DIPLOMARBEIT

Zur Instanzoptimalität adaptiver 2D FEM

ausgeführt am

Institut für
Analysis und Scientific Computing
TU Wien

unter der Anleitung von

Univ.-Prof. Dipl.-Math. Dr.techn. Dirk Praetorius

und

Dipl.-Ing. Dr.techn. Alexander Haberl

durch

Michael Innerberger BSc. BSc.
Matrikelnummer: 01225448
Linke Wienzeile 130
1060 Wien

Wien, am 13.09.2018

Dirk Praetorius	Michael Innerberger

Kurzfassung

Ziel dieser Arbeit ist der Beweis der Instanzoptimalität adaptiver Finite Elemente Methoden (AFEM) für verschiedene Modellprobleme. Aufbauend auf dem Begriff der Populationen, der es erlaubt, bestimmte geometrische Eigenschaften von Gittern und deren Knotenmengen zu beweisen, sowie dem Begriff der Energie, die eng mit der Finite Elemente Lösung auf einem Gitter und deren Approximationsfehler verbunden ist, wird ein abstraktes Framework geschaffen, um Instanzoptimalität einer AFEM zu zeigen. Drei Eigenschaften werden sich hier als hinreichend erweisen: eine Lower Diamond Estimate der Energie, diskrete lokale Äquivalenz von Energie und a posteriori Fehlerschätzer, sowie eine Forderung an den Markierungsschritt.

Diese Eigenschaften werden für zwei Modellprobleme nachgewiesen. Als Erstes werden elliptische Diffusionsprobleme mit gemischten Neumann- und homogenen Dirichlet-Randbedingungen betrachtet, die durch H^1-konforme Finite Elemente beliebiger Ordnung diskretisiert werden. Weiter wird das abstrakte Framework auf zielorientierte adaptive Finite Elemente Methoden (GOAFEM) angewendet, bei denen die interessierende Größe der Funktionalwert eines linearen Funktionals der Lösung ist. Um den Fehler dieses Funktionalwerts abzuschätzen, wird eine modifizierte Fehlergröße eingeführt und für diese Instanzoptimalität gezeigt.

Abschließend werden die Resultate einiger numerischer Experimente angegeben.

Insgesamt verallgemeinert die Diplomarbeit die Arbeit [Diening, Kreuzer, Stevenson; Found. Comput. Math. 16 (2016)], in der Instanzoptimalität für eine adaptive P1-FEM für das Poisson-Problem mit homogenen Dirichlet-Randbedingungen gezeigt wird.

Abstract

This thesis aims to prove instance optimality of adaptive finite element methods (AFEMs) for various model problems. Based on the concept of populations, which allows for the proof of certain geometric properties of meshes and their sets of nodes, as well as the concept of energy, which is closely related to the finite element solution of a mesh and its approximation error, an abstract framework is developed for proving instance optimality of an AFEM for selected problems. Three properties will turn out to be sufficient for instance optimality: a lower diamond estimate of the energy, discrete local equivalence of energy and a posteriori error estimator, as well as an assumption on the marking step.

These properties will be shown to be valid for two model problems. First, elliptic diffusion problems with mixed Neumann and homogeneous Dirichlet boundary conditions will be considered, which are discretised by H^1-conforming finite elements of arbitrary order. Furthermore, the abstract framework will be applied to goal-oriented adaptive finite element methods (GOAFEM), in which the quantity of interest is the value of a linear functional of the solution. To estimate the error of this value, a modified error quantity is introduced, for which instance optimality is shown.

Finally, the theoretical findings are underpinned by numerical experiments.

Overall, the present diploma thesis generalizes the work [Diening, Kreuzer, Stevenson; Found. Comput. Math. 16 (2016)], which proves instance optimality of an adaptive P1-FEM for the Poisson problem with homogeneous Dirichlet boundary conditions.

Danksagung

Zuallererst möchte ich hier meinen Betreuern Professor Dirk Praetorius und Alexander Haberl für die gute Zusammenarbeit und die immer lehrreichen und anregenden Gespräche während der Arbeit an meiner Diplomarbeit danken. Außerdem danke ich meinem Hauptbetreuer Dirk Praetorius für die finanzielle Unterstützung im Rahmen des FWF-Projekts P27005 "Optimal Adaptivity for BEM and FEM-BEM Coupling", die dazu beigetragen hat, dass ich mich ein wenig unbeschwerter auf meine Arbeit konzentrieren konnte.

Ich bedanke mich an dieser Stelle auch bei meinen Studienkollegen für zahllose Stunden, die wir gemeinsam über Übungsbeispielen mehr oder weniger intensiv gebrütet haben, und für viele schöne Erfahrungen in Studium und Freizeit. Des Weiteren will ich meiner Arbeitsgruppe und vor Allem meinen Bürokollegen meinen Dank aussprechen, sowohl für die geistige Anregung weit über das Thema meiner Arbeit hinaus, als auch für die nötige geistige Zerstreuung während anstrengender Tage im Büro.

Mein besonderer Dank gilt meiner Freundin Magdalena. Dafür, dass du immer ein offenes Ohr und einen guten Ratschlag parat hast, wenn ich einmal nicht weiter weiß, und für die schöne Zeit, die ich mit dir in den letzten Jahren verbringen durfte.

Zu guter Letzt danke ich meinen Eltern für die moralische Unterstützung, mit der sie meinen Interessen und Entscheidungen begegnet sind, und die finanzielle Hilfe, um diese zu verwirklichen.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 3. Oktober 2018

Michael Innerberger

Inhaltsverzeichnis

1 Einleitung **1**
- 1.1 Generische Formulierung adaptiver FEM . . . 1
- 1.2 Historischer Überblick zur Optimalität von AFEM . . . 2
- 1.3 Aufbau der Arbeit . . . 3

2 Gitter und Populationen **5**
- 2.1 Newest Vertex Bisection . . . 5
- 2.2 Populationen . . . 8
- 2.3 Genealogie von Populationen . . . 11

3 Allgemeiner Beweis der Energieoptimalität **15**
- 3.1 Voraussetzungen . . . 15
  - 3.1.1 Markierungsstrategie . . . 15
  - 3.1.2 Energie . . . 17
- 3.2 Vorbereitungen . . . 19
- 3.3 Energieoptimalität . . . 24

4 Konforme Finite Elemente für Diffusionsprobleme **35**
- 4.1 Variationsformulierung und Energie . . . 36
  - 4.1.1 Schwache Formulierung und Diskretisierung . . . 36
  - 4.1.2 Energie . . . 37
- 4.2 Lower Diamond Estimate . . . 40
  - 4.2.1 Die Scott-Zhang Projektion als Transferoperator . . . 40
  - 4.2.2 Beweis der LDE . . . 46
- 4.3 Gesamtenergie . . . 48
  - 4.3.1 Lower Diamond Estimate . . . 49
- 4.4 Residualschätzer . . . 51
  - 4.4.1 Diskrete lokale Abschätzungen . . . 52
- 4.5 Markierungsstrategie . . . 63
- 4.6 Instanzoptimalität . . . 64

5 Goal-Oriented FEM **67**
- 5.1 Problemstellung . . . 67
  - 5.1.1 Fehlergröße . . . 68
- 5.2 Energie und AFEM . . . 70
  - 5.2.1 Energie . . . 70
  - 5.2.2 Fehlerschätzer . . . 71
  - 5.2.3 Markierungsschritt . . . 71

5.3 Instanzoptimalität 72
5.3.1 Voraussetzungen für Energieoptimalität 72
5.3.2 Verallgemeinerter Markierungsschritt 75
5.3.3 Instanzoptimalität für die Fehlergröße 77

6 Numerische Resultate **79**
6.1 Homogene Dirichlet-Daten 79
6.2 Gemischte Randbedingungen 82
6.3 GOAFEM 84

Literaturverzeichnis **89**

1 Einleitung

1.1 Generische Formulierung adaptiver FEM

In den letzten drei Jahrzehnten sind adaptive Finite Elemente Methoden (AFEM) zur numerischen Lösung von partiellen Differentialgleichungen in Anwendungen wie Technik und Naturwissenschaften immer wichtiger geworden. Einerseits benötigt eine AFEM, abhängig vom vorliegenden Problem, unter Umständen deutlich weniger Rechenkapazität als klassische FEM mit kleiner Gitterweite. Andererseits eignet sich eine AFEM aufgrund ihrer Robustheit gegenüber dem initialen Gitter besonders gut für Black-Box Löser, bei denen sich ein Anwender keine Gedanken über Approximationsfehler machen muss.

Eine AFEM besteht grob gesagt aus einem Finite Elemente Löser und einer Rückkopplungsschleife, die steuert, an welchen Stellen das der Finite Elemente Methode zugrundeliegende Gitter lokal verfeinert werden muss, um den Fehler der Finite Elemente Approximation *optimal* zu reduzieren. Jedem solchen Verfahren liegt das Schema aus Algorithmus 1 zugrunde, das so lange wiederholt wird, bis eine gewünschte Genauigkeit erreicht ist, oder

Algorithmus 1 Abstrakte AFEM

```
1: while Abbruchkriterium nicht erreicht do
2:     SOLVE
3:     ESTIMATE
4:     MARK
5:     REFINE
6: end while
```

die Rechenkapazität erschöpft ist. Dessen Komponenten wollen wir kurz beleuchten:

SOLVE – Hier wird die Finite Elemente Lösung zu einem gegebenen Gitter berechnet. Dieser Schritt für sich alleine betrachtet bildet die klassische (nicht adaptive) Finite Elemente Methode.

ESTIMATE – In diesem Schritt werden Fehlerschätzer berechnet, die ein Indikator für den (unbekannten) Fehler der Approximation an die exakte Lösung sein sollen und sich der bereits berechneten Approximation bedienen (a posteriori Fehlerschätzer).

MARK – Aufgrund der berechneten Fehlerschätzer werden Teile des Gitters zur Verfeinerung markiert. In den meisten gängigen AFEM werden Elemente markiert, es ist aber auch möglich, wie in dieser Arbeit, Elementränder zu markieren. In zwei Raumdimensionen sind dies die Kanten des Gitters.

REFINE – Mittels einer Verfeinerungsregel werden Teile des Gitters verfeinert. Dies geschieht derart, dass mindestens alle markierten Teile des Gitters verfeinert werden. Um gewisse Gittereigenschaften zu erhalten, können aber auch mehr als die markierten Teile verfeinert werden.

1.2 Historischer Überblick zur Optimalität von AFEM

Ziel der gesamten Prozedur ist es, Gitter zu erhalten, auf denen die Finite Elemente Lösung der exakten möglichst nahe kommt, dabei der Aufwand zu deren Berechnung aber möglichst gering ist. Betrachtet man eine Folge von Gittern, die aus einem initialen Gitter durch sukzessive Anwendung einer festgelegten Verfeinerungsstrategie hervorgehen, so kann man den Approximationsfehler als Funktion der Anzahl der Freiheitsgrade ansehen und erhält eine Folge von Fehlerwerten. Obwohl in numerischen Experimenten beobachtet werden kann, dass eine AFEM hier bessere Resultate erzielt als eine klassische FEM mit uniform feinem Gitter, war lange Zeit keine theoretische Antwort auf die Frage, ob die Folge von Lösungen einer AFEM überhaupt gegen die exakte Lösung konvergiert, bekannt.

Unter gewissen Voraussetzungen an die Schritte ESTIMATE und REFINE, vor allem aber an MARK, konnte in der grundlegenden Arbeit [Dör96] Konvergenz einer AFEM für ein einfaches Modellproblem gezeigt werden. Die dort verwendete Markierungsstrategie wird heute als Dörfler-Markierung bezeichnet und wird von einem Großteil der mathematischen Arbeiten über AFEM verwendet. Dieses Konvergenzresultat wurde im Folgenden immer mehr verallgemeinert und Voraussetzungen abgeschwächt.

Vor rund zehn Jahren konnte in [Ste07] auch das Problem gelöst werden, wie die oben erwähnte Optimalität in einen mathematisch sauberen Rahmen gesetzt und für gewisse AFEM gezeigt werden kann. Dieser mathematische Rahmen ist das Konzept der *Ratenoptimalität*. Dies bedeutet Folgendes: Betrachtet man die Konvergenzraten der Fehler für alle möglichen Folgen von Gittern, ausgehend von einem initialen Gitter, so generiert eine ratenoptimale AFEM eine Folge von Gittern, deren Fehler mit der bestmöglichen Rate konvergiert. Diese Art, Optimalität zu charakterisieren, ist bereits gut verstanden und auf eine Vielzahl an Problemen angewendet worden. In [CFPP14] etwa wird ein abstrakter Beweis der Ratenoptimalität für eine große Klasse von FEM und BEM (Boundary Element Method) für lineare und nichtlineare Probleme gegeben. Eine umfassende Einführung in adaptive Finite Elemente Methoden kann etwa in [Ver13] nachgelesen werden.

Eine andere Art, Optimalität zu charakterisieren, ist *Instanzoptimalität*. Eine instanzoptimale AFEM produziert in jedem Schritt ein Gitter, auf dem der Approximationsfehler bis auf eine Konstante kleiner gleich dem kleinsten Fehler ist, der auf jenen Gittern gemacht werden kann, die bis auf eine Konstante gleich viele oder weniger Elemente haben als das der AFEM. Dies wurde erstmals in [BDD04] für eine bestimmte AFEM gezeigt, wobei der Verfeinerungsschritt zusätzlich noch eine Vergröberung enthält, die in numerischen Experimenten unnötig erscheint. Erst kürzlich ist mit [DKS16] eine instanzoptimale AFEM veröffentlicht worden, die auch in der numerischen Analysis diesen Vergröberungsschritt weglassen kann.

Wir wollen in dieser Arbeit nicht näher auf das Konzept der Ratenoptimalität eingehen. Es sei hier nur angemerkt dass es schwächer ist als Instanzoptimalität in dem Sinn, dass

eine instanzoptimale AFEM schon ratenoptimal ist [DKS16]. Instanzoptimalität wird im Laufe dieser Arbeit genauer definiert und untersucht werden.

1.3 Aufbau der Arbeit

Das Resultat über eine instanzoptimale AFEM ohne Vergröberungsschritt in [DKS16] ist in gewisser Weise sehr restriktiv. Erstens behandelt es nur ein Poisson-Problem mit homogenen Dirichlet-Daten und eine Diskretisierung mit Finiten Elementen niedrigster Ordnung ($p = 1$). Zweitens sind sowohl Markierungsschritt als auch Verfeinerungsregel sehr stark limitiert. Wir werden uns in dieser Arbeit vor allem mit der ersten Restriktion beschäftigen und die Methoden aus [DKS16] auf eine größere Klasse von Problemen übertragen und erweitern.

In Kapitel 2 werden wir deshalb einen kurzen Überblick über die in [DKS16] eingeführten Konzepte geben, die auf der Verfeinerungsregel aufbauen. Dabei sollen die meisten Resultate nur zitiert werden, da hier nicht der Fokus liegt.

Weiter werden wir in Kapitel 3 Instanzoptimalität für eine Energie zeigen, die im Zusammenhang mit dem Approximationsfehler von FEM-Lösungen steht. Die Techniken hierfür werden wir größtenteils aus [DKS16] übernehmen, die Resultate aber in einer Form präsentieren und beweisen, die unabhängig von dem dort betrachteten Problem ist. Auf diese Weise soll ein abstraktes Framework für die folgenden Kapitel geschaffen werden.

Schließlich wollen wir in den Kapiteln 4 und 5 Anwendungen für den abstrakten Rahmen geben, die die Resultate aus [DKS16] erweitern. Kapitel 4 konzentriert sich dabei auf Diffusionsprobleme mit stückweise konstanten Diffusionsmatrizen und gemischten Randbedingungen (inhomogene Neumann- und homogene Dirichlet-Randbedingungen jeweils auf einem Teil des Randes), die mittels Finiten Elementen beliebiger Ordnung ($p \geq 1$) behandelt werden. In Kapitel 5 widmen wir uns schließlich dem Fall, dass uns nicht die gesamte Lösung, sondern nur der Wert eines Zielfunktionals der Lösung interessiert (Goal-Oriented FEM). Hierbei werden allgemeinere rechte Seiten zugelassen, als dies in Kapitel 4 der Fall war. Im Grunde könnten diese Probleme auch gemeinsam behandelt werden, beide Themen verlangen aber jeweils eine etwas andere Behandlung, weshalb sie in dieser Arbeit getrennt vorgestellt werden.

In dieser Arbeit werden wir grundlegende Kenntnisse von Funktionalanalyis ([Yos80]), der Theorie partieller Differentialgleichungen ([Eva10]), numerischer Mathematik und speziell der Finite Elemente Methode ([Bra13]) voraussetzen und einige wohlbekannte Resultate und Sichtweisen ohne genauere Referenz verwenden.

2 Gitter und Populationen

Die gesamte Analysis der in [DKS16] vorgestellten instanzoptimalen Finite-Elemente-Methode stützt sich auf eine neue Sichtweise auf hierarchische Gitter und den dabei eingeführten Begriff der Populationen. Wir werden in diesem Kapitel zunächst einige Eigenschaften von Gittern vorstellen und auf hierarchische Strukturen von solchen Gittern eingehen, die durch eine bestimmte Verfeinerungsstrategie, Newest Vertex Bisection (NVB), erzeugt werden. Danach werden wir den Begriff der Populationen definieren und mit diesen Strukturen in Verbindung bringen. Schließlich sollen hier einige Resultate festgehalten werden, die für den Beweis der Instanzoptimalität benötigt werden.

2.1 Newest Vertex Bisection

Für ein polygonales Gebiet $\Omega \subseteq \mathbb{R}^2$ definieren wir zunächst, wie eine passende Diskretisierung mittels Dreiecken aussieht [Bra13].

Definition 2.1. *Eine Zerlegung $\mathcal{T} = \{T_1, T_2, \dots, T_N\}$ in abgeschlossene Dreiecke $T_i \subseteq \overline{\Omega}$ heißt* reguläres Gitter*, wenn folgende Bedingungen erfüllt sind:*

1. *Das Gitter überdeckt das Gebiet:*

$$\overline{\Omega} = \bigcup_{i=1}^{N} T_i.$$

2. *Besteht für $i \neq j$ die Menge $T_i \cap T_j$ aus genau einem Punkt, so ist dieser ein Eckpunkt sowohl von T_i, als auch von T_j.*

3. *Besteht für $i \neq j$ die Menge $T_i \cap T_j$ aus mehr als einem Punkt, so ist $T_i \cap T_j$ eine Kante sowohl von T_i, als auch von T_j.*

Da wir im Folgenden nur reguläre Gitter verwenden, werden wir diese oft nur Gitter nennen. Weiter bezeichnen wir für ein Gitter $\mathcal{T}$ mit

$$\mathcal{V} = \mathcal{V}(\mathcal{T})$$

die Menge der Eckpunkte (Knoten, Vertices) und mit

$$\mathcal{E} = \mathcal{E}(\mathcal{T})$$

die Menge aller Kanten aller in $\mathcal{T}$ enthaltenen Dreiecke. Außerdem bezeichen wir mit $\mathcal{E}^\Gamma = \mathcal{E}^\Gamma(\mathcal{T})$ alle Kanten $E \in \mathcal{E}$ mit $E \subseteq \partial\Omega$.

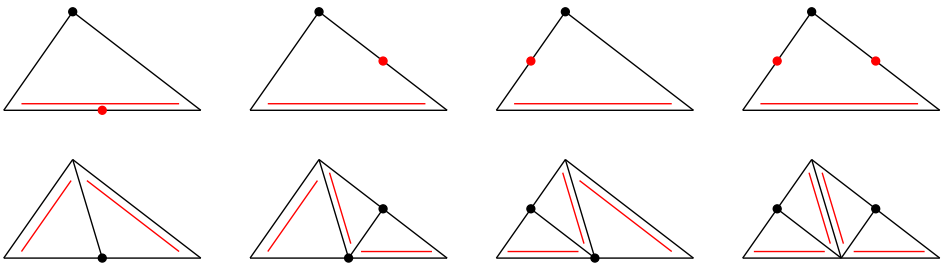


Abbildung 2.1: *Möglichkeiten der Verfeinerung eines Dreiecks durch NVB. In der oberen Reihe sind alle möglichen Markierungen abgebildet, darunter die resultierenden Verfeinerungen. Referenzkanten sind durch rote Linien angedeutet, newest vertices durch einen schwarzen Punkt. Markierte Kanten sind mit einem roten Punkt versehen.*

Eine adaptive Finite-Elemente-Methode benötigt natürlich auch eine Möglichkeit, Gitter lokal zu verfeinern. Hierfür gibt es verschiedene Verfeinerungsstrategien. In dieser Arbeit werden wir uns auf Newest Vertex Bisection beschränken, da sie erlaubt, die erzeugten Gitter als Populationen mit einigen interessanten Eigenschaften aufzufassen, wie wir später sehen werden.

Für ein gegebenes Gitter $\mathcal{T}$ ist bei NVB pro Element eine Kante — die Verfeinerungskante (refinement edge) — ausgewählt, die verfeinert werden soll. Ist nun eine beliebige Kante zur Verfeinerung markiert, werden auch die Verfeinerungskanten der anliegenden Dreiecke markiert. Dies wird rekursiv durchgeführt, bis keine zusätzliche Kante mehr markiert wird (höchstens also $3\,\#\mathcal{T}$ mal, da spätestens dann alle Kanten markiert sind). Dieses rekursive Markieren der Referenzkanten ist nötig, damit das entstehende Gitter wieder regulär ist. Sind alle nötigen Kanten markiert, werden zunächst alle markierten Referenzkanten halbiert, wodurch insgesamt ein neuer Knoten (newest vertex), zwei neue Kanten und zwei neue Elemente entstehen. Für die neu entstandenen Elemente ist die Verfeinerungskante die dem newest vertex gegenüberliegende Kante. Nun werden erneut alle markierten Referenzkanten halbiert, womit insgesamt alle ursprünglich markierten Kanten verfeinert wurden. Die Vorgehensweise von NVB ist in Abbildung 2.1 illustriert. Da die Referenzkante immer zusätzlich markiert wird, sofern eine andere Kante des Dreiecks markiert wird, sind dort bereits alle möglichen Arten, ein Dreieck durch NVB zu verfeinern, abgebildet. [KPP13]

Die Menge aller regulären Gitter, die Verfeinerungen eines initialen Gitters $\mathcal{T}_0$ bezüglich NVB sind, bezeichnen wir als $\mathbb{T} := \mathbb{T}(\mathcal{T}_0)$. Diese hierarchische Struktur trägt eine Halbordnung und, wie wir später sehen werden, auch eine Verbandsstruktur.

Definition 2.2. *Für ein Gitter $\mathcal{T} \in \mathbb{T}$ definieren wir die Menge aller regulären* Verfeinerungen*, respektive* Vergröberungen *von $\mathcal{T}$ als*

$$\begin{aligned} \operatorname{refine}(\mathcal{T}) &:= \left\{\mathcal{T}' \in \mathbb{T} \mid \mathcal{T}' \textit{ ist Verfeinerung von } \mathcal{T}\right\}, \\ \operatorname{coarse}(\mathcal{T}) &:= \left\{\mathcal{T}' \in \mathbb{T} \mid \mathcal{T} \in \operatorname{refine}(\mathcal{T}')\right\}. \end{aligned}$$

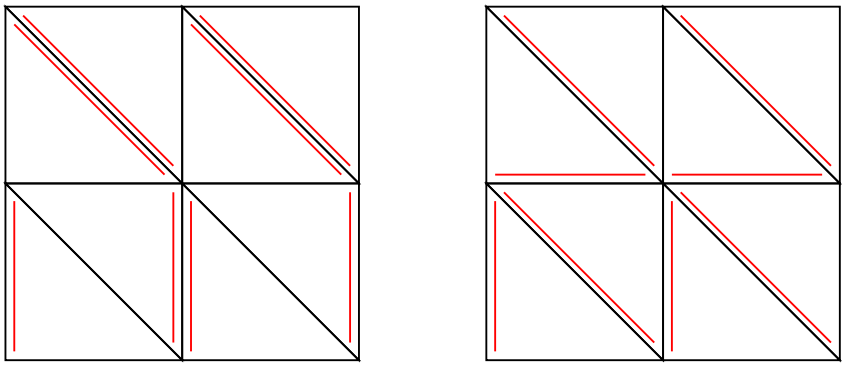


Abbildung 2.2: *Ein Gitter, das die BDD-Bedingung erfüllt (links) und eines, das sie verletzt (rechts).*

Durch die Relationen

$$\mathcal{T} \leq \mathcal{T}' :\Longleftrightarrow \mathcal{T}' \in \operatorname{refine}(\mathcal{T}),$$
$$\mathcal{T} \geq \mathcal{T}' :\Longleftrightarrow \mathcal{T}' \in \operatorname{coarse}(\mathcal{T})$$

wird eine Halbordnung auf $\mathbb{T}$ *definiert.*

Eine Frage bei dieser Methode ist, wie die Verfeinerungskanten des initialen Gitters am besten gewählt werden sollen. Wir beschränken uns in dieser Arbeit auf initiale Gitter $\mathcal{T}_0$, die folgende Bedingung erfüllen.

Definition 2.3 (BDD-Bedingung)**.** *Sei* $\mathcal{T}$ *ein reguläres Gitter.* $\mathcal{T}$ *erfüllt die* BDD-Bedingung*, wenn für alle* $T, T' \in \mathcal{T}$ *gilt:*
Wenn $T \cap T'$ *die Verfeinerungskante von* T *ist, so folgt, dass sie auch Verfeinerungskante von* T' *ist.*

Es kann gezeigt werden, dass für jedes initiale Gitter eine Auswahl an Verfeinerungskanten existiert, sodass die BDD-Bedingung erfüllt ist [BDD04, Lemma 2.1]. Beispiele für Gitter, die die BDD-Bedingung erfüllen, bzw. verletzen, sind in Abbildung 2.2 gegeben.

Schließlich wollen wir noch die zu einer Menge von Dreiecken gehörige Fläche und Patches definieren.

Definition 2.4. *Für eine Menge* $\mathcal{U} \subseteq \bigcup_{\mathcal{T} \in \mathbb{T}} \mathcal{T}$ *von Dreiecken ist die Fläche* $\Omega(\mathcal{U})$ *aller Dreiecke definiert als*

$$\Omega(\mathcal{U}) := \left(\bigcup_{T \in \mathcal{U}} T \right)^{\circ}.$$

Für ein Dreieck $T \in \mathcal{T}$ *und eine Kante* $E \in \mathcal{E}$ *definieren wir* Elementpatch, Kantenpatch

und (im Falle $E \subseteq \partial\Omega$) Randpatch *als*

$$\begin{aligned}\omega_T &:= \left\{T' \in \mathcal{T} \mid T \cap T' \neq \emptyset\right\},\\ \omega_E &:= \left\{T' \in \mathcal{T} \mid E \cap T' \neq \emptyset\right\},\\ \omega_E^\Gamma &:= \left\{E' \in \mathcal{E} \mid E' \subseteq \partial\Omega \text{ und } E \cap E' \neq \emptyset\right\}.\end{aligned}$$

Außerdem bezeichnen wir mit

$$\omega_E^{\mathrm{red}} := \left\{T' \in \mathcal{T} \mid E \subseteq \partial T'\right\}$$

den reduzierten Kantenpatch *von E.*

Gitter, die durch NVB entstehen, haben eine Eigenschaft, die garantiert, dass für jedes Dreieck T die Anzahl an Dreiecken im Patch ω_T uniform beschränkt ist. Diese Eigenschaft wird typischerweise wie folgt definiert, woraus die Beschränktheit der Patches folgt [EGP18].

Proposition 2.5. *Sei $\mathcal{T}_0$ ein Gitter. Alle Gitter $\mathcal{T} \in \mathbb{T}$ sind* formregulär*, d.h. es gibt eine Konstante $C_{\mathrm{sr}} > 0$, sodass*

$$\max_{T\in\mathcal{T}} \frac{\operatorname{diam}(T)}{|T|^{1/2}} \leq C_{\mathrm{sr}}. \tag{2.1}$$

Die Konstante C_{sr} hängt nur vom initialen Gitter $\mathcal{T}_0$ ab. □

2.2 Populationen

Eine andere Sichtweise auf von NVB erzeugte Gitter erhält man durch sogenannte Populationen [DKS16]:

Definition 2.6. *Wir bezeichnen die Menge der Knoten aller durch NVB erzeugten Gitter mit*

$$\mathcal{P}_\infty := \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{V}(\mathcal{T}).$$

Ihre Elemente nennen wir Personen*. Eine Menge $\mathcal{P} \subseteq \mathcal{P}_\infty$ von Personen heißt* Population*, falls es ein Gitter $\mathcal{T} \in \mathbb{T}$ gibt, sodass*

$$\mathcal{P} = \mathcal{V}(\mathcal{T}).$$

Die Menge aller Populationen ausgehend von einer Startpopulation $\mathcal{P}_0 := \mathcal{V}(\mathcal{T}_0)$ bezeichnen wir mit $\mathbb{P}$.

Offensichtlich gibt es eine eindeutige Beziehung zwischen Populationen und Gittern. Wir schreiben daher $\mathcal{P}(\mathcal{T}), \mathcal{T}(\mathcal{P})$ für die zu einem Gitter gehörige Population, respektive das zu einer Population gehörige Gitter. Auf $\mathbb{P}$ definiert die Mengeninklusion eine Halbordnung, die mit der Halbordnung auf $\mathbb{T}$ übereinstimmt ($\mathcal{P}, \mathcal{P}' \in \mathbb{P}$):

$$\mathcal{P} \subseteq \mathcal{P}' \iff \mathcal{T}(\mathcal{P}) \leq \mathcal{T}(\mathcal{P}').$$

Über Vereinigung und Schnitt zweier Populationen kann man darüber hinaus Operationen definieren, die $\mathbb{P}$ zu einem Verband machen, und diese auf Gitter übertragen.

Definition 2.7. *Für Populationen $\mathcal{P}_1, \mathcal{P}_2 \in \mathbb{P}$ definieren wir die* feinste gemeinsame Vergröberung*, beziehungsweise die* gröbste gemeinsame Verfeinerung *als*

$$\mathcal{P}_1 \cap \mathcal{P}_2, \quad \mathcal{P}_1 \cup \mathcal{P}_2.$$

Für Gitter $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ definieren wir die analogen Größen als

$$\mathcal{T}_1 \wedge \mathcal{T}_2 := \mathcal{T}\big(\mathcal{P}(\mathcal{T}_1) \cap \mathcal{P}(\mathcal{T}_2)\big), \quad \mathcal{T}_1 \vee \mathcal{T}_2 := \mathcal{T}\big(\mathcal{P}(\mathcal{T}_1) \cup \mathcal{P}(\mathcal{T}_2)\big).$$

Bemerkung 2.8. Die in Defintion 2.7 vorgestellten Mengen sind tatsächlich wieder Populationen, wie man sich leicht überlegen kann: NVB verfeinert ein Dreieck immer auf dieselbe Art, die von der Wahl der Verfeinerungskanten im initialen Gitter festgelegt wird. Deshalb muss für zwei Dreiecke $T_1 \in \mathcal{T}_1$ und $T_2 \in \mathcal{T}_2$ mit $|\Omega(T_1 \cap T_2)| \neq 0$ gelten, dass entweder $T_1 \subseteq T_2$ oder $T_2 \subseteq T_1$. Um zur gröbsten gemeinsamen Verfeinerung zu gelangen, müssen also nur in jedem Teil des Gebiets die Knoten des lokal feineren Gitters verwendet werden. Da die Knoten des gröberen Gitters aber in dieser Menge auch enthalten sind, entspricht dies einer Vereinigung der Knotenmengen, genau so wie es in Defintion 2.7 definiert ist. //

Da NVB als Verfeinerungsstrategie Kanten verfeinert, wäre eine Korrespondenz zwischen Kanten und Personen wünschenswert. Bezeichnen wir für $\mathcal{T} \in \mathbb{T}$ mit $\mathcal{T}^{++}$ das Gitter, in dem alle Kanten verfeinert wurden, also

$$\mathcal{T}^{++} := \min\left\{\mathcal{T}' \geq \mathcal{T} \mid \mathcal{E}(\mathcal{T}) \cap \mathcal{E}(\mathcal{T}') = \emptyset\right\},$$

und die zugehörigen Personen als $\mathcal{P}^{++} = \mathcal{P}(\mathcal{T}(\mathcal{P})^{++})$. Dann können wir mittels der Abbildung

$$\operatorname{midpt} : \bigcup_{\mathcal{T} \in \mathbb{T}} \mathcal{E}(\mathcal{T}) \to \mathcal{P}_\infty \setminus \mathcal{P}_0,$$

die jeder Kante ihren Mittelpunkt zuordnet, die Kanten eines Gitters mit bestimmten Personen identifizieren:

$$\mathcal{P}^{++} \setminus \mathcal{P} = \operatorname{midpt}(\mathcal{E}(\mathcal{T}(\mathcal{P}))).$$

Wegen der Art, wie NVB Kanten verfeinert, ist die Abbildung midpt außerdem umkehrbar, da jeder Knoten, der nicht aus dem initialen Gitter stammt, eindeutig als Mittelpunkt einer Kante dargestellt werden kann. Dies ist in Abbildung 2.3 dargestellt. Beachte, dass die Eckpunkte des Quadrats in $\mathcal{P}_0$ enthalten sind und sich somit nicht als Mittelpunkt einer Kante darstellen lassen.

Mithilfe der Verbandsoperationen auf $\mathbb{T}$ und $\mathbb{P}$ lassen sich nun Verfeinerungen und Vergröberungen von Gittern bezüglich bestimmter Kanten definieren.

Definition 2.9. *Für $\mathcal{P} \in \mathbb{P}$ und eine endliche Teilmenge $\mathcal{C} \subseteq \mathcal{P}_\infty$ definieren wir die* gröbste Verfeinerung von $\mathcal{P}$*,* die $\mathcal{C}$ enthält*, als*

$$\mathcal{P} \oplus \mathcal{C} := \bigcap \left\{\mathcal{P}' \in \mathbb{P} \mid \mathcal{P} \cup \mathcal{C} \subseteq \mathcal{P}'\right\}.$$

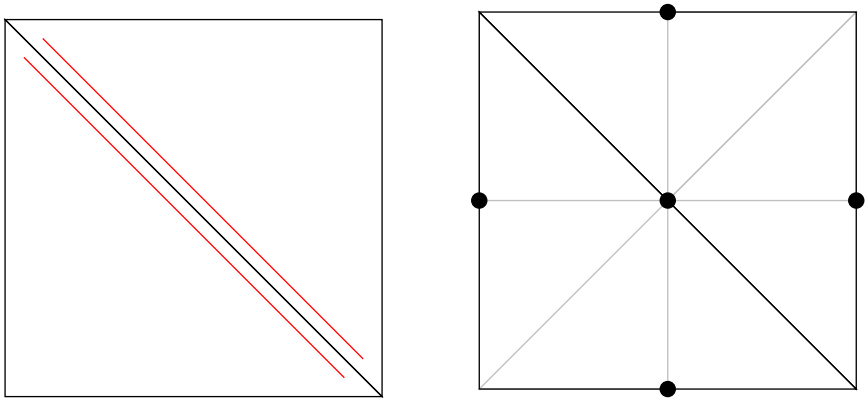


Abbildung 2.3: *Die Kanten eines Gitters (links) sind bei Verfeinerung mittels NVB eindeutig durch ihre Mittelpunkte (rechts) bestimmt. Die eingezeichneten Knoten des rechten Gitters sind in $\mathcal{P}_0^{++}$ enhalten, nicht aber in $\mathcal{P}_0$.*

Ist zusätzlich $\mathcal{C} \subseteq \mathcal{P}_\infty \setminus \mathcal{P}_0$ können wir die feinste Vergröberung von $\mathcal{P}$, die $\mathcal{C}$ nicht enthält, *definieren als*

$$\mathcal{P} \ominus \mathcal{C} := \bigcup \{\mathcal{P}' \in \mathbb{P} \mid \mathcal{P}' \subseteq \mathcal{P}, \mathcal{C} \cap \mathcal{P}' = \emptyset\}.$$

Für ein Gitter $\mathcal{T} \in \mathbb{T}$ und eine Menge von Kanten $\mathcal{M} \subseteq \mathcal{E}(\mathcal{T})$ definieren wir die gröbste Verfeinerung von $\mathcal{T}$, in der alle Kanten aus $\mathcal{M}$ verfeinert wurden, *als*

$$\operatorname{refine}(\mathcal{T}; \mathcal{M}) := \mathcal{T}\big(\mathcal{P}(\mathcal{T}) \oplus \operatorname{midpt}(\mathcal{M})\big).$$

Für eine Menge an Kanten $\mathcal{M} \subseteq \bigcup_{\mathcal{T}' \in \operatorname{coarse}(\mathcal{T})} \mathcal{E}(\mathcal{T}')$ definieren wir durch

$$\operatorname{coarse}(\mathcal{T}; \mathcal{M}) := \mathcal{T}\big(\mathcal{P}(\mathcal{T}) \ominus \operatorname{midpt}(\mathcal{M})\big)$$

das feinste Gitter, das gröber als $\mathcal{T}$ ist und alle Kanten $\mathcal{M}$ enthält.

Der adaptive Algorithmus wird später in jedem Schritt eine Menge von Kanten des aktuellen Gitters markieren und diese verfeinern, um zu einem neuen Gitter zu gelangen. Da hierbei eventuell mehr Kanten verfeinert werden müssen, als markiert waren, um zu einem regulären Gitter zu gelangen, ist es nötig, die Anzahl der verfeinerten Kanten durch die Anzahl der markierten abzuschätzen [BDD04, Theorem 2.4].

Satz 2.10 (Mesh-Closure Estimate)**.** *Für eine Folge $(\mathcal{T}_k)_{k\in\mathbb{N}} \subseteq \mathbb{T}$ mit Startgitter $\mathcal{T}_0$ und $\mathcal{T}_{k+1} = \operatorname{refine}(\mathcal{T}_k; \mathcal{M}_k)$ für $\mathcal{M}_k \subseteq \mathcal{E}(\mathcal{T}_k)$ gilt*

$$\#(\mathcal{T}_k \setminus \mathcal{T}_0) \leq C_{\mathrm{MC}} \sum_{i=0}^{k-1} \#\mathcal{M}_i.$$

Die Konstante $C_{\mathrm{MC}} > 0$ hängt nur vom initialen Gitter $\mathcal{T}_0$ ab. □

Bemerkung 2.11. Da für jeden im Zuge der NVB hinzugefügten Knoten die Anzahl der Dreiecke um mindestens eins (bei Randknoten), höchstens aber um zwei (bei inneren Knoten) steigt, gilt für $\mathcal{P}, \mathcal{P}' \in \mathbb{P}$ mit $\mathcal{P} \leq \mathcal{P}'$

$$\#(\mathcal{P}' \setminus \mathcal{P}) \leq \#(\mathcal{T}(\mathcal{P}') \setminus \mathcal{T}(\mathcal{P})) \leq 2\,\#(\mathcal{P}' \setminus \mathcal{P}),$$

womit die Mesh-Closure Estimate in analoger Form auch für Populationen gilt. //

2.3 Genealogie von Populationen

Nun, da wir die grundlegenden Begriffe eingeführt haben, um Populationen zu beschreiben, wollen wir einige Resultate vorstellen, die für den Beweis der Instanzoptimalität benötigt werden. Die Beweise dieser Resultate sind für unsere Zwecke nebensächlich. Sie werden daher im Folgenden nicht angegeben. Es sei dafür auf [DKS16] verwiesen. Um diese Resultate zu formulieren, benötigen wir jedoch noch einen Generationsbegriff auf der Menge $\mathbb{P}$ aller Populationen ausgehend von einer initialen Population $\mathcal{P}_0$.

Definition 2.12.

(i) Für ein Dreieck $T \in \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{T}$ ist die Generation $\operatorname{gen}(T)$ *des Dreiecks definiert als die Anzahl der Bisektionen, die notwendig sind, um T aus $\mathcal{T}_0$ zu erhalten.*

(ii) Für einen Knoten $P \in \mathcal{P}_0$ ist die Generation *definiert als* $\operatorname{gen}(P) := 0$. *Für einen Knoten $P \in \mathcal{P}_\infty \setminus \mathcal{P}_0$ gibt es ein Dreieck $T \in \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{T}$, sodass P Mittelpunkt der Verfeinerungskante von T ist. Die Generation des Knotens ist definiert als* $\operatorname{gen}(P) := \operatorname{gen}(T) + 1$.

(iii) Gilt für $P' \in \mathcal{P}_\infty \setminus \mathcal{P}_0$ und $T \in \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{T}$, dass P' Mittelpunkt der Verfeinerungskante von T ist, so nennen wir den newest vertex P von T Vater *von P'. Umgekehrt heißt P'* Sohn *von P.*

Bemerkung 2.13. Die Knotengeneration aus Defintion 2.12 (ii) ist für jeden Knoten definiert, da ein neuer Knoten nur entstehen kann, wenn eine Kante verfeinert wird. Damit dies geschieht, muss sie Verfeinerungskante eines angrenzenden Elements sein. Sie ist darüber hinaus eindeutig, wenn das initiale Gitter die BDD-Bedingung aus Defintion 2.3 erfüllt. Die Frage, wie das Konzept der Populationen, deren Eigenschaften hauptsächlich auf der Knotengeneration aufbauen, auf Gitter ohne die BDD-Bedingung verallgemeinert werden kann, ist bis dato noch offen. //

Aufbauend auf dem Generationsbegriff, können nun alle Vorfahren und Nachfahren eines Knoten definiert werden. Dies führt zum wichtigen Begriff der freien Knoten.

Definition 2.14. *Bezeichnen wir für einen Knoten $P \in \mathcal{P}_\infty$ die Menge seiner Väter als $\mathcal{U} := \{P' \in \mathcal{P}_\infty \mid P' \text{ ist Vater von } P\}$, so ist die Menge seiner* Vorfahren *rekursiv definiert als*

$$\operatorname{anc}(P) := \begin{cases} \emptyset, & \operatorname{gen}(P) = 0, \\ \mathcal{U} \cup \bigcup_{Q\in\mathcal{U}} \operatorname{anc}(Q), & \operatorname{gen}(P) \geq 1. \end{cases}$$

Die Menge aller Nachfahren *von P definieren wir als*

$$\operatorname{desc}(P) := \left\{P' \in \mathcal{P}_\infty \mid P \in \operatorname{anc}(P')\right\}.$$

Weiters definieren wir das Urbild von $k \in \mathbb{N}$ unter gen *in naheliegender Weise:*

$$\operatorname{gen}^{-1}(k) := \{P \in \mathcal{P}_\infty \mid \operatorname{gen}(P) = k\}.$$

Das folgende Resultat zeigt, dass für einen Knoten die Anzahl seiner Vorfahren in einer bestimmten Generation gleichmäßig beschränkt werden kann. In [DKS16] wird diese Eigenschaft auch *limited genetic diversity* genannt.

Proposition 2.15. *Die* Konstante der genetischen Diversität *erfüllt*

$$C_{\mathrm{GD}} := \sup_{P\in\mathcal{P}_\infty} \sup_{k\in\mathbb{N}} \#\big(\operatorname{anc}(P) \cap \operatorname{gen}^{-1}(k)\big) < \infty.$$

Diese Konstante hängt nur vom initialen Gitter $\mathcal{T}_0$ ab. □

Mithilfe von Vorfahren und Nachkommen lassen sich nun freie Knoten einer Menge definieren. Diese sind in dem Sinne frei, dass sie keine Nachkommen in der betrachteten Menge haben.

Definition 2.16. *Für eine Teilmenge $\mathcal{U} \subseteq \mathcal{P}_\infty \setminus \mathcal{P}_0$ nennen wir*

$$\operatorname{free}(\mathcal{U}) := \{P \in \mathcal{U} \mid \operatorname{desc}(P) \cap \mathcal{U} = \emptyset\}$$

die Menge der freien Knoten *von $\mathcal{U}$.*

Bemerkung 2.17. Wir haben die freien Knoten einer Menge nur für Teilmengen von $\mathcal{P}_\infty \setminus \mathcal{P}_0$ definiert. In Bemerkung 2.8 haben wir gesehen, dass die Menge $\mathcal{P}_\infty \setminus \mathcal{P}_0$ über die Abbildung midpt mit der Menge aller Kanten, die durch wiederholte Anwendung von NVB auf $\mathcal{T}_0$ entstehen, identifiziert werden kann. Daher ist es naheliegend für eine Menge $\mathcal{U} \subseteq \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{E}(\mathcal{T})$ von Kanten die Menge der freien Kanten als

$$\operatorname{free}(\mathcal{U}) := \operatorname{midpt}^{-1}(\operatorname{free}(\operatorname{midpt}(\mathcal{U})))$$

zu definieren. Dies sind Kanten, deren Mittelpunkte keine Nachfahren in $\operatorname{midpt}(\mathcal{U})$ haben. Das kann für eine Kante $E \in \mathcal{E}$ nur dann zutreffen, falls diese nicht bei Verfeinerung einer Kante aus $\mathcal{U} \setminus \{E\}$ mitverfeinert wird. //

Das folgende Lemma führt einige wichtige Eigenschaften von Populationen an. Aus Gründen der Übersichtlichkeit verzichten wir auch hier auf Beweise, da diese einer genaueren Betrachtung von Populationen bedürften, was nicht das Ziel dieser Arbeit ist. Sie können in [DKS16] nachgelesen werden.

Lemma 2.18. *Für eine Menge von Knoten $\mathcal{U} \subseteq \mathcal{P}_\infty \setminus \mathcal{P}_0$ und eine Population $\mathcal{P} \in \mathbb{P}$ gilt:*

(i) *Wenn* $\#\mathcal{U} < \infty$ *gilt, dann ist*

$$\mathcal{U} \cup \operatorname{anc}(\mathcal{U}) = \operatorname{free}(\mathcal{U}) \cup \operatorname{anc}(\operatorname{free}(\mathcal{U})). \tag{2.2}$$

(ii) *Für eine Menge* $\mathcal{C} \subseteq \mathcal{P}_\infty$ *gilt*

$$\mathcal{P} \oplus \mathcal{C} = \mathcal{P} \cup \mathcal{C} \cup \operatorname{anc}(\mathcal{C}). \tag{2.3}$$

(iii) *Für eine Menge* $\mathcal{C} \subseteq \operatorname{free}(\mathcal{P} \setminus \mathcal{P}_0)$ *gilt*

$$\mathcal{P} \ominus \mathcal{C} = \mathcal{P} \setminus \mathcal{C}. \tag{2.4}$$

(iv) *Für eine Teilmenge* $\mathcal{C} \subseteq \mathcal{U}$ *und* $\#\mathcal{U} < \infty$ *gilt mit der Konstanten* C_{GD} *aus Proposition 2.15, dass*

$$\#\operatorname{free}(\mathcal{C}) \leq C_{\mathrm{GD}} \#\operatorname{free}(\mathcal{U}). \tag{2.5}$$

(v) *Es gilt* $\operatorname{anc}(\mathcal{P}) \subseteq \mathcal{P}$.

Bemerkung 2.19. Die Aussagen von Lemma 2.18 unterstreichen die Wichtigkeit des Konzeptes der freien Knoten. Gleichung (2.3) zeigt, dass nach dem Hinzufügen von Knoten $\mathcal{C}$ zu einer bestehenden Population höchstens noch Knoten aus den Vorfahren von $\mathcal{C}$ hinzukommen müssen, um wieder zu einer Population zu gelangen. In Gleichung (2.2) wird gezeigt, dass es für endliche Mengen ausreicht, dafür nur die freien Knoten und deren Vorfahren zu betrachten. Analog dazu besagt Gleichung (2.4), dass bei der Entfernung freier Knoten aus einer Population keine weiteren Knoten mehr entfernt werden müssen, um wieder zu einer Population zu gelangen.

Die wohl wichtigste Eigenschaft ist Gleichung (2.5). Sie besagt, dass für zwei geschachtelte, endliche Mengen von Knoten die Anzahl der freien Knoten in einem gewissen Sinne immer miteinander vergleichbar ist. Diese Abschätzung wird sich im Beweis der Instanzoptimalität als essentiell erweisen. //

3 Allgemeiner Beweis der Energieoptimalität

Das zentrale Resultat dieser Arbeit ist, dass unter gewissen Voraussetzungen die AFEM instanzoptimal bezüglich einer Energie ist — dies soll im Folgenden Energieoptimalität genannt werden. Wir wollen in diesem Kapitel zunächst alle benötigten Begriffe einführen und dann das Resultat in einem allgemeinen Rahmen formulieren und beweisen.

Das Resultat stammt aus [DKS16], wobei hier herausgearbeitet wurde, welche Voraussetzungen essentiell sind, um ein allgemeines Framework zu schaffen, in dem eine größere Klasse von Problemen betrachtet werden kann, und welche Resultate sich auf Populationen stützen. Die Beweisstruktur ist teilweise aus [Hab] entnommen, da diese dort ausführlicher und übersichtlicher ist, als in der Originalarbeit.

Energieoptimalität benötigt als Voraussetzungen gewisse Bedingungen, auf die wir im Folgenden genauer eingehen werden:

- Der ganze Beweis arbeitet mit Populationen. Diese wurden in Kapitel 2 bereits vorgestellt und die benötigten Resultate angeführt. Bisher ist nicht bekannt, wie dieses Konzept vermieden werden kann. Auch ist nicht bekannt, ob und wie sich Energieoptimalität auf andere Verfeinerungs- und Markierungsstrategien als NVB mit kantenbasierter Markierung übertragen lässt. Dies soll im Rahmen dieser Arbeit nicht betrachtet werden.
- Die Markierungsstrategie muss den gesamten Fehler der mit einer Kante mitverfeinerten Kanten geeignet abschätzen, was, wie wir später sehen werden, durch eine modifizierte Maximumsstrategie erreicht werden kann.
- Die Energie muss mit der hierarchischen Struktur der Gitter und dem Fehlerschätzer in geeigneter Weise kompatibel sein.

3.1 Voraussetzungen

3.1.1 Markierungsstrategie

Zunächst werden wir betrachten, welche Eigenschaften die benutzte Markierungsstrategie aufweisen muss. Eine Markierungsstrategie markiert (in unserem Fall) Kanten aufgrund der Werte eines Fehlerschätzers. Ein *Fehlerschätzer* η^2 ist formal eine Abbildung

$$\eta^2 : \mathbb{T} \times \bigcup_{\mathcal{T}\in\mathbb{T}} \mathcal{E}(\mathcal{T}) \to \mathbb{R}_0^+$$

mit der Einschränkung, dass für Argumente $(\mathcal{T}, E)$ gelten muss, dass die Kante E aus der Kantenmenge $\mathcal{E}(\mathcal{T})$ des Gitters $\mathcal{T}$ stammt. Wir werden aber im Folgenden auch die Abbildung

$$\eta^2_{\mathcal{T}} : \mathcal{E}(\mathcal{T}) \to \mathbb{R}_0^+$$

für fixiertes erstes Argument $\mathcal{T} \in \mathbb{T}$ Fehlerschätzer nennen. Für eine Menge von Kanten $\mathcal{U} \subseteq \mathcal{E}(\mathcal{T})$ definieren wir den gesamten Fehlerschätzer als

$$\eta_{\mathcal{T}}^2(\mathcal{U}) := \sum_{E \in \mathcal{U}} \eta_{\mathcal{T}}^2(E).$$

In jedem Schritt der AFEM soll der Fehlerschätzer nur mithilfe von vorhandenen Daten berechnet werden können und die Kanten bestimmen, die beim Übergang zum nächsten Schritt verfeinert werden sollen. Bei der tatsächlichen Verfeinerung werden aber eventuell mehr Kanten verfeinert, um wieder zu einem regulären Gitter zu gelangen. Um eine energieoptimale AFEM zu erhalten, muss der Fehlerschätzer dies beim Markieren der Kanten berücksichtigen.

Definition 3.1. *Es sei die Menge aller Kanten in $\mathcal{E}(\mathcal{T})$, die bei alleiniger Verfeinerung von $E \in \mathcal{E}(\mathcal{T})$ mitverfeinert werden, mit*

$$\operatorname{tail}_{\mathcal{T}}(E) := \mathcal{E}(\mathcal{T}) \setminus \mathcal{E}(\operatorname{refine}(\mathcal{T}; E))$$

bezeichnet. Wir nennen diese Menge auch den Schweif von E in $\mathcal{T}$. *Für eine Menge $\mathcal{U} \subseteq \mathcal{E}(\mathcal{T})$ von Kanten sei der Schweif definiert als*

$$\operatorname{tail}_{\mathcal{T}}(\mathcal{U}) := \bigcup_{E \in \mathcal{U}} \operatorname{tail}_{\mathcal{T}}(E).$$

Die Menge der markierten Kanten auf einem Gitter $\mathcal{T}$ wollen wir mit $\mathcal{M}$ bezeichnen. Die Markierungsstrategie muss derart gewählt werden, dass es eine Konstante $\mu > 0$ gibt, sodass

$$\mathcal{M} \neq \emptyset \quad \text{und} \quad \eta_{\mathcal{T}}^2\big(\operatorname{tail}_{\mathcal{T}}(\mathcal{M})\big) \geq \mu \, \#\mathcal{M} \, \eta_{\mathcal{T}}^2\big(\operatorname{tail}_{\mathcal{T}}(E)\big) \quad \text{für alle } E \in \mathcal{E}(\mathcal{T}). \tag{E1}$$

Bemerkung 3.2. Das Kriterium (E1) schränkt die Auswahl der verwendeten Markierungsstrategie wesentlich ein, da die üblichen Strategien (Dörfler-Marking, Maximumsstrategie) dieses nicht erfüllen. Formt man diese Ungleichung um, erhält man

$$\frac{\eta_{\mathcal{T}}^2\big(\operatorname{tail}_{\mathcal{T}}(\mathcal{M})\big)}{\#\mathcal{M}} \geq \mu \eta_{\mathcal{T}}^2\big(\operatorname{tail}_{\mathcal{T}}(E)\big) \quad \text{für alle } E \in \mathcal{E}(\mathcal{T}),$$

was eine Art Mittelwert über die Schweife der markierten Kanten darstellt. Hier darf die linke Seite nicht zu klein werden. Dies kann geschehen, indem Kanten markiert werden, deren Beitrag zum Schätzer entweder zu gering ist, oder deren Schweife sich zu viel überschneiden. In diesem Ausdruck wird nicht die Summe über die Schätzer aller Schweife gebildet, sondern es werden zuerst alle Schweife vereinigt und dann der Schätzer bestimmt, wodurch Beiträge von Kanten, die zu mehr als einem Schweif gehören, nur einmal gezählt werden. Die Markierungsstrategie muss also die Schweife der Kanten berücksichtigen. //

3.1.2 Energie

Lower Diamond Estimate

Wir wollen nun den Begriff Energie, den wir bisher rein intuitiv verwendet haben, präzisieren und Eigenschaften angeben, die eine Energie erfüllen muss, damit im Zuge einer AFEM Energieoptimalität bewiesen werden kann.

Definition 3.3. *Wir nennen eine Abbildung* $\mathcal{G} : \mathbb{T} \to \mathbb{R}$ Energie*, wenn sie bezüglich der Ordnungsrelation auf* $\mathbb{T}$ *monoton fallend ist, also*

$$\mathcal{T} \leq \mathcal{T}' \implies \mathcal{G}(\mathcal{T}') \leq \mathcal{G}(\mathcal{T}) \quad \text{für alle } \mathcal{T}, \mathcal{T}' \in \mathbb{T}.$$

Bemerkung 3.4. Da ein Gitter $\mathcal{T}$ durch seine Menge an Knoten $\mathcal{P}$ eindeutig bestimmt ist und umgekehrt, lässt sich der Energiebegriff auch auf Funktionen $\mathbb{P} \to \mathbb{R}$ übertragen. Hierbei gilt

$$\mathcal{G}(\mathcal{P}) := \mathcal{G}(\mathcal{T}(\mathcal{P})),$$

wobei hier dieselbe Bezeichnung für die Energie gewählt wurde, was aber nicht zu Verwirrungen führen sollte. //

Eine Energie muss in Summe zwei Eigenschaften erfüllen, um im Rahmen einer AFEM Energieoptimalität zu erhalten. Um die erste Eigenschaft formulieren zu können, benötigen wir zuerst noch eine spezielle Struktur von Gitterverfeinerungen [DKS16].

Definition 3.5. *Für eine Menge von Gittern* $\{\mathcal{T}_1, \dots, \mathcal{T}_m\} \subseteq \mathbb{T}$ *und*

$$\mathcal{T}^\wedge := \bigwedge_{j=1}^{m} \mathcal{T}_j \in \operatorname{coarse}(\mathcal{T}_j) \quad \text{für alle } j,$$

$$\mathcal{T}_\vee := \bigvee_{j=1}^{m} \mathcal{T}_j \in \operatorname{refine}(\mathcal{T}_j) \quad \text{für alle } j,$$

nennen wir das Tupel $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_1, \dots, \mathcal{T}_m)$ *einen* Lower Diamond*, falls die Flächen* $\Omega(\mathcal{T}_j \setminus \mathcal{T}_\vee)$ *paarweise disjunkt sind. Analog nennen wir es einen* Upper Diamond*, falls die Flächen* $\Omega(\mathcal{T}^\wedge \setminus \mathcal{T}_j)$ *paarweise disjunkt sind.*

Bemerkung 3.6. Die Flächen $\Omega(\mathcal{T}_j \setminus \mathcal{T}_\vee)$ aus obiger Definition geben jene Flächen an, auf denen sich bei der Verfeinerung von $\mathcal{T}_j$ auf $\mathcal{T}_\vee$ Dreiecke geändert haben. Analoges gilt für die Flächen $\Omega(\mathcal{T}^\wedge \setminus \mathcal{T}_j)$ bei Vergröberung von $\mathcal{T}_j$ auf $\mathcal{T}^\wedge$. Die Lower Diamond Struktur ist schematisch in Abbildung 3.1 dargestellt.

Die Lower Diamond Struktur werden wir später dazu verwenden, Abschätzungen für Energiereduktionen zu erhalten. Die Energien werden später typischerweise über Vektorräume, die mit den Gittern korreliert sind, definiert. Man kann die Lower Diamond Struktur abstrakt auf Vektorräumen $\mathcal{X}_1, \dots, \mathcal{X}_m \subseteq \mathcal{X}$ definieren, mittels $\mathcal{X}^\wedge := \bigcap_i \mathcal{X}_i$ und $\mathcal{X}_\vee := \operatorname{span}(\bigcup_i \mathcal{X}_i)$. Die Disjunktheit der Verfeinerungsflächen korrespondiert hier zur

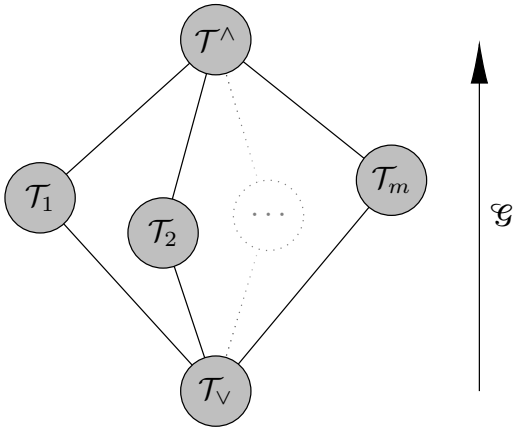


Abbildung 3.1: *Schematische Darstellung eines Lower Diamond aus Definition 3.5. Die Verbindungslinien zwischen den Gittern symbolisieren, dass diese durch die Halbordnung $\leq$ auf $\mathbb{T}$ vergleichbar sein müssen, wobei hier das untere Gitter feiner ist, als das obere.*

Orthogonalität der Quotientenräume $\mathcal{X}_i/\mathcal{X}^\wedge$, die sich mit Unterräumen des Raumes der stückweisen Polynome identifizieren lassen, bezüglich des L^2-Skalarprodukts. Sich hier nicht auf FEM-Diskretisierungen und Gitter zu beziehen führt allerdings eine neue Abstraktionsebene ein, deren Nutzen für unsere Zwecke gering scheint. //

Die folgende Eigenschaft einer Energie besagt, dass die Energiereduktion zwischen den Spitzen eines Lower Diamond äquivalent zur Summe der Energiereduktionen auf dessen unterer Hälfte ist. Anders formuliert kann in einem Lower Diamond, der folgende Eigenschaft erfüllt, von der gesamten Energiereduktion nicht beliebig viel auf einer Hälfte passieren. Ein analoges Resultat, welches wir allerdings im Folgenden nicht benötigen, gilt auch für einen Upper Diamond.

Definition 3.7. *Sei $\mathcal{G} : \mathbb{T} \to \mathbb{R}$ eine Energie. Wir sagen, $\mathcal{G}$ erfüllt die* Lower Diamond Estimate*, wenn es eine Konstante $C_{\mathrm{LD}} > 0$ gibt, die nur von $\mathcal{G}$ und $\mathcal{T}_0$ abhängt, sodass*

$$C_{\mathrm{LD}}^{-1}\big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_\vee)\big) \leq \sum_{j=1}^{m} \big(\mathcal{G}(\mathcal{T}_j) - \mathcal{G}(\mathcal{T}_\vee)\big) \leq C_{\mathrm{LD}}\big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_\vee)\big) \tag{E2}$$

für alle Lower Diamonds $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_1, \dots, \mathcal{T}_m)$ erfüllt ist.

Diskrete lokale Zuverlässigkeit und Effizienz

Die zweite Eigenschaft für Energieoptimalität ist, dass die Energie eine gewisse Beziehung zum Fehlerschätzer haben muss. Die Bedingung, die ein Fehlerschätzer normalerweise erfüllen muss, um eine sinnvolle adaptive FEM zu erhalten, ist Zuverlässigkeit (engl. reliability), also dass der Fehler der FEM Approximation bis auf eine Konstante durch den

Fehlerschätzer beschränkt werden kann. Die umgekehrte Abschätzung heißt Effizienz, da sie verhindert, dass eine bereits exakte Lösung vom Fehlerschätzer nicht als solche erkannt wird [Ver13].

Für konventionelle ratenoptimale AFEM werden diese Abschätzungen nur auf dem gesamten Gitter benötigt. Um Instanzoptimalität zu erhalten, muss der Fehlerschätzer jedoch die Information über den Fehler genauer auflösen. Wir benötigen deshalb obige Abschätzungen erstens für die Differenz zweier diskreter Approximationen und zweitens lokal auf Gittern. Diese nennt man diskrete lokale Zuverlässigkeit, beziehungsweise Effizienz. Wir fordern diese Ungleichungen nun für Energiedifferenzen, da wir diese später mit dem Approximationsfehler in Beziehung setzen wollen.

Definition 3.8. *Seien $\mathcal{G}$ eine Energie und η^2 ein Fehlerschätzer. Wir sagen η^2 ist* diskret lokal äquivalent *zur Energiedifferenz von $\mathcal{G}$, wenn es Konstanten $C_{\mathrm{dle}}, C_{\mathrm{dlr}} > 0$ gibt, sodass*

$$C_{\mathrm{dle}}\,\eta_{\mathcal{T}}^2(\mathcal{E}\setminus\mathcal{E}') \leq \mathcal{G}(\mathcal{T}) - \mathcal{G}(\mathcal{T}') \leq C_{\mathrm{dlr}}\,\eta_{\mathcal{T}}^2(\mathcal{E}\setminus\mathcal{E}'), \tag{E3}$$

für alle Gitter $\mathcal{T}, \mathcal{T}' \in \mathbb{T}$ mit $\mathcal{T}' \in \mathrm{refine}(\mathcal{T})$ und $\mathcal{E} := \mathcal{E}(\mathcal{T})$, sowie $\mathcal{E}' := \mathcal{E}(\mathcal{T}')$.
Die erste Abschätzung heißt dabei diskrete lokale Effizienz*, die zweite heißt diskrete lokale* Zuverlässigkeit*.*

3.2 Vorbereitungen

Wir werden nun einige Hilfsresultate beweisen, die später helfen werden, den Beweis der Energieoptimalität übersichtlicher zu gestalten. Wir werden diesen Beweis in seiner Gesamtheit ohne Populationen formulieren, weshalb hier auch einige Beweisschritte ausgelagert sind, die das Konzept der Populationen verwenden. Dies dient ebenfalls der Übersichtlichkeit und Uniformität des Beweises. Die Resultate sind allesamt [DKS16] entnommen.

Zunächst zeigen wir ein Resultat für freie Kanten, nämlich dass bei der Verfeinerung eines Gitters alle verfeinerten Kanten bereits durch deren freie Kanten erhalten werden können. Dabei wird in [DKS16] nur die Inklusion $\subseteq$ der ersten Gleichheit gezeigt, der Vollständigkeit wegen wird dieses Resultat hier aber ausführlicher gestaltet.

Lemma 3.9. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $\mathcal{T}' \in \mathrm{refine}(\mathcal{T})$. Betrachte die Mengen der Kanten $\mathcal{E} := \mathcal{E}(\mathcal{T})$ und $\mathcal{E}' := \mathcal{E}(\mathcal{T}')$. Dann gilt*

$$\mathcal{E}\setminus\mathcal{E}' \quad = \bigcup_{E\in\mathrm{free}(\mathcal{E}\setminus\mathcal{E}')} \mathrm{tail}_{\mathcal{T}}(E) = \bigcup_{E\in\mathcal{E}\setminus\mathcal{E}'} \mathrm{tail}_{\mathcal{T}}(E). \tag{3.1}$$

Es werden also alle beim Übergang von $\mathcal{T}$ nach $\mathcal{T}'$ verfeinerten Kanten durch die Schweife ihrer (freien) Kanten überdeckt und umgekehrt.

Beweis. Zunächst wollen wir die erste Gleichheit zeigen. Wir verwenden die Identifikation von Kanten und Populationen, um die Werkzeuge aus Lemma 2.18 verwenden zu können.

Dazu setzen wir $\mathcal{C} := \operatorname{midpt}(\mathcal{E} \setminus \mathcal{E}')$, $\mathcal{U} := \operatorname{free}(\mathcal{C})$ und $\mathcal{P} = \mathcal{P}(\mathcal{T})$, sowie $\mathcal{P}' = \mathcal{P}(\mathcal{T}')$. Es gilt mit diesen Bezeichnungen

$$\mathcal{C} = \mathcal{P}' \cap (\mathcal{P}^{++} \setminus \mathcal{P}), \tag{3.2}$$

$$\operatorname{midpt}\left(\bigcup_{E \in \operatorname{free}(\mathcal{E} \setminus \mathcal{E}')} \operatorname{tail}_{\mathcal{T}}(E) \right) = \bigcup_{P \in \mathcal{U}} (\mathcal{P} \oplus \{P\}) \setminus \mathcal{P}. \tag{3.3}$$

Aus Darstellung (3.2) folgt, dass $\mathcal{C}$ sowohl Teilmenge von $\mathcal{P}'$ als auch von $\mathcal{P}^{++}$ ist. Mit der Eigenschaft (v) aus Lemma 2.18, die besagt, dass Populationen ihre eigenen Vorfahren enthalten, gilt

$$\operatorname{anc}(\mathcal{C}) \subseteq \operatorname{anc}(\mathcal{P}^{++}) \subseteq \mathcal{P}^{++}. \tag{3.4}$$

Weiter erhält man mit ebenso elementaren Überlegungen

$$\begin{aligned} \operatorname{anc}(\mathcal{C}) \setminus \mathcal{P} &\overset{(3.4)}{=} (\operatorname{anc}(\mathcal{C}) \cap \mathcal{P}^{++}) \setminus \mathcal{P} \\ &\overset{(3.2)}{\subseteq} \operatorname{anc}(\mathcal{P}') \cap (\mathcal{P}^{++} \setminus \mathcal{P}) \\ &\subseteq \mathcal{P}' \cap (\mathcal{P}^{++} \setminus \mathcal{P}) \overset{(3.2)}{=} \mathcal{C}. \end{aligned} \tag{3.5}$$

Somit enthält $\mathcal{C}$ bereits alle Vorfahren, die nicht in $\mathcal{P}$ enthalten sind. Als nächstes zeigen wir noch

$$\begin{aligned} \mathcal{C} \setminus \mathcal{P} &\overset{(3.5)}{=} (\mathcal{C} \cup \operatorname{anc}(\mathcal{C})) \setminus \mathcal{P} \\ &\overset{2.18(i)}{=} (\operatorname{free}(\mathcal{C}) \cup \operatorname{anc}(\operatorname{free}(\mathcal{C}))) \setminus \mathcal{P} \\ &= (\mathcal{U} \cup \operatorname{anc}(\mathcal{U})) \setminus \mathcal{P}. \end{aligned}$$

Damit lässt sich nun die erste Gleichheit zeigen, indem man den Ausdruck (3.3) noch weiter umformt:

$$\begin{aligned} \bigcup_{P \in \mathcal{U}} (\mathcal{P} \oplus \{P\}) \setminus \mathcal{P} &\overset{2.18(ii)}{=} \bigcup_{P \in \mathcal{U}} (\mathcal{P} \cup \{P\} \cup \operatorname{anc}(\{P\})) \setminus \mathcal{P} \\ &= (\mathcal{P} \cup \mathcal{U} \cup \operatorname{anc}(\mathcal{U})) \setminus \mathcal{P} \\ &= (\mathcal{U} \cup \operatorname{anc}(\mathcal{U})) \setminus \mathcal{P} \\ &\overset{(3.5)}{=} \mathcal{C} \setminus \mathcal{P}. \end{aligned}$$

Nun müssen wir den Ausdruck $\mathcal{C} \setminus \mathcal{P}$ nur noch zurückübersetzen. Da $\operatorname{midpt}(\mathcal{E}) \cap \mathcal{P} = \emptyset$, gilt $\mathcal{C} \setminus \mathcal{P} = \mathcal{C} = \operatorname{midpt}(\mathcal{E} \setminus \mathcal{E}')$. Insgesamt erhalten wir

$$\operatorname{midpt}(\mathcal{E} \setminus \mathcal{E}') = \mathcal{C} \setminus \mathcal{P} = \bigcup_{P \in \mathcal{U}} (\mathcal{P} \oplus \{P\}) \setminus \mathcal{P} = \operatorname{midpt}\left(\bigcup_{E \in \operatorname{free}(\mathcal{E} \setminus \mathcal{E}')} \operatorname{tail}_{\mathcal{T}}(E) \right). \tag{3.6}$$

Mit der Bijektivität von midpt folgt die erste Gleichheit in (3.1).

Der zweite Teil ist mit obigen Überlegungen nun nicht mehr schwer zu zeigen und sehr ähnlich dem ersten:

$$\begin{aligned}\bigcup_{P\in\mathcal{C}}(\mathcal{P}\oplus\{P\})\setminus\mathcal{P} &\overset{2.18(ii)}{=}\bigcup_{P\in\mathcal{C}}(\mathcal{P}\cup\{P\}\cup\operatorname{anc}(\{P\}))\setminus\mathcal{P}\\ &=(\mathcal{P}\cup\mathcal{C}\cup\operatorname{anc}(\mathcal{C}))\setminus\mathcal{P}\\ &=(\mathcal{C}\cup\operatorname{anc}(\mathcal{C}))\setminus\mathcal{P}\\ &\overset{(3.5)}{=}\mathcal{C}\setminus\mathcal{P}.\end{aligned}$$

Über die Identifikation von Kanten und Populationen über die Abbildung midpt folgt nun, wie im ersten Fall, die zweite Gleichheit in (3.1). □

Das nächste Lemma zeigt, dass ein Lower Diamond schon auf sehr einfache Weise durch die Verbandsoperationen $\wedge$ und $\vee$ gebildet werden kann.

Lemma 3.10. *Seien $\mathcal{T}_1,\mathcal{T}_2\in\mathbb{T}$ mit $\mathcal{T}_1\neq\mathcal{T}_2$, $\mathcal{T}_\vee:=\mathcal{T}_1\vee\mathcal{T}_2$ und $\mathcal{T}^\wedge:=\mathcal{T}_1\wedge\mathcal{T}_2$. Dann ist $(\mathcal{T}^\wedge,\mathcal{T}_\vee;\mathcal{T}_1,\mathcal{T}_2)$ ein Lower Diamond.*

Beweis. Wir müssen zeigen, dass die Verfeinerungsflächen disjunkt sind. Angenommen die Flächen $\Omega(\mathcal{T}_1\setminus\mathcal{T}_\vee)$ und $\Omega(\mathcal{T}_2\setminus\mathcal{T}_\vee)$ wären nicht disjunkt. Es gilt $\Omega(\mathcal{T}_1\setminus\mathcal{T}_\vee)=\Omega(\mathcal{T}_\vee\setminus\mathcal{T}_1)$, da jedes zu verfeinernde Dreieck von den daraus entstehenden verfeinerten Dreiecken überdeckt wird und umgekehrt. Deshalb muss es mindestens ein Dreieck T in der Menge

$$(\mathcal{T}_\vee\setminus\mathcal{T}_1)\cap(\mathcal{T}_\vee\setminus\mathcal{T}_2)=\mathcal{T}_\vee\setminus(\mathcal{T}_1\cup\mathcal{T}_2)$$

geben. Andererseits muss nach Definition von $\mathcal{T}_\vee$ aber auch gelten, dass $\mathcal{T}_\vee\subseteq\mathcal{T}_1\cup\mathcal{T}_2$, wie in Bemerkung 2.8 ausgeführt wurde. Damit ist obige Menge leer, was einen Widerspruch zu unserer Annahme darstellt. □

Das nächste Resultat ist etwas technisch und verwendet intensiv den Populationsbegriff und Eigenschaften von Populationen. Auch wenn es sich für Kanten formulieren lässt, was wegen der Beziehung zwischen Gittern und Populationen immer möglich ist, wirkt die Formulierung mit Populationen natürlicher. Da wir aber den Beweis der Energieoptimalität ohne Populationen formulieren und beweisen wollen, werden wir auch das folgende Resultat in diesen Rahmen einbetten.

Lemma 3.11. *Sei $\mathcal{G}$ eine Energie, die die Lower Diamond Estimate erfüllt, $\mathcal{T}^\star\in\mathbb{T}$ ein Gitter, $\mathcal{T}_\star\in\operatorname{refine}(\mathcal{T}^\star)$ und*

$$\mathcal{C}:=\operatorname{free}\left(\bigcup_{\mathcal{T}^\star\leq\mathcal{T}\leq\mathcal{T}_\star}\mathcal{E}(\mathcal{T})\setminus\mathcal{E}(\mathcal{T}_\star)\right)$$

die freie Teilmenge aller beim Übergang von $\mathcal{T}^\star$ zu $\mathcal{T}_\star$ verfeinerten Kanten. Seien außerdem $N,\alpha\in\mathbb{N}$ mit $\alpha N\leq\#\mathcal{C}$ gegeben und $\mathcal{C}_1,\dots,\mathcal{C}_N\subseteq\mathcal{C}$ paarweise disjunkte Teilmengen mit

$\#\mathcal{C}_i \geq \alpha$ für alle $i \in \{1, \dots, N\}$ und $\bigcup_{i=1}^N \mathcal{C}_i = \mathcal{C}$. Dann gilt mit $\mathcal{T}_i := \operatorname{coarse}(\mathcal{T}_\star; \mathcal{C}_i)$, dass

$$\mathcal{G}(\mathcal{T}^\star) - \mathcal{G}(\mathcal{T}_\star) \geq \min\{1, C_{\mathrm{LD}}^{-1}\} \sum_{i=1}^N \big(\mathcal{G}(\mathcal{T}_i) - \mathcal{G}(\mathcal{T}_\star)\big). \tag{3.7}$$

Beweis. Wir überlegen uns zuerst die zu $\mathcal{C}$ korrespondierende Menge von Personen. Wir setzen wie gewohnt $\mathcal{P}_\star := \mathcal{P}(\mathcal{T}_\star)$, $\mathcal{P}^\star := \mathcal{P}(\mathcal{T}^\star)$ und $\mathcal{P}_i := \mathcal{P}(\mathcal{T}_i)$, außerdem sei $\widetilde{\mathcal{C}}_i := \operatorname{midpt}(\mathcal{C}_i)$. Da die Menge aller Populationen (anders als die Menge aller Gitter) durch die Mengeninklusion teilgeordnet ist, gilt für alle $\mathcal{T}^\star \leq \mathcal{T} \leq \mathcal{T}_\star$

$$\operatorname{midpt}\big(\mathcal{E}(\mathcal{T}) \setminus \mathcal{E}(\mathcal{T}_\star)\big) = \mathcal{P}_\star \cap \big(\mathcal{P}(\mathcal{T})^{++} \setminus \mathcal{P}(\mathcal{T})\big) \subseteq \mathcal{P}_\star \setminus \mathcal{P}^\star. \tag{3.8}$$

Ist umgekehrt ein Knoten $P \in \mathcal{P}_\star \setminus \mathcal{P}^\star$ gegeben, so muss dieser durch Bisektion einer Kante von $\mathcal{T}^\star$ oder einem Gitter aus $\operatorname{refine}(\mathcal{T}^\star)$, das gröber als $\mathcal{T}_\star$ ist, hervorgegangen sein. Es gilt somit

$$\mathcal{P}_\star \setminus \mathcal{P}^\star \subseteq \operatorname{midpt}\left(\bigcup_{\mathcal{T}^\star \leq \mathcal{T} \leq \mathcal{T}_\star} \mathcal{E}(\mathcal{T}) \setminus \mathcal{E}(\mathcal{T}_\star)\right).$$

Insgesamt haben wir die korrespondierende Menge von Personen gefunden:

$$\widetilde{\mathcal{C}} := \operatorname{midpt}(\mathcal{C}) = \operatorname{free}\big(\mathcal{P}_\star \setminus \mathcal{P}^\star\big). \tag{3.9}$$

Wir führen nun eine Fallunterscheidung nach der Anzahl der freien Teilmengen $\mathcal{C}_i$ durch.

Fall 1: $N = 1$.
Da hier $\mathcal{T}_1 \geq \mathcal{T}^\star$ gilt, folgt die Behauptung in diesem Fall trivialerweise aus der Monotonie der Energie:

$$\mathcal{G}(\mathcal{T}^\star) - \mathcal{G}(\mathcal{T}_\star) \geq \mathcal{G}(\mathcal{T}_1) - \mathcal{G}(\mathcal{T}_\star).$$

Fall 2: $N > 1$.
Wir wollen die Lower Diamond Estimate auf einen Lower Diamond anwenden, der zwischen $\mathcal{P}^\star$ und $\mathcal{P}_\star$ liegt. Dafür sehen wir uns die Mengen $\mathcal{P}_i$ genauer an. Die Mengen $\widetilde{\mathcal{C}}_i$ sind Teilmengen der freien Knoten von $\mathcal{P}_\star \setminus \mathcal{P}^\star$. Nach Definition von $\operatorname{coarse}(\mathcal{T}_\star; \mathcal{C}_i)$ und der Eigenschaft bei Vergröberungen bezüglich freier Teilmengen aus Lemma 2.18 (iii) gilt somit

$$\mathcal{P}_i = \mathcal{P}_\star \ominus \widetilde{\mathcal{C}}_i = \mathcal{P}_\star \setminus \widetilde{\mathcal{C}}_i. \tag{3.10}$$

Die beiden Enden des Lower Diamonds werden durch die gröbste gemeinsame Verfeinerung und die feinste gemeinsame Vergröberung der $\mathcal{P}_i$ gebildet. Diese lassen sich mit $\mathcal{P}_\star$ und $\mathcal{P}^\star$ in Verbindung bringen:

$$\bigcap_{i=1}^N \mathcal{P}_i \supseteq \mathcal{P}^\star, \quad \bigcup_{i=1}^N \mathcal{P}_i = \mathcal{P}_\star,$$

wie aus der Definition der $\mathcal{P}_i$ ersichtlich ist. Des Weiteren gilt wegen Gleichung (3.10)

$$\mathcal{P}_\star \setminus \mathcal{P}_i = \widetilde{\mathcal{C}}_i,$$

also sind diese Mengen disjunkt nach Voraussetzung. Damit ist $(\bigcap_{i=1}^N \mathcal{P}_i, \bigcup_{i=1}^N \mathcal{P}_i;$ $\mathcal{P}_1, \dots, \mathcal{P}_N)$ ein Lower Diamond und es gilt die Lower Diamond Estimate. Gemeinsam mit der Monotonie der Energie und der Korrespondenz von Gittern und Populationen ergibt sich

$$\begin{aligned}\mathcal{G}(\mathcal{T}^\star) - \mathcal{G}(\mathcal{T}_\star) &\geq \mathcal{G}\left(\mathcal{T}\left(\bigcap_{i=1}^N \mathcal{P}_i\right)\right) - \mathcal{G}(\mathcal{T}_\star) \\ &\geq C_{\mathrm{LD}}^{-1} \sum_{i=1}^N \big(\mathcal{G}(\mathcal{T}_i) - \mathcal{G}(\mathcal{T}_\star)\big).\end{aligned}$$

Fasst man die zwei eben betrachteten Fälle zusammen, erhält man die Behauptung. □

Zu guter Letzt wollen wir noch eine Aussage über die Anzahl der verfeinerten Kanten in einem Lower Diamond beweisen.

Lemma 3.12. *Seien $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ mit $\mathcal{T}_1 \neq \mathcal{T}_2$ gegeben. Seien weiters $\mathcal{T}^\wedge := \mathcal{T}_1 \wedge \mathcal{T}_2$ und $\mathcal{T}_\vee := \mathcal{T}_1 \vee \mathcal{T}_2$ definiert. Für die Menge*

$$\mathcal{C} := \bigcup_{\mathcal{T}^\wedge \leq \mathcal{T} \leq \mathcal{T}_2} \mathcal{E}(\mathcal{T}) \setminus \mathcal{E}(\mathcal{T}_\vee)$$

aller verfeinerten Kanten beim Übergang von $\mathcal{T}^\wedge$ nach $\mathcal{T}_2$ und die Menge

$$\mathcal{U} := \mathcal{E}(\mathcal{T}_1) \setminus \mathcal{E}(\mathcal{T}_\vee)$$

aller verfeinerten Kanten aus $\mathcal{T}_1$ beim Übergang zu $\mathcal{T}_\vee$ gilt:

$$\#\,\mathrm{free}(\mathcal{U}) \leq C_{\mathrm{GD}} \#\,\mathrm{free}(\mathcal{C}).$$

Beweis. Es seien wiederum mit $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}^\wedge$ und $\mathcal{P}_\vee$ die korrespondierenden Populationen bezeichnet. Wie wir bereits im Beweis von Lemma 3.11 gesehen haben, gilt

$$\begin{aligned}\widetilde{\mathcal{U}} &:= \mathrm{midpt}(\mathcal{U}) \overset{(3.8)}{=} \mathcal{P}_\vee \cap \big(\mathcal{P}_1^{++} \setminus \mathcal{P}_1\big), \\ \widetilde{\mathcal{C}} &:= \mathrm{midpt}(\mathcal{C}) \overset{(3.9)}{=} \mathcal{P}_2 \setminus \mathcal{P}^\wedge.\end{aligned} \tag{3.11}$$

Mit den De Morganschen Gesetzen gilt

$$\mathcal{P}_\vee \setminus \mathcal{P}_1 = (\mathcal{P}_1 \cup \mathcal{P}_2) \setminus \mathcal{P}_1 = \mathcal{P}_2 \setminus (\mathcal{P}_1 \cap \mathcal{P}_2) = \mathcal{P}_2 \setminus \mathcal{P}^\wedge.$$

Wegen (3.11) gilt die Inklusion $\widetilde{\mathcal{U}} \subseteq \mathcal{P}_\vee \setminus \mathcal{P}_1$ und mit Lemma 2.18 (iv)

$$\begin{aligned}\#\operatorname{free}(\widetilde{\mathcal{U}}) &\leq C_{\mathrm{GD}} \#\operatorname{free}(\mathcal{P}_\vee \setminus \mathcal{P}_1)\\ &= C_{\mathrm{GD}} \#\operatorname{free}(\widetilde{\mathcal{C}}).\end{aligned}$$

Da die Abbildung midpt bijektiv ist, lässt sie insbesondere die Mächtigkeit einer abgebildeten Menge invariant, somit folgt die Behauptung. □

Das letzte Resultat verwendet sehr stark die Konzepte der Populationen und deren Generation. Diese Konzepte stützen sich intensiv auf die verwendete Verfeinerungsregel (die hier NVB ist). Für andere Verfeinerungsregeln ist aktuell noch nicht klar, ob und wie diese Resultate dort formuliert und bewiesen werden können.

3.3 Energieoptimalität

Wir können nun das zentrale Resultat dieses Kapitels beweisen. Zur Erinnerung: Wir nehmen in der gesamten Arbeit an, dass das initiale Gitter $\mathcal{T}_0$ regulär ist, die initiale Verteilung der Verfeinerungskanten der BDD-Bedingung aus Defintion 2.3 genügt und die Verfeinerung mittels kantenbasierter NVB durchgeführt wird. Unter den gestellten Voraussetzungen an Energie, Fehlerschätzer und Markierungsstrategie aus den vorherigen Abschnitten können wir zeigen, dass die daraus resultierende AFEM instanzoptimal bezüglich der Energie ist. Wir benötigen dafür noch einen weiteren Begriff.

Definition 3.13. *Sei $\mathcal{G}$ eine Energie. Wir definieren die* optimale Energie *mit höchstens $m \in \mathbb{N}$ neuen Dreiecken als*

$$\mathcal{G}_m^{\mathrm{opt}} := \min\left\{\mathcal{G}(\mathcal{T}) \mid \mathcal{T} \in \mathbb{T}, \#(\mathcal{T} \setminus \mathcal{T}_0) \leq m\right\}.$$

Ein Gitter $\mathcal{T}_m^{\mathrm{opt}} \in \{\mathcal{T} \in \mathbb{T} \mid \#(\mathcal{T} \setminus \mathcal{T}_0) \leq m\}$, das $\mathcal{G}_m^{\mathrm{opt}} = \mathcal{G}(\mathcal{T}_m^{\mathrm{opt}})$ erfüllt, nennen wir ein zugehöriges optimales Gitter.

Die optimale Energie ist also die kleinste Energie, die auf Gittern angenommen werden kann, die höchstens m mehr Elemente als $\mathcal{T}_0$ enthalten. Optimale Gitter für eine gegebene Zahl m müssen nicht eindeutig sein. Da aber die folgenden Resultate nicht davon abhängen, welches der optimalen Gitter gewählt wurde, werden wir diese Feinheit im Folgenden übergehen und für ein gegebenes m ein beliebiges, aber festes, zu $\mathcal{G}_m^{\mathrm{opt}}$ gehöriges optimales Gitter auswählen und dieses als *das* optimale Gitter $\mathcal{T}_m^{\mathrm{opt}}$ bezeichnen. Es sei außerdem angemerkt, dass die optimalen Energien natürlich in m fallend sind, die optimalen Gitter jedoch nicht durch die Halbordnung $\leq$ auf $\mathbb{T}$ vergleichbar sein müssen.

Das Hauptresultat dieses Kapitels kann nun formuliert werden.

Satz 3.14 (Energieoptimalität)**.** *Sei $\mathcal{G}$ eine Energie, die die Lower Diamond Estimate (E2) erfüllt und η^2 ein Fehlerschätzer, sodass diskrete lokale Zuverlässigkeit und Effizienz aus Gleichung (E3) gelten. Seien außerdem $(\mathcal{T}_k)_{k\in\mathbb{N}} \subseteq \mathbb{T}$ eine Menge von Gittern, für die $\mathcal{T}_{k+1} = \operatorname{refine}(\mathcal{T}_k; \mathcal{M}_k)$ für alle $k \in \{1, 2, \ldots\}$ gilt, wobei die Mengen der markierten*

Kanten $\mathcal{M}_k \subseteq \mathcal{E}(\mathcal{T}_k)$ die Bedingung (E1) erfüllen.
Dann existiert $C_{\mathrm{mesh}} \geq 1$ mit $C_{\mathrm{mesh}} \in \mathcal{O}(\mu^{-2})$, sodass für die Energie des k-ten Gitters

$$\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}_m^{\mathrm{opt}} \quad \text{für alle } m \text{ mit } \#(\mathcal{T}_k \setminus \mathcal{T}_0) \geq C_{\mathrm{mesh}} m.$$

Die Konstante C_{mesh} hängt nur von C_{LD}, C_{dle}, C_{dlr}, C_{GD}, C_{MC} und μ ab.

Dieser Satz besagt, dass die Energie eines Gitters aus unserer AFEM kleiner ist als eine gewisse optimale Energie, solange bis dahin nur genügend Elemente verfeinert wurden. Daher stammt der Name Energieoptimalität.

Wir werden den Beweis dieses Satzes in verschiedene Teilresultate gliedern, die wir im Folgenden der Übersichtlichkeit halber als Lemmata formulieren. Diese stammen aus [DKS16], die Beweise sind an [Hab] angelehnt. Das erste Lemma hiervon ist etwas technisch und liefert eine Abschätzung für die Energiedifferenz zweier Schritte des adaptiven Verfahrens. Die Voraussetzungen sind hierbei so gewählt, dass dieses Hilfsresultat in den nächsten zwei Lemmata in leicht unterschiedlicher Form angewendet werden kann.

Lemma 3.15. *Es gelten die Voraussetzungen von Satz 3.14. Sei darüber hinaus $\mathcal{T} \in \mathbb{T}$ ein Gitter mit $\mathcal{T} \neq \mathcal{T}_k$, $\mathcal{T}^\wedge := \mathcal{T}_k \wedge \mathcal{T}$, $\mathcal{T}_\vee := \mathcal{T}_k \vee \mathcal{T}$ und $\mathcal{U} := \operatorname{free}\big(\mathcal{E}(\mathcal{T}_k) \setminus \mathcal{E}(\mathcal{T}_\vee)\big)$. Erfüllt das Gitter $\mathcal{T}$ die Bedingung*

$$\mathcal{G}(\mathcal{T}_k) \geq \mathcal{G}(\mathcal{T}) \geq \mathcal{G}(\mathcal{T}_\vee), \tag{3.12}$$

so gibt es eine Konstante $C_{\mathrm{opt}} = C_{\mathrm{LD}}^{-1} \frac{C_{\mathrm{dle}}}{C_{\mathrm{dlr}}} > 0$, sodass

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T})\big).$$

Beweis. Wir bezeichnen mit $\mathcal{E}_k := \mathcal{E}(\mathcal{T}_k)$ die Menge der Kanten des k-ten Gitters und setzen $\eta_k^2 := \eta_{\mathcal{T}_k}^2$. Nach Definition ist $\mathcal{E}_k \setminus \mathcal{E}_{k+1} = \operatorname{tail}_{\mathcal{T}_k}(\mathcal{M}_k) =: \operatorname{tail}_k(\mathcal{M}_k)$. Wegen der diskreten lokalen Effizienz aus Gleichung (E3) und der Voraussetzung an die Menge der markierten Kanten (E1) gilt daher im k-ten Schritt des adaptiven Verfahrens für jede Kante $E \in \mathcal{E}_k$

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq C_{\mathrm{dle}}\, \eta_k^2(\mathcal{E}_k \setminus \mathcal{E}_{k+1}) \geq \mu C_{\mathrm{dle}}\, \#\mathcal{M}_k \eta_k^2\big(\operatorname{tail}_k(E)\big).$$

Summiert man diese Ungleichung über alle Kanten aus $\mathcal{U}$ erhält man

$$\#\mathcal{U}\big(\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1})\big) \geq \mu C_{\mathrm{dle}}\, \#\mathcal{M}_k \sum_{E \in \mathcal{U}} \eta_k^2\big(\operatorname{tail}_k(E)\big).$$

Die Summe über die Schätzerbeiträge der Schweife lässt sich wegen der Positivität des Schätzers nach unten durch den Schätzer der Vereinigung aller Schweife abschätzen. Dies und Umformen der Ungleichung führt mit Lemma 3.9 auf

$$\begin{aligned} \mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) &\geq \mu C_{\mathrm{dle}} \frac{\#\mathcal{M}_k}{\#\mathcal{U}}\, \eta_k^2 \left(\bigcup_{E \in \mathcal{U}} \operatorname{tail}_k(E) \right) \\ &\overset{(3.1)}{=} \mu C_{\mathrm{dle}} \frac{\#\mathcal{M}_k}{\#\mathcal{U}}\, \eta_k^2 \left(\mathcal{E}_k \setminus \mathcal{E}_\vee \right). \end{aligned}$$

Mit der diskreten lokalen Zuverlässigkeit aus Gleichung (E3) folgt

$$\begin{aligned}\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) &\geq \mu C_{\mathrm{dle}} \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \eta_k^2\big(\mathcal{E}(\mathcal{T}_k) \setminus \mathcal{E}(\mathcal{T}_\vee)\big) \\ &\geq \mu \frac{C_{\mathrm{dle}}}{C_{\mathrm{dlr}}} \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_\vee)\big).\end{aligned}$$

Wir wollen nun noch den letzten Ausdruck umformen. Nach Lemma 3.10 ist $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_k, \mathcal{T})$ ein Lower Diamond. Mit der Voraussetzung (3.12) und der Lower Diamond Estimate erhält man für die Differenz im letzten Ausdruck

$$\begin{aligned}\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_\vee) &= \tfrac{1}{2}\big(\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_\vee) + \mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_\vee)\big) \\ &\overset{(3.12)}{\geq} \tfrac{1}{2}\big(\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_\vee) + \mathcal{G}(\mathcal{T}) - \mathcal{G}(\mathcal{T}_\vee)\big) \\ &\overset{\mathrm{LDE}}{\geq} C_{\mathrm{LD}}^{-1}\big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_\vee)\big) \\ &\overset{(3.12)}{\geq} C_{\mathrm{LD}}^{-1}\big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T})\big).\end{aligned}$$

Setzen wir nun alle Schritte zusammen, folgt

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq \mu C_{\mathrm{LD}}^{-1} \frac{C_{\mathrm{dle}}}{C_{\mathrm{dlr}}} \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T})\big)$$

und schließlich mit $C_{\mathrm{opt}} := C_{\mathrm{LD}}^{-1} \frac{C_{\mathrm{dle}}}{C_{\mathrm{dlr}}}$ die Aussage. □

Das nächste Hilfsresultat zeigt, dass in einem Schritt pro verfeinerter Kanten zumindest ein fester Bruchteil des Weges zur nächsten optimalen Energie gemacht wird.

Lemma 3.16. *Es seien die Voraussetzungen von Satz 3.14 erfüllt. Dann existiert eine Konstante $C'_{\mathrm{opt}} = \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} C_{\mathrm{GD}}^{-1} > 0$, sodass die folgende Aussage gilt:*
Wenn für $k, m \in N_0$ gilt, dass $\mathcal{G}(\mathcal{T}_k) \in \big(\mathcal{G}_{m+1}^{\mathrm{opt}}, \mathcal{G}_m^{\mathrm{opt}}\big]$, so gilt für die Energiereduktion beim nächsten Schritt des adaptiven Verfahrens

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq C'_{\mathrm{opt}} \mu \#\mathcal{M}_k \big(\mathcal{G}_m^{\mathrm{opt}} - \mathcal{G}_{m+1}^{\mathrm{opt}}\big). \tag{3.13}$$

Beweis. Wir wählen k, m so, dass $\mathcal{G}(\mathcal{T}_k) \in \big(\mathcal{G}_{m+1}^{\mathrm{opt}}, \mathcal{G}_m^{\mathrm{opt}}\big]$ gilt. Zusätzlich setzen wir $\mathcal{T}_\vee := \mathcal{T}_k \vee \mathcal{T}_{m+1}^{\mathrm{opt}}$ und $\mathcal{T}^\wedge := \mathcal{T}_k \wedge \mathcal{T}_{m+1}^{\mathrm{opt}}$. Da somit $\mathcal{T}_\vee \leq \mathcal{T}_{m+1}^{\mathrm{opt}}$ gilt, gilt auch

$$\mathcal{G}(\mathcal{T}_k) > \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) = \mathcal{G}_{m+1}^{\mathrm{opt}} \geq \mathcal{G}(\mathcal{T}_\vee)$$

und insbesondere $\mathcal{T}_k \neq \mathcal{T}_{m+1}^{\mathrm{opt}}$. Damit lässt sich Lemma 3.15 anwenden und wir erhalten mit $\mathcal{U} = \mathrm{free}\big(\mathcal{E}(\mathcal{T}_k) \setminus \mathcal{E}(\mathcal{T}_\vee)\big)$ die Abschätzung

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}})\big). \tag{3.14}$$

Wir wollen uns die Differenz $\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}})$ nun genauer ansehen. Definiere

$$\mathcal{C} := \operatorname{free}\left(\bigcup_{\mathcal{T}^\wedge \leq \mathcal{T} \leq \mathcal{T}_{m+1}^{\mathrm{opt}}} \mathcal{E}(\mathcal{T}) \setminus \left(\mathcal{E}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right) \right).$$

Da $\mathcal{C}$ endlich ist, sei $\mathcal{C} = \{C_1, \dots, C_N\}$. Damit definieren wir $\mathcal{T}_i' := \operatorname{coarse}(\mathcal{T}_{m+1}^{\mathrm{opt}}; C_i)$. Lemma 3.11 gibt uns mit $\mathcal{T}^* = \mathcal{T}^\wedge$ und $\mathcal{T}_* = \mathcal{T}_{m+1}^{\mathrm{opt}}$ die Abschätzung

$$\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \geq \min\{1, C_{\mathrm{LD}}^{-1}\} \sum_{i=1}^{N} \left(\mathcal{G}(\mathcal{T}_i') - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right). \tag{3.15}$$

Da sich für jede vergröberte Kante die Menge der Dreiecke in einem Gitter um mindestens eins verringert, gilt $\#(\mathcal{T}_i' \setminus \mathcal{T}_0) \leq \#(\mathcal{T}_{m+1}^{\mathrm{opt}} \setminus \mathcal{T}_0) - 1 \leq m$. Damit gilt nach Definition der optimalen Gitter $\mathcal{G}(\mathcal{T}_i') \geq \mathcal{G}_m^{\mathrm{opt}}$. Mit (3.14) und (3.15) erhalten wir also

$$\begin{aligned}
\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) &\geq C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \left(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right) \\
&\geq \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \sum_{i=1}^{N} \left(\mathcal{G}(\mathcal{T}_i') - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right) \\
&\geq \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \sum_{i=1}^{N} \left(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right) \\
&= \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} \mu \frac{\#\mathcal{M}_k \#\mathcal{C}}{\#\mathcal{U}} \left(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right).
\end{aligned}$$

Es muss nun noch das Verhältnis $\#\mathcal{C}/\#\mathcal{U}$ abgeschätzt werden. Erinnern wir uns zunächst daran, wie die beiden Mengen definiert sind. Die Menge $\mathcal{C}$ enthält die freien Kanten, die beim Übergang von $\mathcal{T}^\wedge$ nach $\mathcal{T}_{m+1}^{\mathrm{opt}}$ verfeinert werden. Die Menge $\mathcal{U}$ beinhaltet die freien Kanten aus $\mathcal{T}_k$, die verfeinert werden müssen, will man zu $\mathcal{T}_\vee$ gelangen. Außerdem sind $\mathcal{T}^\wedge$ und $\mathcal{T}_\vee$ als gemeinsame Vergröberung, beziehungsweise Verfeinerung von $\mathcal{T}_k$ und $\mathcal{T}_{m+1}^{\mathrm{opt}}$ definiert, sodass wir Lemma 3.12 benutzen können. Dieses sagt uns, da $\mathcal{U}$ und $\mathcal{C}$ nur freie Kanten beinhalten und somit $\mathcal{U} = \operatorname{free}(\mathcal{U})$ und $\mathcal{C} = \operatorname{free}(\mathcal{C})$ gilt, dass $\#\mathcal{U} \leq C_{\mathrm{GD}} \#\mathcal{C}$ ist. Wir erhalten also abschließend

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} \mu \#\mathcal{M}_k C_{\mathrm{GD}}^{-1} \left(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+1}^{\mathrm{opt}}) \right), \tag{3.16}$$

was mit $C_{\mathrm{opt}}' := \min\{1, C_{\mathrm{LD}}^{-1}\} C_{\mathrm{opt}} C_{\mathrm{GD}}^{-1}$ die Aussage beweist. Diese Konstante hängt nur vom initialen Gitter $\mathcal{T}_0$ und der Struktur der Energie $\mathcal{G}$ ab. □

Das letzte Lemma zeigt, dass bei Verfeinerung hinreichend vieler Kanten in einem Schritt die optimale Energie schon um eine ganze oder sogar mehrere 'Stufen' absteigt.

Lemma 3.17. *Es gelten die Voraussetzungen von Satz 3.14. Dann existiert eine Konstante $C_{\mathrm{opt}}'' \in \mathcal{O}(\mu^{-1})$ mit $C_{\mathrm{opt}}'' \geq 1$, sodass folgende Aussage gilt:*

Wenn für $k, m \in N_0$ gilt, dass $\mathcal{G}(\mathcal{T}_k) \in \left(\mathcal{G}^{\text{opt}}_{m+1}, \mathcal{G}^{\text{opt}}_m\right]$, so gilt für die Energie im nächsten Schritt mit $\alpha := \lfloor \#\mathcal{M}_k / C''_{\text{opt}} \rfloor$:

$$\mathcal{G}(\mathcal{T}_{k+1}) \leq \mathcal{G}^{\text{opt}}_{m+\alpha}.$$

Beweis. Wir gehen ähnlich vor wie im Beweis von Lemma 3.16. Seien dafür k, m so, dass $\mathcal{G}(\mathcal{T}_k) \in \left(\mathcal{G}^{\text{opt}}_{m+1}, \mathcal{G}^{\text{opt}}_m\right]$ gilt. Wir setzen nun die Konstante C''_{opt} fest als

$$C''_{\text{opt}} := \max\left\{1\,,\, \frac{C_{\text{GD}}}{C_{\text{opt}}\mu}\,,\, \frac{2C_{\text{GD}}}{C_{\text{opt}}\mu \min\{1, C_{\text{LD}}^{-1}\}}\right\}.$$

Des Weiteren definieren wir $\mathcal{T}_\vee := \mathcal{T}_k \vee \mathcal{T}^{\text{opt}}_{m+\alpha}$ und $\mathcal{T}^\wedge := \mathcal{T}_k \wedge \mathcal{T}^{\text{opt}}_{m+\alpha}$.

Für den Fall $\alpha = 0$ ist die Aussage trivialerweise erfüllt wegen der Monotonie der Energie:

$$\mathcal{G}(\mathcal{T}_{k+1}) \leq \mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}^{\text{opt}}_m = \mathcal{G}^{\text{opt}}_{m+\alpha},$$

wobei in der zweiten Ungleichung eingegangen ist, dass wir $\mathcal{G}(\mathcal{T}_k)$ im Intervall $\left(\mathcal{G}^{\text{opt}}_{m+1}, \mathcal{G}^{\text{opt}}_m\right]$ voraussetzen. Wir widmen uns im Folgenden dem interessanteren Fall $\alpha \geq 1$.

Wegen $\alpha \geq 1$ und der Monotonie der Energie ist $\mathcal{G}(\mathcal{T}_k) > \mathcal{G}^{\text{opt}}_{m+1} \geq \mathcal{G}^{\text{opt}}_{m+\alpha} = \mathcal{G}(\mathcal{T}^{\text{opt}}_{m+\alpha})$, weshalb $\mathcal{T}_k \neq \mathcal{T}^{\text{opt}}_{m+\alpha}$ gilt. Mit der Menge $\mathcal{U} = \text{free}\left(\mathcal{E}(\mathcal{T}_k) \setminus \mathcal{E}(\mathcal{T}_\vee)\right)$ können wir Lemma 3.15 anwenden und erhalten

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq C_{\text{opt}}\mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \left(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+\alpha})\right). \tag{3.17}$$

Um die Mächtigkeit dieser Menge abzuschätzen, wenden wir wiederum Lemma 3.12 auf die Mengen $\mathcal{U}$ und $\mathcal{C} := \text{free}\left(\bigcup_{\mathcal{T}^\wedge \leq \mathcal{T} \leq \mathcal{T}^{\text{opt}}_{m+\alpha}} \mathcal{E}(\mathcal{T}) \setminus \mathcal{E}(\mathcal{T}_\vee)\right)$ an, wobei wir beachten, dass $\mathcal{U} = \text{free}(\mathcal{U})$ und $\mathcal{C} = \text{free}(\mathcal{C})$ gilt. Dies gibt uns

$$\#\mathcal{U} \leq C_{\text{GD}}\, \#\mathcal{C}. \tag{3.18}$$

Wir führen nun eine Fallunterscheidung nach $N := \lfloor \#\mathcal{C}/\alpha \rfloor$ durch.

Fall 1: $N = 0$.
Nach Definition von $N = \lfloor \#\mathcal{C}/\alpha \rfloor$ gilt $\#\mathcal{C} < \alpha = \lfloor \#\mathcal{M}_k / C''_{\text{opt}} \rfloor \leq \#\mathcal{M}_k / C''_{\text{opt}}$. Dies führt auf

$$\frac{\#\mathcal{M}_k}{\#\mathcal{U}} \overset{(3.18)}{\geq} \frac{\#\mathcal{M}_k}{C_{\text{GD}} \#\mathcal{C}} > \frac{C''_{\text{opt}}}{C_{\text{GD}}}.$$

Da nach Definition $C''_{\text{opt}} \geq \frac{C_{\text{GD}}}{C_{\text{opt}}\mu}$ gilt, erhalten wir insgesamt

$$\begin{aligned}
\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) &\overset{(3.16)}{\geq} C_{\text{opt}}\mu \frac{C''_{\text{opt}}}{C_{\text{GD}}} \left(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}^{\text{opt}}_{m+\alpha})\right) \\
&\geq \mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}^{\text{opt}}_{m+\alpha}) \\
&\geq \mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}^{\text{opt}}_{m+\alpha})
\end{aligned}$$

und schließlich $\mathcal{G}(\mathcal{T}_{k+1}) \leq \mathcal{G}(\mathcal{T}^{\text{opt}}_{m+\alpha}) = \mathcal{G}^{\text{opt}}_{m+\alpha}$.

Fall 2: $N \geq 1$.
Nach Definition von N gilt $\alpha N = \alpha\lfloor \#\mathcal{C}/\alpha \rfloor \leq \#\mathcal{C}$. Wir teilen die Menge $\mathcal{C}$ in paarweise disjunkte Teilmengen $\mathcal{C}_1, \dots, \mathcal{C}_N$ auf, von denen jede mindestens α Elemente enthält. Setzen wir $\mathcal{T}_i' := \operatorname{coarse}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}}; \mathcal{C}_i)$, so werden für jedes i mindestens α Knoten entfernt, um zu $\mathcal{T}_i'$ zu gelangen. Da nach Definition $\#(\mathcal{T}_{m+\alpha}^{\mathrm{opt}} \setminus \mathcal{T}_0) \geq m + \alpha$ gilt und pro entferntem Knoten auch mindestens ein Dreieck entfernt wird, gilt schließlich $\#(\mathcal{T}_i' \setminus \mathcal{T}_0) \geq m$. Daraus folgern wir $\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) \leq \mathcal{G}(\mathcal{T}_i')$. Mit dieser Ungleichung und Lemma 3.11 erhalten wir, dass

$$\begin{aligned}\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}}) &\overset{(3.7)}{\geq} \min\{1, C_{\mathrm{LD}}^{-1}\} \sum_{i=1}^{N} \big(\mathcal{G}(\mathcal{T}_i') - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}})\big) \\ &\geq N \min\{1, C_{\mathrm{LD}}^{-1}\} \big(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}})\big). \end{aligned} \tag{3.19}$$

Zusätzlich bemerken wir noch, dass wegen $N \geq 1$ gilt, dass $N = \lfloor \frac{\#\mathcal{C}}{\alpha} \rfloor \geq \frac{\#\mathcal{C}}{2\alpha}$ ist, und erinnern daran, dass $\alpha = \lfloor \frac{\#\mathcal{M}_k}{C''_{\mathrm{opt}}} \rfloor \leq \frac{\#\mathcal{M}_k}{C''_{\mathrm{opt}}}$. Insgesamt erhalten wir damit

$$N \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \geq \frac{\#\mathcal{C}\#\mathcal{M}_k}{2\alpha\#\mathcal{U}} \geq \frac{C''_{\mathrm{opt}}\#\mathcal{C}}{2\#\mathcal{U}} \overset{(3.18)}{\geq} \frac{C''_{\mathrm{opt}}}{2C_{\mathrm{GD}}}. \tag{3.20}$$

Mit diesen Abschätzungen können wir die Energiereduktion vom k-ten zum $(k+1)$-ten Gitter weiter abschätzen:

$$\begin{aligned}\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) &\overset{(3.17)}{\geq} C_{\mathrm{opt}}\mu \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_{m+\alpha})\big) \\ &\overset{(3.19)}{\geq} C_{\mathrm{opt}}\mu \min\{1, C_{\mathrm{LD}}^{-1}\} N \frac{\#\mathcal{M}_k}{\#\mathcal{U}} \big(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}})\big) \\ &\overset{(3.20)}{\geq} C_{\mathrm{opt}}\mu \min\{1, C_{\mathrm{LD}}^{-1}\} \frac{C''_{\mathrm{opt}}}{2C_{\mathrm{GD}}} \big(\mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}})\big)\end{aligned}$$

Nach Definition ist $C''_{\mathrm{opt}} \geq \frac{2C_{\mathrm{GD}}}{C_{\mathrm{opt}}\mu \min\{1, C_{\mathrm{LD}}^{-1}\}}$, womit sich obige Abschätzung zu

$$\mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{k+1}) \geq \mathcal{G}(\mathcal{T}_m^{\mathrm{opt}}) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}}) \geq \mathcal{G}(\mathcal{T}_k) - \mathcal{G}(\mathcal{T}_{m+\alpha}^{\mathrm{opt}})$$

vereinfachen lässt. Auch in diesem Fall folgt $\mathcal{G}(\mathcal{T}_{k+1}) \leq \mathcal{G}(\mathcal{T}_m^{\mathrm{opt}})$.
Damit ist die Behauptung bewiesen. □

Bemerkung 3.18. Lemma 3.16 zeigt, dass bei Markierung von mindestens einer Kante in jedem Schritt nach höchstens $\lceil 1/C'_{\mathrm{opt}}\mu \rceil$ Schritten die Energie um eine optimale Stufe abgesenkt wird. Dahingegen zeigt Lemma 3.17, dass die Energie in einem Schritt für je $\lceil C''_{\mathrm{opt}} \rceil$ markierte Kanten um eine Stufe abgesenkt wird. Es liegt daher nahe, zu erwarten, dass sich die Energie um mindestens eine Stufe absenkt, wenn entweder in einem einzigen Schritt hinreichend viele Kanten markiert wurden, oder insgesamt schon hinreichend viele Schritte durchgeführt wurden. Diese Überlegungen werden die zentrale Idee des Beweises von Satz 3.14 darstellen. //

Bevor wir nun den Beweis von Satz 3.14 angehen können, benötigen wir noch zwei elementare Eigenschaften der Floor-Funktion, die wir im Folgenden beweisen wollen.

Lemma 3.19. *Seien $a, b \in \mathbb{R}^+$ mit $b \geq 1$. Dann gelten*

$$\lfloor b \rfloor \geq \frac{b}{2}, \tag{3.21}$$

$$\lfloor a \rfloor + \lfloor b \rfloor \geq \left\lfloor a + \frac{b}{2} \right\rfloor. \tag{3.22}$$

Beweis. Wir zeigen zunächst Ungleichung (3.21) für $1 \leq b < 2$. Es gilt hier, dass $\lfloor b \rfloor = 1 > b/2$, also (3.21). Für $b \geq 2$ folgt

$$\lfloor b \rfloor \geq b - 1 \geq \frac{b}{2},$$

womit wir (3.21) insgesamt für $b \geq 1$ gezeigt haben.

Auch Ungleichung (3.22) zeigen wir zunächst für $1 \leq b < 2$. Da hier wieder $\lfloor b \rfloor = 1$ und $b/2 < 1$ gelten, folgt

$$\lfloor a \rfloor + \lfloor b \rfloor = \lfloor a \rfloor + 1 = \lfloor a + 1 \rfloor \geq \left\lfloor a + \frac{b}{2} \right\rfloor.$$

Für $b \geq 2$ haben wir $b/2 - 1 \geq 0$, womit wir

$$\lfloor a \rfloor + \lfloor b \rfloor \geq \lfloor a + b - 1 \rfloor = \left\lfloor a + \frac{b}{2} + \frac{b}{2} - 1 \right\rfloor \geq \left\lfloor a + \frac{b}{2} \right\rfloor$$

erhalten. Somit haben wir auch (3.22) für $y \geq 1$ gezeigt. □

Beweis von Satz 3.14. Wir legen zu Beginn einige Konstanten fest, wobei C'_{opt} und C''_{opt} die Konstanten aus den Lemmata 3.16 und 3.17 sind:

$$M_k := \sum_{i=0}^{k-1} \#\mathcal{M}_i, \tag{3.23}$$

$$R := \left\lceil \frac{1}{C'_{\text{opt}}\mu} \right\rceil, \tag{3.24}$$

$$L := 2(R-1)(C''_{\text{opt}} - 1) + 2C''_{\text{opt}}. \tag{3.25}$$

Beachte, dass $C''_{\text{opt}} \geq 1$ laut Beweis von Lemma 3.17. Wegen $C'_{\text{opt}} < \infty$ gilt ferner $R \geq 1$. Damit ist jedenfalls $L > 0$ sichergestellt. Die Wahl von L mag etwas willkürlich wirken, sie wird im weiteren Verlauf des Beweises aber klarer. Nach Voraussetzung wird in jedem Schritt mindestens eine Kante markiert, weshalb M_k in k streng monoton steigend und für $k > 0$ auch strikt positiv sind. Für $k = 0$ gilt $M_0 = 0$ laut Konvention für die leere Summe.

Wir zeigen nun per Induktion, dass

$$\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}\big(\mathcal{T}^{\mathrm{opt}}_{\lfloor M_k/L \rfloor}\big) \quad \text{für alle } k \in \mathbb{N}_0 \tag{3.26}$$

gilt. Im Induktionsanfang $k = 0$ gilt $\mathcal{T}_0^{\mathrm{opt}} = \mathcal{T}_0$ nach Definition der optimalen Gitter. Es folgt $\mathcal{G}(\mathcal{T}_0) \leq \mathcal{G}(\mathcal{T}_0^{\mathrm{opt}})$ und (3.26) gilt für $k = 0$. Es bleibt also der Induktionsschritt zu zeigen. Wir fixieren dafür ein $k \geq 1$ und nehmen an, dass (3.26) für alle $k' \in \mathbb{N}_0$ mit $k' < k$ gilt. Wir können außerdem annehmen, dass $\mathcal{G}(\mathcal{T}_k) > \inf_{\mathcal{T}\in\mathbb{T}} \mathcal{G}(\mathcal{T})$ gilt, da sonst die Aussage sowieso erfüllt ist.

Für den Induktionsschritt betrachten wir die Menge

$$I := \big\{ j \in \{\max\{k-i, 0\} \mid 1 \leq i \leq R\} \,\big|\, \#\mathcal{M}_j \geq C''_{\mathrm{opt}} \big\}.$$

Sie besteht aus allen Indizes der höchstens R Schritte vor dem k-ten, in denen C''_{opt} oder mehr Kanten markiert wurden. Wir werden im Folgenden eine Fallunterscheidung nach der Mächtigkeit dieser Menge durchführen. Zur Motivation, diese Menge einzuführen und diese Fallunterscheidung zu betrachten, siehe auch Bemerkung 3.18.

Fall 1: $I \neq \emptyset$.
Wir bezeichnen mit ℓ das maximale Element von I. Es gilt nach dieser Definition $\#\mathcal{M}_\ell \geq C''_{\mathrm{opt}}$ und $\#\mathcal{M}_{\ell'} < C''_{\mathrm{opt}}$ für alle ℓ' mit $\ell < \ell' < k$, derer es höchstens $R-1$ viele gibt. Damit und mit der Definition von L erhalten wir

$$\begin{aligned}
\frac{\#\mathcal{M}_\ell}{2C''_{\mathrm{opt}}} = \frac{\#\mathcal{M}_\ell L}{2C''_{\mathrm{opt}} L} &\overset{(3.25)}{=} \frac{\#\mathcal{M}_\ell\big((R-1)(C''_{\mathrm{opt}}-1) + C''_{\mathrm{opt}}\big)}{C''_{\mathrm{opt}} L} \\
&\geq \frac{C''_{\mathrm{opt}}(R-1)(C''_{\mathrm{opt}}-1) + \#\mathcal{M}_\ell C''_{\mathrm{opt}}}{C''_{\mathrm{opt}} L} \qquad (3.27) \\
&= \frac{(R-1)(C''_{\mathrm{opt}}-1) + \#\mathcal{M}_\ell}{L}, \qquad (3.28)
\end{aligned}$$

sowie

$$\begin{aligned}
M_\ell + \#\mathcal{M}_\ell + (R-1)(C''_{\mathrm{opt}}-1) &\geq M_\ell + \#\mathcal{M}_\ell + \sum_{\ell<\ell'<k} \#\mathcal{M}_{\ell'} \\
&\overset{(3.23)}{=} \sum_{\ell'=0}^{\ell-1} \#\mathcal{M}_{\ell'} + \#\mathcal{M}_\ell + \sum_{\ell'=\ell+1}^{k-1} \#\mathcal{M}_{\ell'} \\
&= \sum_{\ell'=0}^{k-1} \#\mathcal{M}_{\ell'} = M_k. \qquad (3.29)
\end{aligned}$$

Wegen $\ell < k$ gilt nach Induktionsvoraussetzung $\mathcal{G}(\mathcal{T}_\ell) \leq \mathcal{G}^{\mathrm{opt}}_{\lfloor M_\ell/L \rfloor}$, weshalb es eine natürliche Zahl

$$m \geq \lfloor M_\ell/L \rfloor$$

gibt, sodass $\mathcal{G}(\mathcal{T}_\ell) \in \big(\mathcal{G}^{\mathrm{opt}}_{m+1}, \mathcal{G}^{\mathrm{opt}}_m\big]$. Dies erlaubt uns, Lemma 3.17 anzuwenden, womit wir $\mathcal{G}(\mathcal{T}_{\ell+1}) \leq \mathcal{G}^{\mathrm{opt}}_{m+\lfloor \#\mathcal{M}_\ell/C''_{\mathrm{opt}} \rfloor}$ erhalten. Den Index der hier auftretenden optimalen Energie

können wir abschätzen, indem wir Lemma 3.19 verwenden, wobei, wie bereits erwähnt, $\#\mathcal{M}_\ell \geq C''_{\text{opt}}$ gilt. Damit lassen sich obige Nebenrechnungen kombinieren und wir erhalten

$$\begin{aligned} m + \left\lfloor \frac{\#\mathcal{M}_\ell}{C''_{\text{opt}}} \right\rfloor &\geq \left\lfloor \frac{M_\ell}{L} \right\rfloor + \left\lfloor \frac{\#\mathcal{M}_\ell}{C''_{\text{opt}}} \right\rfloor \\ &\overset{(3.22)}{\geq} \left\lfloor \frac{M_\ell}{L} + \frac{\#\mathcal{M}_\ell}{2C''_{\text{opt}}} \right\rfloor \\ &\overset{(3.28)}{\geq} \left\lfloor \frac{M_\ell + \#\mathcal{M}_\ell + (R-1)(C''_{\text{opt}} - 1)}{L} \right\rfloor \\ &\overset{(3.29)}{\geq} \left\lfloor \frac{M_k}{L} \right\rfloor. \end{aligned} \tag{3.30}$$

Mit der Monotonie der Energie, der letzten Abschätzung und $\ell < k$ erhalten wir schließlich

$$\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}(\mathcal{T}_{\ell+1}) \leq \mathcal{G}^{\text{opt}}_{m+\lfloor \#\mathcal{M}_\ell / C''_{\text{opt}} \rfloor} \overset{(3.30)}{\leq} \mathcal{G}^{\text{opt}}_{\lfloor M_k/L \rfloor}.$$

Damit ist dieser Fall erledigt.

Fall 2: $I = \emptyset$ und $k < R$.
In diesem Fall wurden noch keine $R = \lceil 1/(C'_{\text{opt}}\mu) \rceil$ Schritte durchgeführt und in keinem der Schritte wurden C''_{opt} oder mehr Kanten markiert. Daher greift die Aussage von Lemma 3.16 nicht so, wie es in Bemerkung 3.18 angedeutet wurde. Da bis jetzt höchstens $R-1$ Schritte mit jeweils höchstens $C''_{\text{opt}} - 1$ markierten Kanten durchgeführt wurden, gilt jedoch

$$M_k \leq (R-1)(C''_{\text{opt}} - 1) < L.$$

Damit gilt $\lfloor M_k/L \rfloor = 0$ und (3.26) gilt in der Form $\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}(\mathcal{T}_0) = \mathcal{G}(\mathcal{T}_0^{\text{opt}})$.

Fall 3: $I = \emptyset$ und $k \geq R$.
In diesem Fall war unter den R letzten Schritten keiner, in dem C''_{opt} oder mehr Kanten markiert wurden. Für $\ell \geq k - R$ lässt sich die Energiedifferenz des $(k-R)$-ten und des $(l+1)$-ten Gitters als Teleskopsumme schreiben:

$$\mathcal{G}(\mathcal{T}_{k-R}) - \mathcal{G}(\mathcal{T}_{\ell+1}) = \sum_{i=0}^{\ell-k+R} \big(\mathcal{G}(\mathcal{T}_{k-R+i}) - \mathcal{G}(\mathcal{T}_{k-R+i+1})\big). \tag{3.31}$$

Nach Induktionsvoraussetzung gibt es außerdem wieder eine natürliche Zahl

$$m \geq \lfloor M_{k-R}/L \rfloor,$$

sodass $\mathcal{G}(\mathcal{T}_{k-R}) \in \big(\mathcal{G}^{\text{opt}}_{m+1}, \mathcal{G}^{\text{opt}}_m\big]$ ist. Mit Lemma 3.16 können wir nun die Energiedifferenzen aller Schritte abschätzen, deren Energie in demselben Intervall liegt. Für alle $\ell \in J := \Big\{\ell' \in \{k-R, \dots, k-1\} \mid \mathcal{G}(\mathcal{T}_{\ell'}) \in \big(\mathcal{G}^{\text{opt}}_{m+1}, \mathcal{G}^{\text{opt}}_m\big]\Big\}$ erhalten wir mit der Darstellung (3.31)

die Abschätzung

$$\begin{aligned}\mathcal{G}(\mathcal{T}_{k-R}) - \mathcal{G}(\mathcal{T}_{\ell+1}) &\overset{(3.13)}{\geq} \sum_{i=0}^{\ell-k+R} C'_{\mathrm{opt}}\mu \#\mathcal{M}_{k-R+i}\big(\mathcal{G}_m^{\mathrm{opt}} - \mathcal{G}_{m+1}^{\mathrm{opt}}\big) \\ &\geq C'_{\mathrm{opt}}\mu(\ell - k + R + 1)\big(\mathcal{G}_m^{\mathrm{opt}} - \mathcal{G}_{m+1}^{\mathrm{opt}}\big) \\ &\geq C'_{\mathrm{opt}}\mu(\ell - k + R + 1)\big(\mathcal{G}(\mathcal{T}_{k-R}) - \mathcal{G}_{m+1}^{\mathrm{opt}}\big), \end{aligned} \tag{3.32}$$

da in jedem Schritt mindestens eine Kante markiert werden muss.

Wir betrachten nun das maximale Element $\ell_{\max}$ der Menge J. Ist $\ell_{\max} < k-1$, gilt $\mathcal{G}(\mathcal{T}_{\ell_{\max}+1}) \leq \mathcal{G}_{m+1}^{\mathrm{opt}}$ nach Definition von J. Ist andererseits $\ell_{\max} = k-1$, erhalten wir für die Konstante in (3.32)

$$C'_{\mathrm{opt}}\mu(\ell_{\max} - k + R + 1) = C'_{\mathrm{opt}}\mu R \overset{(3.24)}{=} C'_{\mathrm{opt}}\mu \left\lceil \frac{1}{C'_{\mathrm{opt}}\mu} \right\rceil \geq 1.$$

Es gilt also auch $\mathcal{G}(\mathcal{T}_{\ell_{\max}+1}) \leq \mathcal{G}_{m+1}^{\mathrm{opt}}$, wie man durch obige Abschätzung und Umformen von (3.32) sieht. Außerdem haben wir

$$L = 2(R-1)(C''_{\mathrm{opt}} - 1) + 2C''_{\mathrm{opt}} = 2R(C''_{\mathrm{opt}} - 1) - 2(C''_{\mathrm{opt}} - 1) + 2C''_{\mathrm{opt}} \geq R(C''_{\mathrm{opt}} - 1)$$

und $\#\mathcal{M}_\ell \leq C''_{\mathrm{opt}} - 1$ für alle Schritte mit Index $\ell \in \{k-R, \dots, k-1\}$, weshalb wir

$$\begin{aligned} m + 1 \geq \left\lfloor \frac{M_{k-R}}{L} \right\rfloor + 1 &= \left\lfloor \frac{M_{k-R}}{L} + \frac{L}{L} \right\rfloor \\ &\geq \left\lfloor \frac{M_{k-R} + R(C''_{\mathrm{opt}} - 1)}{L} \right\rfloor \geq \left\lfloor \frac{M_k}{L} \right\rfloor \end{aligned}$$

erhalten. Da jedenfalls $\ell_{\max} + 1 \leq k$ und $\lfloor M_k/L \rfloor \leq m+1$ gelten, folgt mit der Monotonie der Energie

$$\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}(\mathcal{T}_{\ell_{\max}+1}) \leq \mathcal{G}_{m+1}^{\mathrm{opt}} \leq \mathcal{G}_{\lfloor M_k/L \rfloor}^{\mathrm{opt}}.$$

Damit ist auch in diesem Fall die Induktionsbehauptung gezeigt.

Nachdem wir nun (3.26) gezeigt haben, ist der Weg nicht mehr weit. Nach der Mesh-Closure Estimate gibt es eine Konstante C_{MC}, sodass

$$\#(\mathcal{T}_k \setminus \mathcal{T}_0) \leq C_{\mathrm{MC}} M_k.$$

Setzen wir $C_{\mathrm{mesh}} := 2C_{\mathrm{MC}}L$, so folgt für $m \in \mathbb{N}$ aus $C_{\mathrm{mesh}} m \leq \#(\mathcal{T}_k \setminus \mathcal{T}_0) \leq C_{\mathrm{MC}} M_k$, dass $1 \leq m \leq M_k/(2L)$ und mit (3.21) somit $m \leq \frac{M_k}{2L} \leq \lfloor \frac{M_k}{L} \rfloor$. Mit dieser Abschätzung für m, der Monotonie der Energie und (3.26) gilt schließlich

$$\mathcal{G}(\mathcal{T}_k) \leq \mathcal{G}_{\lfloor \frac{M_k}{L} \rfloor}^{\mathrm{opt}} \leq \mathcal{G}_m^{\mathrm{opt}},$$

womit die Behauptung gezeigt ist.

Da C_{MC} nur vom initialen Gitter $\mathcal{T}_0$ abhängt, folgt die Abhängigkeit von C_{mesh} von μ aus der von L. Die dominanten Beiträge hierbei stammen von $\lceil 1/(C'_{\mathrm{opt}}\mu) \rceil$ und C''_{opt}, womit wir insgesamt $C_{\mathrm{mesh}} \in \mathcal{O}(\mu^{-2})$ erhalten. □

4 Konforme Finite Elemente für Diffusionsprobleme

Wir wollen in diesem Kapitel das allgemeine Framework aus den vorherigen Kapiteln auf konkrete Probleme und deren Finite-Elemente-Diskretisierungen anwenden. In [DKS16] wurde als Modellproblem das Poisson-Problem mit homogenen Dirichlet-Randdaten gewählt und für FE-Diskretisierungen erster Ordnung Instanzoptimalität gezeigt. In diesem Kapitel soll dieses Resultat in drei Richtungen verallgemeinert werden. Erstens werden wir den Differentialoperator verallgemeinern, indem wir stückweise konstante Diffusionsmatrizen zulassen. Zweitens betrachten wir auf einem Teil des Randes inhomogene Neumann-Randdaten. Zu guter Letzt wollen wir für die FE-Diskretisierung beliebige Polynomordnung zulassen.

Im Folgenden soll zunächst unser Modellproblem vorgestellt und gezeigt werden, wie hier auf natürliche Weise eine Energie definiert werden kann, die in unser Framework passt. Danach widmen wir uns dem Residualschätzer für das Modellproblem und zeigen die für Instanzoptimalität nötigen Zusammenhänge mit der Energie. Abschließend formulieren wir den gesamten Algorithmus für die AFEM und zeigen, dass der Markierungsschritt ebenfalls die nötigen Kriterien erfüllt.

Da die in diesem Kapitel verwendeten Resultate weitestgehend bekannt sind, werden wir auf die explizite Ausformulierung der meisten Konstanten verzichten und $A \lesssim B$ schreiben, falls eine Konstante $C > 0$ existiert, sodass $A \leq CB$. Außerdem schreiben wir $A \simeq B$ für $A \lesssim B \lesssim A$.

Wir werden nun kurz die Notation für die im Folgenden verwendeten Funktionenräume einführen. Für ausführliche Definitionen sei auf die Literatur verwiesen [Eva10].

Definition 4.1. *Sei $\Omega \subseteq \mathbb{R}^d$ ein beschränktes Gebiet. Wir bezeichnen die Menge aller quadratintegrablen Funktionen mit*

$$L^2(\Omega) := \left\{ u \mid u \text{ ist messbar und } \|u\|_{L^2(\Omega)} \leq \infty \right\},$$

wobei die L^2-Norm gegeben ist durch

$$\|u\|_{L^2(\Omega)} := \left(\int_\Omega |u|^2 \, \mathrm{d}x \right)^{1/2}.$$

Weiter definieren wir den Raum aller schwach differenzierbaren Funktionen auf Ω als

$$H^1(\Omega) := \left\{ u \in L^2(\Omega) \mid \nabla u \in L^2(\Omega) \right\},$$

wobei die Ableitungen hier schwach aufzufassen sind. Fassen wir Randwerte von H^1-Funktionen im Sinne von Spuren auf, können wir auch

$$H^1_0(\Omega) := \left\{ u \in H^1(\Omega) \mid u|_{\partial\Omega} \equiv 0 \right\}, \quad H^1_D(\Omega) := \left\{ u \in H^1(\Omega) \mid u|_{\Gamma_D} \equiv 0 \right\}$$

definieren.

Wir werden später auf diesen Räumen (Bi-)Linearformen betrachten, weshalb wir dafür ebenfalls einige Begriffe einführen wollen.

Definition 4.2. *Sei $(\mathcal{H}, (\cdot,\cdot)_{\mathcal{H}})$ ein Hilbertraum. Wir nennen den Raum aller stetigen linearen Funktionale $\mathcal{H} \to \mathbb{R}$ den* Dualraum *von $\mathcal{H}$ und schreiben dafür $\mathcal{H}'$.*

Sei zusätzlich $a : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ eine Bilinearform. Wir nennen diese stetig, *falls es eine Konstante $K > 0$ gibt, sodass*

$$|a(u,v)| \leq K \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$$

für alle $u, v \in \mathcal{H}$ und elliptisch, *falls $\kappa > 0$ existiert, sodass*

$$a(u,u) \geq \kappa \|u\|_{\mathcal{H}}^2$$

für alle $u \in \mathcal{H}$ gilt.

4.1 Variationsformulierung und Energie

Wir betrachten ein zusammenhängendes, beschränktes Gebiet $\Omega \subseteq \mathbb{R}^2$ mit polygonalem Rand $\Gamma := \partial\Omega$. Wir teilen den Rand in zwei disjunkte und relativ offene Teilmengen $\Gamma_D, \Gamma_N \subseteq \Gamma$, den Dirichlet- und den Neumannrand, sodass $\overline{\Gamma_D \cup \Gamma_N} = \Gamma$. Die Kanten aus $\mathcal{E}^\Gamma$, die in diesen Teilmengen liegen, bezeichnen wir als $\mathcal{E}^D$ beziehungsweise $\mathcal{E}^N$. Die Daten $f \in L^2(\Omega)$ und $\phi \in L^2(\Gamma_N)$ seien gegeben. Darüber hinaus sei eine matrixwertige Funktion $A \in [L^\infty(\Omega)]^{2\times 2}$ gegeben, zu der es ein Gitter $\mathcal{T}_0$ auf Ω gibt, auf dem A stückweise konstant, symmetrisch und positiv definit ist. Wir werden im Folgenden voraussetzen, dass alle auftretenden Gitter Verfeinerungen von $\mathcal{T}_0$ sind.

Das Modellproblem, mit dem wir uns in diesem Kapitel beschäftigen, ist folgendes Randwertproblem:

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad \text{in } \Omega, \\ A\nabla u \cdot \nu &= \phi \quad \text{auf } \Gamma_N, \\ u &= 0 \quad \text{auf } \Gamma_D. \end{aligned} \tag{4.1}$$

4.1.1 Schwache Formulierung und Diskretisierung

Um die Finite Elemente Methode auf das Problem (4.1) anwenden zu können, müssen wir es erst schwach formulieren: Finde $u \in H^1_D(\Omega)$, sodass

$$a(u,v) := \int_\Omega A\nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x + \int_{\Gamma_N} \phi v \,\mathrm{d}s =: F(v) \quad \text{für alle } v \in H^1_D(\Omega). \tag{4.2}$$

Bemerkung 4.3. Für die Existenz einer eindeutigen Lösung von (4.2) reichen nach dem Lemma von Lax-Milgram (siehe [Eva10, Abschnitt 6.2]) Stetigkeit und Elliptizität der

Bilinearform $a(\cdot,\cdot)$, sowie Stetigkeit des Funktionals F, jeweils auf dem Hilbertraum $H^1_D(\Omega)$. Wir werden diese Eigenschaften im Folgenden immer fordern und somit die eindeutige Lösbarkeit des Problems voraussetzen. //

Ziel der Finite Elemente Methode ist es, das Problem (4.2) auf einem endlichdimensionalen Teilraum zu lösen. Dafür definieren wir nun Räume von stückweisen Polynomen auf einem Gitter $\mathcal{T}$ [Ver13, Abschnitt 3.2].

Definition 4.4. *Sei $\mathcal{T}$ ein Gitter auf Ω. Für ein Dreieck $T \in \mathcal{T}$ definieren wir den Raum aller Polynome vom totalen Grad $p \in \mathbb{N}_0$ als $P^p(T)$. Analog bezeichnen wir für eine Kante $E \in \mathcal{E}$ den Raum aller Polynome vom totalen Grad p mit $P^p(E)$. Weiter definieren wir den Raum aller* stückweisen Polynome vom Grad $p \in \mathbb{N}_0$ auf $\mathcal{T}$ *als*

$$P^p(\mathcal{T}) := \left\{u \in L^2(\Omega) \mid u|_T \in P^p(T) \text{ für alle } T \in \mathcal{T}\right\}.$$

Schließlich definieren wir für $p \in \mathbb{N}$ die Räume aller global stetigen, stückweisen Polynome vom Grad p *ohne bzw. mit Nullrandbedingungen als*

$$\mathcal{S}^p(\mathcal{T}) := P^p(\mathcal{T}) \cap H^1(\Omega), \quad \mathcal{S}^p_D(\mathcal{T}) := P^p(\mathcal{T}) \cap H^1_D(\Omega).$$

Die Randwerte solcher stückweisen Polynome auf einem Randstück $\gamma \subseteq \partial\Omega$ bezeichnen wir mit $\mathcal{S}^p(\gamma) := \{u|_\gamma \mid u \in \mathcal{S}^p(\mathcal{T})\}$.

Für ein Gitter $\mathcal{T}$ auf Ω verwenden wir als endlichdimensionalen Teilraum der Finite Elemente Diskretisierung den eben definierten Raum

$$\mathcal{S}^p_D(\mathcal{T}) \subseteq H^1_D(\Omega).$$

Wegen der Tatsache, dass der diskrete Raum in $H^1_D(\Omega)$ enthalten ist, nennt man die daraus resultierende FEM auch H^1-konform.

Das diskrete Analogon zu (4.2) lässt sich nun wie folgt formulieren: Finde $u_\mathcal{T} \in \mathcal{S}^p_D(\mathcal{T})$, sodass

$$\int_\Omega A\nabla u_\mathcal{T} \cdot \nabla v_\mathcal{T} \,\mathrm{d}x = \int_\Omega f v_\mathcal{T} \,\mathrm{d}x + \int_{\Gamma_N} \phi v_\mathcal{T} \,\mathrm{d}s \quad \text{für alle } v_\mathcal{T} \in \mathcal{S}^p_D(\mathcal{T}). \tag{4.3}$$

Die AFEM generiert, ausgehend vom initialen Gitter $\mathcal{T}_0$ von Ω, eine Folge von Gittern $(\mathcal{T}_k)_{k\in\mathbb{N}_0}$, sodass $\mathcal{T}_i \leq \mathcal{T}_j$ für $i \leq j$. Wir bezeichnen im Folgenden für ein Gitter $\mathcal{T} \in \mathbb{T}$ die Finite Elemente Lösung von (4.3) mit $u_\mathcal{T}$. Die exakte Lösung von (4.2) bezeichnen wir mit u_{ex}.

4.1.2 Energie

Das Problem (4.2) lässt sich in einen allgemeinen Kontext einbetten, in dem auf natürliche Weise eine Energie definiert werden kann.

Definition 4.5. *Sei $(\mathcal{H}, (\cdot,\cdot)_\mathcal{H})$ ein Hilbertraum, $a : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ eine symmetrische, stetige und elliptische Bilinearform und $F \in \mathcal{H}'$. Für $u \in \mathcal{H}$ bezeichnen wir mit*

$$\mathscr{E}(u) := \tfrac{1}{2}a(u,u) - F(u)$$

das Energiefunktional *zu* $a(\cdot,\cdot)$ *und* F*. Durch die Bilinearform lässt sich außerdem die sogenannte* Energienorm

$$\|\!|u|\!\| := a(u,u)^{1/2}$$

definieren.

Setzen wir $\mathcal{H} := H^1_D(\Omega)$ und für $u \in \mathcal{H}$ und $\omega \subseteq \Omega$

$$\|\!|u|\!\|^2_\omega := \int_\omega A\nabla u \cdot \nabla u \,\mathrm{d}x,$$

so können wir für das Problem (4.1) mit Definition 4.5 eine Energie $\mathscr{E}$ definieren. Die folgende Proposition zeigt, dass eine Lösung von (4.2) diese Energie minimiert. Für einen Beweis verweisen wir auf [Bra13, Charakterisierungssatz 2.2].

Proposition 4.6. *Sei* $\mathcal{H}$ *ein Hilbertraum,* $a(\cdot,\cdot)$ *eine symmetrische, stetige und elliptische Bilinearform und* $F \in \mathcal{H}'$*. Weiter sei* $\mathscr{E}$ *die zu* $a(\cdot,\cdot)$ *und* F *gehörige Energie aus Definition 4.5. Ein Element* $u \in \mathcal{H}$ *ist eine Lösung von*

$$a(u,v) = F(v) \quad \textit{für alle } v \in \mathcal{H}$$

genau dann, wenn es das Minimierungsproblem

$$\mathscr{E}(u) = \min_{v\in\mathcal{H}} \mathscr{E}(v) \tag{4.4}$$

löst. □

Wir wollen nun noch eine Motivation für den Namen Energienorm geben.

Proposition 4.7. *Seien* $\mathcal{H}_1, \mathcal{H}_2$ *Hilberträume mit* $\mathcal{H}_1 \subseteq \mathcal{H}_2$*. Weiter seien* u_i *mit* $i = 1,2$ *Lösungen der Variationsprobleme*

$$a(u_i,v) = F(v) \quad \textit{für alle } v \in \mathcal{H}_i. \tag{4.5}$$

Dann gilt für die Differenz der Energien der Lösungen

$$\mathscr{E}(u_1) - \mathscr{E}(u_2) = \frac{1}{2}\|\!|u_1 - u_2|\!\|^2. \tag{4.6}$$

Beweis. Setzen wir in die Definition der Energie ein und benutzen, dass $\|\!|u_1 - u_2|\!\|^2 = \|\!|u_1|\!\|^2 - \|\!|u_2|\!\|^2 - 2a(u_1 - u_2, u_2)$ gilt, erhalten wir für die Energiedifferenz

$$\begin{aligned}\mathscr{E}(u_1) - \mathscr{E}(u_2) &= \frac{1}{2}(\|\!|u_1|\!\|^2 - \|\!|u_2|\!\|^2) - (F(u_1) - F(u_2)) &(4.7)\\ &= \frac{1}{2}\|\!|u_1 - u_2|\!\|^2 + a(u_1 - u_2, u_2) - (F(u_1) - F(u_2)). &(4.8)\end{aligned}$$

Aufgrund der Definition der u_i als Lösung auf $\mathcal{H}_i$ gilt weiter, dass $a(u_2, u_2) = F(u_2)$. Außerdem erhalten wir mit der Symmetrie der Bilinearform und der Tatsache, dass $\mathcal{H}_1 \subseteq \mathcal{H}_2$ ist, $a(u_1, u_2) = a(u_2, u_1) = F(u_1)$. Daraus folgt nun mit der Bilinearität von $a(\cdot, \cdot)$, dass

$$\mathscr{E}(u_1) - \mathscr{E}(u_2) = \frac{1}{2} \|\!| u_1 - u_2 |\!\|^2 + a(u_1 - u_2, u_2) - (a(u_1, u_2) - a(u_2, u_2)) \tag{4.9}$$

$$= \frac{1}{2} \|\!| u_1 - u_2 |\!\|^2, \tag{4.10}$$

was die zu beweisende Identität ist. □

Die eben bewiesene Proposition impliziert, dass der Fehler der Finite Elemente Diskretisierung zur exakten Lösung u_{ex} in der Energienorm durch die Differenz der Energien gegeben ist:

$$\tfrac{1}{2} \|\!| u - u_{\text{ex}} |\!\|^2 = \mathscr{E}(u) - \mathscr{E}(u_{\text{ex}}).$$

Diese Beziehung wird uns später erlauben, die Energieoptimalität aus Kapitel 3 auf Instanzoptimalität des Fehlers zu übertragen.

Bemerkung 4.8. Um das Energiefunktional in diesem Abschnitt mit dem Energiebegriff aus Definition 3.3 in Einklang zu bringen, müssen wir dieses auch auf Gittern definieren und Monotonie zeigen. Wir setzen dazu

$$\mathscr{E}(\mathcal{T}) := \mathscr{E}(u_{\mathcal{T}}),$$

wobei $u_{\mathcal{T}}$ die Finite Elemente Lösung auf $\mathcal{T}$ ist. Dann folgt aus der Minimierungseigenschaft (4.4), dass für $\mathcal{T} \leq \mathcal{T}'$ wegen $\mathcal{S}^p_D(\mathcal{T}) \subseteq \mathcal{S}^p_D(\mathcal{T}')$ auch $\mathscr{E}(\mathcal{T}) \geq \mathscr{E}(\mathcal{T}')$ gilt. //

Zu guter Letzt wollen wir noch eine Normäquivalenz für die Energienorm zeigen.

Lemma 4.9. *Für das Problem (4.1) ist die Energienorm äquivalent zur H^1-Seminorm:*

$$\|\!| v |\!\| \simeq |v|_{H^1(\Omega)} \quad \textit{für alle } v \in H^1(\Omega). \tag{4.11}$$

Beweis. Setzt man in die Definition der Energienorm ein, erhält man

$$\|\!| v |\!\|^2 = a(v, v) = \sum_{T \in \mathcal{T}_0} \int_T A \nabla v \cdot \nabla v \, \mathrm{d}x = \sum_{T \in \mathcal{T}_0} \left\| A^{1/2} \nabla v \right\|^2_{L^2(T)}, \tag{4.12}$$

wobei wir $A^{1/2}$ elementweise verstehen. Dies ist möglich, weil nach Definition von $\mathcal{T}_0$ für $T \in \mathcal{T}_0$ gilt, dass $A|_T \in \mathbb{R}^{2\times 2}$ symmetrisch und positiv definit ist. Damit existieren durch Stetigkeit und positive Definitheit elementweise Konstanten $C_T, c_T > 0$ für $T \in \mathcal{T}_0$:

$$c_T \, |v|^2_{H^1(T)} \leq \left\| A^{1/2} \nabla v \right\|^2_{L^2(T)} \leq C_T \, |v|^2_{H^1(T)}. \tag{4.13}$$

Da $\#\mathcal{T}_0$ endlich ist, existieren Maximum beziehungsweise Minimum dieser Konstanten. Diese sind jedenfalls positiv, womit die Behauptung gezeigt ist. □

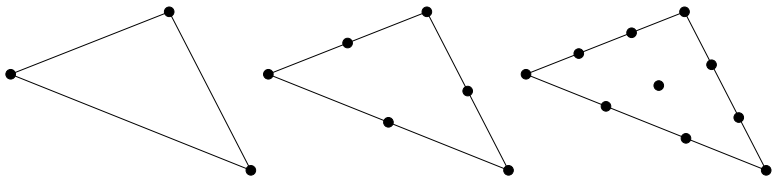


Abbildung 4.1: *Schematische Darstellung der Auswertungspunkte der Finite Elemente Polynome für $p = 1, 2, 3$ (von links nach rechts).*

4.2 Lower Diamond Estimate

Nun wollen wir zeigen, dass die eben definierte Energie $\mathscr{E}$ auch die Lower Diamond Estimate (E2) erfüllt. Wir benötigen ein Werkzeug, um die Trennungseigenschaften der Lower Diamond Struktur (die ja laut Definition 3.5 auf Gittern definiert sind) auf Funktionenräume zu übertragen. Dies ermöglicht uns ein der Scott-Zhang Projektion sehr ähnlicher Operator. Wir werden im Folgenden diesen Operator definieren und auch die klassische Scott-Zhang Projektion ein wenig genauer betrachten.

4.2.1 Die Scott-Zhang Projektion als Transferoperator

Für die Beweise von Lower Diamond Estimate und diskreter lokaler Zuverlässigkeit benötigen wir einen Operator, der in geeigneter Weise die diskreten Lösungen auf zwei Gittern in Verbindung bringt. Dies wird durch einen Clement-Operator erreicht, der sehr ähnlich zur Scott-Zhang Projektion ist [DKS16]. Die Scott-Zhang Projektion ist ein Quasiinterpolationsoperator, der das erste Mal in [SZ90] vorgestellt wurde. Wir werden nun zunächst allgemeine Scott-Zhang–artige Projektionen definieren und einige wichtige Eigenschaften dieser zeigen. Dazu müssen wir uns die Konstruktion der Finite Elemente Räume genauer ansehen. Aus dieser allgemeinen Definition werden wir die klassische Variante der Scott-Zhang Projektion als Spezialfall herleiten. Danach werden wir als einen weiteren Spezialfall einen Transferoperator definieren und auf die Unterschiede zur klassischen Scott-Zhang Projektion aus [SZ90] eingehen. Im Gegensatz zu [DKS16] definieren wir diese Operatoren hier für Finite Elemente mit beliebiger Polynomordnung und den Fall $\Gamma_D \subseteq \partial\Omega$, wobei Gleichheit nicht zwingend gelten muss.

Für ein Dreieck $T \in \mathcal{T}$ kann jedes Polynom des zugehörigen Polynomraumes $P^p(T)$, bestehend aus Polynomen p-ter Ordnung, durch die Auswertung an $(p+1)(p+2)/2$ verschiedenen Punkten aus T eindeutig bestimmt werden (siehe zum Beispiel [Bra13, II §5] und Abbildung 4.1). Wir bezeichnen die Menge aller dieser Auswertungspunkte in $\mathcal{T}$ mit $\mathcal{Z}(\mathcal{T})$. Für die folgende Definition Scott-Zhang–artiger Projektionen wird benötigt, dass zu jedem dieser Punkte entweder ein Dreieck, oder eine Kante des Gitters assoziiert ist, wobei diese Wahl bestimmten Einschränkungen unterliegt.

Definition 4.10. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter. Für einen Auswertungspunkt $z \in \mathcal{Z}(\mathcal{T})$ nennen wir $\sigma_z \in \mathcal{T} \cup \mathcal{E}$ das* zu z assoziierte Simplex*, wenn es folgende Eigenschaften erfüllt:*

- *Falls z ein innerer Knoten eines Dreiecks $T \in \mathcal{T}$ ist, gilt $\sigma_z = T$.*

- *Falls z ein innerer Knoten einer Kante $E \in \mathcal{E}$ ist, gilt $\sigma_z = E$.*
- *Falls $z \in \mathcal{V}$ ist, so ist $\sigma_z = E \in \mathcal{E}$ mit $z \in E$ beliebig.*

Nun beleuchten wir genauer, wie der Finite Elemente Raum $\mathcal{S}^p(\mathcal{T})$ aufgebaut ist. Durch die Wahl der Auswertungspunkte und des Polynomgrades p gibt es eine Basis $\Phi(\mathcal{T}) := \{\varphi_z \mid z \in \mathcal{Z}(\mathcal{T})\}$ aus stückweisen Polynomen, die

$$\varphi_z(z') = \delta_{zz'} \quad \text{für alle } z, z' \in \mathcal{Z}(\mathcal{T}) \tag{4.14}$$

erfüllt, wobei $\delta_{zz'}$ das Kronecker-Delta bezeichnet. Diese Basis wird auch *nodale Basis* genannt. Die zu $\Phi(\mathcal{T})$ gehörige duale Basis von $(\mathcal{S}^p(\mathcal{T}))'$ wird nach (4.14) durch die Punktauswertungsfunktionale an allen Punkten aus $\mathcal{Z}(\mathcal{T})$ gebildet. Die Einschränkung eines Polynoms aus $\mathcal{S}^p(\mathcal{T})$ auf ein Simplex σ_z ist wieder ein Polynom vom Grad höchstens p und liegt in $L^2(\sigma_z)$, wobei dies im Fall $\sigma_z \in \mathcal{E}$ im Sinne von Spuren aufzufassen ist. Da auch die auf σ_z eingeschränkten Basisfunktionen $\varphi_{z'}|_{\sigma_z}(z'') = \delta_{z'z''}$ erfüllen, falls $z', z'' \in \sigma_z$, und $L^2(\sigma_z)$ ein Hilbertraum ist, lässt sich über den Satz von Riesz eine Menge von Funktionen $\{\psi_{z,z'} \mid z' \in \mathcal{Z}(\mathcal{T}) \cap \sigma_z\} \subseteq P^p(\sigma_z)$ finden, die

$$\int_{\sigma_z} \psi_{z,z''}(x)\phi_{z'}|_{\sigma_z}(x)\,\mathrm{d}x = \delta_{z'z''} \quad \text{für alle } z', z'' \in \mathcal{Z}(\mathcal{T}) \cap \sigma_z$$

erfüllt. Diese Menge kann als Menge der Riesz-Repräsentanten der dualen Basis der Polynome auf σ_z bezüglich $L^2(\sigma_z)$ aufgefasst werden. Nimmt man nun für jeden Auswertungspunkt $z \in \mathcal{Z}(\mathcal{T})$ die Funktion $\psi_{z,z} =: \psi_z$ und fasst diese zur Menge $\Psi(\mathcal{T}) := \{\psi_z \mid z \in \mathcal{Z}(\mathcal{T})\}$ zusammen, gilt

$$\int_{\sigma_z} \psi_z(x)\varphi_{z'}(x)\,\mathrm{d}x = \delta_{zz'} \quad \text{für alle } z, z' \in \mathcal{Z}(\mathcal{T}). \tag{4.15}$$

Aufgrund der Interpretation der Funktionen aus $\Psi(\mathcal{T})$ als lokale Riesz-Repräsentanten der dualen Basis bezüglich $L^2(\sigma_z)$, werden wir diese Menge vereinfachend *die duale Basis bezüglich der Menge* $\{\sigma_z \mid z \in \mathcal{Z}(\mathcal{T})\}$ nennen.

Mit diesen Begriffen können wir schließlich Scott-Zhang–artige Projektionen definieren.

Definition 4.11. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $p \in \mathbb{N}$, $\Phi(\mathcal{T})$ die nodale Basis von $\mathcal{S}^p(\mathcal{T})$ und $\Psi(\mathcal{T})$ die bezüglich $\{\sigma_z\}$ duale Basis. Wir definieren eine* Scott-Zhang–artige Projektion *$\mathcal{J}_\mathcal{T} : H^1(\Omega) \to \mathcal{S}^p(\mathcal{T})$ für $v \in H^1(\Omega)$ als*

$$\mathcal{J}_\mathcal{T} v := \sum_{z \in \mathcal{Z}(\mathcal{T})} \varphi_z \int_{\sigma_z} \psi_z(x)v(x)\,\mathrm{d}x. \tag{4.16}$$

Im Grunde hängen die eben definierten Projektionen auch vom Polynomgrad p ab, wir werden im Folgenden diese Abhängigkeit jedoch nicht explizit anschreiben. Zur Vereinfachung der Notation verzichten wir, sofern aus dem Zusammenhang klar ersichtlich, auf den Index $\mathcal{T}$. Wir schreiben dann schlicht $\mathcal{J}$.

Wir wollen nun zeigen, dass jeder Operator, der Definition 4.11 erfüllt, tatsächlich eine Projektion ist.

Lemma 4.12. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $p \in \mathbb{N}$. Dann gilt für eine Scott-Zhang–artige Projektion $\mathcal{J} : H^1(\Omega) \to \mathcal{S}^p(\mathcal{T})$, dass $\mathcal{J}v = v$ für alle $v \in \mathcal{S}^p(\mathcal{T})$.*

Beweis. Wir entwickeln eine Funktion $v \in \mathcal{S}^p(\mathcal{T})$ nach der nodalen Basis mit Koeffizienten $(c_z)_{z\in\mathcal{Z}(\mathcal{T})}$:

$$v(x) = \sum_{z\in\mathcal{Z}(\mathcal{T})} c_z \varphi_z(x). \tag{4.17}$$

Setzen wir diese Darstellung in die Definition der Scott-Zhang–artigen Projektionen (4.16) ein, so erhalten wir mit der Dualitätseigenschaft von ψ_z und $\varphi_{z'}$, dass

$$\begin{aligned}
\mathcal{J}v &= \sum_{z\in\mathcal{Z}(\mathcal{T})} \varphi_z \int_{\sigma_z} \psi_z(x) v(x)\,\mathrm{d}x \\
&\overset{(4.17)}{=} \sum_{z\in\mathcal{Z}(\mathcal{T})} \sum_{z'\in\mathcal{Z}(\mathcal{T})} \varphi_z \int_{\sigma_z} \psi_z(x) c_{z'} \varphi_{z'}(x)\,\mathrm{d}x \\
&\overset{(4.15)}{=} \sum_{z\in\mathcal{Z}(\mathcal{T})} \sum_{z'\in\mathcal{Z}(\mathcal{T})} \varphi_z c_{z'} \delta_{zz'} = \sum_{z\in\mathcal{Z}(\mathcal{T})} c_z \varphi_z = v.
\end{aligned}$$

Damit ist die Projektionseigenschaft nachgewiesen. □

Wir benötigen außerdem noch einige Standardresultate über Stabilität und Approximationsgüte von Scott-Zhang–artigen Projektionen in verschiedenen Normen. Hier bezeichnet $h_T := |T|^{1/2}$ für $T \in \mathcal{T}$ und $h_E := |E|$ für $E \in \mathcal{E}$. Für einen Beweis verweisen wir bezüglich der ersten beiden Stabilitätseigenschaften auf [SZ90, section 3], bezüglich der letzten beiden Approximationseigenschaften auf [SZ90, section 4].

Proposition 4.13. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $T \in \mathcal{T}$ und $E \in \mathcal{E}$. Für eine beliebige Scott-Zhang–artige Projektion $\mathcal{J}$ aus Definition 4.11 gibt es Konstanten $C_i > 0$ mit $i \in \{1,2,3,4\}$, sodass für alle $v \in H^1(\Omega)$ folgende Eigenschaften gelten:*

(i) $|\mathcal{J}v|_{H^1(T)} \le C_1 |v|_{H^1(\omega_T)}$.

(ii) $|\mathcal{J}v|_{H^1(\Omega)} \le C_2 |v|_{H^1(\Omega)}$.

(iii) $\|(1-\mathcal{J})v\|_{L^2(T)} \le C_3 h_T |v|_{H^1(\omega_T)}$.

(iv) $\|(1-\mathcal{J})v\|_{L^2(E)} \le C_4 h_E^{1/2} |v|_{H^1(\omega_E)}$. □

Die Konstanten C_i mit $i \in \{1,2,3,4\}$ hängen nur von der Formregularitätskonstante C_{sr} und dem Polynomgrad p ab.

Die klassische Scott-Zhang Projektion aus [SZ90] ist ein Spezialfall der eben vorgestellten Scott-Zhang–artigen Projektionen.

Definition 4.14. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $p \in \mathbb{N}$. Zusätzlich zu Definition 4.10 gelte für die simplizes $\{\sigma_z \mid z \in \mathcal{Z}(\mathcal{T})\}$ folgende Einschränkung: Für $z \in \mathcal{V}$ muss $\sigma_z = E \subseteq \overline{\Gamma_D}$ mit $z \in E$ sein, falls $z \in \overline{\Gamma_D}$, und $\sigma_z = E \subseteq \overline{\Gamma_N}$ mit $z \in E$, falls $z \in \Gamma_N$.*

Als Scott-Zhang–artige Projektion erbt die Scott-Zhang Projektion die Eigenschaften aus Lemma 4.12 und Proposition 4.13. Zusätzlich erhält sie homogene Dirichlet-Randdaten, wie im nächsten Lemma gezeigt wird.

Lemma 4.15. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $p \in \mathbb{N}$. Dann gilt für die Scott-Zhang Projektion, dass $\mathcal{J}v \in \mathcal{S}^p_D(\mathcal{T})$ für alle $v \in H^1_D(\Omega)$.*

Beweis. Nach der Wahl der Simplizes σ_z gilt für jeden Auswertungspunkt $z \in \mathcal{Z}(\mathcal{T})$ mit $z \in \overline{\Gamma_D}$, dass auch $\sigma_z \subseteq \overline{\Gamma_D}$ gilt. Damit verschwindet v auf σ_z, womit das Integral $\int_{\sigma_z} \psi_z(x)v(x)\,\mathrm{d}x = 0$ wird. Werten wir die Projektion von v an $z \in \Gamma_D$ aus, erhalten wir mit der Definition der nodalen Basis

$$\begin{aligned}\big(\mathcal{J}v\big)(z) &= \sum_{z' \in \mathcal{Z}(\mathcal{T})} \varphi_{z'}(z) \int_{\sigma_{z'}} \psi_{z'}(x)v(x)\,\mathrm{d}x \\ &= \sum_{z' \in \mathcal{Z}(\mathcal{T})} \delta_{z'z} \int_{\sigma_{z'}} \psi_{z'}(x)v(x)\,\mathrm{d}x \\ &= \int_{\sigma_z} \psi_z(x)v(x)\,\mathrm{d}x = 0.\end{aligned}$$

Da die Randwerte einer Funktion aus $\mathcal{S}^p(\mathcal{T})$ auf einer Kante eindeutig durch die Werte an den Auswertungspunkten dieser Kante bestimmt werden, verschwindet somit die Projektion $\mathcal{J}v$ auf Γ_D. □

Aufbauend auf der Definition und den Eigenschaften der Scott-Zhang–artigen Projektionen werden wir nun einen Transferoperator als Spezialfall konstruieren, der eine Funktion von einem feineren auf ein gröberes Gitter projiziert und dabei die Funktion auf den nicht vergröberten Dreiecken unverändert lässt.

Definition 4.16. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $\mathcal{T}' \geq \mathcal{T}$ und $p \in \mathbb{N}$. Wir definieren $\mathcal{U}_1 := \mathcal{T} \cap \mathcal{T}'$ und $\mathcal{U}_2 := \mathcal{T} \setminus \mathcal{T}'$. Für einen Auswertungspunkt $z \in \mathcal{Z}(\mathcal{T})$ erfülle das zu z assoziierte Simplex σ_z zusätzlich zu Definition 4.10 folgende Eigenschaften:*

- *Falls $z \in \mathcal{V}$ ist, so ist $\sigma_z = E \in \mathcal{E}$ mit $z \in E$ beliebig, wobei $E \subseteq \partial[\Omega(\mathcal{U}_i)]$ gelten muss, sofern $z \in \partial[\Omega(\mathcal{U}_i)]$ für $i = 1, 2$.*
- *Falls $z \in \mathcal{V}$ mit $z \in \Gamma_D$ und $z \notin \partial[\Omega(\mathcal{T} \cap \mathcal{T}')]$, soll zusätzlich $\sigma_z \subseteq \partial[\Omega(\mathcal{T} \setminus \mathcal{T}')] \cap \overline{\Gamma_D}$ gelten.*

Wir definieren den Transferoperator

$$\mathcal{Q}^{\mathcal{T}'}_{\mathcal{T}} : \mathcal{S}^p(\mathcal{T}') \to \mathcal{S}^p(\mathcal{T}) \tag{4.18}$$

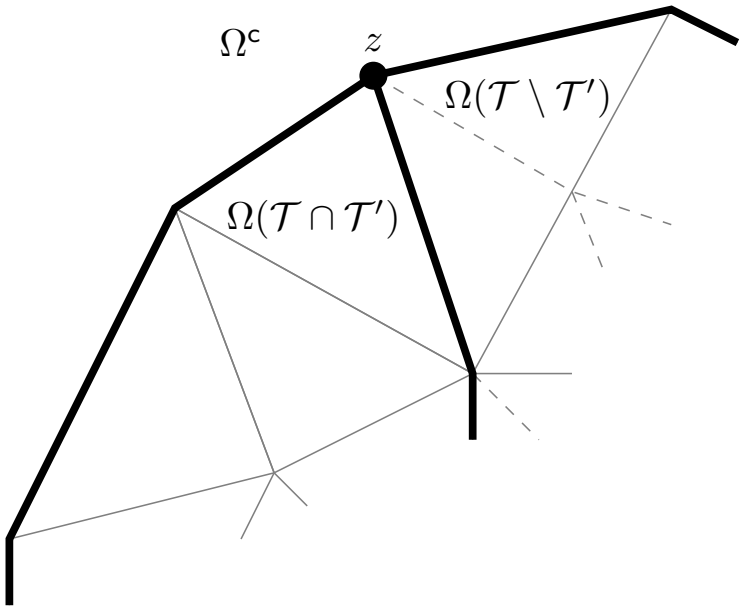


Abbildung 4.2: *Aufteilung des Gitters in $\mathcal{T} \cap \mathcal{T}'$ und $\mathcal{T} \setminus \mathcal{T}'$. Das zum Auswertungspunkt $z \in \mathcal{Z}(\mathcal{T})$ assoziierte Simplex kann nach Definition 4.16 nicht in $\partial\Omega$ liegen, obwohl z am Rand liegt.*

als eine Scott-Zhang–artige Projektion mit den Simplizes $\{\sigma_z \mid z \in \mathcal{Z}(\mathcal{T})\}$, die obige Bedingungen erfüllen.

Bemerkung 4.17. Beachte, dass die Wahl der Simplizes in Definition 4.16 wohldefiniert ist. Da $\mathcal{T} \cap \mathcal{T}'$ und $\mathcal{T} \setminus \mathcal{T}'$ eine Partition von $\mathcal{T}$ bilden und das Gitter selbst keine isolierten Punkte hat, können sich diese beiden Mengen nicht bloß an einem Punkt berühren. Somit ist der erste Punkt aus obiger Definition erfüllbar.

Für einen Punkt $z \in \mathcal{V}$ mit $z \in \Gamma_D$ und $z \notin \partial[\Omega(\mathcal{T} \cap \mathcal{T}')]$ gilt jedenfalls $z \in \partial[\Omega(\mathcal{T} \setminus \mathcal{T}')]$. Da die Ränder $\partial\Omega$ und $\partial[\Omega(\mathcal{T} \setminus \mathcal{T}')]$ geschlossen sind, muss es jedenfalls zwei Kanten $E_\ell, E_r \subseteq \partial\Omega \cap \partial[\Omega(\mathcal{T} \setminus \mathcal{T}')]$ geben. Weil auch Γ_D keine isolierten Punkte enthalten kann, ist zumindest eine dieser beiden Kanten auch in $\overline{\Gamma_D}$ enthalten. Deshalb ist auch der zweite Punkt der vorigen Definition erfüllbar. //

Aufgrund der Definition als Scott-Zhang Projektion können wir einige Eigenschaften von dieser übernehmen (siehe dazu die folgende Proposition). Im Grunde ist der Transferoperator als solcher auch für Funktionen aus $H^1(\Omega)$ definiert. In diesem Fall ist allerdings nicht mehr gegeben, dass er homogene Randdaten erhält.

Betrachtet man die Unterschiede in den Definitionen 4.14 und 4.16, sticht ins Auge, dass die zu Gitterknoten $z \in \mathcal{V}$ assoziierten Simplizes beim Transferoperator nicht den Rand von Ω berücksichtigen, sondern die beiden Ränder der Mengen $\Omega(\mathcal{T} \setminus \mathcal{T}')$ und $\Omega(\mathcal{T} \cap \mathcal{T}')$. Von diesen Mengen enthält die erste die Dreiecke, die sich beim Übergang von $\mathcal{T}'$ auf $\mathcal{T}$

vergröbert haben, die zweite die Dreiecke, die sich nicht geändert haben. Dies führt zu Situationen, wie etwa in Abbildung 4.2 dargestellt. Hier liegt der Auswertungspunkt z zwar auf $\partial\Omega$, aber auch auf $\partial[\Omega(\mathcal{T} \setminus \mathcal{T}')]$ und $\partial[\Omega(\mathcal{T} \cap \mathcal{T}')]$. Daher wird erzwungen, dass das Simplex σ_z die einzige an z angrenzende, gemeinsame Kante der Mengen $\Omega(\mathcal{T} \setminus \mathcal{T}')$ und $\Omega(\mathcal{T} \cap \mathcal{T}')$ ist und damit nicht in $\partial\Omega$ liegt. Wir zeigen in folgender Proposition, dass diskrete Randdaten durch den Transferoperator trotzdem erhalten werden.

Proposition 4.18. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $\mathcal{T}' \geq \mathcal{T}$ und $p \in \mathbb{N}$. Seien weiter $\mathcal{U}_1 := \mathcal{T} \cap \mathcal{T}'$ und $\mathcal{U}_2 := \mathcal{T} \setminus \mathcal{T}'$. Der Transferoperator $\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'}$ hat folgende Eigenschaften:*

(i) Die Werte von $(\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v)|_{\Omega(\mathcal{U}_i)}$ für $v \in \mathcal{S}^p(\mathcal{T}')$ hängen nur von $v|_{\Omega(\mathcal{U}_i)}$ ab.

(ii) Für $v \in \mathcal{S}^p(\mathcal{T}')$ und $T \in \mathcal{T} \cap \mathcal{T}'$ gilt $\big((\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v) - v\big)|_T = 0$.

(iii) Der Transferoperator erhält diskrete homogene Randdaten:

$$\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'}\big(\mathcal{S}_D^p(\mathcal{T}')\big) \subseteq \mathcal{S}_D^p(\mathcal{T}).$$

Beweis. Wir übernehmen die Beweisidee aus [DKS16].

Ad (i): Durch die Wahl der Simplizes in Definition 4.16 ist für jeden Auswertungspunkt $z \in \Omega(\mathcal{U}_i)$ für $i = 1, 2$ auch das assoziierte Simplex $\sigma_z \subseteq \Omega(\mathcal{U}_i)$. Der Wert von $(\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v)(z)$ hängt aber nur von den Werten von v auf σ_z ab. Daraus folgt die Behauptung.

Ad (ii): Die Menge $\mathcal{T} \cap \mathcal{T}'$ besteht aus allen Dreiecken, die sich beim Übergang von $\mathcal{T}'$ zu $\mathcal{T}$ nicht geändert haben. Dies gilt dann natürlich auch für deren Kanten. Daher liegt für jeden Auswertungspunkt $z \in \mathcal{T} \cap \mathcal{T}'$ auch das assoziierte Simplex in $\mathcal{T} \cap \mathcal{T}'$ und hat sich deshalb nicht geändert. Somit können wir für alle Dreiecke $T \in \mathcal{T} \cap \mathcal{T}'$ die Rechnung aus Lemma 4.12 wiederholen und erhalten $(\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v)|_T = v|_T$.

Ad (iii): Wir betrachten einen Auswertungspunkt $z \in \partial\Gamma_D$ am Dirichletrand und eine Funktion $v \in \mathcal{S}_D^p(\mathcal{T}')$. Für den Fall, dass $z \in \Gamma_D \cap \partial[\Omega(\mathcal{T} \cap \mathcal{T}')]$ wissen wir aus Punkt (ii), dass $(\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v)(z) = v(z) = 0$. Ist andererseits $z \in \Gamma_D$ und $z \notin \partial[\Omega(\mathcal{T} \cap \mathcal{T}')]$, so ist nach dem letzten Kriterium in Definition 4.16, wie die Simplizes zu wählen sind, $\sigma_z \subseteq \Gamma_D \cap \partial[\Omega(\mathcal{T} \setminus \mathcal{T}')]$. Daher gilt $v|_{\sigma_z} = 0$, weshalb der Beweis von Lemma 4.15 hier analog funktioniert.

Damit ist insgesamt gezeigt, dass $\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'} v \equiv 0$ auf Γ_D. □

Bemerkung 4.19. Als Scott-Zhang–artige Projektion erfüllt der eben definierte Transferoperator die Eigenschaften von Proposition 4.13. Aufgrund von Proposition 4.18 (i) gelten diese Eigenschaften für ein Dreieck $T \in \mathcal{T}$ sogar mit $\omega_T \cap \mathcal{U}_i$ anstelle von ω_T, sofern $T \in \mathcal{U}_i$. Eine Analoge Aussage gilt auch für ω_E für eine Kante $E \in \mathcal{E}$. Aus dem Beweis des letzten Punktes der vorigen Proposition, dass diskrete Randdaten erhalten werden, ist auch ersichtlich, weshalb das letzte Kriterium in Definition 4.16 notwendig ist. Sei dazu ein Knoten $z \in \mathcal{V}$ am Dirichlet-Rand Γ_D gegeben. Solange z und das assoziierte Simplex σ_z beide in $\Omega(\mathcal{T} \cap \mathcal{T}')$ liegen, ist egal, ob σ_z am Rand liegt. Ist jedoch $z \in \Omega(\mathcal{T} \setminus \mathcal{T}') \setminus \Omega(\mathcal{T} \cap \mathcal{T}')$, so muss das assoziierte Simplex σ_z ebenfalls in Γ_D liegen, um diskrete Randdaten zu erhalten. Dies wird durch ebendiese Bedingung erzwungen. //

4.2.2 Beweis der LDE

Mit dem Transferoperator aus Definition 4.16 können wir uns daran machen, die Lower Diamond Estimate für die Energie $\mathscr{E}$ nachzuweisen. Wir werden dafür zunächst zwei Lemmata beweisen. Das erste hiervon besagt, dass der Fehler der Finite Elemente Approximationen zweier Gitter äquivalent zum Fehler ist, den der Transferoperator bedingt.

Lemma 4.20. *Seien $\mathcal{X}_1, \mathcal{X}_2 \subseteq H^1_D(\Omega)$ lineare Unterräume mit $\mathcal{X}_1 \subseteq \mathcal{X}_2$. Weiter sei $\Pi : \mathcal{X}_2 \to \mathcal{X}_1$ eine lineare Projektion mit $|\Pi v|_{H^1(\Omega)} \lesssim |v|_{H^1(\Omega)}$ für alle $v \in \mathcal{X}_2$. Dann gilt*

$$|u_2 - u_1|_{H^1(\Omega)} \simeq |u_2 - \Pi u_2|_{H^1(\Omega)} \tag{4.19}$$

mit den Lösungen u_1, u_2 von (4.3) auf $\mathcal{X}_1$ und $\mathcal{X}_2$.

Beweis. Da u_1 die Bestapproximation von u_2 in $\mathcal{X}_1$ bezüglich $\|\!|\cdot|\!\|$ ist, gilt mit der Normäquivalenz aus Lemma 4.9

$$|u_2 - u_1|_{H^1(\Omega)} \simeq \|\!|u_2 - u_1|\!\| \leq \|\!|u_2 - \Pi u_2|\!\| \simeq |u_2 - \Pi u_2|_{H^1(\Omega)} . \tag{4.20}$$

Für die andere Richtung erhalten wir mit Linearität, H^1-Stetigkeit und Projektionseigenschaft von Π

$$\begin{aligned} |u_2 - \Pi u_2|_{H^1(\Omega)} &\leq |u_2 - u_1|_{H^1(\Omega)} + |u_1 - \Pi u_2|_{H^1(\Omega)} && (4.21)\\ &= |u_2 - u_1|_{H^1(\Omega)} + |\Pi(u_1 - u_2)|_{H^1(\Omega)} \lesssim |u_2 - u_1|_{H^1(\Omega)} . && (4.22) \end{aligned}$$

Beide Abschätzungen zusammengenommen ergeben die Behauptung. □

Das zweite Lemma kombiniert für einen Lower Diamond die Transferoperatoren auf jedem Ast zu einem Transferoperator zwischen den Spitzen der Diamantstruktur (siehe auch Abbildung 3.1). Hierbei ist wichtig, dass die Transferoperatoren jeweils die Menge der vergröberten Dreiecke berücksichtigen und diese Mengen in einem Lower Diamond disjunkt sind.

Lemma 4.21. *Sei $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_1, \ldots, \mathcal{T}_m)$ ein Lower Diamond und $p \in \mathbb{N}$. Seien weiters $\mathcal{Q}_i := \mathcal{Q}_{\mathcal{T}_i}^{\mathcal{T}_\vee}$ und $\Omega_i := \Omega(\mathcal{T}_i \setminus \mathcal{T}_\vee)$ für $i = 1, \ldots, m$. Dann kommutieren die $\mathcal{Q}_i$ paarweise und der Operator $\mathcal{Q} := \mathcal{Q}_1 \circ \ldots \circ \mathcal{Q}_m$ bildet $\mathcal{S}^p_D(\mathcal{T}_\vee)$ auf $\mathcal{S}^p_D(\mathcal{T}^\wedge)$ ab.*

Weiter gibt es eine Konstante $C > 0$, die nur von $\mathcal{T}^\wedge$ und p abhängt, sodass die Stabilitätsabschätzung

$$|\mathcal{Q}v|_{H^1(\Omega)} \leq C\, |v|_{H^1(\Omega)}$$

für alle $v \in \mathcal{S}^p(\mathcal{T}_\vee)$ gilt. Der Operator $\mathcal{Q}$ lässt sich auf den Ω_i folgendermaßen darstellen:

$$\mathcal{Q}v = \begin{cases} \mathcal{Q}_i v & \text{auf } \Omega_i, \\ v & \text{auf } \Omega \setminus \bigcup_{i=1}^m \Omega_i. \end{cases} \tag{4.23}$$

Beweis. Wir weisen zunächst die Kommutatorrelation nach. Dazu überlegen wir uns als Erstes, dass die Verkettung von zwei Projektionen $\mathcal{Q}_i$ wohldefiniert ist. Dies folgt mit

$\mathcal{T}_i \leq \mathcal{T}_\vee$ für alle $i = 1, \dots, m$, da laut Proposition 4.18 (iii) die $\mathcal{Q}_i$ den Raum $\mathcal{S}^p_D(\mathcal{T}_\vee)$ auf $\mathcal{S}^p_D(\mathcal{T}_i) \subseteq \mathcal{S}^p_D(\mathcal{T}_\vee)$ abbilden.

Nach Proposition 4.18 (ii) ist der Transferoperator $\mathcal{Q}_i$ auf Ω_i^c die Identität. Für fixierte Indizes $i \neq j$ betrachten wir daher zwei Fälle, nämlich die Einschränkung der Operatoren auf eine Teilmenge $\omega \subseteq \Omega$ der Bauart Ω_i^c.

Fall 1: $\omega = (\Omega_i \cup \Omega_j)^\mathsf{c}$.
Hier wirken sowohl $\mathcal{Q}_i$ als auch $\mathcal{Q}_j$ wie die Identität. Damit gilt

$$\mathcal{Q}_i\mathcal{Q}_j v = \mathcal{Q}_j v = v = \mathcal{Q}_i v = \mathcal{Q}_j\mathcal{Q}_i v \quad \text{auf } \omega = (\Omega_i \cup \Omega_j)^\mathsf{c}.$$

Fall 2: $\omega = \Omega_i$.
Da $\Omega_i \subseteq \operatorname{int}(\Omega_j)^\mathsf{c}$ wirkt hier nur $\mathcal{Q}_j$ als Identität. Mit Proposition 4.18 (i), wonach die Werte von $(\mathcal{Q}_i v)|_{\Omega_i}$ nur von $v|_{\Omega_i} = (\mathcal{Q}_j v)_{\Omega_i}$ abhängen, folgt

$$\mathcal{Q}_i\mathcal{Q}_j v = \mathcal{Q}_i v = \mathcal{Q}_j\mathcal{Q}_i v \quad \text{auf } \omega = \Omega_i.$$

Wendet man die Überlegungen aus den eben genannten beiden Fällen auf die Verkettung $\mathcal{Q}$ von m verschiedenen solchen Transferoperatoren an, folgt sofort, dass das Bild von $\mathcal{S}^p_D(\mathcal{T}_\vee)$ die Menge

$$\bigcap_{i=1}^m \mathcal{S}^p_D(\mathcal{T}_i) = \mathcal{S}^p_D\left(\bigwedge_{i=1}^m \mathcal{T}_i\right) = \mathcal{S}^p_D(\mathcal{T}^\wedge)$$

ist. Außerdem folgt Gleichung (4.23), da die Ω_i paarweise disjunkt sind.

Es bleibt noch die Stabilitätsabschätzung nachzuweisen. Diese folgt wiederum aus der Disjunktheit der Ω_i und Proposition 4.13 (ii), wobei Proposition 4.18 (i) garantiert, dass diese Abschätzung auch für die Ω_i gilt. Für $v \in \mathcal{S}^p(\mathcal{T}_\vee)$ erhalten wir damit

$$\begin{aligned}
|\mathcal{Q}v|^2_{H^1(\Omega)} &= |\mathcal{Q}v|^2_{H^1(\Omega\setminus\bigcup_{i=1}^m \Omega_i)} + \sum_{i=1}^m |\mathcal{Q}v|^2_{H^1(\Omega_i)} \\
&\overset{(4.23)}{=} |v|^2_{H^1(\Omega\setminus\bigcup_{i=1}^m \Omega_i)} + \sum_{i=1}^m |\mathcal{Q}_i v|^2_{H^1(\Omega_i)} \\
&\overset{4.13(ii)}{\lesssim} |v|^2_{H^1(\Omega\setminus\bigcup_{i=1}^m \Omega_i)} + \sum_{i=1}^m |v|^2_{H^1(\Omega_i)} = |v|^2_{H^1(\Omega)}\,.
\end{aligned}$$

Damit sind alle Behauptungen gezeigt. □

Mit diesen Hilfsmitteln können wir eine Abschätzung zeigen, die uns über Proposition 4.7 die Lower Diamond Estimate für $\mathfrak{E}$ liefert.

Satz 4.22. *Sei $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_1, \dots, \mathcal{T}_m)$ ein Lower Diamond und $u_\wedge$, $u_\vee$ und u_i die Lösungen von (4.3) auf $\mathcal{T}^\wedge$, $\mathcal{T}_\vee$ und $\mathcal{T}_i$. Dann gilt*

$$|u_\vee - u_\wedge|_{H^1(\Omega)} \simeq \sum_{i=1}^m |u_\vee - u_i|_{H^1(\Omega)}\,. \tag{4.24}$$

Beweis. Für $i = 1, \dots, m$ seien Ω_i und $\mathcal{Q}_i$ definiert wie in Lemma 4.21. Benutzen wir Lemma 4.20, die Eigenschaft (4.23) für $\mathcal{Q}$ und $\mathcal{Q}_i u_\vee = u_\vee$ auf Ω_i^c erhalten wir

$$\begin{aligned} |u_\vee - u_\wedge|^2_{H^1(\Omega)} &\overset{(4.19)}{\simeq} |u_\vee - \mathcal{Q} u_\vee|^2_{H^1(\Omega)} \overset{(4.23)}{=} \sum_{i=1}^m |u_\vee - \mathcal{Q}_i u_\vee|^2_{H^1(\Omega_i)} \\ &= \sum_{i=1}^m |u_\vee - \mathcal{Q}_i u_\vee|^2_{H^1(\Omega)} \overset{(4.19)}{\simeq} \sum_{i=1}^m |u_\vee - u_i|^2_{H^1(\Omega)}\,, \end{aligned}$$

was zu zeigen war. □

Korollar 4.23. *Es gelten die Voraussetzungen von Satz 4.22. Dann erfüllt die Energie $\mathscr{E}$ aus Definition 4.5 die Lower Diamond Estimate*

$$\mathscr{E}(\mathcal{T}^\wedge) - \mathscr{E}(\mathcal{T}_\vee) \simeq \sum_{i=1}^m \big(\mathscr{E}(\mathcal{T}_i) - \mathscr{E}(\mathcal{T}_\vee)\big). \tag{4.25}$$

Beweis. Dies folgt aus Satz 4.22 mit der Äquivalenz von Energienorm und H^1-Seminorm aus Lemma 4.9 und Proposition 4.7. Betrachtet man die linke Seite von (4.24) folgt

$$\tfrac{1}{2}\,|u_\wedge - u_\vee|^2_{H^1(\Omega)} \overset{(4.11)}{\simeq} \tfrac{1}{2}\,\|\!|u_\wedge - u_\vee|\!\|^2 \overset{(4.6)}{=} \mathscr{E}(\mathcal{T}^\wedge) - \mathscr{E}(\mathcal{T}_\vee). \tag{4.26}$$

Für die andere Seite gilt eine analoge Äquivalenz, was die Behauptung zeigt. □

4.3 Gesamtenergie

Die in Definition 4.5 definierte Energie passt noch nicht in den allgemeinen Rahmen von Kapitel 3, da diskrete lokale Effizienz und Zuverlässigkeit zum Residualschätzer, den wir im nächsten Abschnitt betrachten wollen, nicht gezeigt werden kann. Wir führen daher eine leicht modifizierte *Gesamtenergie* ein, die diese Eigenschaften erfüllt. Dies hat allerdings den Nachteil, dass wir auch hier wieder die Lower Diamond Estimate nachweisen müssen.

Definition 4.24 (Oszillationen)**.** *Sei $T \in \mathcal{T} \in \mathbb{T}$ ein Dreieck eines Gitters, $E \in \mathcal{E}$ eine Kante und $p \in \mathbb{N}$. Weiter seien Π_T, Π_E die L^2-Orthogonalprojektionen auf den Raum der Polynome vom Grad $p-2$ auf T, beziehungsweise Polynome vom Grad $p-1$ auf E. Für ein Problem der Bauart (4.1) definieren wir als* (Daten-)Oszillationen*, beziehungsweise* Neumann-Oszillationen

$$\begin{aligned} \operatorname{osc}^2(T) &:= \begin{cases} h_T^2\,\|f\|^2_{L^2(T)} & \text{falls } p = 1, \\ h_T^2\,\|(1-\Pi_T)f\|^2_{L^2(T)} & \text{falls } p \geq 2. \end{cases} \\ \operatorname{osc}_N^2(T) &:= \sum_{E \in \mathcal{E}^N(T)} h_T\,\|(1-\Pi_E)\phi\|^2_{L^2(E)}\,. \end{aligned}$$

Für eine Menge $\mathcal{U} \subseteq \mathcal{T}$ von Dreiecken definieren wir die entsprechenden Größen als quadratische Summe $\operatorname{osc}(\mathcal{U}) := (\sum_{T\in\mathcal{U}} \operatorname{osc}^2(T))^{1/2}$ *und* $\operatorname{osc}_N(\mathcal{U}) := (\sum_{T\in\mathcal{U}} \operatorname{osc}_N^2(T))^{1/2}$.

Definition 4.25 (Gesamtenergie)**.** *Sei $\mathscr{E}$ die Energie aus Definition 3.3 und $p \in \mathbb{N}$. Wir definieren die* Gesamtenergie $\mathscr{G} : \mathbb{T} \to \mathbb{R}$ *als*

$$\mathscr{G}(\mathcal{T}) := \mathscr{E}(\mathcal{T}) + \operatorname{osc}_N^2(\mathcal{T}) + \operatorname{osc}^2(\mathcal{T}).$$

Bemerkung 4.26. Es ist die Gesamtenergie eine Energie im Sinne von Definition 3.3. Um dies einzusehen, müssen wir wieder Monotonie zeigen. Wir betrachten dazu zwei Gitter $\mathcal{T} \leq \mathcal{T}'$. Laut Bemerkung 4.8 gilt die Monotonie bereits für $\mathscr{E}$. Für die Oszillationsterme folgt dies aus lokalen Abschätzungen:

Sehen wir uns zunächst den Term $\operatorname{osc}(\mathcal{T})$ für $p = 1$ an. Für $T \in \mathcal{T}$ gilt

$$\begin{aligned} \operatorname{osc}^2(T) = h_T^2 \, \|f\|_{L^2(T)}^2 &= \sum_{T'\in\mathcal{T}',\, |T\cap T'|\neq 0} h_T^2 \, \|f\|_{L^2(T')}^2 \\ &\geq \sum_{T'\in\mathcal{T}',\, |T\cap T'|\neq 0} h_{T'}^2 \, \|f\|_{L^2(T')}^2 = \operatorname{osc}^2\big(\{T' \in \mathcal{T}' \mid |\neq| 0\}\big). \end{aligned} \tag{4.27}$$

Durch Vereinigung über alle $T \in \mathcal{T}$ erhält man die gewünschte Beziehung.

Für $p \geq 2$ folgt dies auf ähnliche Weise. Für $T \in \mathcal{T}$ gilt aufgrund der Bestapproximationseigenschaft von Π_T und $\Pi_{T'}$

$$\|(1-\Pi_T)f\|_{L^2(T)}^2 = \sum_{T'\in\mathcal{T}',\, |T\cap T'|\neq 0} \|(1-\Pi_T)f\|_{L^2(T')}^2 \geq \sum_{T'\in\mathcal{T}',\, |T\cap T'|\neq 0} \|(1-\Pi_{T'})f\|_{L^2(T')}^2 .$$

Zusammen mit obiger Abschätzung ergibt sich die Monotonie für osc und analog auch für osc_N. //

4.3.1 Lower Diamond Estimate

Wie bereits erwähnt, müssen wir auch für die Gesamtenergie die Lower Diamond Estimate nachweisen. Dafür benötigen wir ein Hilfsresultat.

Lemma 4.27. *Seien $\mathcal{T} \leq \mathcal{T}'$ zwei Gitter in $\mathbb{T}$. Dann gilt*

$$\begin{aligned} \operatorname{osc}^2(\mathcal{T}) - \operatorname{osc}^2(\mathcal{T}') &\simeq \operatorname{osc}^2(\mathcal{T}\setminus\mathcal{T}'), \\ \operatorname{osc}_N^2(\mathcal{T}) - \operatorname{osc}_N^2(\mathcal{T}') &\simeq \operatorname{osc}_N^2(\mathcal{T}\setminus\mathcal{T}'). \end{aligned}$$

Beweis. Wir beginnen mit der Abschätzung für osc und $p = 1$. Da wir die Gitterweite h_T als $|T|^{1/2}$ definiert haben, reduziert jede Bisektion h_T^2 um den Faktor $1/2$. Jedes Dreieck

$T \in \mathcal{T} \setminus \mathcal{T}'$ wird beim Übergang auf $\mathcal{T}$ zumindest einmal verfeinert, weshalb $h_T^2 \geq 2h_{T'}^2$ für alle $T' \in \mathcal{T}'$ mit $|T \cap T'| \neq 0$ gilt. Eine Rechnung analog zu (4.27) liefert die Abschätzung

$$\operatorname{osc}^2(\mathcal{T}' \setminus \mathcal{T}) \leq \tfrac{1}{2}\operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}').$$

Kombiniert man diese mit der Identität

$$\operatorname{osc}^2(\mathcal{T}) - \operatorname{osc}^2(\mathcal{T}') = \sum_{T\in\mathcal{T}} h_T^2 \|f\|_{L^2(T)}^2 - \sum_{T'\in\mathcal{T}'} h_{T'}^2 \|f\|_{L^2(T')}^2 = \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}') - \operatorname{osc}^2(\mathcal{T}' \setminus \mathcal{T}),$$

erhält man schließlich

$$\operatorname{osc}^2(\mathcal{T}) - \operatorname{osc}^2(\mathcal{T}') \leq \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}'),$$

sowie

$$\operatorname{osc}^2(\mathcal{T}) - \operatorname{osc}^2(\mathcal{T}') \geq \tfrac{1}{2}\operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}').$$

Wie in Bemerkung 4.26 folgt auch die Abschätzung für osc_N aufgrund von $h_T \geq \sqrt{2}h_{T'}$ analog mit Konstante $(\sqrt{2}-1)/\sqrt{2}$ anstatt $1/2$. □

Wir können nun die Lower Diamond Estimate für die Gesamtenergie $\mathcal{G}$ nachweisen.

Satz 4.28. *Sei $(\mathcal{T}^\wedge, \mathcal{T}_\vee; \mathcal{T}_1, \ldots, \mathcal{T}_m)$ ein Lower Diamond und $\mathcal{G}$ die Energie aus Definition 4.25. Dann erfüllt diese Energie die Lower Diamond Estimate*

$$\mathcal{G}(\mathcal{T}^\wedge) - \mathcal{G}(\mathcal{T}_\vee) \simeq \sum_{i=1}^{m} \big(\mathcal{G}(\mathcal{T}_i) - \mathcal{G}(\mathcal{T}_\vee)\big). \tag{4.28}$$

Beweis. Die Gesamtenergie hat die Struktur $\mathcal{G}(\mathcal{T}) = \mathcal{E}(\mathcal{T}) + R(\mathcal{T})$, wobei R ein Term ist, der nach der Definition der Oszillationen und Lemma 4.27 die beiden Eigenschaften

$$\begin{aligned} &R(\mathcal{T}) - R(\mathcal{T}') \simeq R(\mathcal{T} \setminus \mathcal{T}') \quad \text{für } \mathcal{T}' \geq \mathcal{T} \text{ und} \\ &R(\mathcal{U}_1) + R(\mathcal{U}_2) = R(\mathcal{U}_1 \cup \mathcal{U}_2) \text{ für disjunkte Teilmengen } \mathcal{U}_1, \mathcal{U}_2 \subseteq \mathcal{T} \end{aligned}$$

erfüllt. Für die Energie $\mathcal{E}$ haben wir die Lower Diamond Estimate schon in Korollar 4.23 gesehen, es bleibt noch die Abschätzung für R zu zeigen. Diese folgt aus obigen Eigenschaften. Da die Flächen $\Omega(\mathcal{T}_i \setminus \mathcal{T}_\vee)$ für $i = 1, \ldots, m$ nach Definition disjunkt sind, gilt $\mathcal{T}^\wedge \setminus \mathcal{T}_\vee = \bigcup_{i=1}^m (\mathcal{T}_i \setminus \mathcal{T}_\vee)$ und damit

$$R(\mathcal{T}^\wedge) - R(\mathcal{T}_\vee) \simeq R(\mathcal{T}^\wedge \setminus \mathcal{T}_\vee) = R\Big(\bigcup_{i=1}^{m} (\mathcal{T}_i \setminus \mathcal{T}_\vee)\Big) = \sum_{i=1}^{m} R(\mathcal{T}_i \setminus \mathcal{T}_\vee) \simeq \sum_{i=1}^{m} \big(R(\mathcal{T}_i) - R(\mathcal{T}_\vee)\big).$$

Damit ist die Behauptung gezeigt. □

4.4 Residualschätzer

Wir wollen in diesem Abschnitt einen Fehlerschätzer für unser Problem einführen. Für diesen müssen wir die diskreten lokalen Abschätzungen aus Definition 3.8 zeigen. Dafür definieren wir zuerst Kantensprünge einer unstetigen Funktion [Ver13, Abschnitt 1.3.5].

Definition 4.29. *Für ein Dreieck $T \in \mathcal{T} \in \mathbb{T}$ bezeichnen wir den* äußeren normierten Normalvektor *mit $\nu_T(x)$, wobei $x \in \partial T$. Auf einer Kante $E \in \mathcal{E}$ definieren wir für eine Funktion $v \in [P^p(\mathcal{T})]^2$ mit $p \in \mathbb{N}$ weiter den* (Normal-)Sprung *von v als*

$$[\![v]\!](x) := \lim_{t \to 0+} (\nu_T \cdot v)(x - t\nu_T) - \lim_{t \to 0+} (\nu_T \cdot v)(x + t\nu_T),$$

wobei $x \in E$ und T so, dass $E \subseteq \partial T$.

Wir bedienen uns des Kantensprunges bei der folgenden Definition des Residualschätzers [Fei, section 5.4].

Definition 4.30 (Residualschätzer)**.** *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und A, f und ϕ die Daten aus (4.1). Sei weiter $u_\mathcal{T}$ die Lösung des diskreten Problems (4.3) auf $\mathcal{T}$. Wir definieren das* Residuum $\operatorname{res}_\mathcal{T} : \mathcal{T} \cup \mathcal{E} \to \mathbb{R}$ *als*

$$\operatorname{res}_\mathcal{T}(T) := h_T \left\| \operatorname{div}(A\nabla u_\mathcal{T}) + f \right\|_{L^2(T)},$$

$$\operatorname{res}_\mathcal{T}(E) := \begin{cases} h_E^{1/2} \left\| [\![A\nabla u_\mathcal{T}]\!] \right\|_{L^2(E)} & \text{für } E \subseteq \Omega, \\ h_E^{1/2} \left\| \phi - A\nabla u_\mathcal{T} \cdot \nu_T \right\|_{L^2(E)} & \text{für } E \in \mathcal{E}^N, \\ 0 & \text{für } E \in \mathcal{E}^D. \end{cases}$$

Damit können wir für $E \in \mathcal{E}$ den kantenbasierten Residualschätzer *als*

$$\eta_\mathcal{T}^2(E) := \operatorname{res}_\mathcal{T}^2(E) + \sum_{T \in \omega_E^{\mathrm{red}}} \operatorname{res}_\mathcal{T}^2(T)$$

definieren. Für eine Menge $\mathcal{U} \subseteq \mathcal{T} \cup \mathcal{E}$ definieren wir außerdem das Residuum als quadratische Summe $\operatorname{res}_\mathcal{T}(\mathcal{U}) := \left(\sum_{U \in \mathcal{U}} \operatorname{res}_\mathcal{T}^2(U) \right)^{1/2}$.

Bemerkung 4.31. Beachte, dass in der Definition 4.25 der Gesamtenergie diese für $p = 1$ als Summe der Energie des Variationsproblems $\mathscr{E}$ und des Residuums $\operatorname{res}_\mathcal{T}^2(\mathcal{T})$ definiert ist (der Term $\operatorname{div}(A\nabla u_\mathcal{T})$ verschwindet für $u_\mathcal{T} \in \mathcal{S}^1_D(\mathcal{T})$). Für $p \geq 2$ jedoch ist die Gesamtenergie definiert als Summe von Energie und Oszillationen. Dies hat den Grund, dass für $p \geq 2$ das elementweise Residuum durch die Oszillationen abschätzbar ist, wie wir im nächsten Abschnitt sehen werden, für $p = 1$ jedoch nicht. Im Fall $p \geq 2$ ist das Residuum jedoch nicht zwingend monoton, da der Term $\operatorname{div}(A\nabla u_\mathcal{T})$ nicht mehr verschwindet, und deshalb nicht als Energie geeignet. //

Bemerkung 4.32. Im Großteil der Literatur zu adaptiven Finite Elemente Methoden werden Fehlerschätzer elementbasiert definiert. Wir definieren den Residualschätzer hier wie in [DKS16], also kantenbasiert. Dies hat folgende Gründe.

Wird anstatt der Kante ein Dreieck für die Bisektion markiert, so kann die a posteriori Analysis nicht berücksichtigen, welche Kante des Dreiecks verfeinert wird. Somit müssen auch Dreiecke und Kanten in diese Analysis einfließen, die gar nicht verfeinert werden. Damit können die diskreten lokalen Abschätzungen (3.8) nicht gelten.

Behebt man dieses Problem, indem alle Kanten eines markierten Dreiecks verfeinert werden (wie etwa bei der am weitesten rechts abgebildeten Verfeinerung in Abbildung 2.1), so gibt es zwei Möglichkeiten, den Gitterabschluss zu bilden. Die erste hiervon ist, einfach benachbarte Dreiecke zu markieren und auf analoge Weise zu verfeinern. Dies würde sich jedoch auf alle Dreiecke eines Gitters ausbreiten, was einer uniformen Verfeinerung entspricht. Die zweite Möglichkeit ist, den Gitterabschluss durch einfache Bisektion zu bilden. Dabei ist aber nicht klar, ob das Markierungskriterium (E1) erfüllbar ist, da kantenbasierte und elementbasierte Fehlerschätzer nur schwer in Verbindung zu bringen sind. //

4.4.1 Diskrete lokale Abschätzungen

Wir machen uns daran, die benötigten diskreten lokalen Abschätzungen aus Definition 3.8 zu beweisen. Diese zeigen wir jedoch zuerst für den sogenannten *totalen Fehler*

$$\|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!\|^2 + \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}') + \operatorname{osc}_N^2(\mathcal{T} \setminus \mathcal{T}'),$$

wobei $\mathcal{T} \leq \mathcal{T}'$ Gitter sind und $u_{\mathcal{T}}, u_{\mathcal{T}'}$ die Lösungen von (4.3) auf den beiden Gittern bezeichnen. Wir werden später sehen, dass der totale Fehler bis auf eine Konstante gleich der Energiedifferenz der beiden Gitter ist. Der Beweis der diskreten lokalen Abschätzungen wird in mehreren Schritten geschehen, die teilweise die Resultate aus [DKS16] auf $p > 1$ verallgemeinern und aus [EGP18] übernommen sind.

Als Erstes werden wir die diskrete lokale Zuverlässigkeit zeigen (siehe [EGP18, Lemma 6] und vergleiche mit [DKS16, Lemma 4.3]).

Satz 4.33. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $\mathcal{T}' \geq \mathcal{T}$ mit dazugehörigen Kantenmengen $\mathcal{E}$ und $\mathcal{E}'$. Seien weiter $p \in \mathbb{N}$ und $u_{\mathcal{T}} \in \mathcal{S}^p_D(\mathcal{T})$ und $u_{\mathcal{T}'} \in \mathcal{S}^p_D(\mathcal{T}')$ die Lösungen von (4.3) auf $\mathcal{T}$ und $\mathcal{T}'$. Dann existiert eine Konstante $C_{\mathrm{rel}} > 0$, sodass*

$$\|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!\|^2 \leq C_{\mathrm{rel}} \eta^2_{\mathcal{T}}(\mathcal{E} \setminus \mathcal{E}'). \tag{4.29}$$

Die Konstante C_{rel} hängt nur von $\mathcal{T}_0$, p und der Diffusionsmatrix A ab.

Beweis. Für $v \in \mathcal{S}^p_D(\mathcal{T}) \subseteq \mathcal{S}^p_D(\mathcal{T}')$ gilt wegen der schwachen Formulierung (4.3) die Galerkin-Orthogonalität:

$$\int_\Omega A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) \cdot \nabla v \,\mathrm{d}x = \int_\Omega (f - f)v \,\mathrm{d}x + \int_{\Gamma_N} (\phi - \phi)v \,\mathrm{d}s = 0. \tag{4.30}$$

Es bezeichne nun $\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'}$ den Transferoperator von $\mathcal{S}^p_D(\mathcal{T}')$ nach $\mathcal{S}^p_D(\mathcal{T})$ aus Definition 4.16. Wir definieren $v := (1 - \mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'})(u_{\mathcal{T}'} - u_{\mathcal{T}})$. Wegen Proposition 4.18 (iii) gilt $\mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'}(u_{\mathcal{T}'} - u_{\mathcal{T}}) \in$

$\mathcal{S}^p_D(\mathcal{T})$, weshalb mit Galerkin-Orthogonalität und der schwachen Formulierung gilt, dass

$$\begin{aligned}
\|\!|u_{\mathcal{T}'} - u_{\mathcal{T}}|\!\|^2 &= \int_\Omega A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) \cdot \nabla(u_{\mathcal{T}'} - u_{\mathcal{T}})\,\mathrm{d}x \\
&\overset{(4.30)}{=} \int_\Omega A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) \cdot \big(\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) - \nabla\mathcal{Q}^{\mathcal{T}'}_{\mathcal{T}}(u_{\mathcal{T}'} - u_{\mathcal{T}})\big)\,\mathrm{d}x \\
&= \int_\Omega A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) \cdot \nabla v\,\mathrm{d}x \\
&\overset{(4.3)}{=} \int_\Omega fv\,\mathrm{d}x + \int_{\Gamma_N} \phi v\,\mathrm{d}s - \int_\Omega A\nabla u_{\mathcal{T}} \cdot \nabla v\,\mathrm{d}x.
\end{aligned}$$

Integrieren wir den letzten Term des letzten Ausdrucks partiell auf jedem Dreieck in $\mathcal{T}$, erhalten wir

$$-\int_\Omega A\nabla u_{\mathcal{T}} \cdot \nabla v\,\mathrm{d}x = \sum_{T\in\mathcal{T}} \left(\int_T \operatorname{div}(A\nabla u_{\mathcal{T}}) v\,\mathrm{d}x - \int_{\partial T\setminus\Gamma_D} A\nabla u_{\mathcal{T}} \cdot \nu v\,\mathrm{d}s \right).$$

Beachtet man, dass wegen Proposition 4.18 (ii) $v = 0$ auf $\mathcal{T}\cap\mathcal{T}'$ gilt, ergibt sich durch Umordnen der Kanten und Zusammenfassen der Integrale über gleiche Mengen

$$\begin{aligned}
\|\!|u_{\mathcal{T}'} - u_{\mathcal{T}}|\!\|^2 = &\sum_{T\in\mathcal{T}\setminus\mathcal{T}'} \int_T \big(f + \operatorname{div}(A\nabla u_{\mathcal{T}})\big) v\,\mathrm{d}x \\
&+ \sum_{\substack{E\in\mathcal{E}\setminus\mathcal{E}' \\ E\subseteq\Omega}} \int_E [\![A\nabla u_{\mathcal{T}}]\!]\, v\,\mathrm{d}s + \sum_{\substack{E\in\mathcal{E}\setminus\mathcal{E}' \\ E\subseteq\Gamma_N}} \int_E \big(\phi - A\nabla u_{\mathcal{T}}\cdot\nu\big) v\,\mathrm{d}s.
\end{aligned} \tag{4.31}$$

Die drei hier auftretenden Summanden bezeichnen wir der Reihe nach mit B_1, B_2 und B_3.

B_1 : Wir wenden hier die Cauchy-Schwarz Ungleichung einerseits in der kontinuierlichen Version elementweise auf die Integrale an und andererseits in der diskreten Form auf die Summe. Verwenden wir zuerst die kontinuierliche Version und schätzen die entstehende L^2-Norm von v mit Proposition 4.13 (iii) weiter ab, so erhalten wir

$$\begin{aligned}
\sum_{T\in\mathcal{T}\setminus\mathcal{T}'} \int_T \big(f + \operatorname{div}(A\nabla u_{\mathcal{T}})\big) v\,\mathrm{d}x &\le \sum_{T\in\mathcal{T}\setminus\mathcal{T}'} \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\, \|v\|_{L^2(T)} \\
&\overset{4.13}{\lesssim} \sum_{T\in\mathcal{T}\setminus\mathcal{T}'} h_T\, \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\, |u_{\mathcal{T}'} - u_{\mathcal{T}}|_{H^1(\omega_T)}\,.
\end{aligned}$$

Die diskrete Cauchy-Schwarz Ungleichung und die Normäquivalenz aus Lemma 4.9 können wir nun verwenden, um mit der Definition des Elementresiduums folgende Abschätzung zu

erhalten:

$$\begin{aligned} B_1 &\leq \sum_{T\in\mathcal{T}\setminus\mathcal{T}'} h_T \left\|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\right\|_{L^2(T)} |u_{\mathcal{T}'} - u_{\mathcal{T}}|_{H^1(\omega_T)} \\ &\leq \left(\sum_{T\in\mathcal{T}\setminus\mathcal{T}'} h_T^2 \left\|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\right\|_{L^2(T)}^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}\setminus\mathcal{T}'} |u_{\mathcal{T}'} - u_{\mathcal{T}}|_{H^1(\omega_T)}^2\right)^{1/2} \\ &\overset{(*)}{\lesssim} \left(\sum_{T\in\mathcal{T}\setminus\mathcal{T}'} h_T^2 \left\|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\right\|_{L^2(T)}^2\right)^{1/2} |u_{\mathcal{T}'} - u_{\mathcal{T}}|_{H^1(\Omega)} \\ &\overset{(4.11)}{\lesssim} \operatorname{res}_{\mathcal{T}}(\mathcal{T}\setminus\mathcal{T}') \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| . \end{aligned}$$

Hier wurde für $(*)$ außerdem benutzt, dass wegen der Formregularität von $\mathcal{T}$ die Anzahl der Dreiecke im Patch ω_T für jedes T durch eine uniforme Konstante beschränkt ist, die nur von $\mathcal{T}_0$ abhängt.

B_2 : Wir verwenden dieselbe Technik, wie für den Term B_1. Zuerst wenden wir die Cauchy-Schwarz Ungleichung in der kontinuierlichen Version an. Danach schätzen wir jedoch die auftretenden L^2-Normen mit Punkt (iv) aus Proposition 4.13 ab, wodurch wir

$$\|v\|_{L^2(E)} \overset{4.13}{\lesssim} h_E^{1/2} \|u_{\mathcal{T}'} - u_{\mathcal{T}}\|_{L^2(\omega_E)}$$

erhalten. Normäquivalenz und die Definition des Kantenresiduums liefern schließlich

$$B_2 \lesssim \operatorname{res}_{\mathcal{T}}\big(\{E \in \mathcal{E}\setminus\mathcal{E}' \mid E \subseteq \Omega\}\big) \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| .$$

B_3 : Dieser Term lässt sich analog zu B_2 abschätzen und liefert

$$B_3 \lesssim \operatorname{res}_{\mathcal{T}}\big(\{E \in \mathcal{E}\setminus\mathcal{E}' \mid E \subseteq \Gamma_N\}\big) \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| .$$

Insgesamt erhalten wir durch die drei eben betrachteten Terme, wegen der Äquivalenz der Normen $\sum|\cdot|$ und $(\sum|\cdot|^2)^{1/2}$ am $\mathbb{R}^n$ und der Definition des Residualschätzers in Definition 4.30, dass

$$\begin{aligned} \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right|^2 &\lesssim \big(\operatorname{res}_{\mathcal{T}}(\mathcal{T}\setminus\mathcal{T}') + \operatorname{res}_{\mathcal{T}}\big(\{E \in \mathcal{E}\setminus\mathcal{E}' \mid E \subseteq \Omega\}\big) \\ &\qquad + \operatorname{res}_{\mathcal{T}}\big(\{E \in \mathcal{E}\setminus\mathcal{E}' \mid E \subseteq \Gamma_N\}\big)\big) \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| \\ &\lesssim \big(\operatorname{res}_{\mathcal{T}}(\mathcal{T}\setminus\mathcal{T}') + \operatorname{res}_{\mathcal{T}}(\mathcal{E}\setminus\mathcal{E}')\big) \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| \\ &\lesssim \eta_{\mathcal{T}}^2(\mathcal{E}\setminus\mathcal{E}') \left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right| . \end{aligned}$$

Durch Division durch den Faktor $\left|\!\left|\!\left| u_{\mathcal{T}'} - u_{\mathcal{T}} \right|\!\right|\!\right|$ haben wir die Behauptung gezeigt. □

Wir widmen uns nun der diskreten lokalen Effizienz. Der Beweis dieser Eigenschaft verwendet die Bubble Function Technik, wie sie etwa in [Ver13, Abschnitt 3.8] behandelt wird. Dazu definieren wir zwei Bubble Funktionen. Hier bezeichnet $\widehat{\phi}_z \in \mathcal{S}^1(\mathcal{T})$ für $z \in \mathcal{V}$ eine stückweise affine Hutfunktion, die $\widehat{\phi}_z(z') = \delta_{zz'}$ für alle $z' \in \mathcal{V}$ erfüllt. Für eine Kante $E \in \mathcal{E}$

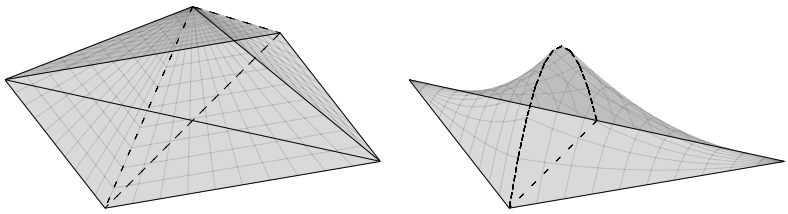


Abbildung 4.3: *Die Funktionen b_E (links) und b_T (rechts). Die durchgängigen Linien sind die Dreiecke auf denen die jeweilige Bubble-Funktion definiert ist, die strichlierten Linien geben die notwendigen Verfeinerungen dieser Dreiecke an.*

ist $\mathcal{T}' := \operatorname{refine}(\mathcal{T}; E)$ das gröbste Gitter, in dem E verfeinert wurde. Wir definieren die Bubble Funktion zu E als

$$b_E := \widehat{\phi}_{\operatorname{midpt}(E)} \in \mathcal{S}^1(\mathcal{T}'). \tag{4.32}$$

Für ein Element $T \in \mathcal{T}$ mit Referenzkante E und Newest Vertex z ist $\mathcal{T}' := \operatorname{refine}(\mathcal{T}; E)$ das gröbste Gitter, in dem T verfeinert wurde. Die Bubble Funktion für T definieren wir als

$$b_T := \widehat{\phi}_z \cdot \widehat{\phi}_{\operatorname{midpt}(E)} \in \mathcal{S}^2(\mathcal{T}'). \tag{4.33}$$

Diese beiden Funktionen sind in Abbildung 4.3 dargestellt.

Wir werden nun zwei Hilfsresultate über lokale Gitterweiten, beziehungsweise über mit diesen Bubble Funktionen gewichtete L^2-Normen zeigen.

Lemma 4.34. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $T \in \mathcal{T}$ und $E \in \mathcal{E}$ mit $E \in \partial T$. Dann gilt für die lokale Gitterweite*

$$C_{\mathrm{sr}}^{-1} h_T \leq h_E \leq C_{\mathrm{sr}} h_T.$$

Beweis. Wir beweisen zunächst die Ungleichung $h_T \lesssim h_E$. Der Flächeninhalt eines Dreiecks T lässt sich darstellen durch das Produkt der Längen der Kante E und der Höhe H_E auf diese Kante:

$$h_T^2 = |T| = h_E |H_E|.$$

Wir wissen außerdem aus Proposition 2.5, dass Netze, die durch NVB entstanden sind, formregulär sind. Da wegen der Dreiecksungleichung immer eine Kante des Dreiecks existiert, die länger als H_E ist, lässt sich damit der Flächeninhalt abschätzen zu

$$h_T^2 = h_E |H_E| \leq h_E \operatorname{diam}(T) \overset{(2.1)}{\leq} C_{\mathrm{sr}} h_E h_T.$$

Durch Division durch h_T folgt die erste Richtung der Ungleichung.

Die andere Richtung folgt aus analogen Überlegungen wie für die Höhe auf eine Kante. Da jede Kante höchstens so lang ist, wie der Elementdurchmesser, folgt aus der Formregularität

$$h_E \leq \operatorname{diam}(T) \overset{(2.1)}{\leq} C_{\mathrm{sr}} h_T.$$

Damit ist die Behauptung bewiesen. □

Lemma 4.35. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $T \in \mathcal{T}$, $E \in \mathcal{E}$ und $p \in \mathbb{N}$. Für alle Funktionen $\gamma \in P^p(E)$ gibt es eine Fortsetzung $\widehat{\gamma} \in P^p(\omega_E^{\mathrm{red}})$, sodass mit einer Konstanten $C_1 > 0$*

$$\widehat{\gamma}|_E = \gamma \quad \textit{und} \quad \|\widehat{\gamma}\|_{L^2(\omega_E^{\mathrm{red}})} \leq C_1 h_E^{1/2} \|\gamma\|_{L^2(E)}. \tag{4.34}$$

Außerdem gibt es Konstanten $C_2, C_3 > 0$, sodass für alle $\gamma \in P^p(E)$ und $v \in P^p(T)$

$$\|\gamma\|_{L^2(E)} \leq C_2 \left\| b_E^{1/2} \gamma \right\|_{L^2(E)} \quad \textit{und} \quad \|v\|_{L^2(T)} \leq C_3 \left\| b_T^{1/2} v \right\|_{L^2(T)} \tag{4.35}$$

gilt. Die Konstanten C_1, C_2, C_3 hängen dabei nur von $\mathcal{T}_0$ ab.

Beweis. Wir beginnen mit der Behauptung, dass eine Fortsetzung existiert, die (4.34) erfüllt. Dies folgt durch Transformation auf ein Referenzelement und ein Skalierungsargument. Da dies Standardtechniken in der Analysis der FEM sind, werden wir nicht allzu genau auf Transformation und Skalierung eingehen, sondern den Fokus auf die anderen Beweisschritte legen. Für $T \in \omega_E^{\mathrm{red}}$ transformieren wir das Dreieck durch eine affine Transformation F_T auf das Referenzelement $T_{\mathrm{ref}} := \operatorname{conv}\{(0,0),(0,1),(1,0)\}$. Auf dem Bild der betrachteten Kante $F_T(E)$ ist durch $\gamma \circ F_T^{-1}$ ein Polynom vom Grad p gegeben. Dieses können wir konstant in Normalenrichtung zu dieser Kante zu einem Polynom γ_{ref} vom Grad p auf dem gesamten Referenzelement fortsetzen. Mit $\widehat{\gamma}|_T := \gamma_{\mathrm{ref}} \circ F_T \in P^p(T)$ erhalten wir ein Polynom, dass die erste Eigenschaft von (4.34) erfüllt.

Für die zweite Eigenschaft verwenden wir ein Skalierungsargument. Wir können die Funktion γ_{ref} immer auf ein Rechteck Q fortsetzen, dessen eine Seite $F_T(E)$ ist und dessen andere Seite normal darauf steht und Länge 1 hat. Damit hat γ_{ref} auf Q Tensorstruktur (da es ja normal zu einer Seite des Rechtecks konstant fortgesetzt wurde) und es gilt

$$\|\gamma_{\mathrm{ref}}\|_{L^2(T_{\mathrm{ref}})}^2 = \int_{T_{\mathrm{ref}}} \gamma_{\mathrm{ref}}^2 \,\mathrm{d}x \leq \int_Q \gamma_{\mathrm{ref}}^2 \,\mathrm{d}x = \int_{F_T(E)} \gamma_{\mathrm{ref}}^2 \,\mathrm{d}s = \|\gamma_{\mathrm{ref}}\|_{L^2(F_T(E))}^2.$$

Durch Skalierung entsteht bei den L^2-Normen jeweils ein Faktor h_T für T und $h_E^{1/2}$ für E, sowie jeweils generische Konstanten, die nur vom Referenzelement abhängen. Insgesamt gilt damit

$$\|\widehat{\gamma}\|_{L^2(T)} \lesssim h_T h_E^{-1/2} \|\gamma\|_{L^2(E)}. \tag{4.36}$$

Wegen der Äquivalenz von Element- und Kantengröße aus Lemma 4.34 erhalten wir

$$h_T h_E^{-1/2} \lesssim h_E h_E^{-1/2} = h_E^{1/2}, \tag{4.37}$$

woraus zusammen mit (4.36) der zweite Teil von (4.34) folgt.

Für (4.35) sehen wir uns zunächst die Aussage für ein Dreieck $T \in \mathcal{T}$ an. Wir wenden wieder eine Transformation und ein Skalierungsargument an. Transformiert auf das Referenzelement T_{ref} gilt für die Bubble Funktion $b_{T_{\text{ref}}} := b_T \circ F_T^{-1}$, dass $0 \leq b_{T_{\text{ref}}} \leq 1$, wobei die Funktion nur am Rand verschwindet. Somit ist die gewichtete L^2-Norm

$$\|\cdot\|_{T_{\text{ref}}} := \left\| b_{T_{\text{ref}}}^{1/2} (\cdot) \right\|_{L^2(T_{\text{ref}})}$$

eine Norm auf $L^2(T_{\text{ref}})$ und insbesondere auch auf dem endlichdimensionalen Teilraum $P^p(T_{\text{ref}})$. Auf endlichdimensionalen Räumen sind aber alle Normen äquivalent, weshalb

$$\|\cdot\|_{T_{\text{ref}}} \simeq \|\cdot\|_{L^2(T_{\text{ref}})}$$

gilt, wobei die Konstanten hier nur von T_{ref} abhängen. Durch ein Skalierungsargument analog zu oben, bei dem wir diesmal auf beiden Seiten den Faktor h_T bekommen, gilt schließlich die Aussage für T in (4.35). Die Aussage für Kanten folgt analog. □

Wir benötigen noch eine inverse Abschätzung, um die Energienorm lokal durch die L^2-Norm abschätzen zu können.

Proposition 4.36. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $v \in \mathcal{S}_D^p(\mathcal{T})$. Dann existiert eine Konstante $C > 0$, sodass*

$$|\!|\!| v |\!|\!|_T \leq C h_T^{-1} \|v\|_{L^2(T)}, \tag{4.38}$$

wobei C nur von $\mathcal{T}_0$ und A abhängt.

Beweis. Aus der Normäquivalenz in Lemma 4.9 folgt

$$|\!|\!| v |\!|\!|_T \simeq |v|_{H^1(T)}.$$

Laut [Bra13, Satz II.6.8] gilt zudem, dass

$$|v|_{H^1(T)} \lesssim h_T^{-1} \|v\|_{L^2(T)},$$

womit insgesamt die Behauptung folgt. □

Diese Hilfsmittel reichen aus, um die diskrete lokale Effizienz zu zeigen. Dieses Resultat ist aus [EGP18, Lemma 9] entnommen und verallgemeinert [DKS16] auf beliebige Polynomgrade.

Satz 4.37. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $\mathcal{T}' \geq \mathcal{T}$ mit dazugehörigen Kantenmengen $\mathcal{E}$ und $\mathcal{E}'$. Seien weiter $p \in \mathbb{N}$ und $u_{\mathcal{T}} \in \mathcal{S}_D^p(\mathcal{T})$, sowie $u_{\mathcal{T}'} \in \mathcal{S}_D^p(\mathcal{T}')$ die Lösungen von (4.3) auf $\mathcal{T}$, respektive $\mathcal{T}'$. Dann gilt für $E \in \mathcal{E} \setminus \mathcal{E}'$, dass*

$$\operatorname{res}_{\mathcal{T}}(E) \lesssim |\!|\!| u_{\mathcal{T}} - u_{\mathcal{T}'} |\!|\!|_{\omega_E^{\text{red}}} + \operatorname{res}_{\mathcal{T}}(\omega_E^{\text{red}}) + \operatorname{osc}_N(\omega_E^{\text{red}}). \tag{4.39}$$

Falls $p \geq 2$, gilt für $T \in \mathcal{T} \setminus \mathcal{T}'$ außerdem, dass

$$\operatorname{res}_{\mathcal{T}}(T) \lesssim |||(u_{\mathcal{T}} - u_{\mathcal{T}'})|||_T + \operatorname{osc}(T). \tag{4.40}$$

Insgesamt gilt die diskrete lokale Effizienz: Es gibt eine Konstante $C_{\text{eff}} > 0$, sodass

$$\eta_{\mathcal{T}}^2(\mathcal{E} \setminus \mathcal{E}') \leq C_{\text{eff}}\big(|||u_{\mathcal{T}} - u_{\mathcal{T}'}|||^2 + \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}') + \operatorname{osc}_N^2(\mathcal{T} \setminus \mathcal{T}')\big). \tag{4.41}$$

Die auftretenden Konstanten und insbesondere C_{eff} hängen nur von den Problemdaten und $\mathcal{T}_0$ ab.

Beweis. Wir teilen den Beweis in mehrere Schritte.

Schritt 1: Abschätzung (4.39) für das Kantenresiduum.
Da am Dirichletrand laut Definition $\operatorname{res}_{\mathcal{T}}(E) = 0$ gilt, ist im Fall $E \in \mathcal{E}^D$ nichts zu zeigen und wir unterteilen die verbleibenden Beweisschritte in zwei Fälle.

Fall 1: (4.39) für $E \in \mathcal{E} \setminus \mathcal{E}'$, sodass $E \notin \mathcal{E}^\Gamma$.
Wir erinnern an die Bubble Funktion $b_E \in \mathcal{S}_D^1(\mathcal{T}')$. Laut Lemma 4.35 gibt es eine Funktion in $\mathcal{S}^{p-1}(\mathcal{T})$, die den Sprungterm $[\![A\nabla u_{\mathcal{T}}]\!]$ auf ω_E^{red} fortsetzt, wir werden diese der Einfachheit halber auch mit $[\![A\nabla u_{\mathcal{T}}]\!]$ bezeichnen. Definieren wir das Produkt $v := [\![A\nabla u_{\mathcal{T}}]\!]\, b_E \in \mathcal{S}_D^p(\mathcal{T}')$, so können wir die L^2-Norm des Sprungterms abschätzen durch

$$\|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)}^2 \overset{(4.34)}{\lesssim} \left\|[\![A\nabla u_{\mathcal{T}}]\!]\, b_E^{1/2}\right\|_{L^2(E)}^2 = \int_E [\![A\nabla u_{\mathcal{T}}]\!]\, v \,\mathrm{d}s.$$

Dieses Integral können wir aufgrund der Definition des Kantensprunges auf die beiden benachbarten Dreiecke aufteilen. Beachte hierbei, dass $v = 0$ auf $\mathcal{E} \setminus \{E\}$. Damit erhalten wir durch partielle Integration

$$\|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)}^2 \lesssim \sum_{T \in \omega_E^{\text{red}}} \int_{\partial T} A\nabla u_{\mathcal{T}} \cdot \nu_T v \,\mathrm{d}s = \sum_{T \in \omega_E^{\text{red}}} \int_T A\nabla u_{\mathcal{T}} \nabla v + \operatorname{div}(A\nabla u_{\mathcal{T}}) v \,\mathrm{d}x.$$

Betrachtet man den ersten Term im letzten Integral, so kann man die Summe hier wegen $v = 0$ auf $\Omega(\mathcal{T} \setminus \omega_E^{\text{red}})$ auf ganz $\mathcal{T}$ fortsetzen. Da $v \in \mathcal{S}_D^p(\mathcal{T}')$ ist, können wir einen Term $A\nabla u_{\mathcal{T}'} \nabla v$ einschieben und die schwache Formulierung (4.3) verwenden, wobei wir beachten, dass $v = 0$ auf allen Kanten in $\mathcal{E} \setminus \{E\}$:

$$\begin{aligned}
\sum_{T \in \omega_E^{\text{red}}} \int_T A\nabla u_{\mathcal{T}} \nabla v \,\mathrm{d}x = \int_\Omega A\nabla u_{\mathcal{T}} \nabla v \,\mathrm{d}x &= \int_\Omega A\nabla (u_{\mathcal{T}} - u_{\mathcal{T}'}) \nabla v \,\mathrm{d}x + \int_\Omega A\nabla u_{\mathcal{T}'} \nabla v \,\mathrm{d}x \\
&\overset{(4.3)}{=} \int_\Omega A\nabla (u_{\mathcal{T}} - u_{\mathcal{T}'}) \nabla v \,\mathrm{d}x + \int_\Omega f v \,\mathrm{d}x + \int_{\Gamma_N} \phi v \,\mathrm{d}s \\
&= \sum_{T \in \omega_E^{\text{red}}} \int_T A\nabla (u_{\mathcal{T}} - u_{\mathcal{T}'}) \nabla v \,\mathrm{d}x + \int_T f v \,\mathrm{d}x.
\end{aligned}$$

Fasst man die Abschätzungen der einzelnen Terme zusammen, ergibt sich mit der Cauchy-Schwarz Ungleichung

$$\begin{aligned}\|[\![A\nabla u_{\mathcal{T}}]\!]\|^2_{L^2(E)} &\lesssim \sum_{T\in\omega_E^{\mathrm{red}}} \left(\int_T A\nabla(u_{\mathcal{T}} - u_{\mathcal{T}'})\nabla v \,\mathrm{d}x + \int_T (f + \operatorname{div}(A\nabla u_{\mathcal{T}}))v \,\mathrm{d}x \right) \\ &\leq \sum_{T\in\omega_E^{\mathrm{red}}} \left(|\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T \, |\!|\!| v |\!|\!|_T + \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)} \, \|v\|_{L^2(T)} \right). \quad (4.42)\end{aligned}$$

Energienorm und L^2-Norm von v können wir noch weiter abschätzen. Durch die inverse Ungleichung in Proposition 4.36 und Lemma 4.35 erhalten wir

$$h_T \, |\!|\!| v |\!|\!|_T \overset{(4.38)}{\lesssim} \|v\|_{L^2(T)} \overset{(4.34)}{\lesssim} h_E^{1/2} \, \|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)} \,. \quad (4.43)$$

Wir können schließlich die Abschätzung (4.43) in (4.42) verwenden und erhalten mit der Definition des Kantenresiduums und $h_T \simeq h_E$

$$\begin{aligned}h_E^{1/2} \, \|[\![A\nabla u_{\mathcal{T}}]\!]\|^2_{L^2(E)} &\lesssim \left(\sum_{T\in\omega_E^{\mathrm{red}}} |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!| + h_E \, \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)} \right) \|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)} \\ &\lesssim \left(|\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_{\omega_E^{\mathrm{red}}} + \operatorname{res}_{\mathcal{T}}(\omega_E^{\mathrm{red}}) \right) \|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)} \,.\end{aligned}$$

Durch Division durch den Faktor $\|[\![A\nabla u_{\mathcal{T}}]\!]\|_{L^2(E)}$ folgt im Fall einer inneren Kante (4.39).

Fall 2: (4.39) für $E \in \mathcal{E} \setminus \mathcal{E}'$, sodass $E \in \mathcal{E}^N$.
Wir betrachten wiederum die Bubble Funktion für die Kante E. Außerdem sei hier an die Projektion $\Pi_E : L^2(E) \to P^{p-1}(E)$ aus Definition 4.24 erinnert. Sehen wir uns wieder den Residualterm an, so folgt mit der Dreiecksungleichung, dass

$$h_E^{1/2} \, \|\phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(E)} \leq h_E^{1/2} \, \|\phi - \Pi_E\phi\|_{L^2(E)} + h_E^{1/2} \, \|\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(E)} \,. \quad (4.44)$$

Hierbei entspricht der erste Term wegen $h_T \simeq h_E$ für $E \subseteq \partial T$ den Neumann-Oszillationen $\operatorname{osc}_N(T)$, den zweiten Term wollen wir nun abschätzen. Da $\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu$ wegen der Projektion im Raum $P^{p-1}(E)$ liegt, können wir analog zum ersten Fall vorgehen und eine geeignete Bubble Funktion verwenden, um Lemma 4.35 anzuwenden. Mit der Definition

$$v := (\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu) b_E$$

folgt

$$\|\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu\|^2_{L^2(E)} \overset{(4.34)}{\lesssim} \int_E (\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s. \quad (4.45)$$

Durch Einschieben des Terms ϕv in das Integral erhalten wir

$$\int_E (\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s = \int_E (\Pi_E\phi - \phi) v \,\mathrm{d}s + \int_E (\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s.$$

Den ersten Term können wir mithilfe der Cauchy-Schwarz Ungleichung, der Definition von v und der Tatsache, dass $b_E \leq 1$ ist, durch die Neumann-Oszillationen abschätzen:

$$\int_E (\Pi_E\phi - \phi) v \,\mathrm{d}s \leq \|(1-\Pi_E)\phi\|_{L^2(E)} \, \|v\|_{L^2(E)} \leq \|(1-\Pi_E)\phi\|_{L^2(E)} \, \|\Pi_E\phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(E)} \,.$$

Den zweiten Term müssen wir etwas eingehender betrachten. Für $E \subseteq \Gamma_N$ ist $\#\omega_E^{\mathrm{red}} = 1$, wir bezeichnen das entsprechende Dreieck mit T. Dieses ist eine Obermenge des Trägers von v. Da $v \in \mathcal{S}_D^p(\mathcal{T}')$ ist, können wir die schwache Formulierung für das Produkt ϕv verwenden, den Rest integrieren wir partiell, womit wir

$$\begin{aligned}\int_E (\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s &= \int_{\Gamma_N} \phi v \,\mathrm{d}s - \int_E A\nabla u_{\mathcal{T}} \cdot \nu v \,\mathrm{d}s \\ &\overset{(4.3)}{=} \int_\Omega A\nabla u_{\mathcal{T}'} \cdot \nabla v \,\mathrm{d}x - \int_\Omega f v \,\mathrm{d}x - \int_T A\nabla u_{\mathcal{T}} \cdot \nabla v + \operatorname{div}(A\nabla u_{\mathcal{T}}) v \,\mathrm{d}x \\ &= \int_T A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}})\nabla v \,\mathrm{d}x - \int_T \big(f + \operatorname{div}(A\nabla u_{\mathcal{T}})\big) v \,\mathrm{d}x\end{aligned}$$

erhalten. Anwenden der Cauchy-Schwarz Ungleichung und der inversen Ungleichung aus Proposition 4.36 liefert

$$\begin{aligned}\int_E (\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s &\leq |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T \, \|v\|_T + \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)} \, \|v\|_{L^2(T)} \\ &\overset{(4.38)}{\lesssim} \big(h_T^{-1} |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T + \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\big) \, \|v\|_{L^2(T)}. \end{aligned} \tag{4.46}$$

Schätzen wir nun noch den Term $\|v\|_{L^2(T)}$ in obiger Gleichung mit (4.34) durch $\|v\|_{L^2(T)} \lesssim h_E^{1/2} B := h_E^{1/2} \|\Pi_E \phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(T)}$ ab, so erhalten wir

$$\begin{aligned} B^2 &\overset{(4.45)}{\lesssim} \int_E (\phi - A\nabla u_{\mathcal{T}} \cdot \nu) v \,\mathrm{d}s \\ &\overset{(4.46)}{\lesssim} \big(h_T^{-1} |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T + \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\big) \, \|v\|_{L^2(T)} \\ &\lesssim \big(h_E^{-1/2} |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T + h_E^{1/2} \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\big) B. \end{aligned} \tag{4.47}$$

Kombiniert man alle obigen Abschätzungen, folgt insgesamt

$$\begin{aligned}\operatorname{res}_{\mathcal{T}}(E) = h_E^{1/2} \|\phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(E)} &\overset{(4.44)}{\leq} \operatorname{osc}_N(T) + h_E^{1/2} \|\Pi_E \phi - A\nabla u_{\mathcal{T}} \cdot \nu\|_{L^2(E)} \\ &\overset{(4.47)}{\lesssim} |\!|\!| u_{\mathcal{T}'} - u_{\mathcal{T}} |\!|\!|_T + \operatorname{osc}_N(T) + \operatorname{res}_{\mathcal{T}}(T),\end{aligned}$$

womit auch für diesen Fall (4.39) gezeigt ist.

Schritt 2: Abschätzung (4.40) für das Elementresiduum.
Der Beweis läuft hier sehr ähnlich wie der des zweiten Falles des ersten Schritts, denn auch hier haben wir es mit dem Fehler der Lösung zu den gegebenen Daten zu tun. Wir verlangen laut Voraussetzung $p \geq 2$ und rufen uns noch einmal die Projektion $\Pi_T : L^2(T) \to P^{p-2}(T)$ ins Gedächtnis (die Projektion auf Polynome vom Grad $p-2$ ist hier der Grund für die Einschränkung auf $p \geq 2$). Wir erhalten durch Einfügen von $\Pi_T f$

$$\operatorname{res}_{\mathcal{T}}(T) = h_T \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)} \lesssim \operatorname{osc}(T) + h_T \|\Pi_T f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}. \tag{4.48}$$

Mit der Bubble Funktion $b_T \in \mathcal{S}^2_D(\mathcal{T}')$, die laut Lemma 4.35 für jedes $T \in \mathcal{T} \setminus \mathcal{T}'$ existiert, setzen wir $v := (\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T}))b_T \in \mathcal{S}^p_D(\mathcal{T}')$. Eine Abschätzung mit (4.35) gibt uns analog zum Beweis des zweiten Falles von (4.39)

$$\|\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\|^2_{L^2(T)} \lesssim \int_T \big(\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\big) v \,\mathrm{d}x. \tag{4.49}$$

Es bleibt also noch dieser Term abzuschätzen. Wir schieben erneut einen Term fv ein. Da $v \in \mathcal{S}^p_D(\mathcal{T}')$, können wir für fv wiederum die schwache Formulierung einsetzen (wobei die Randterme wegfallen, da $v = 0$ auf ∂T) und für $\operatorname{div}(A\nabla u_\mathcal{T})v$ partiell integrieren. Damit erhalten wir

$$\begin{aligned}\int_T \big(\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\big) v \,\mathrm{d}x &\leq \|(1-\Pi_T)f\|_{L^2(T)} \|v\|_{L^2(T)} + \int_T \big(f + \operatorname{div}(A\nabla u_\mathcal{T})\big) v \,\mathrm{d}x \\ &\overset{\text{part.Int.}}{\overset{(4.3)}{=}} \|(1-\Pi_T)f\|_{L^2(T)} \|v\|_{L^2(T)} + \int_T A\nabla(u_{\mathcal{T}'} - u_\mathcal{T})\nabla v \,\mathrm{d}x.\end{aligned}$$

Mithilfe der Cauchy-Schwarz Ungleichung und der inversen Ungleichung aus Proposition 4.36 ergibt sich

$$\int_T \big(\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\big) v \,\mathrm{d}x \overset{(4.38)}{\lesssim} \left(\|(1-\Pi_T)f\|_{L^2(T)} + h_T^{-1} \left|\!\left|\!\left| u_{\mathcal{T}'} - u_\mathcal{T} \right|\!\right|\!\right| \right) \|v\|_{L^2(T)}. \tag{4.50}$$

Benutzen wir, dass $b_T \leq 1$, gilt außerdem $\|v\|_{L^2(T)} \leq \|\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\|_{L^2(T)}$. Zusammen führt dies mit (4.49), (4.50) und Division durch $\|\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\|_{L^2(T)}$ auf

$$\|\Pi_T f + \operatorname{div}(A\nabla u_\mathcal{T})\|_{L^2(T)} \lesssim \|(1-\Pi_T)f\|_{L^2(T)} + h_T^{-1} \left|\!\left|\!\left| u_{\mathcal{T}'} - u_\mathcal{T} \right|\!\right|\!\right|.$$

Gemeinsam mit (4.48) beweist dies (4.40).

Schritt 3: Beweis von (4.41).
Um diese Abschätzung zu erhalten, fassen wir die ersten beiden Schritte zusammen. Nach Definition des Residualschätzers gilt, dass

$$\eta^2_\mathcal{T}(\mathcal{E} \setminus \mathcal{E}') = \sum_{E \in \mathcal{E} \setminus \mathcal{E}'} \left(\operatorname{res}^2_\mathcal{T}(E) + \sum_{T \in \omega_E^{\mathrm{red}}} \operatorname{res}^2_\mathcal{T}(T) \right).$$

Aufgrund der Fallunterscheidung in der Definition der Oszillationen osc müssen wir diese Unterscheidung auch hier berücksichtigen.

Fall 1: $p = 1$.
In diesem Fall gilt, da $u_\mathcal{T} \in \mathcal{S}^1_D(\mathcal{T})$, dass $\operatorname{div}(A\nabla u_\mathcal{T}) = 0$ elementweise. Deshalb gilt für $T \in \mathcal{T}$

$$\operatorname{res}_\mathcal{T}(T) = h_T \|f + \operatorname{div}(A\nabla u_\mathcal{T})\|_{L^2(T)} = h_T \|f\|_{L^2(T)} = \operatorname{osc}(T).$$

Da jedes Dreieck in $\mathcal{T}$ in höchstens drei reduzierten Kantenpatches enthalten ist und da für jeden reduzierten Kantenpatch ω_E^{red} mit E aus $\mathcal{E} \setminus \mathcal{E}'$ gilt, dass $\omega_E^{\mathrm{red}} \subseteq \mathcal{T} \setminus \mathcal{T}'$, folgt mit (4.39) schließlich die zu zeigende Ungleichung (4.41).

Fall 2: $p \geq 2$.
In diesem Fall lässt sich der Term $\operatorname{res}_{\mathcal{T}}(\omega_E^{\mathrm{red}})$, der in (4.39) auftritt, elementweise durch (4.40) abschätzen. Mit ähnlichen Argumenten wie im vorigen Fall, lassen sich diese schließlich global zu (4.41) zusammenfassen.
Damit ist der Beweis vollständig. □

Bemerkung 4.38. Der Beweis der Abschätzung (4.40) für die diskrete lokale Effizienz verwendet eine Projektion auf einen Polynomraum kleineren Grades. Hätte man in der Problemformulierung noch einen zusätzlichen Term u^2 (wie er bei der Helmholtz-Gleichung auftritt), müsste in diesem Beweis noch ein Term der Form $\|(1-\Pi)u\|_{L^2(T)}$ zu einem der Terme auf der rechten Seite von (4.40) abgeschätzt werden. Dies ist allerdings ein bis dato offenes Problem. Deshalb wurde hier eine Problemformulierung gewählt, die keinen solchen Term beinhaltet.
Ähnliches gilt auch für Robin-Randbedingungen. Hier taucht in der Abschätzung (4.39) ein Term der Form $\|(1-\Pi_E)u\|_{L^2(E)}$ auf, den wir bis dato noch nicht in den Griff bekommen haben. //

Wir können die Ergebnisse aus diesem Abschnitt noch zusammenfassen und auf die Gesamtenergie übertragen, um die Voraussetzung aus Definition 3.8 nachzuweisen.

Korollar 4.39. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, $\mathcal{T}' \geq \mathcal{T}$ und $p \in \mathbb{N}$. Seien weiter $\mathcal{G}$ die Gesamtenergie aus Definition 4.25 und $\eta_{\mathcal{T}}^2$ der Residualschätzer aus Definition 4.30. Dann gilt*

$$\mathcal{G}(\mathcal{T}) - \mathcal{G}(\mathcal{T}') \simeq \eta_{\mathcal{T}}^2(\mathcal{E} \setminus \mathcal{E}'), \tag{4.51}$$

wobei die Konstanten nur von den gegebenen Daten und $\mathcal{T}_0$ abhängen.

Beweis. Wir beginnen damit, die Oszillationsterme durch Residualterme abzuschätzen. Für $p = 1$ und $T \in \mathcal{T}$ gilt offensichtlich $\operatorname{osc}(T) = \operatorname{res}_{\mathcal{T}}(T)$. Für $p \geq 2$ ist die L^2-Orthogonalprojektion $\Pi_T : L^2(T) \to P^{p-2}(T)$ involviert. Aufgrund deren Bestapproximationseigenschaft und der Tatsache, dass elementweise $\operatorname{div}(A\nabla u_{\mathcal{T}}) \in P^{p-2}(T)$ gilt, erhalten wir

$$\begin{aligned}\operatorname{osc}(T) &= h_T \|(1-\Pi_T)f\|_{L^2(T)} = h_T \inf_{v \in P^{p-2}(T)} \|f - v\|_{L^2(T)} \\ &\leq h_T \|f + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)} = \operatorname{res}_{\mathcal{T}}(T).\end{aligned}$$

Analog folgt auch für die Neumann-Oszillationen $\operatorname{osc}_N(T) \leq \operatorname{res}_{\mathcal{T}}(E)$ für alle Randkanten E mit $E \subseteq \partial T$.
Zusammen mit der diskreten lokalen Zuverlässigkeit (Satz 4.33) und der diskreten lokalen Effizienz (Satz 4.37), folgt also

$$|||u_{\mathcal{T}} - u_{\mathcal{T}'}|||^2 + \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}') + \operatorname{osc}_N^2(\mathcal{T} \setminus \mathcal{T}') \simeq \eta_{\mathcal{T}}^2(\mathcal{E} \setminus \mathcal{E}'). \tag{4.52}$$

Wegen der Gleichheit von Energiedifferenz und Energienorm aus Proposition 4.7 gilt

$$\mathcal{G}(\mathcal{T}) - \mathcal{G}(\mathcal{T}') = \tfrac{1}{2} |||u_{\mathcal{T}} - u_{\mathcal{T}'}|||^2 + \operatorname{osc}^2(\mathcal{T} \setminus \mathcal{T}') + \operatorname{osc}_N^2(\mathcal{T} \setminus \mathcal{T}').$$

Zusammen mit (4.52) folgt schließlich die Behauptung. □

4.5 Markierungsstrategie

In diesem Abschnitt wollen wir uns die Markierungsstrategie anschauen, die in der AFEM benutzt werden soll. Wie bereits in Kapitel 3 besprochen, muss diese Strategie die Schweife der Kanten bei Verfeinerung berücksichtigen und ist in Algorithmus 2 angegeben. Diese Strategie sowie die folgende Proposition sind aus [DKS16] entnommen.

Algorithmus 2 Markierungsstrategie

Input: Kantenmenge $\mathcal{E}$ und Fehlerschätzer $\eta_{\mathcal{T}}^2(E)$ für jede Kante $E \in \mathcal{E}$, Markierungsparameter $\vartheta \in (0,1)$.
Output: Menge $\mathcal{M} \subseteq \mathcal{E}$ der markierten Kanten.

1: $\overline{\varrho} := \max\{\eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(E)) \mid E \in \mathcal{E}\}$
2: $\mathcal{M} := \emptyset$, $\widetilde{\mathcal{M}} := \emptyset$ und $\mathcal{U} := \mathcal{E}$
3: **while** $\mathcal{U} \neq \emptyset$ **do**
4: wähle $E \in \mathcal{U}$
5: berechne $\varrho(E) := \eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(E) \setminus \widetilde{\mathcal{M}})$
6: **if** $\varrho(E) \geq \vartheta\overline{\varrho}$ **then**
7: $\mathcal{M} := \mathcal{M} \cup \{E\}$
8: $\widetilde{\mathcal{M}} := \widetilde{\mathcal{M}} \cup \operatorname{tail}_{\mathcal{T}}(E)$
9: **end if**
10: $\mathcal{U} := \mathcal{U} \setminus \operatorname{tail}_{\mathcal{T}}(E)$
11: **end while**

Wir wollen kurz Algorithmus 2 erläutern. In einem Schritt der while-Schleife bezeichnen $\mathcal{M}$ alle bereits markierten Kanten, $\widetilde{\mathcal{M}}$ alle Kanten im Schweif zumindest einer markierten Kante und $\mathcal{U}$ die Kanten, die noch betrachtet werden müssen. Der Algorithmus berechnet für jede Kante $E \in \mathcal{E}$, die noch nicht betrachtet wurde, einen Indikator $\varrho(E)$, der aus den Schätzerbeiträgen der Kanten im Schweif von E linearkombiniert wird (Zeile 5). Hierbei werden wegen $\operatorname{tail}_{\mathcal{T}}(E) \setminus \widetilde{\mathcal{M}}$ nur die Beiträge der Kanten verwendet, die nicht im Schweif einer bereits markierten Kante gehören.

In Zeile 6 wird der eben berechnete Indikator mit dem maximalen Indikator $\overline{\varrho}$ verglichen. Falls das Kriterium erfüllt ist, wird die Kante E markiert (Zeile 7) und ihr Schweif von der Berechnung zukünftiger Indikatoren ausgenommen (Zeile 8). Schließlich wird der gesamte Schweif der Kante von zukünftigen Betrachtungen ausgenommen (Zeile 10), da $\operatorname{tail}_{\mathcal{T}}(E') \subseteq \operatorname{tail}_{\mathcal{T}}(E)$ für alle $E' \in \operatorname{tail}_{\mathcal{T}}(E)$ gilt. Somit wird die Kante E' entweder ohnehin verfeinert (falls E markiert wurde), oder aber es gilt $\varrho(E') \leq \varrho(E) < \vartheta\overline{\varrho}$.

Wir zeigen nun, dass der angegebene Algorithmus auch das Markierungskriterium (E1) erfüllt.

Proposition 4.40. *Sei $\eta_{\mathcal{T}}^2$ der Fehlerschätzer zu einer Finite Elemente Lösung von (4.1) auf $\mathcal{T} \in \mathbb{T}$. Dann gilt für die Menge der markierten Kanten aus Algorithmus 2, dass $\mathcal{M} \neq \emptyset$, und es erfüllt jede Teilmenge $\emptyset \neq M \subseteq \mathcal{M}$ das Markierungskriterium (E1) mit $\mu = \vartheta$.*

Beweis. Wir zeigen zuerst, dass $\mathcal{M} \neq \emptyset$. Solange $\mathcal{M} = \emptyset$ gilt, gilt wegen der gleichzeitigen Ausführung der Zeilen 7 und 8 auch $\widetilde{\mathcal{M}} = \emptyset$. An irgendeinem Punkt wird aber in Zeile 4 die

Kante $E \in \mathcal{E}$ erwischt, für die gilt, dass $\varrho(E) = \overline{\varrho} \geq \vartheta\overline{\varrho}$. Somit wird diese Kante markiert, weshalb $\mathcal{M} \neq \emptyset$.

Für die zweite Forderung in (E1) beachten wir, dass in Zeile 5 wegen $\operatorname{tail}_{\mathcal{T}}(E') \subseteq \operatorname{tail}_{\mathcal{T}}(E)$ für $E' \in \operatorname{tail}_{\mathcal{T}}(E)$ keine Kante doppelt zur Berechnung der Indikatoren herangezogen wird. Deshalb gilt in einem beliebigen Schritt des Algorithmus für die gerade ausgewählte Kante E, dass

$$\eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(\mathcal{M} \cup \{E\})) = \eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(\mathcal{M})) + \varrho(E). \tag{4.53}$$

Beachten wir weiter, dass für jede markierte Kante die Abschätzung $\varrho(E) \geq \vartheta\overline{\varrho}$ in Zeile 6 gilt, folgt durch iteratives Anwenden von (4.53), dass bei Terminierung des Algorithmus

$$\eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(\mathcal{M})) \geq \vartheta\, \#\mathcal{M}\overline{\varrho} = \vartheta\, \#\mathcal{M} \max_{E\in\mathcal{E}} \eta_{\mathcal{T}}^2(\operatorname{tail}_{\mathcal{T}}(E))$$

gilt. Vergleicht man dies mit (E1), ist mit $\mu = \vartheta$ die Behauptung bewiesen. Wir haben außerdem nirgends verwendet, dass $\mathcal{M}$ alle Kanten enthält, die der Algorithmus vorschlägt, weshalb auch die Behauptung für eine Teilmenge $M \subseteq \mathcal{M}$ erfüllt ist. □

Bemerkung 4.41. Es sei hier angemerkt, dass Algorithmus 2 wegen der wiederholten Berechnung der Beiträge der Schweife zum Markierungsindikator $\varrho(E)$ in Zeile 5 im Allgemeinen nicht in $\mathcal{O}(\#\mathcal{E})$ arbeiten kann. Wegen seiner iterativen Form ist der Algorithmus jedoch leicht verständlich und parallelisierbar. In [DKS16] ist auch ein Algorithmus angegeben, dessen Aufwand linear in der Anzahl der Kanten ist. Dieser ist aber rekursiv und somit nicht für parallele Berechnungen geeignet. //

4.6 Instanzoptimalität

Mit den Resultaten aus den vorigen Abschnitten ist es ein leichtes zu zeigen, dass die AFEM aus diesem Kapitel in das Framework von Kapitel 3 passt. Damit können wir Instanzoptimalität für den totalen Fehler formulieren und zeigen.

Satz 4.42. *Sei $(\mathcal{T}_k)_{k\in\mathbb{N}_0}$ die Folge der Gitter, die von der AFEM (Algorithmus 1) mit dem Fehlerschätzer η^2 aus Definition 4.30 und dem Markierungsschritt aus Algorithmus 2 erzeugt werden. Dann gilt mit der Konstante C_{mesh} aus Satz 3.14: Für alle $\mathcal{T} \in \mathbb{T}$ mit $\#(\mathcal{T} \setminus \mathcal{T}_0) \leq \#(\mathcal{T}_k \setminus \mathcal{T}_0)/C_{\mathrm{mesh}}$ gilt*

$$|||u_{\mathrm{ex}} - u_{\mathcal{T}_k}|||^2 + \operatorname{osc}^2(\mathcal{T}_k) + \operatorname{osc}_N^2(\mathcal{T}_k) \leq 2\big(|||u_{\mathrm{ex}} - u_{\mathcal{T}}|||^2 + \operatorname{osc}^2(\mathcal{T}) + \operatorname{osc}_N^2(\mathcal{T})\big). \tag{4.54}$$

Beweis. Wir weisen die Voraussetzungen von Satz 3.14 nach. Diese sind die das Markierungskriterium (E1), die Lower Diamond Estimate (E2) und diskrete lokale Äquivalenz (E3). Nach Proposition 4.40 erfüllt die Markierungsstrategie (E1), mit Satz 4.37 erhalten wir (E2) Korollar 4.39 gibt uns (E3).

Damit können wir die Aussage von Satz 3.14 anwenden. Da nach Voraussetzung $\#(\mathcal{T} \setminus \mathcal{T}_0) \leq \#(\mathcal{T}_k \setminus \mathcal{T}_0)/C_{\text{mesh}}$ gilt, können wir $m := \#(\mathcal{T} \setminus \mathcal{T}_0)$ setzen und erhalten

$$\mathscr{G}(\mathcal{T}_k) \overset{3.14}{\leq} \mathscr{G}_m^{\text{opt}} \leq \mathscr{G}(\mathcal{T}).$$

Ziehen wir von der linken und der rechten Seite der Ungleichungskette $\mathscr{E}(u_{\text{ex}})$ ab, so erhalten wir mit Proposition 4.7

$$\tfrac{1}{2} \left|\!\left|\!\left| u_{\text{ex}} - u_{\mathcal{T}_k} \right|\!\right|\!\right|^2 + \text{osc}^2(\mathcal{T}_k) + \text{osc}_N^2(\mathcal{T}_k) \leq \tfrac{1}{2} \left|\!\left|\!\left| u_{\text{ex}} - u_{\mathcal{T}} \right|\!\right|\!\right|^2 + \text{osc}^2(\mathcal{T}) + \text{osc}_N^2(\mathcal{T}).$$

Die Vorfaktoren der Oszillationsterme der linken Seite können durch $1/2$ abgeschätzt werden, der Vorfaktor der Energienorm auf der rechten seite durch 1. Damit folgt die Behauptung. □

5 Goal-Oriented FEM

Im letzten Kapitel haben wir versucht, durch adaptive Verfahren den Fehler einer FEM Approximation in der Energienorm zu minimieren. Wir wollen uns in diesem Kapitel mit einer etwas anderen Problemstellung beschäftigen, nämlich der Minimierung des Wertes eines Funktionals der Lösung. Dieses Funktional wird auch Zielfunktional genannt, weshalb die Methode, die wir hier vorstellen, zielorientierte, oder auch Goal-Oriented AFEM (GOAFEM) genannt wird. Diese ist in Anwendungen sehr verbreitet, da hierbei noch mehr Fokus auf problemspezifisch relevante Größen gelegt werden kann [FPvdZ16].

Es werden hier zunächst das Modellproblem und einige wichtige Begriffe, welche sehr ähnlich zu denen aus dem letzten Kapitel sind, definiert. Das wesentliche Unterscheidungsmerkmal des hier vorgestellten Algorithmus gegenüber dem aus dem letzten Kapitel wird der Markierungsschritt sein. Da sowohl die Genauigkeit der FEM Lösung, als auch die der Auswertung des Funktionals gesteuert wird, muss ein solcher Markierungsschritt diese beiden Aspekte berücksichtigen. Wir werden uns hier auf sogenanntes separiertes Makieren beschränken (siehe [FPvdZ16]), bei dem für die beiden Aspekte zunächst separat Kanten markiert werden, welche dann geeignet zusammengeführt werden. Wir zeigen schließlich, dass dieser Markierungsschritt in das Framework aus Kapitel 3 passt und beweisen damit Instanzoptimalität der hier vorgestellten GOAFEM.

5.1 Problemstellung

Wir betrachten ein Modellproblem mit homogenen Randdaten. Allerdings lassen wir hier die Beschränkung fallen, dass die rechte Seite der Gleichung die Form $\int_\Omega fv\,\mathrm{d}x$ mit $f \in L^2(\Omega)$ hat. Es sei dazu $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit polygonalem Rand $\Gamma_D = \partial\Omega$. Weiter sei $A \in [L^\infty(\Omega)]^{2\times 2}$ eine matrixwertige Funktion, zu der es ein Gitter $\mathcal{T}_0$ auf Ω gibt, auf dem A stückweise konstant, symmetrisch und positiv definit ist. Es seien zusätzlich Funktionen $f_1 \in L^2(\Omega)$ und $f_2 \in [L^2(\Omega)]^2$ gegeben. An f_2 stellen wir außerdem die Forderung, dass $\operatorname{div}(f_2) \in L^2(T)$ und $[\![f_2]\!] \in L^2(\partial T)$ für alle $T \in \mathcal{T}_0$ ist.

Das in diesem Kapitel betrachtete Modellproblem ist dann

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f_1 + \operatorname{div}(f_2) && \text{in } \Omega, \\ u &= 0 && \text{auf } \partial\Omega. \end{aligned} \tag{5.1}$$

Die Divergenz $F_2 := \operatorname{div}(f_2) \in H^{-1}(\Omega) := (H^1_0(\Omega))'$ auf der rechten Seite dieser Gleichung ist im Sinne von Distributionen aufzufassen (siehe dazu [Yos80]). Um diesen Divergenzterm in die schwache Formulierung überzuführen betrachtet man eine beliebige glatte Funktion $v \in C^\infty_0(\Omega)$ mit kompaktem Träger in Ω und wendet die Definition der distributionellen Ableitung an:

$$F_2(v) = -\int_\Omega f_2 \cdot \nabla v\,\mathrm{d}x. \tag{5.2}$$

Da $C_0^\infty(\Omega)$ dicht in $H_0^1(\Omega)$ ist, gilt diese Darstellung auch für Funktionen $v \in H_0^1(\Omega)$. Beachte, dass die homogenen Randbedingungen auf dem gesamten Rand hier essentiell sind, da ohne sie die Umformung in Gleichung (5.2) nicht zulässig wäre.

In der schwachen Form lautet das Problem daher wie folgt: Finde $u \in H_0^1(\Omega)$, sodass

$$\int_\Omega A\nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f_1 v - f_2 \cdot \nabla v \,\mathrm{d}x \quad \text{für alle } v \in H_0^1(\Omega). \tag{5.3}$$

Die Finite Elemente Lösung ist analog zu (4.3) definiert. Ebenso sind Bilinearform beziehungsweise Linearform analog zum vorigen Kapitel definiert:

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad F(v) := \int_\Omega f_1 v - f_2 \cdot \nabla v \,\mathrm{d}x.$$

Wir werden die Resultate in diesem Kapitel für Finite Elemente Diskretisierungen mit konformen Elementen beliebiger Ordnung beweisen, also für $u_\mathcal{T} \in \mathcal{S}_0^p(\mathcal{T})$.

Im Gegensatz zum vorigen Kapitel ist aber noch ein weiteres Funktional G aus dem Dualraum $H^{-1}(\Omega)$ des Lösungsraumes $H_0^1(\Omega)$ gegeben. Analog zur rechten Seite von Gleichung (5.1) werden wir voraussetzen, dass Funktionen $g_1 \in L^2(\Omega)$ und $g_2 \in [L^2(\Omega)]^2$ existieren, sodass dieses Funktional für $u \in H_0^1(\Omega)$ als

$$G(u) = \int_\Omega g_1 u - g_2 \cdot \nabla u \,\mathrm{d}x$$

geschrieben werden kann. Für g_2 verlangen wir die selben Einschränkungen, wie für f_2: $\operatorname{div}(g_2) \in L^2(T)$ und $[\![g_2]\!] \in L^2(\partial T)$ für alle $T \in \mathbb{T}$.

Bezeichnen wir mit u_ex die exakte schwache Lösung von (5.1), so ist der Wert, den wir zu approximieren suchen, $G(u_\mathrm{ex})$. Dieser Wert wird in der Literatur auch *quantity of interest* genannt, wir werden ihn im Folgenden schlicht Zielwert nennen.

Bemerkung 5.1. Wie zuvor erwähnt, sind die homogenen Randdaten auf dem gesamten Rand essentiell. Deshalb können wir hier auch keinen Neumannrand zulassen, wie wir es im vorigen Kapitel getan haben. Im Fall, dass $f_2, g_2 \equiv 0$, lassen sich die Methoden aus diesem Kapitel jedoch auch auf Probleme mit inhomogenen Neumann Randdaten auf einem Teil des Randes andwenden. //

5.1.1 Fehlergröße

Für eine Finite Elemente Lösung $u_\mathcal{T}$ von (5.1) auf einem Gitter $\mathcal{T} \in \mathbb{T}$ ist der Fehler, den wir bei bei der Berechnung des Zielwertes machen,

$$|G(u_\mathrm{ex}) - G(u_\mathcal{T})|\,. \tag{5.4}$$

Dieser Fehler lässt sich durch übliche Methoden der a posteriori Fehlerschätzung jedoch nur schwer direkt abschätzen. Es ist deshalb zweckdienlich, hier eine Fehlergröße zu betrachten, die den Fehler (5.4) nach oben abschätzt [FPvdZ16].

Dazu müssen wir zunächst das sogenannte *duale Problem* definieren. Ist das Problem (5.1) in der schwachen Form gegeben, also eine Funktion $u \in H^1_0(\Omega)$ gesucht, sodass

$$a(u,v) = F(v) \quad \text{für alle } v \in H^1_0(\Omega), \tag{5.5}$$

und zusätzlich ein Funktional G gegeben, so sagen wir $w \in H^1_0(\Omega)$ löst das duale Problem, wenn es

$$a(v,w) = G(v) \quad \text{für alle } v \in H^1_0(\Omega) \tag{5.6}$$

erfüllt. Beachte, dass im dualen Problem die Argumente der Bilinearform vertauscht sind, was in unserem Fall aber keinen Unterschied macht, da die in dieser Arbeit betrachteten Bilinearformen ohnehin symmetrisch sind. Wir werden im Folgenden das ursprüngliche Problem (5.5) auch primales Problem nennen, um es vom dualen Problem (5.6) zu unterscheiden.

Seien nun für ein Gitter $\mathcal{T} \in \mathbb{T}$ die Funktionen $w_{\text{ex}} \in H^1_0(\Omega)$ und $w_{\mathcal{T}} \in \mathcal{S}^p_0(\mathcal{T})$ die exakte schwache Lösung des dualen Problems (5.6) beziehungsweise dessen Finite Elemente Lösung auf $\mathcal{T}$. Außerdem seien $u_{\text{ex}}, u_{\mathcal{T}}$ in analoger Weise für das primale Problem (5.5) definiert. Es gilt die Galerkin-Orthogonalität

$$a(u_{\text{ex}} - u_{\mathcal{T}}, w_{\mathcal{T}}) = a(u_{\text{ex}}, w_{\mathcal{T}}) - a(u_{\mathcal{T}}, w_{\mathcal{T}}) \overset{(5.5)}{=} F(w_{\mathcal{T}}) - F(w_{\mathcal{T}}) = 0 \tag{5.7}$$

Damit lässt sich mithilfe der Cauchy-Schwarz–Ungleichung der Fehler (5.4) folgendermaßen nach oben abschätzen:

$$\begin{aligned} |G(u_{\text{ex}}) - G(u_{\mathcal{T}})| \overset{(5.6)}{=} |a(u_{\text{ex}} - u_{\mathcal{T}}, w_{\text{ex}})| &\overset{(5.7)}{=} |a(u_{\text{ex}} - u_{\mathcal{T}}, w_{\text{ex}} - w_{\mathcal{T}})| \\ &\leq \left|\!\left|\!\left| u_{\text{ex}} - u_{\mathcal{T}} \right|\!\right|\!\right| \left|\!\left|\!\left| w_{\text{ex}} - w_{\mathcal{T}} \right|\!\right|\!\right|. \end{aligned}$$

Im vorigen Kapitel hat die dortige Fehlergröße Oszillationsterme enthalten. Dies ist auch hier nötig. Dazu müssen wir eine zusätzliche Notation einführen, um die Oszillationen der Daten von primalem und dualem Problem zu unterscheiden.

Definition 5.2. *Sei $T \in \mathcal{T} \in \mathbb{T}$ ein Dreieck eines Gitters und $p \in \mathbb{N}$. Weiter seien Π_T und Π_E die L^2-Orthogonalprojektionen auf den Raum der Polynome vom Grad $p-2$ auf T beziehungsweise $p-1$ auf E. Für ein Problem der Bauart (5.5) und (5.6) definieren wir als Oszillationen*

$$\operatorname{osc}_F^2(T) := h_T \sum_{E \in \partial T} \|(1-\Pi_E)\,[\![f_2]\!]\|^2_{L^2(E)} + \begin{cases} h_T^2 \,\|f_1 + \operatorname{div}(f_2)\|^2_{L^2(T)} & \text{für } p = 1, \\ h_T^2 \,\|(1-\Pi_T)(f_1 + \operatorname{div}(f_2))\|^2_{L^2(T)} & \text{für } p \geq 2, \end{cases}$$

$$\operatorname{osc}_G^2(T) := h_T \sum_{E \in \partial T} \|(1-\Pi_E)\,[\![g_2]\!]\|^2_{L^2(E)} + \begin{cases} h_T^2 \,\|g_1 + \operatorname{div}(g_2)\|^2_{L^2(T)} & \text{für } p = 1, \\ h_T^2 \,\|(1-\Pi_T)(g_1 + \operatorname{div}(g_2))\|^2_{L^2(T)} & \text{für } p \geq 2. \end{cases}$$

Damit können wir die in diesem Kapitel betrachtete Fehlergröße anschreiben als

$$\left(\left|\!\left|\!\left| u_{\text{ex}} - u_{\mathcal{T}} \right|\!\right|\!\right|^2 + \operatorname{osc}_F^2(\mathcal{T})\right)\left(\left|\!\left|\!\left| w_{\text{ex}} - w_{\mathcal{T}} \right|\!\right|\!\right|^2 + \operatorname{osc}_G^2(\mathcal{T})\right). \tag{5.8}$$

Hier steht das Produkt zweier Fehler der Bauart, wie wir sie schon in Kapitel 4 betrachtet haben und für die wir dort Instanzoptimalität gezeigt haben. In der Tat werden wir die Instanzoptimalität der beiden einzelnen Fehler verwenden, um durch einen modifizierten Markierungsschritt Instanzoptimalität für das Produkt zu zeigen.

5.2 Energie und AFEM

5.2.1 Energie

Wir werden zwei verschiedene Energien definieren und daraufhin deren Produkt betrachten. Um die Monotonie dieses Produkts zu garantieren, müssen wir sichergehen, dass dessen Faktoren positiv sind. Dies erreichen wir durch eine Translation einer Energie der Bauart aus Definition 4.25 um eine feste Konstante. Solche eine Konstante finden wir, da alle hier betrachteten Energien nach unten beschränkt sind.

Rufen wir uns außerdem die Resultate aus Abschnitt 4.1.2 ins Gedächtnis, so bemerken wir, dass die ausschlaggebende Größe überall eine Differenz zweier Energien war. Beim Bilden solch einer Differenz fällt jedoch eine Konstante weg, die zu beiden Energien addiert wurde. Deshalb gelten die Eigenschaften der Energie aus Definition 4.5 uneingeschränkt auch für Energien, die um einen festen Betrag verschoben wurden.

Beachte in der folgenden Definition außerdem, dass die Oszillationen immer nichtnegativ sind.

Definition 5.3. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $p \in \mathbb{N}$. Weiter sei $u_\mathcal{T} \in \mathcal{S}^p_0(\mathcal{T})$ die schwache Lösung des primalen Problems (5.5), sowie $w_\mathcal{T} \in \mathcal{S}^p_0(\mathcal{T})$ die schwache Lösung des dualen Problems (5.6). Wir definieren*

$$\begin{aligned}\mathfrak{F}_\infty &:= \inf_{v \in H^1_0(\Omega)} \tfrac{1}{2}a(v,v) - F(v),\\ \mathfrak{G}_\infty &:= \inf_{v \in H^1_0(\Omega)} \tfrac{1}{2}a(v,v) - G(v).\end{aligned}$$

Damit definieren wir die primale und duale Energie als

$$\begin{aligned}\mathfrak{F}(\mathcal{T}) &:= \tfrac{1}{2}a(u_\mathcal{T},u_\mathcal{T}) - F(u_\mathcal{T}) + \operatorname{osc}^2_F(\mathcal{T}) - \mathfrak{F}_\infty,\\ \mathfrak{G}(\mathcal{T}) &:= \tfrac{1}{2}a(w_\mathcal{T},w_\mathcal{T}) - G(w_\mathcal{T}) + \operatorname{osc}^2_G(\mathcal{T}) - \mathfrak{G}_\infty.\end{aligned}$$

Bemerkung 5.4. Nach Konstruktion von $\mathfrak{F}_\infty$ und $\mathfrak{G}_\infty$ als Infimum der ersten beiden Terme der Energien $\mathfrak{F}$, respektive $\mathfrak{G}$, gilt für jedes Gitter $\mathcal{T} \in \mathbb{T}$, dass $\mathfrak{F}(\mathcal{T}), \mathfrak{G}(\mathcal{T}) \geq 0$. Außerdem gilt für die exakten Lösungen $u_\text{ex}, w_\text{ex} \in H^1_0(\Omega)$ der Probleme (5.5) und (5.6) nach Proposition 4.6, dass

$$\begin{aligned}\inf_{v \in H^1_0(\Omega)} \tfrac{1}{2}a(v,v) - F(v) &= \tfrac{1}{2}a(u_\text{ex},u_\text{ex}) - F(u_\text{ex}),\\ \inf_{v \in H^1_0(\Omega)} \tfrac{1}{2}a(v,v) - G(v) &= \tfrac{1}{2}a(w_\text{ex},w_\text{ex}) - G(w_\text{ex}).\end{aligned}$$

Darüber hinaus gelten, wie eingangs erwähnt, die Resultate aus Abschnitt 4.1.2. Es sei hier jedoch darauf hingewiesen, dass die dortigen Konstanten vom initialen Gitter und der Energie abhängen, die in diesem Fall konkret von der Bilinearform $a(\cdot,\cdot)$ und den Daten f_1, f_2, g_1 und g_2 gebildet wird. //

5.2.2 Fehlerschätzer

Ähnlich zur Energie werden wir auch die Fehlerschätzer zweifach definieren, für das primale und für das duale Problem. Diese Schätzer sind grundsätzlich die im letzten Kapitel verwendeten Residualschätzer, jedoch muss dem Umstand Rechnung getragen werden, dass bei den beiden betrachteten Problemen die rechten Seiten unterschiedlich sind. Deshalb führen wir nun zusätzliche Notation ein.

Definition 5.5. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter, sowie A, f_1, f_2, g_1 und g_2 die Daten aus (5.5) beziehungsweise (5.6). Seien weiter $u_\mathcal{T}, w_\mathcal{T} \in \mathcal{S}^p_0(\mathcal{T})$ die Lösungen dieser Probleme auf $\mathcal{T}$. Wir definieren für $T \in \mathcal{T}$*

$$\begin{aligned} \operatorname{res}_{\mathcal{T};F}(T) &:= h_T \left\| \operatorname{div}(A\nabla u_\mathcal{T}) + f_1 + \operatorname{div}(f_2) \right\|_{L^2(T)}, \\ \operatorname{res}_{\mathcal{T};G}(T) &:= h_T \left\| \operatorname{div}(A\nabla w_\mathcal{T}) + g_1 + \operatorname{div}(g_2) \right\|_{L^2(T)}, \end{aligned}$$

sowie für $E \in \mathcal{E}$

$$\operatorname{res}_{\mathcal{T};F}(E) := \begin{cases} h_E^{1/2} \left\| [\![A\nabla u_\mathcal{T} + f_2]\!] \right\|_{L^2(E)} & \text{für } E \subseteq \Omega, \\ 0 & \text{für } E \subseteq \partial\Omega, \end{cases}$$

$$\operatorname{res}_{\mathcal{T};G}(E) := \begin{cases} h_E^{1/2} \left\| [\![A\nabla w_\mathcal{T} + g_2]\!] \right\|_{L^2(E)} & \text{für } E \subseteq \Omega, \\ 0 & \text{für } E \subseteq \partial\Omega. \end{cases}$$

5.2.3 Markierungsschritt

In diesem Abschnitt wollen wir uns dem Herzstück der hier vorgestellten GOAFEM widmen, dem Markierungsschritt. Wie eingangs erwähnt wird hier separiert markiert werden. Dazu werden zunächst für das primale und für das duale Problem mittels Algorithmus 2 separat Kanten markiert, die dann derart zu einer endgültigen Menge an markierten Kanten zusammengeführt werden, dass für beide Probleme immer gleich viele Kanten markiert werden. Das genaue Vorgehen ist in Algorithmus 3 angeführt.

Algorithmus 3 Markierungsstrategie GOAFEM

Input: Kantenmenge $\mathcal{E}$ und Fehlerschätzer $\eta^2_{\mathcal{T};F}(E)$, $\eta^2_{\mathcal{T};G}(E)$ für jede Kante $E \in \mathcal{E}$, Markierungsparameter $\vartheta \in (0,1)$.
Output: Menge $\mathcal{M} \subseteq \mathcal{E}$ der markierten Kanten.

1: definiere $\overline{\mathcal{M}}_F$ als Output von Algorithmus 2 für $(\mathcal{E}, \eta^2_{\mathcal{T};F}, \vartheta)$
2: definiere $\overline{\mathcal{M}}_G$ als Output von Algorithmus 2 für $(\mathcal{E}, \eta^2_{\mathcal{T};G}, \vartheta)$
3: $n := \min\{\#\overline{\mathcal{M}}_F, \#\overline{\mathcal{M}}_G\}$
4: wähle $\mathcal{M}_F \subseteq \overline{\mathcal{M}}_F$, sodass $\#\mathcal{M}_F = n$
5: wähle $\mathcal{M}_G \subseteq \overline{\mathcal{M}}_G$, sodass $\#\mathcal{M}_G = n$
6: $\mathcal{M} = \mathcal{M}_F \cup \mathcal{M}_G$

Durch die Wahl von n als Minimum in Zeile 3 wird garantiert, dass $\overline{\mathcal{M}}_F$ und $\overline{\mathcal{M}}_G$ mindestens n Elemente enthalten und deshalb Mengen, wie sie in den Zeilen 4 und 5 gewählt werden, auch tatsächlich existieren. Beachte jedoch, dass die Mengen $\mathcal{M}_F$ und $\mathcal{M}_G$ nicht disjunkt sein müssen, weshalb im Allgemeinen nur $\#\mathcal{M} \leq 2n$ und nicht Gleichheit gilt. Es kann darüber hinaus sowohl für das primale, als auch für das duale Problem passieren, dass mehr oder auch weniger Kanten markiert werden, als dies in den Zeilen 1 und 2 durch Anwenden von Algorithmus 2 vorgeschlagen wird. Es ist daher nicht trivial, dass (E1) gilt und wie diese Bedingung überhaupt zu formulieren ist. Dem werden wir uns im folgenden Abschnitt widmen.

5.3 Instanzoptimalität

5.3.1 Voraussetzungen für Energieoptimalität

Wir werden im Folgenden Energieoptimalität für das Produkt

$$(\mathfrak{F} \cdot \mathfrak{G})(\mathcal{T}) := \mathfrak{F}(\mathcal{T})\mathfrak{G}(\mathcal{T})$$

zeigen. Dies birgt jedoch einige Tücken. Zum Einen werden die optimalen Energien für Gitter mit höchstens $m \in \mathbb{N}$ mehr Elementen als in $\mathcal{T}_0$ im Allgemeinen nicht mehr simultan angenommen:

$$\mathfrak{F}_m^{\mathrm{opt}}\mathfrak{G}_m^{\mathrm{opt}} \leq \inf\{(\mathfrak{F} \cdot \mathfrak{G})(\mathcal{T}) \mid \mathcal{T} \in \mathbb{T}, \#(\mathcal{T} \setminus \mathcal{T}_0) \leq m\} =: (\mathfrak{F} \cdot \mathfrak{G})_m^{\mathrm{opt}}, \tag{5.9}$$

wobei die Gleichheit im Allgemeinen nicht gilt. Zum Anderen macht es die Produktstruktur schwierig, das Framework aus Kapitel 3 direkt anzuwenden. Wir werden daher dieses Framework für die Energien $\mathfrak{F}$ und $\mathfrak{G}$ separat anwenden und danach in geeigneter Weise zur Energieoptimalität des Produkts verbinden.

Wir zeigen zunächst das Markierungskriterium (E1). Wir erinnern uns, dass Proposition 4.40 für Algorithmus 2 auch gültig ist, falls nur eine Teilmenge der durch den Algorithmus vorgeschlagenen Kanten markiert wird. Damit können wir zeigen, dass (E1) mit beiden Schätzern sogar für die Menge der insgesamt markierten Kanten $\mathcal{M}$ gilt.

Proposition 5.6. *Seien $\eta^2_{\mathcal{T};F}$ und $\eta^2_{\mathcal{T};G}$ die Fehlerschätzer definiert in Definition 5.5. Dann erfüllt die Menge $\mathcal{M}$ der markierten Kanten aus Algorithmus 3 das Markierungskriterium (E1) für $\eta^2_{\mathcal{T};F}$ mit $\mu = \vartheta/2$:*

$$\eta^2_{\mathcal{T};F}\big(\mathrm{tail}_{\mathcal{T}}(\mathcal{M})\big) \geq \tfrac{\vartheta}{2}\,\#\mathcal{M}\,\eta^2_{\mathcal{T};F}\big(\mathrm{tail}_{\mathcal{T}}(E)\big) \quad \text{für alle } E \in \mathcal{E}(\mathcal{T}). \tag{5.10}$$

Analoges gilt für den Fehlerschätzer $\eta^2_{\mathcal{T};G}$.

Beweis. Da die Mengen $\mathcal{M}_F$ und $\mathcal{M}_G$ in Algorithmus 3 von Algorithmus 2 erzeugt wurden und $\mathcal{M} = \mathcal{M}_F \cup \mathcal{M}_G$ gilt, gilt jedenfalls $\mathcal{M} \neq \emptyset$.

Die Ungleichung (5.10) sehen wir folgendermaßen. Aus der Definition von $\mathcal{M}$ folgt, dass

$$\#\mathcal{M}_F \leq \#\mathcal{M} \leq 2\#\mathcal{M}_F. \tag{5.11}$$

Diese Ungleichungskette können wir nun in Verbindung mit Proposition 4.40 verwenden, da $\mathcal{M}_F \subseteq \overline{\mathcal{M}}_F$ gilt. Wir erhalten somit für alle Kanten $E \in \mathcal{E}$

$$\begin{aligned}\eta_{\mathcal{T};F}^2\big(\operatorname{tail}_{\mathcal{T}}(\mathcal{M})\big) &\geq \eta_{\mathcal{T};F}^2\big(\operatorname{tail}_{\mathcal{T}}(\mathcal{M}_F)\big)\\ &\overset{4.40}{\geq} \vartheta\,\#\mathcal{M}_F\,\eta_{\mathcal{T};F}^2\big(\operatorname{tail}_{\mathcal{T}}(E)\big)\\ &\geq \tfrac{\vartheta}{2}\,\#\mathcal{M}\,\eta_{\mathcal{T};F}^2\big(\operatorname{tail}_{\mathcal{T}}(E)\big).\end{aligned}$$

Der Beweis für $\eta_{\mathcal{T};G}^2$ folgt analog. □

Als nächstes widmen wir uns der Lower Diamond Estimate (E2). Wir vergleichen dazu die Oszillationsterme osc_F^2 und osc_G^2 mit den Oszillationen aus Definition 4.24. Es fällt dabei auf, dass der Kantenterm in Definition 5.2 analog zu den Neumann-Oszillationen in Kapitel 4 und der Elementterm hier analog zum Elementterm in Kapitel 4 definiert ist. Damit gelten analog zu Kapitel 4 die Eigenschaften

$$\begin{aligned}&R(\mathcal{T}) - R(\mathcal{T}') \simeq R(\mathcal{T}\setminus\mathcal{T}') \quad \text{für } \mathcal{T}' \geq \mathcal{T} \text{ und}\\ &R(\mathcal{U}_1) + R(\mathcal{U}_2) = R(\mathcal{U}_1 \cup \mathcal{U}_2) \text{ für disjunkte Teilmengen } \mathcal{U}_1, \mathcal{U}_2 \subseteq \mathcal{T}\end{aligned}$$

für $R = \operatorname{osc}_F^2, \operatorname{osc}_G^2$. Es lässt sich also der Beweis von Satz 4.28 auf die Energien in diesem Kapitel übertragen. Somit gilt für $\mathcal{F}$ und $\mathcal{G}$ jeweils die Lower Diamond Estimate (E2).

Schließlich wenden wir uns den diskreten lokalen Abschätzungen (E3) zu. Auch diese werden wir separat für das primale und das duale Problem beweisen. Die Beweise sind sehr ähnlich zu denen aus Kapitel 4, weshalb hier nur auf die wesentlichen Unterschiede eingegangen wird. Für Details sei auf die Beweise der entsprechenden Resultate in Kapitel 4 verwiesen. Wir beginnen mit der diskreten lokalen Zuverlässigkeit.

Satz 5.7. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $\mathcal{T}' \geq \mathcal{T}$ mit dazugehörigen Kantenmengen $\mathcal{E}$ und $\mathcal{E}'$, sowie $p \in \mathbb{N}$. Seien weiter $u_{\mathcal{T}} \in \mathcal{S}_0^p(\mathcal{T})$ und $u_{\mathcal{T}'} \in \mathcal{S}_0^p(\mathcal{T}')$ die Lösungen des primalen Problems (5.5) auf $\mathcal{T}$ und $\mathcal{T}'$. Dann existiert eine Konstante $C_{\mathrm{rel}} > 0$, sodass*

$$\|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!\|^2 \leq C_{\mathrm{rel}}\,\eta_{\mathcal{T};F}^2(\mathcal{E}\setminus\mathcal{E}'). \tag{5.12}$$

Die Konstante C_{rel} hängt nur von $\mathcal{T}_0$, p und der Diffusionsmatrix A ab. Ebenso gilt die selbe Aussage für die analogen Größen $w_{\mathcal{T}}, w_{\mathcal{T}'}$ und $\eta_{\mathcal{T};G}^2$ des dualen Problems (5.6).

Beweis. Es gilt auch hier die Galerkin-Orthogonalität für $v \in \mathcal{S}_0^p(\mathcal{T}) \subseteq \mathcal{S}_0^p(\mathcal{T}')$:

$$\int_\Omega A\nabla(u_{\mathcal{T}'} - u_{\mathcal{T}}) \cdot \nabla v \,\mathrm{d}x = \int_\Omega (f_1 - f_1)v \,\mathrm{d}x - \int_\Omega (f_2 - f_2)\cdot\nabla v \,\mathrm{d}s = 0.$$

Mit der Galerkin-Orthogonalität und der schwachen Formulierung des primalen Problems, sowie mit der Definition $v := (1 - \mathcal{Q}_{\mathcal{T}}^{\mathcal{T}'})(u_{\mathcal{T}'} - u_{\mathcal{T}})$ erhalten wir

$$\|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!\|^2 = \int_\Omega f_1 v \,\mathrm{d}x - \int_\Omega f_2 \cdot \nabla v \,\mathrm{d}x - \int_\Omega A\nabla u_{\mathcal{T}} \cdot \nabla v \,\mathrm{d}x.$$

Partielle Integration der letzten beiden Terme auf jedem Dreieck in $\mathcal{T}$ liefert

$$\begin{aligned}\|\!|u_{\mathcal{T}'} - u_{\mathcal{T}}|\!\|^2 = \sum_{T\in\mathcal{T}\setminus\mathcal{T}'} \int_T \big(f_1 + \operatorname{div}(f_2) + \operatorname{div}(A\nabla u_{\mathcal{T}})\big) v \,\mathrm{d}x \\ + \sum_{\substack{E\in\mathcal{E}\setminus\mathcal{E}'\\ E\subseteq\Omega}} \int_E [\![A\nabla u_{\mathcal{T}} + f_2]\!]\, v \,\mathrm{d}s.\end{aligned}$$

Diese beiden Terme können analog zu denen in Gleichung (4.31) behandelt werden, womit schließlich die Behauptung für das primale Problem folgt. Die Behauptung für das duale Problem folgt analog. □

Auch die diskrete lokale Effizienz folgt analog zu Kapitel 4, wobei hier etwas mehr Vorsicht geboten ist, da diese auch die Oszillationen involviert, die in diesem Kapitel zwar auf die selbe Weise definiert wurden, aber aufgrund der allgemeineren Problemstellung eine etwas andere Gestalt annehmen.

Satz 5.8. *Sei $\mathcal{T} \in \mathbb{T}$ ein Gitter und $\mathcal{T}' \geq \mathcal{T}$ mit dazugehörigen Kantenmengen $\mathcal{E}$ und $\mathcal{E}'$, sowie $p \in \mathbb{N}$. Seien weiter $u_{\mathcal{T}} \in \mathcal{S}^p_0(\mathcal{T})$ und $u_{\mathcal{T}'} \in \mathcal{S}^p_0(\mathcal{T}')$ die Lösungen von (5.5) auf $\mathcal{T}$, respektive $\mathcal{T}'$. Dann gilt für $E \in \mathcal{E} \setminus \mathcal{E}'$, dass*

$$\operatorname{res}_{\mathcal{T};F}(E) \lesssim |\!|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!|\!|_{\omega_E^{\mathrm{red}}} + \operatorname{res}_{\mathcal{T};F}(\omega_E^{\mathrm{red}}) + \operatorname{osc}_F(\omega_E^{\mathrm{red}}). \tag{5.13}$$

Falls $p \geq 2$, gilt für $T \in \mathcal{T} \setminus \mathcal{T}'$ außerdem, dass

$$\operatorname{res}_{\mathcal{T};F}(T) \lesssim |\!|\!|(u_{\mathcal{T}} - u_{\mathcal{T}'})|\!|\!|_T + \operatorname{osc}_F(T). \tag{5.14}$$

Insgesamt gilt die diskrete lokale Effizienz: Es gibt eine Konstante $C^F_{\mathrm{eff}} > 0$, sodass

$$\eta^2_{\mathcal{T};F}(\mathcal{E} \setminus \mathcal{E}') \leq C^F_{\mathrm{eff}}\big(|\!|\!|u_{\mathcal{T}} - u_{\mathcal{T}'}|\!|\!|^2 + \operatorname{osc}^2_F(\mathcal{T} \setminus \mathcal{T}')\big). \tag{5.15}$$

Die auftretenden Konstanten und insbesondere C^F_{eff} hängen nur von den Problemdaten und $\mathcal{T}_0$ ab. Außerdem gibt es eine Konstante C^G_{eff}, sodass analoge Aussagen für die entsprechenden Größen $w_{\mathcal{T}}, w_{\mathcal{T}'}, \eta^2_{\mathcal{T};G}$ und osc_G des dualen Problems (5.6) gelten.

Beweis. Wir zeigen die drei Behauptungen für das primale Problem einzeln.

Schritt 1: Abschätzung (5.13) für das Kantenresiduum..
Für eine Kante am Rand des Gebietes ist nichts zu tun, da nach Definition $\operatorname{res}_{\mathcal{T}:F}(E) = 0$ für alle $E \in \mathcal{E}^\Gamma$ gilt.

Sei also eine Kante $E \in \mathcal{E} \setminus \mathcal{E}'$ gegeben, sodass $E \notin \mathcal{E}^\Gamma$. Es bezeichne Π_E die $L^2(E)$-Orthogonalprojektion auf den Raum $P^{p-1}(E)$. Da nun ein potentiell nicht polynomieller Term f_2 im Kantenresiduum steht und wir gerne die in Kapitel 4 vorgestellte Bubble-Funktion Technik anwenden würden, müssen wir den Term f_2 erst analog zum Schritt 2 im Beweis von Satz 5.8 behandeln. Wir schätzen also das Residuum mittels Dreiecksungleichung ab:

$$\|[\![A\nabla u_{\mathcal{T}} + f_2]\!]\|_{L^2(E)} \leq \|(1 - \Pi_E)\,[\![f_2]\!]\|_{L^2(E)} + \|[\![A\nabla u_{\mathcal{T}} + \Pi_E f_2]\!]\|_{L^2(E)}\,.$$

Den ersten Term auf der rechten Seite können wir mit den Oszillationen $\operatorname{osc}_F(\omega_E^{\mathrm{red}})$ abschätzen. Da $[\![A\nabla u_{\mathcal{T}} + \Pi_E f_2]\!] \in P^{p-1}(E)$ ist, können wir diesen Ausdruck mit einer Bubble-Funktion multiplizieren und erhalten durch Fortsetzung gemäß Lemma (4.35) eine geeignete Testfunktion $v := [\![A\nabla u_{\mathcal{T}} + \Pi_E f_2]\!]\, b_E \in \mathcal{S}_0^p(\mathcal{T}')$. Wir können nun auch den zweiten Term der rechten Seite in obiger Ungleichung abschätzen:

$$\begin{aligned}\|[\![A\nabla u_{\mathcal{T}} + \Pi_E f_2]\!]\|_{L^2(E)}^2 &\overset{(4.34)}{\lesssim} \int_E [\![A\nabla u_{\mathcal{T}} + \Pi_E f_2]\!]\, v \,\mathrm{d}s \\ &= \int_E [\![A\nabla u_{\mathcal{T}} + f_2]\!]\, v \,\mathrm{d}s + \int_E (1-\Pi_E)\,[\![f_2]\!]\, v \,\mathrm{d}s\end{aligned}$$

Der zweite Term auf der rechten Seite lässt sich wiederum zu einem Oszillationsterm $\operatorname{osc}_F(\omega_E^{\mathrm{red}})\,\|v\|_{L^2(E)}$ abschätzen. Mit dem ersten verfahren wir analog zu Schritt 1 Fall 1 aus dem Beweis von Satz 4.37. Damit folgt dieser Schritt.

Schritt 2: Abschätzung (5.14) für das Elementresiduum.
Dieser Schritt ist analog zu Schritt 2 aus dem Beweis von Satz 4.37. Der einzige Unterschied ist, dass wir die potentiell nicht polynomiellen Terme $f_1 + \operatorname{div}(f_2)$ durch eine polynomielle Version ersetzen müssen (mit der $L^2(T)$-Orthogonalprojektion Π_T auf $P^{p-2}(T)$):

$$\operatorname{res}_{\mathcal{T};F} \le h_T\,\|(1-\Pi_T)(f_1+\operatorname{div}(f_2))\|_{L^2(T)} + h_T\,\|\Pi_T f_1 + \Pi_T \operatorname{div}(f_2) + \operatorname{div}(A\nabla u_{\mathcal{T}})\|_{L^2(T)}\,.$$

Wir haben jedoch die Oszillationen $\operatorname{osc}_F(T)$ so gewählt, dass sie den ersten Term aufnehmen können. Mit dem zweiten Term verfahren wir wie im Beweis von Satz 4.37 und erhalten schließlich (5.14).

Schritt 3: Beweis von (5.15).
Da wir die Abschätzungen (5.13) und (5.14) nachgewiesen haben, folgt dieser Schritt analog zu Schritt 3 im Beweis von Satz 4.37.

Für das duale Problem folgen die Aussagen analog. □

5.3.2 Verallgemeinerter Markierungsschritt

In den Zeilen 4 und 5 von Algorithmus 3 wird sichergestellt, dass gleich viele Kanten für das primale und das duale Problem markiert werden. Dies ist hier nur zur Vereinfachung der Darstellung gewählt. Es kann jedoch auch eine Konstante $C_{\mathrm{mark}} \in (0,\infty)$ gewählt werden, sodass bei o.B.d.A $n = \#\overline{\mathcal{M}}_F \le \#\overline{\mathcal{M}}_G$ in $\overline{\mathcal{M}}_G$ in etwa $C_{\mathrm{mark}} n$ Kanten markiert werden. Konkret kann in Zeile 4 eine Menge $\mathcal{M}_G$ gewählt werden, die

$$\#\mathcal{M}_G = \min\left\{\#\overline{\mathcal{M}}_G, \max\left\{1, \left\lfloor C_{\mathrm{mark}}\#\overline{\mathcal{M}}_F\right\rfloor\right\}\right\} \tag{5.16}$$

erfüllt. Diese etwas sperrige Definition ist dem Umstand geschuldet, dass die Menge $\mathcal{M}_G$ mindestens ein Element enthalten muss, höchstens aber so viele, wie von Algorithmus 2 vorgeschlagen wurden ($\#\overline{\mathcal{M}}_G$) enthalten darf.

Um zu zeigen, dass das Markierungskriterium aus Proposition 5.6 auch für diesen modifizierten Markierungsschritt erfüllt ist, muss eine Äquivalenz wie (5.11) gelten. Diese wollen wir nun kurz betrachten.

Aufgrund der Definition von $\mathcal{M}$ gilt jedenfalls

$$\#\mathcal{M}_F, \#\mathcal{M}_G \leq \#\mathcal{M}.$$

Es bleibt also nur die umgekehrte Abschätzung zu zeigen. Sei, wie oben erwähnt, o.B.d.A. $\#\overline{\mathcal{M}}_F = n$ mit $\#\overline{\mathcal{M}}_F \leq \#\overline{\mathcal{M}}_G$. Wir teilen den Beweis in mehrere Schritte.

Schritt 1: Abschätzung für $\mathcal{M}_F$.
Da die Mächtigkeit von $\mathcal{M}_G$ in (5.16) als Minimum definiert ist, gilt jedenfalls

$$\#\mathcal{M}_G \leq \max\left\{1, \left\lfloor C_{\text{mark}} \#\overline{\mathcal{M}}_F \right\rfloor\right\}.$$

Damit gilt nach der Definition von $\mathcal{M}$ in Zeile 6

$$\begin{aligned}\#\mathcal{M} &\leq \#\mathcal{M}_F + \#\mathcal{M}_G \\ &\leq \#\mathcal{M}_F + \max\left\{1, \left\lfloor C_{\text{mark}} \#\overline{\mathcal{M}}_F \right\rfloor\right\} \leq \max\{2, 1 + C_{\text{mark}}\} \#\mathcal{M}_F.\end{aligned}$$

Für $\mathcal{M}_F$ erhalten wir also insgesamt die Abschätzung

$$\#\mathcal{M}_F \geq \big(\max\{2, 1 + C_{\text{mark}}\}\big)^{-1} \#\mathcal{M}. \tag{5.17}$$

Schritt 2: Abschätzung für $\mathcal{M}_G$.
Hier müssen wir zwei Fälle unterscheiden, je nachdem welches Argument des Minimums in (5.16) das kleinere ist.

Fall 1: $\#\overline{\mathcal{M}}_G \leq \max\left\{1, \left\lfloor C_{\text{mark}} \#\overline{\mathcal{M}}_F \right\rfloor\right\}$.
In diesem Fall ist nicht viel zu tun. Nach der obigen Abschätzung für $\#\mathcal{M}_F$ und der Voraussetzung $\#\overline{\mathcal{M}}_F \leq \#\overline{\mathcal{M}}_G$ erhalten wir

$$\#\mathcal{M}_G = \#\overline{\mathcal{M}}_G \geq \#\overline{\mathcal{M}}_F = \#\mathcal{M}_F \overset{(5.17)}{\geq} \big(\max\{2, 1 + C_{\text{mark}}\}\big)^{-1} \#\mathcal{M}.$$

Fall 2: $\#\overline{\mathcal{M}}_G \geq \max\left\{1, \left\lfloor C_{\text{mark}} \#\overline{\mathcal{M}}_F \right\rfloor\right\}$.
In diesem Fall ist ein wenig mehr zu tun. Es gilt $\#\mathcal{M}_G \geq 1$, weshalb wir

$$\#\mathcal{M}_G \geq \left\lfloor C_{\text{mark}} \#\overline{\mathcal{M}}_F \right\rfloor \geq C_{\text{mark}} \#\overline{\mathcal{M}}_F - 1 \geq C_{\text{mark}} \#\overline{\mathcal{M}}_F - \#\mathcal{M}_G.$$

erhalten. Wir bringen den Term $\#\mathcal{M}_G$ auf die andere Seite der Ungleichung und nutzen aus, dass $\overline{\mathcal{M}}_F = \mathcal{M}_F$ ist, womit

$$2\#\mathcal{M}_G \geq C_{\text{mark}} \#\mathcal{M}_F$$

folgt. Mit der Definition von $\mathcal{M}$ können wir die Mächtigkeit dieser Menge deshalb folgendermaßen abschätzen:

$$\#\mathcal{M} \leq \#\mathcal{M}_F + \#\mathcal{M}_G \leq \left(\tfrac{2}{C_{\text{mark}}} + 1\right)\mathcal{M}_G.$$

In diesem Fall folgt also

$$\#\mathcal{M}_G \geq \left(1 + \tfrac{2}{C_{\text{mark}}}\right)^{-1} \mathcal{M}$$

Fassen wir alle Fälle für $\mathcal{M}_F$ und $\mathcal{M}_G$ zusammen, erhalten wir insgesamt

$$\#\mathcal{M}_F, \#\mathcal{M}_G \geq \min\left\{\big(\max\{2, 1 + C_{\text{mark}}\}\big)^{-1}, \left(1 + \tfrac{2}{C_{\text{mark}}}\right)^{-1}\right\} \#\mathcal{M}.$$

Es folgt also auch für den hier beschriebenen modifizierten Markierungsschritt eine Aussage analog zu Proposition 5.6, womit Instanzoptimalität gezeigt werden kann, wie wir im nächsten Abschnitt sehen werden.

5.3.3 Instanzoptimalität für die Fehlergröße

Mit den Vorbereitungen aus dem vorletzten Abschnitt können wir Instanzoptimalität zeigen.

Satz 5.9. *Sei $(\mathcal{T}_k)_{k\in\mathbb{N}_0}$ die Folge der Gitter, die von der AFEM (Algorithmus 1) mit den Fehlerschätzern $\eta^2_{\cdot;f}$ und $\eta^2_{\cdot;g}$ aus Definition 5.5 und dem Markierungsschritt aus Algorithmus 3 erzeugt werden. Dann existiert eine Konstante $C \geq 0$, sodass gilt: Für alle $\mathcal{T} \in \mathbb{T}$ mit $\#(\mathcal{T} \setminus \mathcal{T}_0) \leq \#(\mathcal{T}_k \setminus \mathcal{T}_0)/C$ gilt*

$$\begin{aligned}\big(\,|||u_{\mathrm{ex}} - u_{\mathcal{T}_k}|||^2 + \mathrm{osc}_F^2(\mathcal{T}_k)\big)\big(\,|||w_{\mathrm{ex}} - w_{\mathcal{T}_k}|||^2 + \mathrm{osc}_G^2(\mathcal{T}_k)\big)&\\ \leq 4\big(\,|||u_{\mathrm{ex}} - u_{\mathcal{T}}|||^2 + \mathrm{osc}_F^2(\mathcal{T})\big)\big(\,|||w_{\mathrm{ex}} - w_{\mathcal{T}}|||^2 + \mathrm{osc}_G^2(\mathcal{T})\big).&\end{aligned} \tag{5.18}$$

Die Konstante C hängt nur von ϑ und $\mathcal{T}_0$, sowie den Daten A, f_1, f_2, g_1 und g_2 ab.

Beweis. Wir haben in Abschnitt 5.3.1 die Voraussetzungen von Satz 3.14 für die Energien $\mathfrak{F}$ und $\mathfrak{G}$ gezeigt. Dieser liefert uns somit zwei Konstanten C^F_{mesh} und C^G_{mesh}, sodass

$$\begin{aligned}\mathfrak{F}(\mathcal{T}_k) &\leq \mathfrak{F}_m^{\mathrm{opt}} \quad \text{falls } \#(\mathcal{T}_k \setminus \mathcal{T}_0) \geq C^F_{\mathrm{mesh}} m,\\ \mathfrak{G}(\mathcal{T}_k) &\leq \mathfrak{G}_m^{\mathrm{opt}} \quad \text{falls } \#(\mathcal{T}_k \setminus \mathcal{T}_0) \geq C^G_{\mathrm{mesh}} m.\end{aligned} \tag{5.19}$$

Wählen wir $C := \max\{C^F_{\mathrm{mesh}}, C^G_{\mathrm{mesh}}\}$, so gelten für $\#(\mathcal{T}_k \setminus \mathcal{T}_0) \geq Cm$ beide Ungleichungen in (5.19). Für beliebiges $\mathcal{T} \in \mathbb{T}$ mit $m := \#(\mathcal{T} \setminus \mathcal{T}_0) \leq \#(\mathcal{T}_k \setminus \mathcal{T}_0)/C$ erhalten wir daher

$$(\mathfrak{F}\cdot\mathfrak{G})(\mathcal{T}_k) \leq \mathfrak{F}_m^{\mathrm{opt}}\mathfrak{G}_m^{\mathrm{opt}} \overset{(5.9)}{\leq} (\mathfrak{F}\cdot\mathfrak{G})_m^{\mathrm{opt}} \leq (\mathfrak{F}\cdot\mathfrak{G})(\mathcal{T}). \tag{5.20}$$

Laut Bemerkung 5.4 gilt für die Konstanten, die wir in Definition 5.3 von den Energien abgezogen haben, dass

$$\begin{aligned}\mathfrak{F}_\infty &= \tfrac{1}{2}a(u_{\mathrm{ex}}, u_{\mathrm{ex}}) - F(u_{\mathrm{ex}}),\\ \mathfrak{G}_\infty &= \tfrac{1}{2}a(w_{\mathrm{ex}}, w_{\mathrm{ex}}) - G(w_{\mathrm{ex}}).\end{aligned}$$

Die Energien in diesem Kapitel enthalten somit schon in geeigneter Weise die exakte Lösung, sind also bereits als Energiedifferenzen aufzufassen. Anwenden von Proposition 4.7 liefert daher

$$\begin{aligned}\mathfrak{F}(\mathcal{T}) &= \tfrac{1}{2}\,|||u_{\mathrm{ex}} - u_{\mathcal{T}}|||^2 + \mathrm{osc}_f^2(\mathcal{T}),\\ \mathfrak{G}(\mathcal{T}) &= \tfrac{1}{2}\,|||w_{\mathrm{ex}} - w_{\mathcal{T}}|||^2 + \mathrm{osc}_g^2(\mathcal{T}).\end{aligned}$$

Dies zeigt die Behauptung. Die behauptete Abhängigkeit der Konstanten C folgt aus der von C^F_{mesh} und C^G_{mesh} aus Satz 3.14. □

6 Numerische Resultate

Wir wollen in diesem Kapitel kurz einige numerische Resultate vorstellen. In der Arbeit [DKS16], die die Grundlage für die hier vorgestellten AFEM darstellt, gibt es keine numerischen Ergebnisse, weshalb hier zunächst der ursprüngliche Algorithmus für homogene Dirichlet-Daten und FEM-Diskretisierungen niedrigster Ordnung betrachtet wird. Hier wird der Fokus auf die durch den Algorithmus erzeugten Gitter und den Aufwand des Markierungsschrittes gelegt. Danach sehen wir uns noch ein Beispiel mit gemischten Randbedingungen an. Beide Beispiele beinhalten eine Geometrie mit einspringender Ecke, da hier Singularitäten in der Ableitung der Lösung auftreten, die die Stärke einer AFEM gegenüber FEM mit uniformer Verfeinerung verdeutlicht.

Der einzige Baustein, der die hier vorgestellte AFEM von konventionellen Methoden unterscheidet, ist der Markierungsschritt. Dieser wurde in das existierende Matlab-Paket p1afem implementiert [FPW11].

6.1 Homogene Dirichlet-Daten

Das hier betrachtete Modellproblem ist

$$\begin{aligned} -\Delta u &= 1 \quad \text{auf } \Omega, \\ u &= 0 \quad \text{auf } \partial\Omega, \end{aligned} \tag{6.1}$$

wobei $\Omega = (-1,1)^2 \setminus ([0,1] \times [-1,0])$ das Gebiet in Abbildung 6.1a ist. Zur Diskretisierung wurde das Setting aus Kapitel 4 mit $p = 1$ verwendet.

Dieses Problem wurde mittels dreier verschiedener (adaptiver) Verfahren gelöst:

(Unif) Das Gitter wird in jedem Schritt uniform verfeinert. Das heißt, dass jedes Gitter $\mathcal{T}_{k+1}$ für $k \in \mathbb{N}_0$ aus dem vorhergehenden Gitter $\mathcal{T}$ durch

$$\mathcal{T}_{k+1} = \operatorname{refine}(\mathcal{T}_k; \mathcal{E}_k)$$

entsteht.

(DKS) Die Methode aus Kapitel 4 wird angewendet. Jedes Gitter $\mathcal{T}_{k+1}$ entsteht aus dem vorhergehenden durch Verfeinerung einer Menge $\mathcal{M}_k \subseteq \mathcal{E}_k$ von Kanten:

$$\mathcal{T}_{k+1} = \operatorname{refine}(\mathcal{T}_k; \mathcal{M}_k).$$

Die Menge $\mathcal{M}_k$ stammt aus dem Algorithmus 2 mit Parameter $\vartheta = 1/4$.

(Std) Eine herkömmliche AFEM wird angewendet (siehe [FPW11]). Hierbei wird der Fehler über einen elementbasierten Residualschätzer

$$\widetilde{\eta}_k^2 : \mathcal{T}_k \to \mathbb{R} : T \mapsto \operatorname{res}_k^2(T) + \sum_{E \subseteq \partial T} \operatorname{res}_k^2(E)$$

geschätzt. Über Dörfler-Markierung wird eine Menge $\mathcal{U}_k \subseteq \mathcal{T}_k$ von Dreiecken markiert. Dies bedeutet, dass $\mathcal{U}_k$ die kleinste Menge ist, die

$$\widetilde{\eta}_k^2(\mathcal{U}_k) \geq \vartheta \widetilde{\eta}_k^2(\mathcal{T}_k)$$

erfüllt. Die Menge der markierten Dreiecke lässt sich mit

$$\mathcal{M}_k := \bigcup_{T \in \mathcal{U}_k} \{E \in \mathcal{E}_k \mid E \subseteq \partial T\}$$

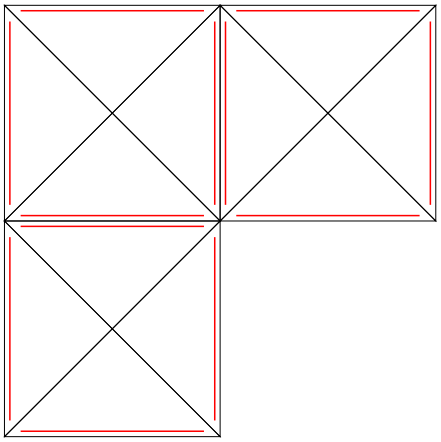
in eine Menge von markierten Kanten übersetzen. Es wird also jede Kante des Dreiecks markiert, sodass im Verfeinerungsschritt drei neue Kanten und vier neue Dreiecke entstehen. Daraus entsteht mittels $\mathcal{T}_{k+1} = \operatorname{refine}(\mathcal{T}_k; \mathcal{M}_k)$ das nächste Gitter. Der Markierungsparameter ist auch hier $\vartheta = 1/4$.

Bemerkung 6.1. Die Berechnung des Schweifs $\operatorname{tail}_E(\mathcal{T})$ für eine Kante $E \in \mathcal{E}$ wurde hier durch die bereits in p1afem implementierte Verfeinerungsroutine vorgenommen. Diese berechnet im ersten Schritt für eine Menge an markierten Kanten alle Kanten, die verfeinert werden müssen, um wieder ein zulässiges Gitter zu erhalten. Dieser Schritt wurde für jede Kante des Gitters durchgeführt, wodurch eine Liste der Schweife für jede Kante erhalten wird. Diese muss, um in Zeile 5 von Algorithmus 2 den Indikator zu berechnen, ständig um alle Kanten bereinigt werden, die in $\widetilde{\mathcal{M}}$ aufgenommen werden.

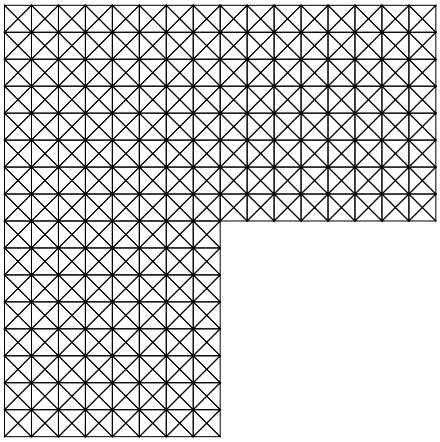
Dass diese Methode im Vergleich zum herkömmlichen Markierungsschritt nicht sehr effizient ist, zeigt Abbildung 6.6. //

Die nach einigen Verfeinerungsschritten entstandenen Gitter sind in Abbildung 6.1 gezeigt. Hier ist deutlich ersichtlich, dass sowohl der Standardalgorithmus (Abbildung 6.1d), als auch der Algorithmus (DKS) (Abbildung 6.1c) zu der einspringenden Ecke hin verfeinern. Allerdings erkennt man beim Algorithmus (DKS), dass auch vermehrt im Inneren des Gebiets verfeinert wurde. Dies entsteht durch die Einbeziehung des Schweifs einer Kante in den Markierungsschritt (siehe Algorithmus 2). Dieser Effekt ist auch nach weiteren Verfeinerungsschritten erkennbar, wie in Abbildung 6.3 ersichtlich ist.

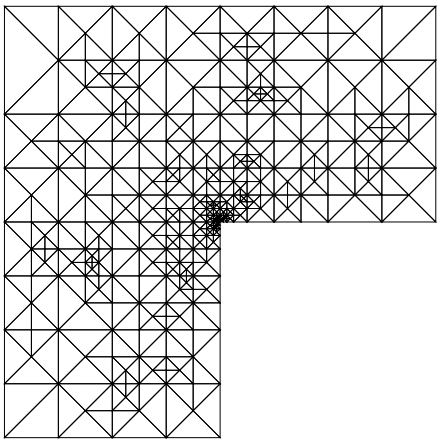
Die Werte der Fehlerschätzer der drei vorgestellten Verfahren ist in Abbildung 6.2 dargestellt. Man beobachtet für die beiden adaptiven Verfahren eine Konvergenzrate $-1/2$ und für uniforme Verfeinerung eine etwas schlechtere Konvergenzrate. Die Rate für uniforme Verfeinerung wird sich für eine höhere Elementanzahl $-1/3$ annähern, wie wir im folgenden Abschnitt sehen werden.



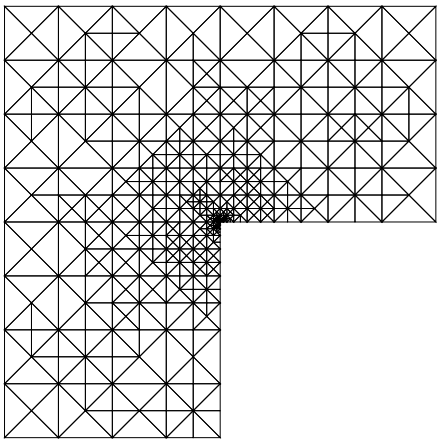
(a) *Initiales Gitter - 12 Elemente*



(b) *(Unif) - 768 Elemente*



(c) *(DKS) - 657 Elemente*



(d) *(Std) - 685 Elemente*

Abbildung 6.1: *Vergleich von Gittern aus verschiedenen AFEM für das Problem* (6.1). *Die Verfeinerungskanten des initialen Gitters sind in (a) mit roten Linien markiert.*

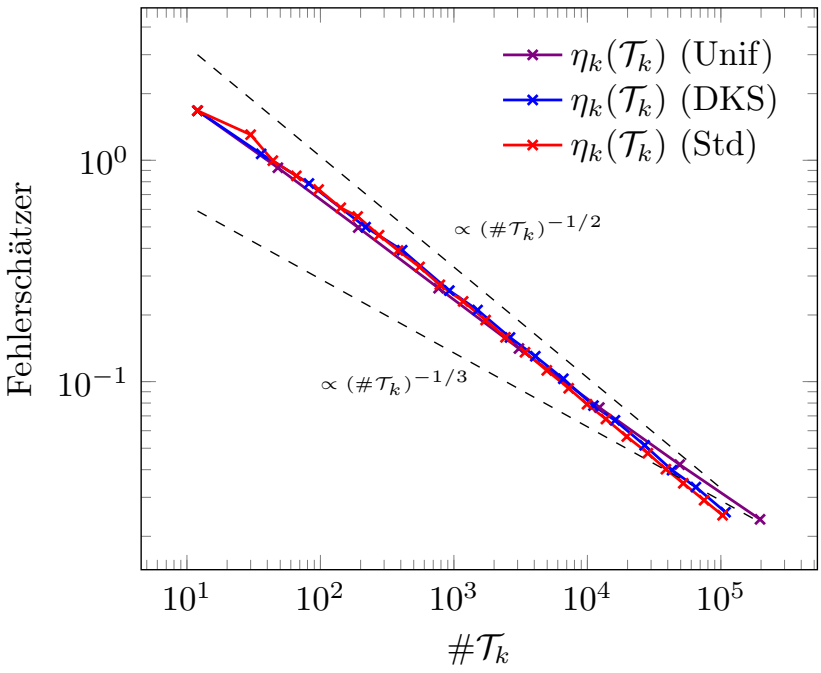


Abbildung 6.2

6.2 Gemischte Randbedingungen

Wir betrachten nun ein Modellproblem mit etwas allgemeineren Randbedingungen auf dem Gebiet Ω aus dem vorigen Abschnitt. Dafür geben wir die exakte Lösung vor:

$$u_{\mathrm{ex}}(x,y) = r(x,y)^{2/3} \sin\left(\tfrac{2}{3}\varphi(x,y)\right),$$

wobei $r(x,y), \varphi(x,y)$ die Polarkoordinaten des Punktes (x,y) sind. Durch Nachrechnen sieht man, dass

$$\Delta u_{\mathrm{ex}} = 0.$$

Weiter teilen wir den Rand auf in die zwei Teilmengen

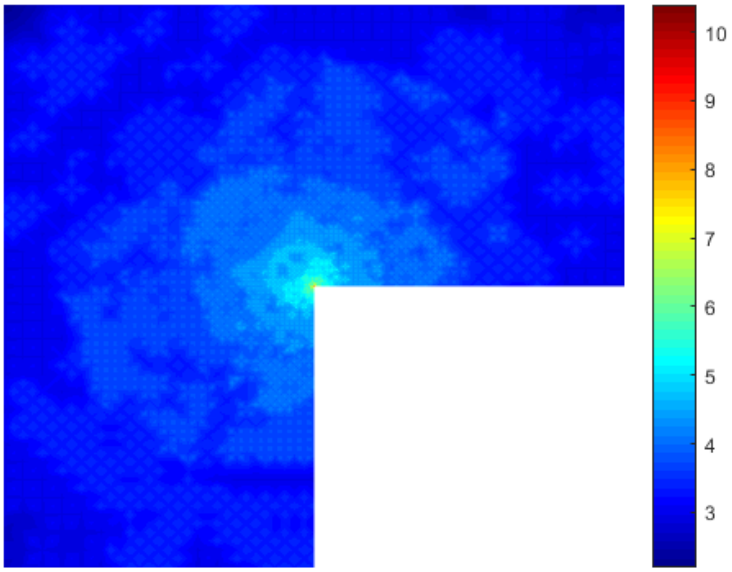
$$\Gamma_D := \big(\{0\} \times [-1,0]\big) \cup \big([0,1] \times \{0\}\big), \quad \Gamma_N := \partial\Omega \setminus \Gamma_D.$$

Es gilt auf Γ_D, dass $\varphi = 0$, bzw. $\varphi = \frac{3}{2}\pi$, weshalb $u_{\mathrm{ex}} \equiv 0$ auf Γ_D.

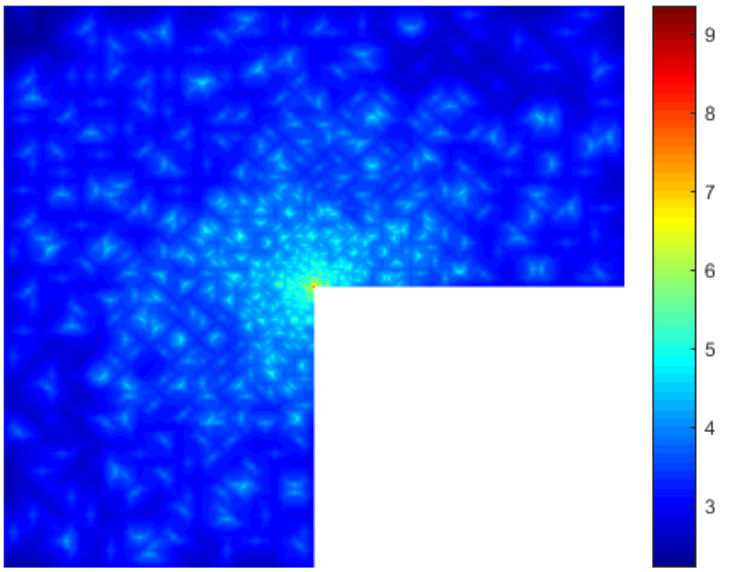
Damit ergibt sich das folgende Modellproblem:

$$\begin{aligned} -\Delta u &= 0 && \text{auf } \Omega, \\ u &= 0 && \text{auf } \Gamma_D, \\ \nabla u \cdot \nu &= \nabla u_{\mathrm{ex}} \cdot \nu && \text{auf } \Gamma_N. \end{aligned} \tag{6.2}$$

Als initiales Gitter der Diskretisierung dieses Problems wählen wir dasjenige Gitter, das aus dem Gitter aus Abbildung 6.1a durch viermalige uniforme Verfeinerung hervor geht.



(a) *(Std) - 11000 Elemente*



(b) *(DKS) - 10000 Elemente*

Abbildung 6.3: *Lokale Gitterweite der verschiedenen AFEM ($\vartheta = 1/4$) für das Problem (6.2). Aufgetragen ist hier der Wert von $-\log_2(h_E)$ für alle Kanten. Beim herkömmlichen Algorithmus (Std) wird das Gitter gleichmäßig zur einspringenden Ecke hin feiner. Bei (DKS) jedoch sind auch innerhalb des Gebiets Regionen mit sehr kleiner lokaler Gitterweite erkennbar.*

Dieses Problem wurde mittels (DKS) mit der Ordnung $p = 1$ gelöst. Dies wurde mit dem herkömmlichen Algorithmus (Std) und uniformer Verfeinerung (Unif) verglichen, die Fehlerschätzer sind in Abbildung 6.4 aufgetragen. Man erkennt deutlich die verschiedenen Konvergenzraten der Fehlerschätzer. Für uniforme Verfeinerung ist diese $-\frac{1}{3}$, wohingegen sie für beide adaptive Verfahren $-\frac{1}{2}$ ist. Der besseren Vergleichbarkeit wegen wurde dazu auch für die herkömmliche AFEM in jedem Schritt der kantenbasierte Fehlerschätzer η_k aus Defintion 4.30 und für (DKS) der elementbasierte Fehlerschätzer $\widetilde{\eta}_k$ berechnet. Da sich die Fehlerchätzer η_k und $\widetilde{\eta}_k$ nur durch die Summation ihrer Beiträge unterscheiden, unterscheiden sich deren Werte für große Werte von $\#\mathcal{T}_k$ nur um einen (annähernd) konstanten Faktor.

In Abbildung 6.5 sind außerdem die Fehlerschätzer des Algorithmus (DKS) für verschiedene Werte von ϑ aufgetragen. Für jeden Wert von $\vartheta > 0$ stellt sich bei hinreichend großer Elementzahl die Konvergenzrate $-1/2$ ein. Für $\vartheta = 0$ entspricht das Verfahren uniformer Verfeinerung, da in jedem Schritt alle Kanten des Gitters markiert werden. Hier beobachtet man erwartungsgemäß die Rate $-1/3$.

Schließlich zeigt Abbildung 6.6 einen Vergleich der Zeiten, die ein Markierungsschritt von (DKS), beziehungsweise (Std) benötigt. Man erkennt hier einen quadratischen Verlauf der Zeit von (DKS), wie aus den theoretischen Überlegungen in Bemerkung 4.41 zu erwarten ist. Die vorliegende Implementierung ist lediglich ein proof-of-concept und nicht auf Effizienz ausgelegt. Den größten Teil der Zeit benötigt hierbei das Berechnen des Schweifs $\operatorname{tail}_{\mathcal{T}}(E)$ einer Kante $E \in \mathcal{E}$. Dieser wird gemäß seiner Definition als Vergleich von $\mathcal{T}$ und $\operatorname{refine}(\mathcal{T}; \{E\})$ berechnet.

6.3 GOAFEM

Zuletzt wollen wir die Methoden aus Kapitel 5 numerisch betrachten. Dazu sehen wir uns das Modellproblem (vergleiche [FPvdZ16, section 4.5])

$$\begin{aligned} -\Delta u &= \operatorname{div}(f_2) && \text{auf } \Omega, \\ u &= 0 && \text{auf } \Gamma_D \end{aligned} \tag{6.3}$$

an. Hier ist $\Omega = (0,1)^2$ und $f_2 \in L^2(\Omega)$ mit $T_F := \operatorname{conv}\{(0,0), (0, \frac{1}{2}), (\frac{1}{2}, 0)\}$ gegeben als

$$f_2(x) := \begin{cases} (1,0)^T & \text{für } x \in T_F, \\ (0,0)^T & \text{sonst .} \end{cases}$$

Die schwache Formulierung von (6.3) lautet: Finde $u \in H^1_0(\Omega)$, sodass

$$\int_\Omega \nabla u \nabla v \, \mathrm{d}x = -\int_{T_F} \frac{\partial v}{\partial x_1} \, \mathrm{d}x \quad \text{für alle } v \in H^1_0(\Omega).$$

Wir definieren außerdem ein Zielfunktional $G \in H^{-1}(\Omega)$ ähnlich zu F mit der Menge $T_G := \operatorname{conv}\{(1,1), (1, \frac{1}{2}), (\frac{1}{2}, 1)\}$ als

$$G(u) := -\int_{T_G} \frac{\partial v}{\partial x_1} \, \mathrm{d}x$$

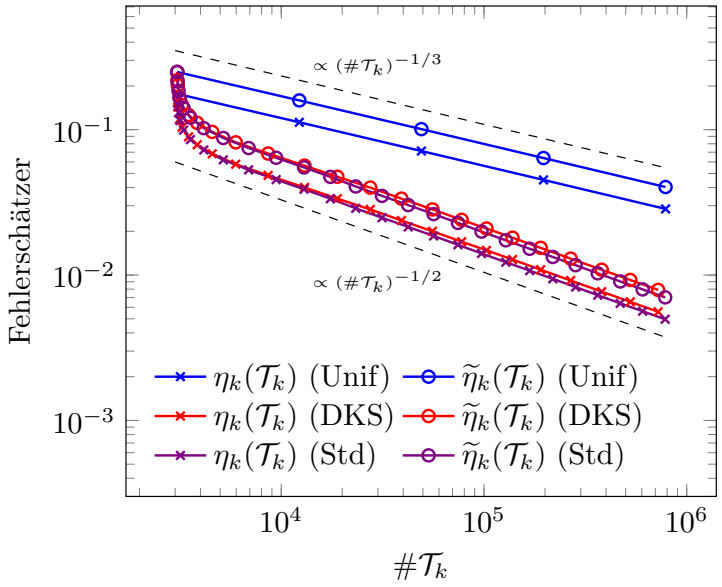


Abbildung 6.4: *Fehlerschätzer für verschiedene AFEM ($\vartheta = 1/4$) für das Problem (6.2). Durch das in Abschnitt 6.2 gewählte Problem ergibt sich für uniforme Verfeinerung eine schlechtere Rate ($-1/3$), als für adaptive Verfahren ($-1/2$). Hierbei bezeichnet η_k den kantenbasierten und $\widetilde{\eta}_k$ den elementbasierten Fehlerschätzer.*

für $u \in H_0^1(\Omega)$. Wir wollen den Wert von G an der exakten Lösung u_{ex} von (6.3) approximieren. Der interessante Punkt an diesem Beispiel sind die Singularitäten, die sich durch die Sprünge von rechter Seite und Zielfunktional ergeben. Das initiale Gitter ist in Abbildung 6.7 gezeigt, zusammen mit den Flächen T_F und T_G.

Wie wir in Kapitel 5 gesehen haben, ist für ein Gitter $\mathcal{T} \in \mathbb{T}$ das Produkt $\eta_{\mathcal{T};F}\eta_{\mathcal{T};G}$ ein geeigneter Schätzer für den Fehler $|G(u_{\text{ex}}) - G(u_{\mathcal{T}})|$. Wir haben das Problem auf drei verschiedene Arten mit Finiten Elementen der Ordnung $p = 1$ gelöst und den eben genannten Schätzer berechnet:

1. Es wurden nur Kanten verfeinert, die von Algorithmus 2 basierend auf dem primalen Fehlerschätzer $\eta^2_{\mathcal{T};F}$ markiert wurden.

2. Es wurden nur Kanten verfeinert, die von Algorithmus 2 basierend auf dem dualen Fehlerschätzer $\eta^2_{\mathcal{T};G}$ markiert wurden.

3. Es wurden alle Kanten verfeinert, die von Algorithmus 3 basierend auf dem primalen und dem dualen Fehlerschätzer markiert wurden.

Die Ergebnisse sind in Abbildung 6.8 aufgetragen. Man sieht in den ersten beiden Zeilen, dass nur jeweils die Singularitäten entweder vom primalen, oder vom dualen Problem durch die Markierungsstrategie erfasst worden sind. Dies führt im Fehlerschätzer, der zur Markierung der Kanten herangezogen wurde, zur optimalen Rate $-1/2$, im jeweils anderen

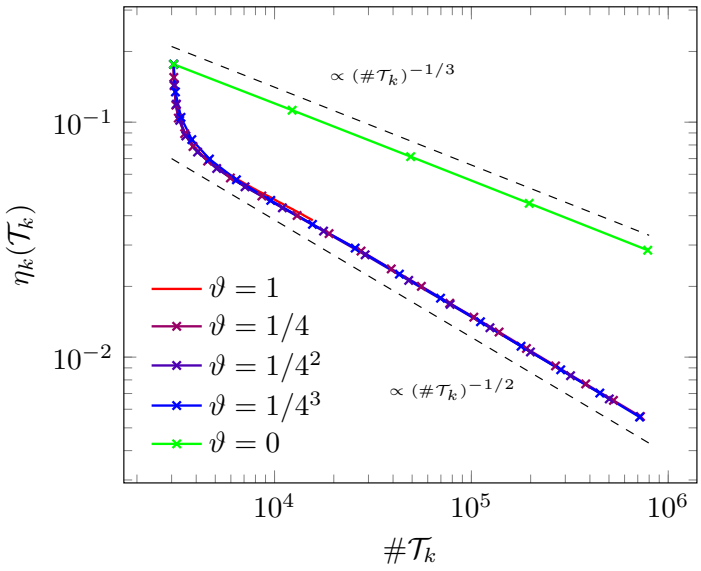


Abbildung 6.5: *Fehlerschätzer für verschiedene Werte von ϑ in (DKS) für das Problem (6.2). Beachte, dass nur für $\vartheta \in (0, 1]$ ein echtes adaptives Verfahren vorliegt. Für $\vartheta = 0$ werden alle Kanten markiert, was uniformer Verfeinerung entspricht.*

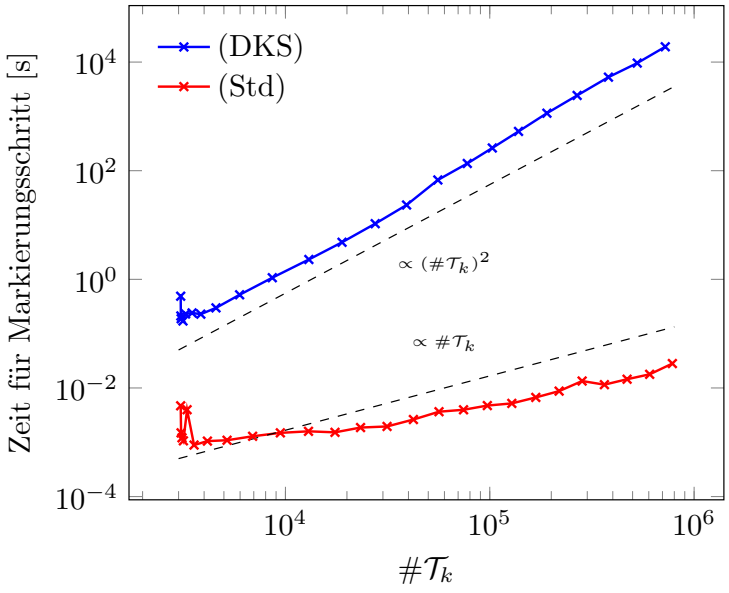


Abbildung 6.6: *Laufzeiten der Markierungsschritte von (Std) und (DKS) (6.2). Nach einer kurzen vorasymptotischen Phase zeigt der Schritt aus (DKS) quadratisches Wachstum.*

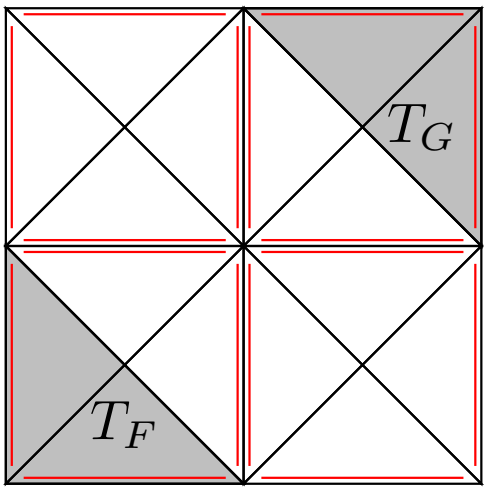


Abbildung 6.7: *Initiales Gitter zur numerischen Lösung von* (6.3). *Die Referenzkanten sind rot eingezeichnet.*

aber zu einer schlechteren. In Summe erhalten wir bei beiden Methoden die Rate $-5/6$ für den Gesamtfehlerschätzer $\eta_{\mathcal{T};F}\eta_{\mathcal{T};G}$.

Markierung mit Algorithmus 3 löst jedoch die Singularitäten beider Probleme geeignet auf. Dies führt bei beiden Fehlerschätzern auf die optimale Rate $-1/2$ (was in Kapitel 5 durch die separate Instanzoptimalität von primalem und dualem Problem bewiesen wurde). Folglich ist auch die Rate des Gesamtfehlerschätzers $\eta_{\mathcal{T};F}\eta_{\mathcal{T};G}$ mit -1 höher als bei Markierung mit nur einem der beiden Fehlerschätzer.

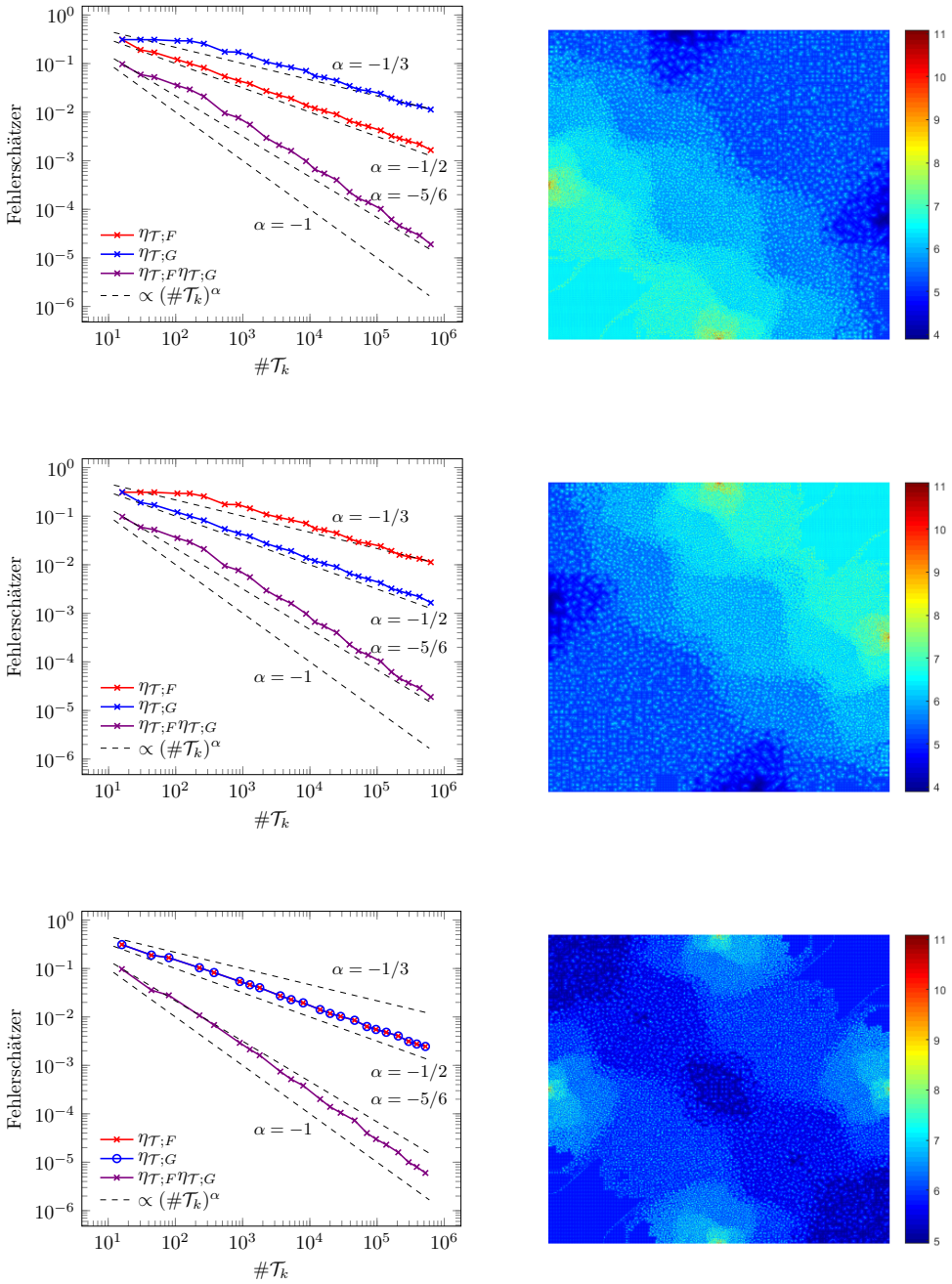


Abbildung 6.8: *Lokale Gitterweite und Konvergenzraten von verschiedenen AFEM ($\vartheta = 1/4$) für das Problem (6.3). Als Gitterweite aufgetragen ist hier der Wert von $-\log_2(h_E)$ für alle Kanten, die Gitter haben jeweils etwa $6 \cdot 10^5$ Elemente. Es wurden hier adaptive Verfeinerungen des Gitters durchgeführt, basierend auf den Fehlerschätzern des primalen Problems (oben) und des dualen Problems (mittig), sowie basierend auf beiden Fehlerschätzern (unten), wobei die Menge der markierten Kanten durch Algorithmus 3 bestimmt wurde.*

Literaturverzeichnis

[BDD04] Peter Binev, Wolfgang Dahmen, and Ron DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97:219–268, 2004.

[Bra13] Dietrich Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer Spektrum, Berlin, 5. Auflage, 2013.

[CFPP14] Carsten Carstensen, Michael Feischl, Markus Page, and Dirk Praetorius. Axioms of adaptivity. *Comput. Math. Appl.*, 67(6):1195–1253, 2014.

[DKS16] Lars Diening, Christian Kreuzer, and Rob Stevenson. Instance optimality of the adaptive maximum strategy. *Found. Comput. Math.*, 16(1):33–68, 2016.

[Dör96] Willy Dörfler. A Convergent Adaptive Algorithm for Poisson's Equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.

[EGP18] Christoph Erath, Gregor Gantner, and Dirk Praetorius. Optimal convergence behavior of adaptive FEM driven by simple (h-$h/2$)-type error estimators. *Preprint, arXiv:1805.00715*, 2018.

[Eva10] Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, RI, 2nd edition, 2010.

[Fei] Michael Feischl. Rate optimality of adaptive algorithms. Dissertation, TU Wien, Institut für Analysis und Scientific Computing, 2015.

[FPvdZ16] Michael Feischl, Dirk Praetorius, and Kristoffer van der Zee. An abstract analysis of optimal goal-oriented adaptivity. *SIAM J. Numer. Anal.*, 54(3):1423–1448, 2016.

[FPW11] Stefan Funken, Dirk Praetorius, and Philip Wissgott. Efficient implementation of adaptive P1-FEM in matlab. *Comput. Methods Appl. Math.*, 11(4):460–490, 2011.

[Hab] Alexander Haberl. Instanz-Optimalität adaptiver FEM. Diplomarbeit, TU Wien, Institut für Analysis und Scientific Computing, 2015.

[KPP13] Michael Karkulik, David Pavlicek, and Dirk Praetorius. On 2D newest vertex bisection: Optimality of mesh-closure and H^1-stability of L_2-projection. *Constr. Approx.*, 38(2):213–234, 2013.

[Ste07] Rob Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[SZ90] L. Ridgway Scott and Shangyou Zhang. Finite element interpolations of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[Ver13] Rüdiger Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods*. Oxford University Press, Oxford, 1st edition, 2013.

[Yos80] Kôsaku Yosida. *Functional Analysis*. Springer-Verlag, Berlin, 6th edition, 1980.