk Praetorius. On 2d newest vertex bisection: Optimality of mesh-closure and h 1-stability of l 2-projection. *Constructive Approximation*, 38(2):213–234, 2013.

[KPP13b] Michael Karkulik, Carl-Martin Pfeiler, and Dirk Praetorius. L2-orthogonal projections onto finite elements on locally refined meshes are h1-stable. *arXiv preprint arXiv:1307.0917*, 2013.

[Pra15] Dirk Praetorius. Vorlesungsskript zur Numerik partieller Differentialgleichungen: stationäre Probleme, 2014–2015.

[Ste08] Rob Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77:227–241, 2008.

[Tra97] C. T. Traxler. An algorithm for adaptive mesh refinement in n dimensions. *Computing*, 59:115–137, 1997.