Skript zur Vorlesung im SS 2008

Adaptive Approximations- und Diskretisierungsverfahren

Thorsten Raasch

15. Dezember 2009

Inhaltsverzeichnis

Einleitung

Bei der effizienten numerischen Simulation realistischer Probleme aus der Praxis, insbesondere bei der Modellierung durch Differentialgleichungen, ist der Einsatz adaptiver Approximationsmethoden unerlässlich. Während bei nichtadaptiven Verfahren üblicherweise global mit der gleichen Diskretisierungsfeinheit gearbeitet wird, verwenden adaptive Verfahren nur in solchen Bereichen eine feinere Diskretisierung, wo lokale Besonderheiten der unbekannten Lösung des Problems aufgelöst werden müssen. Bei der konkreten Durchführung wird dabei die Diskretisierung mit Hilfe sogenannter a posteriori-Fehlerschätzer dynamisch an die unbekannte Lösung des Problems angepasst. Adaptivität dient vorwiegend zur Reduktion der Größe des diskretisierten Problems bei vorgegebener Genauigkeit und erlaubt so überhaupt erst die Durchführbarkeit einer Approximation realistischer Probleme.

Von besonderem mathematischen Interesse sind dabei adaptive Verfahren mit beweisbaren Konvergenz- und Komplexitätseigenschaften. Es gilt hierbei, die folgenden zentralen Fragen zu klären:

(I) Ist die gegebene adaptive Approximationsmethode überhaupt konvergent, d.h. erlaubt sie, eine Approximation der unbekannten Lösung mit einer vorgegebenen Zielgenauigkeit zu bestimmen? Unter welchen Voraussetzungen, z.B. an die unbekannte Lösung oder an gewisse Problem- oder Verfahrensparameter, ist diese Konvergenz gesichert?

(II) Welches Ergebnis liefert die Analyse der Konvergenz*rate*, d.h. des Verhältnisses zwischen Anzahl benutzter Freiheitsgrade und der dabei erreichten Genauigkeit, im Vergleich mit dem "Benchmark" der besten N-Term-Approximation? Unter welchen Voraussetzungen kann das Verfahren in diesem Sinne ggf. als "optimal" nachgewiesen werden?

(III) Zahlen sich für die vorgegebene Problemklasse adaptive Verfahren überhaupt aus, d.h. sind hierdurch im Vergleich zu "naiven" uniformen Diskretisierungsstrategien wirklich höhere Konvergenzraten unter den gleichen Voraussetzungen erzielbar?

Diese Grundfragen werden im Folgenden anhand mehrerer Beispiele genauer erörtert. Neben strukturell einfacheren Verfahren wie etwa der Spline-Approximation einer gegebenen Funktion werden vor allem adaptive Wavelet- und Finite Elemente-Methoden für lineare elliptische Probleme diskutiert und ihre Konvergenz- und Optimalitätseigenschaften analysiert.

1 Grundlagen

1.1 Konvergente adaptive Verfahren

1.1.1 Definition, Allgemeines

In normierten Räumen $(X, \|\cdot\|_X)$ und $(Y, \|\cdot\|_Y)$ sei das Problem gegeben, zu festen Daten $x \in X$ das Bild

$$y = f(x) \in Y \tag{1.1}$$

unter einer Abbildung $f : X \to Y$ zu berechnen. Hierbei stehe f abstrakt für den Lösungsoperator eines gegebenen mathematischen Problems. Dieser kann einfach strukturiert sein, etwa wenn es sich um die Auswertung von Funktionalen $f : X \to \mathbb{R}$, z.B. eines Integrals, handelt. In vielen praktischen Anwendungen ist die exakte Abbildung f allerdings nur implizit gegeben. Dies ist z.B. der Fall bei der Modellierung mit Differentialgleichungen durch die Lösung eines entsprechenden Rand- oder Anfangswertproblems. Da es im Allgemeinen nicht möglich sein wird, y wirklich störungsfrei zu berechnen. werden wir uns im Folgenden darauf beschränken, Näherungslösungen y_ϵ mit einem garantierten *absoluten* Fehler von

$$\|y - y_\epsilon\|_Y \leq \epsilon \tag{1.2}$$

zu bestimmen.

Die konkrete Berechnung einer Näherung y_ϵ geschehe dabei deterministisch, d.h. es gebe eine Abbildung $F : X \times \mathbb{R}_+ \to Y$ mit $y_\epsilon = F(x, \epsilon)$. Wir nehmen darüberhinaus an, dass die Abbildung F durch einen Algorithmus realisiert werde.

Definition 1.1. *Ein* Algorithmus *ist eine aus endlich vielen, wohldefinierten Operationen bestehende Handlungsvorschrift zu Lösung eines Problems. Ein Algorithmus g überführt einen* Eingabewert $(x_1, \ldots, x_n)$ *in einen* Ausgabewert $(y_1, \ldots, y_m)$*, wir schreiben hierfür im Folgenden $g[x_1, \ldots, x_n] \to (y_1, \ldots, y_m)$.*

Hiermit kann direkt definiert werden, was unter einem adaptiven Verfahren sowie unter dessen Konvergenz zu verstehen ist.

Definition 1.2. *Ein Algorithmus $F[x, \epsilon, \ldots] \to y_\epsilon$ heißt* adaptives Verfahren *zur Berechnung von $f(x)$, wenn zur Berechnung von y_ϵ höchstens endlich viele Rechenoperationen notwendig sind, d.h. wenn F für alle zugelassenen Eingabewerte (x, ϵ) terminiert. Wir nennen F* konvergent *(oder auch adaptiv im engeren Sinne), wenn für jeden Ausgabewert y_ϵ die Fehlertoleranz* (1.2) *eingehalten wird.*

Um für einen gegebenen Algorithmus F nachzuweisen, dass er ein konvergentes adaptives Verfahren realisiert, ist also neben der Überprüfung der Ausgabefehlertoleranz auch

das Terminierungsverhalten zu überprüfen. Es sei bereits an dieser Stelle betont, dass bei der Konvergenzanalyse adaptiver Verfahren Rundungsfehler üblicherweise vernachlässigt werden.

Bemerkung 1.3. *Diese Definition von* Konvergenz *ist ganz natürlich und deckt sich mit anderen Konvergenzdefinitionen in der Numerischen Mathematik. Gemeinhin heißt ein numerisches Verfahren konvergent, wenn die von ihm erzeugten Näherungslösungen für immer "feiner" werdende Diskretisierungsparameter gegen die exakte Lösung konvergieren. Im Rahmen adaptiver Verfahren liegt aufgrund der gewünschten Fehlerabschätzung* (1.2) *obige Konvergenzdefinition auf der Hand.*

Um die Berechnung einer hinreichend genauen Näherung y_ϵ sicherstellen zu können, sind in einem Algorithmus mindestens die folgenden drei Grundbausteine notwendig:

1. ein *numerisches Grundverfahren* zur Berechung einer Näherung $\tilde{y}$;
2. ein *Fehlerschätzer* $\mathcal{E}$ zur Kontrolle von $\|y - \tilde{y}\|_Y$;
3. eine *Verfeinerungsstrategie*, um auf der Basis einer noch unzureichenden Näherung $\tilde{y}$ eine verbesserte Näherung $\overline{y}$ zu berechnen.

Typischerweise bestehen adaptive Verfahren genau aus einer Iteration dieser drei Schritte, wobei nach dem Fehlerschätzungsschritt ein passendes Stopkriterium abgefragt wird. Die Konvergenzanalyse besteht dann im wesentlichen daraus, das Zusammenwirken zwischen einem gegebenenfalls "lokalen" Fehlerschätzer und der darauf aufsetzenden Verfeinerungsstrategie zu studieren. An das numerische Grundverfahren wird zunächst nur die Forderung gestellt, implementierbar zu sein. Der dafür benötigte Rechenaufwand geht dann allerdings entscheidend in die Optimalitätsanalyse des Gesamtverfahrens ein.

Für die Konvergenzanalyse ist der Fehlerschätzer $\mathcal{E}$ von zentraler Bedeutung. Typischerweise wird in einem adaptiven Verfahren eine a posteriori-Abschätzung benutzt, d.h. zur Kontrolle von $\|y - \tilde{y}\|$ gehen neben $\tilde{y}$ ggf. noch weitere, im Lauf des Verfahrens gewonnene Informationen ein. Wir schreiben daher im Folgenden auch $\mathcal{E} = \mathcal{E}(\tilde{y})$. Im Gegensatz zu a priori-Fehlerschätzern, die keine konkreten Rechenergebnisse benutzen, sind a posteriori-Schätzer meist genauer, siehe auch Abschnitt 1.1.2. Als wichtig erweisen werden sich zwei Grundeigenschaften eines Fehlerschätzers:

Definition 1.4. *Ein Fehlerschätzer $\mathcal{E}(\tilde{y})$ heißt* verlässlich*, wenn*

$$\|y - \tilde{y}\|_Y \leq C\mathcal{E}(\tilde{y}) \tag{1.3}$$

gilt, mit einer von y und $\tilde{y}$ unabhängigen Konstanten $C > 0$. $\mathcal{E}(\tilde{y})$ heißt effizient*, wenn*

$$\|y - \tilde{y}\|_Y \geq C'\mathcal{E}(\tilde{y}) \tag{1.4}$$

gilt, mit einer weiteren, von y und $\tilde{y}$ unabhängigen Konstanten $C' > 0$.

Mit anderen Worten: ein verlässlicher Fehlerschätzer unterschätzt den exakten Fehler nicht, vorhandene Fehler werden also angezeigt. Umgekehrt ist ein effizienter Fehlerschätzer auch dann klein, wenn der exakte Fehler klein ist, d.h. angezeigte Fehler sind auch wirklich vorhanden.

1.1.2 Beispiele

Fixpunktiteration

Zur Herleitung eines konvergenten adaptiven Verfahrens zur Berechnung von Fixpunkten sei an den folgenden wichtigen Satz erinnert (Beweis: Analysis I/II oder Numerik I).

Satz 1.5 (Fixpunktsatz von Banach)**.** *Sei $(B, \|\cdot\|_B)$ ein Banachraum, $\Omega \subset B$ und $g : B \to B$ mit $g(\Omega) \subset \Omega$ und der Kontraktionsbedingung*

$$\|g(u) - g(v)\|_B \leq q\|u - v\|_B, \quad u, v \in \Omega \tag{1.5}$$

für ein $0 < q < 1$. Dann besitzt g einen eindeutigen Fixpunkt $z = g(z) \in \Omega$ und die Iteration $z^{(k+1)} = g(z^{(k)})$ konvergiert für einen beliebigen Startwert $z^{(0)} \in \Omega$ gegen z mit den Fehlerabschätzungen

$$\|z^{(k)} - z\|_B \leq \frac{q}{1-q}\|z^{(k)} - z^{(k-1)}\|_B \qquad \textit{(a posteriori)} \tag{1.6}$$

$$\leq \frac{q^k}{1-q}\|z^{(1)} - z^{(0)}\|_B \qquad \textit{(a priori).} \tag{1.7}$$

Auf der Basis von Satz 1.5 liegt es nahe, den folgenden Algorithmus zur adaptiven Berechnung des Fixpunktes z zu formulieren.

Algorithmus 1 FIXPOINT$[g, q, \epsilon] \to z_\epsilon$

Wähle $z^{(0)} \in B$ beliebig.
$n := 0$
repeat
  $z^{(n+1)} := g(z^{(n)})$
  $n := n + 1$
until $\frac{q}{1-q}\|z^{(n)} - z^{(n-1)}\|_B \leq \epsilon$
$z_\epsilon := z^{(n)}$

Der Algorithmus **FIXPOINT** ist offenbar deterministisch und er terminiert, da $(z^{(n)})_n$ nach dem Fixpunktsatz von Banach konvergiert und somit für ein $n \in \mathbb{N}$ die Bedingung $\|z^{(n)} - z^{(n-1)}\| \leq \frac{(1-q)\epsilon}{q}$ gilt. Ferner ist wegen (1.6) die Fehlertoleranz für die Rückgabe z_ϵ erfüllt, es handelt sich damit um einen konvergenten adaptiven Algorithmus. Natürlich geht in diese kurze Überlegung in starkem Maße das Vorwissen über die Abbildungs- und Kontraktionseigenschaften von g ein.

Adaptive Spline-Approximation

Als zweites Beispiel betrachten wir einen adaptiven Algorithmus, der bei der Approximation einer gegebenen Funktion $f \in C[0, 1]$ durch stückweise konstante Funktionen in der Maximumnorm $\|v\|_\infty := \|v\|_{[0,1],\infty} := \sup_{x\in[0,1]} |v(x)|$ entsteht. Jede solche Approximation f_ϵ lässt sich zunächst offenbar schreiben als

$$f_\epsilon = \sum_{I \in \mathcal{G}_\epsilon} c_I \chi_I. \tag{1.8}$$

Hierbei bilden die Intervalle aus $\mathcal{G}_\epsilon$ (offen, halboffen oder abgeschlossen) eine *Partition* von $[0,1]$, d.h. $\bigcup_{I\in\mathcal{G}_\epsilon} I = [0,1]$ und verschiedene Intervalle $I, J \in \mathcal{G}_\epsilon$ sind disjunkt. Der Algorithmus zur Berechnung von f_ϵ zerfällt dadurch in zwei Schritte:

1. Berechnung einer geeigneten Partition $\mathcal{G}_\epsilon$
2. Berechnung der Koeffizienten c_I für alle $I \in \mathcal{G}_\epsilon$

Beginnen wir mit dem zweiten Punkt. Um die Argumentation zu erleichtern, nehmen wir an, es existiere eine Koeffizientenabbildung $(f, I) \mapsto c_I(f)$ mit

$$\|f - c_I(f)\|_{I,\infty} = \inf_{c\in\mathbb{R}} \|f - c\|_{I,\infty}, \tag{1.9}$$

d.h. $c_I(f)\chi_I$ sei genau die Bestapproximierende auf I. Nachfolgendes Lemma zeigt, dass diese starke Forderung zumindest bei monotonen Funktionen f realisierbar ist.

Lemma 1.6. *Für $f \in C[0,1]$ und jedes Intervall $I \subset [0,1]$ gilt*

$$\inf_{c\in\mathbb{R}} \|f - c\|_{I,\infty} = \Big\| f - \frac{1}{2}\big(\sup_{x\in I} f(x) + \inf_{x\in I} f(x)\big)\Big\|_{I,\infty}. \tag{1.10}$$

Beweis: Wegen $f \in C[0,1]$ und $I \subset [0,1]$ sind $a := \inf_{x\in I} f(x)$ und $b := \sup_{x\in I} f(x)$ reelle Zahlen mit $a \le f(x) \le b$ für alle $x \in I$. Mit $|y| = \max\{y, -y\}$ für $y \in \mathbb{R}$ gilt für jedes feste $c \in \mathbb{R}$

$$\|f - c\|_{I,\infty} = \max\big\{ \sup_{x\in I}(f(x) - c), \sup_{x\in I}(c - f(x))\big\} = \max\{b - c, c - a\}.$$

Dieses Maximum wird minimal, wenn beide Terme übereinstimmen. Für das Minimum c^* folgt daher $b - c^* = c^* - a$, also $c^* = \frac{1}{2}(a + b)$. □

Bemerkung 1.7. *Falls f nicht monoton ist, wird man in der Praxis mehrere Funktionswerte $f(x_i)$ bei Punkten $x_i \in I$ berechnen und die entsprechenden Maxima und Minima hierüber zur Approximation von $c_I(f)$ heranziehen. In diesem Fall ist allerdings eine weitere Fehleranalyse notwendig.*

Durch die Annahme an die Existenz der Routine $(f, I) \mapsto c_I(f)$ ist nun noch eine passende Partition $\mathcal{G}_\epsilon$ zu ermitteln. Die Grundidee des folgenden Algorithmus ist es, ein Teilintervall I einer gegebenen Partition in zwei Unterintervalle ("Kinder") von I zu unterteilen, falls auf I der Fehler noch zu groß ist. Für diese Strategie benötigen wir offenbar einen lokalen Fehlerschätzer $\mathcal{E} : \mathcal{I} \to \mathbb{R}$, wobei $\mathcal{I}$ die Menge aller Teilintervalle von $[0,1]$ sei. Die Abbildung $\mathcal{E}$ besitze mindestens die folgenden beiden Eigenschaften:

$$\mathcal{E}(I) \ge \|f - c_I(f)\|_{I,\infty}, \tag{1.11}$$

$$\lim_{h\to 0} \sup\{\mathcal{E}(I) : |I| = h\} = 0. \tag{1.12}$$

Algorithmus 2 ADAPTIVE$[f, \epsilon] \to f_\epsilon$

```
if E([0,1]) ≤ ε then
  G := {[0,1]}
else
  G := ∅
  B := {[0,1]}
  repeat
    for all I ∈ B do
      for all children J of I do
        if E(J) ≤ ε then
          G := G ∪ {J}
        else
          B := B ∪ {J}
        end if
      end for
      B := B \ {I}
    end for
  until B = ∅
end if
f_ε := Σ_{I∈G} c_I(f) χ_I
```

Hierbei besagt (1.11) gerade, dass es sich bei $\mathcal{E}$ um einen verlässlichen Fehlerschätzer auf I handelt. (1.12) drückt zu einem gewissen Grad die Effizienz von $\mathcal{E}$ aus. Unter diesen Voraussetzungen lässt sich der Algorithmus 2 **ADAPTIVE** angeben und seine Konvergenz nachweisen, siehe auch [9]. $\mathcal{G}$ und $\mathcal{B}$ stehen dabei für "good" bzw. "bad invervals".

Satz 1.8. ADAPTIVE *ist ein konvergenter adaptiver Algorithmus zur stückweise konstanten Approximation von* $f \in C[0,1]$.

Beweis: Zunächst zur Terminierung des Algorithmus. Bei jedem Durchlauf in der äußeren repeat-Schleife wird die aktuelle Menge $\mathcal{B}$ komplett abgearbeitet, jedes $I \in \mathcal{B}$ wird dabei höchstens durch seine beiden Kind-Intervalle ersetzt. Also halbiert sich in jedem Durchlauf die Länge der Intervalle in $\mathcal{B}$. Wegen (1.12) gibt es daher eine äußere Iterationstiefe, bei der keine Kinder mehr in $\mathcal{B}$ erzeugt werden, woraufhin der Algorithmus terminiert.

Aufgrund der Partitionseigenschaft von $\mathcal{G}_\epsilon$, (1.11) und der Terminierungsbedingung in **ADAPTIVE** rechnet man nun leicht Konvergenz nach:

$$\|f - f_\epsilon\|_\infty = \max_{I \in \mathcal{G}_\epsilon} \|f - c_I(f)\|_{I,\infty} \leq \max_{I \in \mathcal{G}_\epsilon} \mathcal{E}(I) \leq \epsilon.$$

□

Es bleibt zu klären, für welche Funktionenklassen $\mathcal{F} \ni f$ man überhaupt einen berechenbaren Fehlerschätzer $\mathcal{E}$ mit (1.11) und (1.12) angeben kann.

Lemma 1.9. *Sei $f \in C^1[0,1]$ und sei $c_I(f)$ gemäß (1.9) berechnet. Dann erfüllt*

$$\mathcal{E}(I) = \int_I \big|f'(x)\big| \,\mathrm{d}x \tag{1.13}$$

die Bedingungen (1.11) und (1.12).

Beweis: Für alle $x, y \in I$ gilt zunächst

$$\big|f(x) - f(y)\big| = \Big| \int_x^y f'(z) \,\mathrm{d}z \Big| \le \int_I \big|f'(z)\big| \,\mathrm{d}z.$$

Mit $a := \inf_{x\in I} f(x)$ und $b := \sup_{x\in I} f(x)$ existieren zu $\epsilon > 0$ Punkte $x_\epsilon, y_\epsilon \in I$ mit $f(x_\epsilon) \le a + \epsilon$ und $f(y_\epsilon) \ge b - \epsilon$. Es folgt für jedes $x \in I$

$$\begin{aligned}
\Big|f(x) - \frac{1}{2}(a+b)\Big| &\le \frac{1}{2}\Big(\big|f(x) - a\big| + \big|f(x) - b\big|\Big) \\
&\le \frac{1}{2}\Big(\big|f(x) - f(x_\epsilon)\big| + \big|f(x) - f(y_\epsilon)\big|\Big) + \epsilon \\
&\le \int_I \big|f'(z)\big| \,\mathrm{d}z + \epsilon,
\end{aligned}$$

so dass mit $\epsilon \to 0$ und dem Supremum über $x \in I$ die Abschätzung (1.11) folgt. Da f' stetig auf $[0,1]$ ist, folgt (1.12) via

$$\mathcal{E}(I) = \int_I \big|f'(x)\big| \,\mathrm{d}x \le \|f'\|_\infty |I|.$$

□

Der Algorithmus **ADAPTIVE** ist hiermit als konvergent nachgewiesen. Allerdings ist die Anzahl der in einer Splineapproximation f_ϵ benutzten Intervalle in Abhängigkeit der vorgegebenen Genauigkeit ϵ bis jetzt noch unklar. Folgender Satz liefert eine asymptotische Aussage für $\epsilon \to 0$.

Satz 1.10. *Sei $\mathcal{E}$ der Fehlerschätzer aus (1.13). Ferner sei zu $f \in C^1[0,1]$ und $0 < \epsilon < \mathcal{E}([0,1])$ die Approximation $f_\epsilon := \mathbf{ADAPTIVE}[f, \epsilon]$ gegeben. Ist $\mathcal{G}_\epsilon$ die zu f_ϵ gehörende Partition, so gilt die Abschätzung*

$$\#\mathcal{G}_\epsilon \le \frac{2}{\epsilon} \int_0^1 M_{f'}(x) \,\mathrm{d}x \le \frac{2}{\epsilon} \|f'\|_\infty, \tag{1.14}$$

mit der Hardy-Littlewood-Maximalfunktion für $g \in L_1(0,1)$,

$$M_g(x) := \sup_{[0,1] \supset I \ni x} \frac{1}{|I|} \int_I |g(y)| \,\mathrm{d}y. \tag{1.15}$$

Beweis: Wegen $\epsilon < \mathcal{E}([0,1])$ ist der triviale Fall $\mathcal{G}_\epsilon = \{[0,1]\}$ mit $\#\mathcal{G}_\epsilon = 1$ ausgeschlossen. Folglich ist jedes $I \in \mathcal{G}_\epsilon$ Kind eines dyadischen Intervalls $J \neq \mathcal{G}_\epsilon$, so dass für alle $x \in J$ gilt

$$\epsilon < \mathcal{E}(J) = \int_J \big|f'(y)\big| \,\mathrm{d}y \leq |J| M_{f'}(x).$$

Mit dem Infimum über $x \in I \subset J$ und $f' \in L_1(0,1)$ folgt

$$\epsilon \leq |J| \inf_{x \in I} M_{f'}(x) \leq \frac{|J|}{|I|} \int_I M_{f'}(y) \,\mathrm{d}y = 2 \int_I M_{f'}(y) \,\mathrm{d}y,$$

so dass Aufsummation über alle $I \in \mathcal{G}_\epsilon$ und die Partitionseigenschaft schließlich liefert

$$\#\mathcal{G}_\epsilon \cdot \epsilon \leq 2 \sum_{I \in \mathcal{G}_\epsilon} \int_I M_{f'}(y) \,\mathrm{d}y = 2 \int_0^1 M_{f'}(y) \,\mathrm{d}y.$$

Für $g \in C[0,1]$ folgt für alle $x \in [0,1]$

$$M_g(x) = \sup_{[0,1] \supset I \ni x} \frac{1}{|I|} \int_I |g(y)| \,\mathrm{d}y \leq \|g\|_\infty \sup_{[0,1] \supset I \ni x} \frac{1}{|I|} \int_I \mathrm{d}y = \|g\|_\infty,$$

also die rechte Ungleichung in (1.14). □

Es ist also mit **ADAPTIVE** möglich, mit $N := C\epsilon^{-1}$ Freiheitsgraden einen Approximationsfehler von $\epsilon = CN^{-1}$ zu realisieren. Hiermit lässt sich eine gewisse Konvergenz*rate* des adaptiven Algorithmus ablesen.

Definition 1.11. *Ein adaptives Verfahren $F[x, \epsilon, \ldots] \to y_\epsilon$ heißt konvergent* mit Rate $s > 0$*, falls die Approximationen y_ϵ in dem zugrunde liegenden Erzeugendensystem $(\psi_\lambda)_{\lambda \in \nabla}$ von Y jeweils eine Darstellung $y_\epsilon = \sum_{\lambda \in \nabla} c_{\epsilon,\lambda} \psi_\lambda$ mit höchstens endlich vielen nichttrivialen Koeffizienten $\mathbf{c}_\epsilon = (c_{\epsilon,\lambda})_{\lambda \in \nabla}$ besitzen und $C > 0$ existiert mit*

$$\|y - y_\epsilon\|_Y \leq C(\#\mathbf{c}_\epsilon)^{-s}, \quad \textit{für } \epsilon \to 0. \tag{1.16}$$

Das Erzeugendensystem ist in diesem Fall die Menge der charakteristischen Funktionen χ_I, wobei I dyadische Teilintervalle von $[0,1]$ durchläuft. Unter den Voraussetzungen von Satz 1.10 ist also **ADAPTIVE** konvergent mit Rate $s = 1$.

Bemerkung 1.12. *Sowohl für die Durchführung von* **ADAPTIVE** *als auch für die Komplexitätsanalyse haben wir bis jetzt $f \in C^1[0,1]$ angenommen. Wir werden später sehen, dass unter dieser starken Voraussetzung auch eine äquidistante Approximation von f mit Approximationsrate 1 existiert, d.h. adaptive Verfahren bringen unter der Voraussetzung $f \in C^1[0,1]$ zunächst keinen Vorteil. Allerdings lassen sich die hinreichenden Konvergenzvoraussetzungen für* **ADAPTIVE** *bei genauerer Analyse stark abschwächen. So konvergiert das adaptive Verfahren bei gleicher Wahl von $\mathcal{E}$ mit Rate 1 auch für $f \in W^1(L_p(0,1))$, $p > 1$. Dies ist nicht der Fall bei äquidistanter Approximation.*

Demo

1.2 Funktionenräume

Die bei der Lösung partieller Differentialgleichungen auftretenden Funktionen leben statt in Räumen wie C^k meistens in Banachräumen *schwach* differenzierbarer Funktionen. Diese Funktionenräume und darauf aufbauende Hilfsmittel werden im Folgenden eingeführt, siehe z.B. auch [12, Anhang A]. Falls nicht anderweitig eingeschränkt, sei im Folgenden immer $\Omega \subset \mathbb{R}^n$ ein Gebiet, d.h. offen und zusammenhängend. Mit $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^n x_i y_i$, $\mathbf{x} = (x_i), \mathbf{y} = (y_i) \in \mathbb{R}^n$, bezeichnen wir das Euklidische Skalarprodukt auf $\mathbb{R}^n$, $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ sei die Euklidische Norm. Weiter sei zu $\mathbf{x} \in \mathbb{R}^n$ und $r > 0$ der offene Ball $B(\mathbf{x}, r) := \big\{ \mathbf{y} \in \mathbb{R}^n \big| \|\mathbf{x} - \mathbf{y}\| < r \big\}$ erklärt. Schließlich bezeichnen wir mit $|A| := \int_A \mathrm{d}x$ das Lebesgue-Maß einer Menge $A \subset \mathbb{R}^n$.

1.2.1 L_p-Räume

Für $1 \leq p < \infty$ sei

$$L_p(\Omega) := \Big\{ f : \Omega \to \mathbb{R} \Big| \|f\|_{L_p(\Omega)} := \Big(\int_\Omega |f(\mathbf{x})|^p \mathrm{d}\mathbf{x} \Big)^{1/p} < \infty \Big\}, \tag{1.17}$$

für $p = \infty$ setzen wir

$$L_\infty(\Omega) := \Big\{ f : \Omega \to \mathbb{R} \Big| \|f\|_{L_\infty(\Omega)} := \underbrace{\inf_{\substack{N \subset \Omega \\ |N|=0}} \sup_{\in \Omega \setminus N} |f(\mathbf{x})|}_{= \operatorname{ess\,sup}_{\mathbf{x} \in \Omega} |f(x)|} < \infty \Big\}. \tag{1.18}$$

Dann ist $(L_p(\Omega), \| \cdot \|_{L_p(\Omega)}$ für alle $1 \leq p \leq \infty$ ein Banachraum. Falls Ω ein endliches Lebesgue-Maß $|\Omega|$ besitzt (z.B. wenn Ω beschränkt ist), gilt für $1 \leq p \leq q \leq \infty$ die stetige Einbettung $L_q(\Omega) \hookrightarrow L_p(\Omega)$ mit

$$\|f\|_{L_p(\Omega)} \leq |\Omega|^{1/p - 1/q} \|f\|_{L_q(\Omega)}. \tag{1.19}$$

1.2.2 C^k- und Hölder-Räume, Testfunktionen

Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet, d.h. es existiere eine Konstante $M > 0$ mit $\|\mathbf{x}\| \leq M$ für alle $\mathbf{x} \in \Omega$. Dann definieren wir den Raum stetiger, beschränkter Funktionen

$$C^0(\Omega) := \big\{ f : \Omega \to \mathbb{R} \big| f \text{ stetig und } \|f\|_{C^0(\overline{\Omega})} := \sup_{\mathbf{x} \in \overline{\Omega}} |f(\mathbf{x})| < \infty \big\}, \tag{1.20}$$

sowie für jedes $k \in \mathbb{N}$ den Raum aller k-mal stetig differenzierbaren Funktionen

$$C^k(\Omega) := \Big\{ f : \Omega \to \mathbb{R} \Big| \|f\|_{C^k(\overline{\Omega})} := \sum_{|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|_{C^0(\overline{\Omega})} < \infty \Big\}. \tag{1.21}$$

Für $k \in \mathbb{N}_0$ und $0 < \alpha \leq 1$ sei

$$C^{k,\alpha}(\Omega) := \Big\{ f \in C^k(\Omega) \Big| \|f\|_{C^{k,\alpha}(\overline{\Omega})} := \|f\|_{C^k(\overline{\Omega})} + \underbrace{\sum_{|\boldsymbol{\beta}|=k} |D^{\boldsymbol{\beta}} f|_{C^{0,\alpha}(\overline{\Omega})}}_{=: |f|_{C^{k,\alpha}(\overline{\Omega})}} < \infty \Big\} \tag{1.22}$$

der Raum aller k-mal *Hölder-stetig* (bzw. für $\alpha = 1$ *Lipschitz-stetig*) differenzierbaren Funktionen, wobei

$$\|f\|_{C^{0,\alpha}} := \sup_{\substack{\mathbf{x},\mathbf{y}\in\Omega \\ \mathbf{x}\neq\mathbf{y}}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^\alpha}. \tag{1.23}$$

Alle C^k- bzw. $C^{k,\alpha}$-Räume sind Banachräume mit ihrer jeweiligen Norm. Es gelten die stetigen Einbettungen

$$C^{k,\alpha}(\Omega) \hookrightarrow C^{k',\alpha}, \quad C^{k,\alpha}(\Omega) \hookrightarrow C^{k,\alpha'} \text{ für } k \geq k',\ \alpha \geq \alpha'. \tag{1.24}$$

Im Folgenden schreiben wir auch $C^{k,0} := C^k$.

Beispiel 1.13. *Die Funktion $f(x) = \sqrt{x}$ liegt genau in $C^{0,\alpha}((0,1))$ für $0 < \alpha \leq \frac{1}{2}$.*

Weiter habe eine Funktion $f : \Omega \to \mathbb{R}$ *kompakten Träger* in Ω, wenn eine kompakte Menge $K \subset \Omega$ existiert (d.h. K berührt $\partial\Omega$ nicht), so dass $f(\mathbf{x}) = 0$ für alle $\mathbf{x} \in \Omega \setminus K$. Die kleinste kompakte Menge mit dieser Eigenschaft bezeichnen wir mit $\operatorname{supp} f$. Wir definieren für $k \in \mathbb{N}$

$$C_0^k(\Omega) := \{f \in C^k(\Omega) | \operatorname{supp} f \subset \Omega\} \tag{1.25}$$

den Raum aller k-mal stetig differenzierbaren Funktionen auf Ω mit Nullrandwerten k-ter Ordnung, $\mathcal{D}(\Omega) := C_0^\infty(\Omega) := \bigcap_{k\in\mathbb{N}} C_0^k(\Omega)$ ist die Menge der *Testfunktionen*. $\mathcal{D}(\Omega)$ ist dicht in allen $L_p(\Omega)$, $1 \leq p < \infty$.

1.2.3 Randintegrale

Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Gebiet. Für die Einführung schwacher Ableitungen, aber auch als Grundvoraussetzung für verschiedene Sätze aus der Vektoranalysis benötigen wir Zusatzannahmen an die Regularität des Gebietsrandes $\partial\Omega$.

Definition 1.14. *Sei $k \in \mathbb{N}$ und $0 \leq \alpha \leq 1$. Dann ist Ω ein $C^{k,\alpha}$-*Gebiet*, falls sich $\partial\Omega$ lokal als Graph einer $C^{k,\alpha}$-Funktion schreiben lässt. Genauer: es existieren für $1 \leq j \leq m$ Orthonormalbasen $(\mathbf{e}_i^j)_{i=1}^n$ des $\mathbb{R}^n$, $r_j, h_j > 0$ und Funktionen $g_j \in C^{k,\alpha}(B(\mathbf{0}, r_j))$, so dass die Mengen*

$$U_j := \left\{\mathbf{x} = \sum_{i=1}^{n-1} y_i \mathbf{e}_i^j + t\mathbf{e}_n^j \in \mathbb{R}^n \middle| \|\mathbf{y}\| < r_j, |t - g_j(\mathbf{y})| < h_j\right\} \tag{1.26}$$

eine offene Überdeckung von Ω bilden, wobei für $U_j \ni \mathbf{x} = \sum_{i=1}^{n-1} y_i \mathbf{e}_i^j + t\mathbf{e}_n^j$ gilt

(i) $t = g_j(\mathbf{y}) \Rightarrow \mathbf{x} \in \partial\Omega$,

(ii) $0 < t - g_j(\mathbf{y}) < h_j \Rightarrow \mathbf{x} \in \Omega$,

(iii) $0 > t - g_j(\mathbf{y}) > -h_j \Rightarrow \mathbf{x} \notin \overline{\Omega}$.

Speziell heißt ein $C^{0,1}$-Gebiet auch Lipschitzgebiet.

Definition 1.15. *Sei $(U_j)_{j=0}^m$ eine offene Überdeckung von Ω. Dann heißt $(\eta_j)_{j=0}^m \subset \mathcal{D}(\Omega)$* Teilung der Eins *bzgl. $(U_j)_{j=0}^m$, falls $\eta_j \geq 0$ und $\sum_{j=0}^m \eta_j = 1$ auf $\overline{\Omega}$.*

Wir nehmen an, dass zu jeder Überdeckung von Ω eine solche Teilung der Eins existiert. Dann lässt sich wie folgt im Fall von Lipschitzgebieten ein Randintegralbegriff einführen.

Definition 1.16. *Sei Ω ein Lipschitzgebiet und $(U_j)_{j=1}^m$ eine offene Überdeckung von $\partial\Omega$ wie in* (1.26)*. Dann heißt $f : \partial\Omega \to \mathbb{R}$* messbar *(bzw.* integrierbar*), falls die Funktion*

$$\mathbb{R}^{n-1} \supset B(0, r_j) \ni \mathbf{y} \mapsto f\Big(\sum_{i=1}^{n-1} y_i \mathbf{e}_i^j + g_j(\mathbf{y})\mathbf{e}_n^j\Big)$$

messbar bzw. integrierbar ist. Das Randintegral *über f ist definiert als $\int_{\partial\Omega} f \,\mathrm{d}S := \sum_{j=1}^m \int_{\partial\Omega} \eta_j f \,\mathrm{d}S$, wobei für $g : \partial\Omega \to \mathbb{R}$ mit $\operatorname{supp} g \subset \partial\Omega \cap U_j$ gesetzt wird*

$$\int_{\partial\Omega} g \,\mathrm{d}S := \int_{B(0,h_j)} g\Big(\sum_{i=1}^{n} y_i \mathbf{e}_i^j + g_j(\mathbf{y})\mathbf{e}_n^j\Big)\sqrt{1 + \|\nabla g_j(\mathbf{y})\|^2}\,\mathrm{d}\mathbf{y}. \tag{1.27}$$

Diese Definition des Randintegrals ist unabhängig von der speziellen Wahl der Überdeckung (U_j) und von den Funktionen g_j. Ferner ist das Integral (1.27) überhaupt sinnvoll, da aufgrund der Lipschitz-Stetigkeit der g_j fast überall deren Gradient ∇g_j existiert, messbar und wesentlich beschränkt ist, siehe [10, Kapitel 5.8]:

Satz 1.17 (Rademacher)**.** *Sei $A \subset \mathbb{R}^n$ ein C^1-Gebiet und sei $g \in C^{0,1}(A)$. Dann ist g fast überall differenzierbar mit $\|\frac{\partial}{\partial x_i} g\|_{L_\infty(A)} \leq \|g\|_{C^{0,1}(A)}$ für $1 \leq i \leq n$.*

Mit eben diesem Satz von Rademacher lässt sich für ein Lipschitzgebiet Ω an fast allen Randpunkten $\mathbf{x} \in U_j \cap \partial\Omega$ ein *äußerer Normalenvektor*

$$\mathbf{n}(\mathbf{x}) := \big(1 + \|\nabla g_j\|^2\big)^{-1/2}\Big(\sum_{i=1}^{n-1} \frac{\partial}{\partial y_i} g_j(\mathbf{y})\mathbf{e}_i^j - \mathbf{e}_n^j\Big) \in \mathbb{R}^n \tag{1.28}$$

erklären. Es gilt $\|\mathbf{n}(\mathbf{x})\| = 1$ und $\langle \mathbf{x} - \mathbf{y}, \mathbf{n}(\mathbf{x})\rangle \geq 0$ für alle $\mathbf{y} \in \Omega \cap U_j$.

Weiter gilt auf Lipschitzgebieten der Satz von Gauß für glatte Integranden:

Satz 1.18 (Gauss)**.** *Sei Ω ein Lipschitzgebiet und $u \in C^1(\Omega)$ mit $\|u\|_{C^1(\overline{\Omega})} < \infty$. Dann gilt für u und die äußere Normale $\mathbf{n}$*

$$\int_\Omega \frac{\partial}{\partial x_i} u \,\mathrm{d}\mathbf{x} = \int_{\partial\Omega} u\mathbf{n} \cdot \mathbf{e}_i \,\mathrm{d}S, \quad 1 \leq i \leq n. \tag{1.29}$$

Ist auch $v \in C^1(\Omega)$ mit $\|v\|_{C^1(\overline{\Omega})} < \infty$, dann gilt die Regel der partiellen Integration

$$\int_\Omega \Big(u\frac{\partial}{\partial x_i} v + v\frac{\partial}{\partial x_i} u\Big)\,\mathrm{d}x = \int_{\partial\Omega} uv\mathbf{n} \cdot \mathbf{e}_i \,\mathrm{d}S, \quad 1 \leq i \leq n. \tag{1.30}$$

1.2.4 Sobolevräume

Definition 1.19. *Sei* $1 \leq p \leq \infty$, $f \in L_p(\Omega)$ *und* $\boldsymbol{\alpha} \in \mathbb{N}_0^n$. f *besitzt eine* schwache Ableitung $g = D^{\boldsymbol{\alpha}} f$ *der Ordnung* $|\boldsymbol{\alpha}|$, *falls* $g \in L_p(\Omega)$ *die Regel der partiellen Integration erfüllt,*

$$\langle f, D^{\boldsymbol{\alpha}} \varphi \rangle = (-1)^{|\boldsymbol{\alpha}|} \langle g, \varphi \rangle, \quad \text{für alle } \varphi \in \mathcal{D}(\Omega). \tag{1.31}$$

Aufgrund der Dichtheit der Testfunktionen in $L_1(\Omega)$ sind schwache Ableitungen bis auf Mengen vom Lebesguemaß 0 eindeutig. Starke, d.h. klassische, Ableitungen sind auch schwach, dies folgt für Lipschitzgebiete Ω aus Satz 1.18.

Definition 1.20. *Zu* $1 \leq p \leq \infty$ *und* $k \in \mathbb{N}_0$ *sei*

$$W^k(L_p(\Omega)) := \big\{ f \in L_p(\Omega) \big| D^{\boldsymbol{\alpha}} f \in L_p(\Omega) \text{ für alle } |\boldsymbol{\alpha}| \leq k \big\} \tag{1.32}$$

der L_p-Sobolevraum *aller* k*-mal schwach differenzierbaren Funktionen, mit der Norm*

$$\|f\|_{W^k(L_p(\Omega))} := \begin{cases} \Big(\sum\limits_{|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|_{L_p(\Omega)}^p \Big)^{1/p} & , \ 1 \leq p < \infty, \\ \max\limits_{|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|_{L_\infty(\Omega)} & , \ p = \infty. \end{cases} \tag{1.33}$$

$(W^k(L_p(\Omega)), \|\cdot\|_{W^k(L_p(\Omega))})$ ist ein Banachraum, für $p = 2$ ist $H^k(\Omega) := W^k(L_2(\Omega))$ ein Hilbertraum mit Skalarprodukt

$$\langle f, g \rangle_{H^k(\Omega)} := \sum_{|\boldsymbol{\alpha}| \leq k} \langle D^{\boldsymbol{\alpha}} f, D^{\boldsymbol{\alpha}} g \rangle. \tag{1.34}$$

Wir werden im Weiteren auch Sobolevräume gebrochener Ordnung benötigen, sie sind wie folgt definiert.

Definition 1.21. *Zu* $1 \leq p \leq \infty$, $k \in \mathbb{N}_0$ *und* $0 < \theta < 1$ *sei*

$$W^{k+\theta}(L_p(\Omega)) := \big\{ f \in W^k(L_p(\Omega)) \big| \|f\|_{W^{k+\theta}(L_p(\Omega))} < \infty \big\}, \tag{1.35}$$

mit der Norm

$$\|f\|_{W^{k+\theta}(L_p(\Omega))} := \begin{cases} \Big(\|f\|_{W^k(L_p(\Omega))}^p + \sum\limits_{|\boldsymbol{\alpha}|=k} \int_\Omega \int_\Omega \dfrac{|D^{\boldsymbol{\alpha}} f(\mathbf{x}) - D^{\boldsymbol{\alpha}} f(\mathbf{y})|^p}{\|\mathbf{x} - \mathbf{y}\|^{n+\theta p}} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \Big)^{1/p} & , \ 1 \leq p < \infty, \\ \|f\|_{W^k(L_\infty(\Omega))} + \max\limits_{|\boldsymbol{\alpha}|=k} \operatorname*{ess\,sup}\limits_{\substack{\mathbf{x},\mathbf{y} \in \Omega \\ \mathbf{x} \neq \mathbf{y}}} \dfrac{|D^{\boldsymbol{\alpha}} f(\mathbf{x}) - D^{\boldsymbol{\alpha}} f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^{n+\theta}} & , \ p = \infty. \end{cases} \tag{1.36}$$

Schwach differenzierbare Funktionen lassen sich durch stark differenzierbare Funktionen approximieren, denn der Raum $C^\infty(\Omega)$ aller beliebig oft differenzierbaren Funktionen in Ω ist dicht in $W^k(L_p(\Omega))$ für alle $k \in \mathbb{N}_0$ und $1 \leq p < \infty$. Anders ausgedrückt, für jedes $f \in W^k(L_p(\Omega))$ existiert eine Folge $(f_i)_{i=1}^\infty \subset C^\infty(\Omega)$ mit $\|f - f_i\|_{W^k(L_p(\Omega))} \to 0$

für $i \to \infty$. Diese Dichtheitsaussage lässt sich oft in Beweisen nutzen. Allerdings ist sie falsch für $p = \infty$.

Um die Sätze der Vektoranalysis auch auf Funktionen aus Sobolevräumen zu verallgemeinern, ist es notwendig, deren Einschränkung auf niederdimensionale Teilmengen von Ω sinnvoll zu definieren. Man kann zeigen, dass hierfür die Existenz einer schwachen Ableitung genügt.

Satz 1.22. *Sei Ω ein Lipschitzgebiet und $1 \le p \le \infty$. Dann existiert eine eindeutige, stetige lineare Abbildung $\gamma_0 : W^1(L_p(\Omega)) \to L_p(\partial\Omega)$, der* Spuroperator*, so dass $\gamma_0 f = f|_{\partial\Omega}$ für alle Funktionen $f \in C^0(\Omega) \cap W^1(L_p(\Omega))$.*

γ_0 ist allerdings nicht surjektiv. Für $p = 2$ gilt $\gamma_0(H^1(\Omega)) = H^{1/2}(\partial\Omega)$, bei allgemeinem p sieht das Bild von γ_0 komplizierter aus. Offenbar gilt $\gamma_0(\varphi) = 0$ für jede Testfunktion $\varphi \in \mathcal{D}(\Omega)$. Wir definieren daher den Raum $W_0^k(L_p(\Omega)) := \mathrm{clos}_{W^k(L_p(\Omega))}\mathcal{D}(\Omega)$ aller k-mal schwach differenzierbaren Funktionen mit Nullrandbedingungen, entsprechend den Raum $H_0^k(\Omega)$ für den Speziallfall $p = 2$. Folgende wichtige Abschätzung für Funktionen aus $W_0^1(L_p(\Omega))$ wird häufig benutzt:

Satz 1.23 (Poincaré/Friedrichs)**.** *Sei Ω ein beschränktes Lipschitzgebiet und $1 \le p \le \infty$. Dann gilt für jedes $f \in W_0^1(L_p(\Omega))$*

$$\|f\|_{L_p(\Omega)} \le \operatorname{diam}\Omega \Big\| \frac{\partial}{\partial x_i} f \Big\|_{L_p(\Omega)}, \quad 1 \le i \le n, \tag{1.37}$$

mit dem Durchmesser $\operatorname{diam}\Omega := \sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\mathbf{x} - \mathbf{y}\|$.

<u>*Beweis:*</u> Sei o.B.d.A. $f \in C_0^1(\Omega)$ und $i = 1$, ansonsten argumentiere über die Dichtheit von $C_0^1(\Omega)$ in $W_0^1(L_p(\Omega))$. Ferner betrachten wir nur $p < \infty$, denn der Fall $p = \infty$ folgt mit analoger Argumentation. Wegen $f \in C_0^1(\Omega)$ können wir f normgleich durch 0 auf ganz $\mathbb{R}^n$ fortsetzen, d.h. $\|D^{\boldsymbol{\alpha}} f\|_{L_p(\mathbb{R}^n)} = \|D^{\boldsymbol{\alpha}} f\|_{L_p(\Omega)}$ für $|\alpha| \le 1$. Sei $R := \operatorname{diam}\Omega < \infty$ und $\mathbf{x} \in \Omega$ beliebig. Dann gilt $\mathbf{x} + R\mathbf{e}_1 \notin \Omega$. Denn ansonsten gäbe es $\mathbf{x}, \mathbf{y} \in \Omega$ mit $\|\mathbf{x} - \mathbf{y}\| = R$. Wählt man dann aufgrund der Offenheit von Ω ein $\epsilon > 0$ mit $B(\mathbf{x}, \epsilon) \subset \Omega$, so folgt $\mathbf{z} := \mathbf{x} + \frac{\epsilon}{2}\frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|} \in B(\mathbf{x}, \epsilon) \subset \Omega$ mit dem Widerspruch $\|\mathbf{y}-\mathbf{z}\| = \|\mathbf{x}-\mathbf{y}\| + \epsilon/2 > R$. Also gilt doch $\mathbf{x} + R\mathbf{e}_1 \notin \Omega$ und damit

$$f(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x} + R\mathbf{e}_1) = \int_{x_1}^{x_1+R} \frac{\partial}{\partial x_1} f(s, x_2, \ldots, x_n)\, \mathrm{d}\mathbf{x}, \quad \text{für alle } \mathbf{x} \in \Omega.$$

Die Hölder-Ungleichung liefert wegen $\frac{1}{p} + \frac{1}{p/(p-1)} = 1$ daraus

$$|f(\mathbf{x})|^p \le \int_{\mathbb{R}} \Big| \frac{\partial}{\partial x_1} f(s, x_2, \ldots, x_n) \Big|^p \,\mathrm{d}s \underbrace{\Big(\int_0^R \mathrm{d}s \Big)^{p-1}}_{=R^{p-1}}, \quad \text{für alle } \mathbf{x} \in \Omega.$$

Mit Aufintegration in der ersten Koordinate über $\mathbb{R}$ folgt wegen $\operatorname{diam}\Omega = R$ zunächst

$$\int_{\mathbb{R}} \big| f(s, x_2, \ldots, x_n) \big|^p \,\mathrm{d}s \le R^p \int_{\mathbb{R}} \Big| \frac{\partial}{\partial x_1} f(s, x_2, \ldots, x_n) \Big|^p \,\mathrm{d}s,$$

so dass Integration in den anderen Koordinaten die Behauptung ergibt. □

Eine wichtige Folgerung aus der Poincaré-Friedrichs-Ungleichung ist, dass für Funktionen aus $W_0^k(L_p(\Omega))$ die Sobolev-Norm $\|\cdot\|_{W^k(L_p(\Omega))}$ äquivalent zur Halbnorm $|\cdot|_{W^k(L_p(\Omega))}$ ist, in welche nur Ableitungen der höchsten Ordnung eingehen. Die Poincaré-Friedrichs-Ungleichung kann auf den Fall verallgemeinert werden, dass die Funktionen nicht auf dem ganzen Rand $\partial\Omega$ verschwinden, sondern nur auf einer Teilmenge Γ_D mit $|\Gamma_D| > 0$. Dann existiert immer noch eine Konstante $C = C(\Omega, |\Gamma_D|)$, so dass (1.37) für alle $f \in W^1(L_p(\Omega))$ mit $\gamma_0(f) = 0$ auf Γ_D gilt.

Die folgenden Rechenregeln und Einbettungsresultate werden im weiteren Verlauf benutzt:

Satz 1.24. *Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Lipschitzgebiet, $1 \le p, q \le \infty$ mit $\frac{1}{p} + \frac{1}{q} = 1$ (wobei $\frac{1}{0} = \infty$) und seien $u \in W^1(L_p(\Omega))$, $v \in W^1(L_q(\Omega))$. Dann gilt der Satz von Gauß*

$$\int_\Omega \Big(u\frac{\partial}{\partial x_j}v + v\frac{\partial}{\partial x_j}u\Big)\,\mathrm{d}\mathbf{x} = \int_{\partial\Omega} uv\mathbf{n}\cdot\mathbf{e}_j\,\mathrm{d}S, \tag{1.38}$$

vgl. Satz 1.18. Weiter gilt für $u \in W^2(L_p(\Omega))$ und $v \in W^1(L_q(\Omega))$ die Greensche Formel

$$\int_\Omega \Delta uv\,\mathrm{d}x = \int_{\partial\Omega} v\nabla u\cdot\mathbf{n}\,\mathrm{d}S - \int_\Omega \nabla u\nabla v\,\mathrm{d}\mathbf{x}. \tag{1.39}$$

Satz 1.25. *Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Lipschitzgebiet.*

(i) Falls $1 \le p \le \infty$ und $k, m \in \mathbb{N}_0$ mit $k \le m$, dann gilt $W^m(L_p(\Omega)) \hookrightarrow W^k(L_p(\Omega))$. Die Einbettung ist für $k < m$ sogar kompakt (Satz von Rellich), d.h. beschränkte Folgen in $W^m(L_p(\Omega))$ haben konvergente Teilfolgen in $W^k(L_p(\Omega))$.

(ii) Falls $1 \le p \le q \le \infty$ und $k \in \mathbb{N}_0$, dann gilt $W^k(L_q(\Omega)) \hookrightarrow W^k(L_p(\Omega))$.

(iii) Falls $1 \le p, q \le \infty$ und $k, m \in \mathbb{N}$ mit $m \ge k$ und $m - k > n(\frac{1}{p} - \frac{1}{q})$, dann gilt $W^m(L_p(\Omega)) \hookrightarrow W^k(L_q(\Omega))$ (Einbettungssatz von Sobolev). Insbesondere gilt für $p > n$ und $k \in \mathbb{N}$ die Einbettung $W^k(L_p(\Omega)) \to C^{k-1}(\Omega)$, mit der Norm aus (1.21).

Funktionen aus dem Sobolevraum $W_0^k(L_p(\Omega))$ lassen sich nach Definition stetig durch Null auf ganz $\mathbb{R}^n$ fortsetzen, d.h. es gilt $\|E_0 f\|_{W^k(L_p(\mathbb{R}^n))} = \|f\|_{W^k(L_p(\Omega))}$ für alle $f \in W_0^k(L_p(\Omega))$. Auf Lipschitzgebieten existieren auch für Funktionen mit nichttrivialen Randwerten stetige Fortsetzungsoperatoren von $W^k(L_p(\Omega))$ nach $W^k(L_p(\mathbb{R}^n))$, siehe [13].

Satz 1.26. *Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Lipschitzgebiet, $k \in \mathbb{N}_0$ und $1 \le p \le \infty$. Dann existiert ein Operator $E \in \mathcal{L}(W^k(L_p(\Omega)), W^k(L_p(\mathbb{R}^n)))$ mit $Ev|_\Omega = v$ für alle $v \in W^k(L_p(\Omega))$.*

1.3 Variationsformulierung elliptischer Probleme

Viele durch partielle Differentialgleichungen gegebene Probleme liegen in einer sogenannten *Variationsformulierung* vor. Hiermit meinen wir im Folgenden, dass die gesuchte Lösung u in einem (reellen) Hilbertraum X liegt und die Variationsgleichung

$$a(u, v) = F(v), \quad \text{für alle } v \in X \tag{1.40}$$

erfüllt. Hierbei sind $a : X \times X \to \mathbb{R}$ eine Bilinearform und $F : X \to \mathbb{R}$ eine Linearform. Variationsformulierungen wie (1.40) sind in mehrerer Hinsicht wichtig:

- Zum einen sind viele physikalische Probleme, etwa aus der Mechanik, in einer Variationsformulierung gegeben. Dies ist insbesondere der Fall, wenn die gesuchte Größe u ein geeignetes Energiefunktional minimiert.
- Variationsformulierungen gestatten einen einfachen mathematischen Zugang im Hinblick auf Existenz, Eindeutigkeit und Stabilität von Lösungen.
- Variationsformulierungen erlauben die Herleitung numerischer und sogar adaptiver Methoden, wo (1.40) nur auf endlich-dimensionalen Teilräumen X_h erfüllt ist.

1.3.1 Poisson-Gleichung

Als Beispiel zur Herleitung einer Variationsformulierung betrachten wir die *Poisson-Gleichung* in starker Formulierung: Für ein Lipschitz-Gebiet $\Omega \subset \mathbb{R}^n$ und $f \in C^0(\Omega)$ sei eine Funktion $u \in C^2(\Omega)$ gesucht mit

$$-\Delta u = f \text{ in } \Omega, \quad u|_{\partial\Omega} = 0. \tag{1.41}$$

Lösungen dieser Art nennen wir im Folgenden *klassisch*. Mit Hilfe der Greenschen Formel (1.39) gilt für alle $v \in C_0^1(\Omega)$

$$\underbrace{\int_\Omega fv \,\mathrm{d}\mathbf{x}}_{=:F(v)} = -\int_\Omega \Delta u v \,\mathrm{d}\mathbf{x} = \underbrace{\int_\Omega \nabla u \nabla v \,\mathrm{d}\mathbf{x}}_{=:a(u,v)} - \underbrace{\int_{\partial\Omega} v \nabla u \cdot \mathbf{n} \,\mathrm{d}S}_{=0}, \tag{1.42}$$

also löst jede klassische Lösung von (1.41) auch das Variationsproblem (1.42). Es stellt sich daher die Frage, ob auch Äquivalenz gilt. Sei dazu $u \in C^2(\Omega) \cap C_0^1(\Omega)$ eine Lösung von (1.42). Dann ist diese eindeutig, denn für eine weitere Lösung $\tilde{u}$ gälte $a(u - \tilde{u}, v) = 0$ für alle $v \in C_0^1(\Omega)$, woraus mit $v = \tilde{u} - u$ folgt $\nabla(u - \tilde{u}) = 0$. Daher ist $u - \tilde{u}$ konstant und wegen der Nullrandwerte konstant Null. Weiter löst u auch (1.41). Denn für alle $v \in C_0^1(\Omega)$ ist

$$\int_\Omega (f + \Delta u)v \,\mathrm{d}\mathbf{x} = \int_\Omega (fv - \nabla u \nabla v) \,\mathrm{d}\mathbf{x} = 0,$$

also $f + \Delta u = 0$ fast überall und aufgrund der Stetigkeit auch punktweise. Im Allgemeinen kann man die Existenz stetig differenzierbarer Lösungen allerdings nicht voraussetzen, selbst für unendlich oft differenzierbares f. Stattdessen sind die Lösungen von Variationsproblemen in Räumen schwach differenzierbarer Funktionen zu suchen.

1.3.2 Existenz und Eindeutigkeit

Die Existenz und Eindeutigkeit von Lösungen eines Variationsproblems klärt folgender zentraler Satz.

Satz 1.27 (Lax/Milgram)**.** *Sei X ein Hilbertraum, $F \in X'$, und $a : X \times X \to \mathbb{R}$ eine Bilinearform. Falls a* stetig*, d.h. für ein $C_a > 0$ gilt*

$$|a(v,w)| \leq C_a \|v\|_X \|w\|_X, \quad \textit{für alle } v, w \in X, \tag{1.43}$$

und elliptisch*, d.h. für ein $C_e > 0$ gilt*

$$a(v,v) \geq C_e \|v\|_X^2, \quad \textit{für alle } v \in X, \tag{1.44}$$

dann hat das Variationsproblem (1.40) *genau eine Lösung $u_0 \in X$. Es gilt die a priori-Abschätzung*

$$\|u_0\|_X \leq \frac{\|F\|_{X'}}{C_e}, \tag{1.45}$$

wobei $\|F\|_{X'} = \sup_{0 \neq v \in X} \frac{|F(v)|}{\|v\|_X}$.

<u>*Beweis:*</u> Zunächst zur Eindeutigkeit von u_0. Falls $u, \tilde{u} \in X$ Lösungen von (1.40) sind, folgt $a(u - \tilde{u}, v) = 0$ für alle $v \in X$. Für $v = u - \tilde{u} \in X$ ergibt sich mit der Elliptizität (1.43) insbesondere $0 = a(u - \tilde{u}, u - \tilde{u}) \geq C_e \|u - \tilde{u}\|_X^2$, also $u = \tilde{u}$.

Zur Existenz von u_0 beobachte zunächst, dass für festes $x \in X$ aufgrund der Stetigkeit (1.43) $a(x, \cdot) \in X'$ gilt. Aufgrund des Darstellungssatzes von Riesz existiert daher ein eindeutiges $y \in X$ mit $a(x,v) = \langle y, v \rangle_X$ für alle $v \in X$. Wir definieren einen linearen Operator $A : X \to X$ durch $x \mapsto Ax := y$, dieser ist stetig wegen

$$\|Ax\|_X = \|y\|_X = \sup_{v \in X} \frac{\langle y, v \rangle_X}{\|v\|_X} = \sup_{v \in X} \frac{a(x,v)}{\|v\|_X} \leq C_a \|x\|_X.$$

Ferner ist A wegen

$$\|Ax\|_X = \sup_{v \in X} \frac{\langle y, v \rangle_X}{\|v\|_X} \geq \frac{\langle y, x \rangle_X}{\|x\|_X} = \frac{a(x,x)}{\|x\|_X} \geq C_e \|x\|_X$$

injektiv und hat abgeschlossenes Bild. Angenommen, es gäbe ein $z \in \mathrm{Im}(A)^\perp$, also $\langle Av, z \rangle_X = 0$ für alle $v \in X$. Insbesondere gilt $\langle Az, z \rangle_X = 0$, woraus folgt

$$C_e \|z\|_X^2 \leq a(z,z) = \langle Az, z \rangle_X = 0,$$

also $z = 0$ und A ist surjektiv, also stetig invertierbar. Zu jedem $F \in X'$ existiert ein eindeutiges $f \in X$ mit $F(v) = \langle f, v \rangle_X$ für alle $v \in X$. Wir setzen $u_0 := A^{-1} f$, was $a(u_0, v) = \langle f, v \rangle_X = F(v)$ für alle $v \in X$ erfüllt. (1.45) folgt dann mit

$$\|u_0\|_X \leq \frac{a(u_0,u_0)}{C_e \|u_0\|_X} \leq \frac{1}{C_e} \sup_{v \in X} \frac{a(u_0,v)}{\|v\|_X} = \frac{1}{C_e} \sup_{v \in X} \frac{F(v)}{\|v\|_X} = \frac{\|F\|_{X'}}{C_e}.$$

□

Hierbei bedeutet (1.45), dass die Lösung u_0 Lipschitz-stetig vom Funktional F abhängt.

Beispiel 1.28. *Für die Poisson-Gleichung ist* $X = H_0^1(\Omega)$, $a(v,w) = \int_\Omega \nabla v \nabla w \,\mathrm{d}\mathbf{x}$ *und* $F(v) = \int_\Omega fv \,\mathrm{d}\mathbf{x}$. *Die Bilinearform* a *ist stetig wegen*

$$|a(v,w)| \le \sum_{i=1}^n \int_\Omega \Big|\frac{\partial v}{\partial x_i}\frac{\partial w}{\partial x_i}\Big| \,\mathrm{d}\mathbf{x} \le \sum_{i=1}^n \Big\|\frac{\partial v}{\partial x_i}\Big\|_{L_2(\Omega_i)} \Big\|\frac{\partial w}{\partial x_i}\Big\|_{L_2(\Omega_i)} \le \|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)},$$

die H^1*-Elliptizität folgt aus der Poincaré-Friedrichs-Ungleichung* (1.37) *mit*

$$\begin{aligned}(1+\operatorname{diam}\Omega^2)a(v,v) &= (1+\operatorname{diam}\Omega^2)\sum_{|\boldsymbol{\alpha}|=1} \|D^{\boldsymbol{\alpha}} v\|^2_{L_2(\Omega)} \\ &\ge \|v\|^2_{L_2(\Omega)} + \sum_{|\boldsymbol{\alpha}|=1} \|D^{\boldsymbol{\alpha}} v\|^2_{L_2(\Omega)} \\ &= \|v\|^2_{H^1(\Omega)}.\end{aligned}$$

Das Funktional F *ist stetig auf* $H_0^1(\Omega)$, *da wegen* $f \in C^0(\Omega)$ *und* (1.19) *gilt*

$$|F(v)| \le \|f\|_{L_2(\Omega)}\|v\|_{L_2(\Omega)} \le |\Omega|^{1/2}\|f\|_{L_\infty(\Omega)}\|v\|_{H^1(\Omega)}.$$

1.3.3 Galerkin-Verfahren

Um eine numerische Approximation der Lösung u einer Variationsgleichung (1.40) zu erhalten, betrachtet man abgeschlossene, meist endlich-dimensionale Teilräume X_h, und das "projizierte" Problem

$$a(u_h, v) = F(v), \quad \text{für alle } v \in X_h. \tag{1.46}$$

Da die Räume X_h abgeschlossen sind, existiert eine eindeutige Lösung $u_h \in X_h$ von (1.46) nach Satz 1.27, u_h heißt *Galerkin*-Lösung. Für $X_h = X$ stimmt u_h mit der Lösung u von (1.40) überein. Bei echten Teilräumen lässt sich immerhin noch *Quasi-Optimalität* zeigen, d.h. die Galerkin-Lösung u_h hat bis auf eine Konstante den gleichen Fehler wie die Bestapproximation in X_h bzgl. $\|\cdot\|_X$.

Satz 1.29 (Céa)**.** *Sei* $a : X \times X \to \mathbb{R}$ *eine stetige, elliptische Bilinearform auf* X *und sei der lineare Teilraum* $X_h \subset X$ *abgeschlossen mit der Galerkin-Lösung* $u_h \in X_h$ *von* (1.46) *und der Lösung* u *von* (1.40). *Dann gilt*

$$\|u - u_h\|_X \le \frac{C_a}{C_e} \inf_{v \in X_h} \|u - v\|_X. \tag{1.47}$$

Beweis: Für jedes $v \in X_h$ gilt nach (1.40) und (1.46) $a(u-u_h, v) = 0$, also insbesondere $a(u-u_h, u-u_h) = a(u-u_h, u-v)$. Aufgrund der Stetigkeit und Elliptizität von a folgt

$$\|u-u_h\|_X \le \frac{a(u-u_h,u-u_h)}{C_e\|u-u_h\|_X} = \frac{a(u-u_h,u-v)}{C_e\|u-u_h\|_X} \le \frac{C_a}{C_e}\|u-v\|_X, \quad \text{für alle } v \in X_h,$$

die Behauptung ergibt sich mit dem Infimum über $v \in X_h$. □

Die Konvergenz von u_h gegen u ist also mit der generellen Approximierbarkeit von Elementen aus X durch Elemente aus X_h gekoppelt.

Falls die elliptische Bilinearform a *symmetrisch* ist, also $a(v,w) = a(w,v)$ für alle $v, w \in X$, dann definiert a ein Skalarprodukt auf X mit zugehöriger Norm $\|x\|_a := \sqrt{a(x,x)}$, der *Energienorm*. Diese ist aufgrund der Stetigkeit und Elliptizität von a äquivalent zu $\|\cdot\|_X$, d.h. es existieren Konstanten $c, C > 0$ mit $c\|x\|_X \leq \|x\|_a \leq C\|x\|_X$ für alle $x \in X$. Genauer: hier ist $c = \sqrt{C_e}$, $C = \sqrt{C_a}$. Dann kann die Konstante in (1.47) noch auf $\sqrt{C_a/C_e}$ verbessert werden. Denn aufgrund der Symmetrie von a gilt $a(u-u_h, v) = a(v, u-u_h) = 0$ für alle $v \in X_h$ und somit durch Einsetzen von $u - v = (u - u_h) + (u_h - v)$ der Satz von Pythagoras

$$a(u-v, u-v) = a(u-u_h, u-u_h) + a(u_h - v, u_h - v), \quad \text{für alle } v \in X_h. \tag{1.48}$$

Aus (1.48) und der Elliptizität von a folgt sofort $\|u-v\|_a \geq \|u-u_h\|_a$ für alle $v \in X_h$ und somit

$$\|u-u_h\|_X^2 \leq \frac{1}{C_e}\|u-u_h\|_a^2 = \frac{1}{C_e}\inf_{v\in X_h}\|u-v\|_a^2 \leq \frac{C_a}{C_e}\inf_{v\in X_h}\|u-v\|_X^2.$$

Zur konkreten Durchführung eines Galerkin-Verfahrens wählt man endlich-dimensionale Teilräume X_h und eine Basis $\{\psi_1, \ldots, \psi_m\}$ von X_h. Dann besitzt die Galerkin-Lösung in X_h eine Darstellung $u_h = \sum_{k=1}^m c_k \psi_k$, und das lineare Gleichungssystem

$$\mathbf{Ac} = \mathbf{F} \quad \leftrightarrow \quad \sum_{k=1}^m c_k a(\psi_k, \psi_l) = F(\psi_l), \quad 1 \leq l \leq m \tag{1.49}$$

für den Koeffizientenvektor $\mathbf{c} = (c_1, \ldots, c_m)^\top \in \mathbb{R}^m$ ist wegen der Basiseigenschaft der ψ_k äquivalent zu (1.46). Die Matrix $\mathbf{A} = (a(\psi_k, \psi_l))_{1\leq l,k\leq m}$ heißt *Steifigkeitsmatrix*, der Vektor $\mathbf{F} = (F(\psi_l))_{1\leq l\leq m}$ heißt *Lastvektor*. In den meisten Fällen ist das Funktional F durch Integration gegen eine Funktion $f \in L_2(\Omega)$ gegeben, d.h. die konkrete Durchführung eines Galerkin-Verfahrens erfolgt durch

- Aufstellen des Lastvektors $\mathbf{F}$ (Quadratur),
- Aufstellen der Steifigkeitsmatrix $\mathbf{A}$ (Quadratur) und
- Lösen eines linearen Gleichungssystems $\mathbf{Ac} = \mathbf{F}$ (meist iterativ).

Aus der Elliptiziät der Bilinearform a folgt die positive Definitheit von $\mathbf{A}$. Denn zu beliebigem $\mathbf{0} \neq \mathbf{v} = (v_j)_{1\leq j\leq m} \in \mathbb{R}^m$ und $0 \neq v := \sum_{j=1}^m v_j \psi_j \in X_h$ rechnet man

$$\langle \mathbf{Av}, \mathbf{v}\rangle = \sum_{i,j=1}^m v_i v_j a(\psi_j, \psi_i) = a(v,v) \geq C_e\|v\|_X^2 > 0. \tag{1.50}$$

Ist a zusätzlich symmetrisch, so ist $\mathbf{A}$ symmetrisch und positiv definit, was bei der Anwendung von Iterationsverfahren auf das lineare Gleichungssystem (1.49) ausgenutzt werden kann. Hierbei ist es auch wichtig, dass die Besetzungsstruktur der Matrix $\mathbf{A}$ dünn ist. Dies kann durch Ansatzfunktionen $\{\psi_1, \ldots, \psi_m\}$ mit kleinen Trägern erreicht werden, d.h. pro Punkt $\mathbf{x} \in \Omega$ existieren nur wenige ψ_i mit $\mathbf{x} \in \operatorname{supp}\psi_i$. Konkrete Beispiele für günstige Ansatzsysteme sind Finite Elemente oder auch Wavelet-Systeme.

2 Adaptive Finite Elemente-Verfahren

2.1 Finite Elemente

Sei $\Omega \subset \mathbb{R}^n$ ein beschränktes Lipschitz-Gebiet. Zur Konstruktion eines geeigneten Ansatzraums für ein Galerkin-Verfahren betrachten wir eine *Triangulierung* $\mathcal{T}$ von Ω mit *Simplizes* $T \in \mathcal{T}$.

Definition 2.1. *Zu $k \in \{1, \dots, n\}$ seien $\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(k)} \in \mathbb{R}^n$ derart, dass $\{\mathbf{x}^{(j)} - \mathbf{x}^{(0)}\}_{j=1,\dots,k}$ linear unabhängig sind. Dann heißt*

$$S = \Big\{ \mathbf{x} = \sum_{j=0}^{k} \lambda_j \mathbf{x}^{(j)} \in \mathbb{R}^n \Big| \lambda_j \geq 0, \sum_{j=0}^{k} \lambda_j = 1 \Big\} = \operatorname{conv}\{\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(k)}\} \tag{2.1}$$

*(nicht degeneriertes) k-dimensionales oder auch k-*Simplex *im $\mathbb{R}^n$. Die Koeffizienten λ_j heißen* baryzentrische Koordinaten *von $\mathbf{x} \in S$.*

Für $n = 1$ sind n-dimensionale Simplizes Intervalle, für $n = 2$ Dreiecke und für $n = 3$ Tetraeder. Geeignete Sammlungen derartiger Simplizes werden zu Triangulierungen zusammengefasst.

Definition 2.2. *$\mathcal{T} = \{T_j\}_{1 \leq j \leq m}$ heißt* (simpliziale) konforme Triangulierung *von Ω, falls*

(i) Jedes $T \in \mathcal{T}$ ist ein n-Simplex im $\mathbb{R}^n$.

(ii) $\overline{\Omega} = \bigcup_{T \in \mathcal{T}} T$ und $\operatorname{int} S \cap \operatorname{int} T = \emptyset$ für $S, T \in \mathcal{T}$, $S \neq T$.

(iii) Falls $G := S \cap T \neq \emptyset$ für $S, T \in \mathcal{T}$, dann ist G eine niederdimensionale Seitenfläche von S und von T.

Für ein k-dimensionales Simplex S benutzen wir im Folgenden den *Durchmesser*

$$h(S) := \operatorname{diam}(S) = \max \big\{ \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\| \big| 0 \leq i, j \leq k \big\} \tag{2.2}$$

sowie den *Inkugeldurchmesser*

$$\rho(S) := \sup \big\{ 2R \big| \overline{B(\mathbf{x}, R)} \subset S \text{ für ein } \mathbf{x} \in S \big\}. \tag{2.3}$$

Den Quotienten bezeichnen wir mit $\sigma(S) := \frac{h(S)}{\rho(S)} > 1$. Eine Triangulierung $\mathcal{T}$, bei der für ein $\sigma > 0$ die Abschätzung $\sigma(T) \leq \sigma_0$ für alle $T \in \mathcal{T}$ gilt, wird als *quasi-uniform* bezeichnet.

Zur Existenz und Eindeutigkeit der baryzentrischen Koordinaten notieren wir:

Lemma 2.3. *Sei $S = \operatorname{conv}\{\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k)}\}$ ein k-Simplex im $\mathbb{R}^n$. Dann existiert für jedes $\mathbf{x}$ aus dem k-dimensionalen affinen Teilraum $\mathbf{x}^{(0)} + \operatorname{span}\{\mathbf{x}^{(1)} - \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k)} - \mathbf{x}^{(0)}\}$ von $\mathbb{R}^n$ ein eindeutiger Vektor $(\lambda_0, \ldots, \lambda_{k-1}) \in \mathbb{R}^k$, so dass mit $\lambda_k := 1 - \sum_{j=0}^{k-1} \lambda_j$ die Entwicklung $\mathbf{x} = \sum_{j=0}^{k} \lambda_j \mathbf{x}^{(j)}$ gilt. Die Abbildung $\mathbf{x} \mapsto \boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_k)$ ist affin-linear und bijektiv. Es ist genau dann $\mathbf{x} \in S$, wenn $\boldsymbol{\lambda} \geq \mathbf{0}$.*

Beweis: Sei $\mathbf{x} - \mathbf{x}^{(0)} \in \operatorname{span}\{\mathbf{x}^{(1)} - \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(k)} - \mathbf{x}^{(0)}\}$. Es existieren also Koeffizienten λ_j, $1 \leq j \leq k$, mit $\mathbf{x} - \mathbf{x}^{(0)} = \sum_{j=1}^{k} \lambda_j (\mathbf{x}^{(j)} - \mathbf{x}^{(0)})$. Aufgrund der linearen Unabhängigkeit der Differenzvektoren $\mathbf{x}^{(j)} - \mathbf{x}^{(0)}$ sind die Koeffizienten λ_j und damit die Darstellung

$$\mathbf{x} = \underbrace{(1 - \sum_{j=1}^{k} \lambda_j)}_{=:\lambda_0} \mathbf{x}^{(0)} + \sum_{j=1}^{k} \lambda_j \mathbf{x}^{(j)}$$

eindeutig. Nach Definition gilt $\sum_{j=0}^{k} \lambda_j = 1$, also $\lambda_k = 1 - \sum_{j=0}^{k-1} \lambda_j$. Damit ist $\boldsymbol{\lambda}$ eindeutige Lösung des (i.a. überbestimmten) Gleichungssystems $\left(\begin{smallmatrix} \mathbf{x}_1^{(0)} & \cdots & \mathbf{x}_1^{(k)} \\ 1 & \cdots & 1 \end{smallmatrix}\right) \boldsymbol{\lambda} = \left(\begin{smallmatrix} \mathbf{x} \\ 1 \end{smallmatrix}\right)$.

Zur affin-Linearität der Abbildung $\mathbf{x} \mapsto \boldsymbol{\lambda}$ überlegt man sich, dass aufgrund der linearen Unabhängigkeit der Differenzvektoren $\mathbf{x}^{(j)} - \mathbf{x}^{(0)}$ die Matrix

$$\begin{pmatrix} \mathbf{x}^{(0)} & \mathbf{x}^{(1)} - \mathbf{x}^{(0)} & \cdots & \mathbf{x}^{(k)} - \mathbf{x}^{(0)} \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

und damit

$$\begin{pmatrix} \mathbf{x}^{(0)} & \mathbf{x}^{(1)} & \cdots & \mathbf{x}^{(k)} \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

vollen Rang $k+1$ hat. Folglich ist für jede Wahl von k Indizes $1 \leq i_1 \leq \cdots \leq i_k \leq n$ das quadratische Teilsystem

$$\begin{pmatrix} x_{i_1}^{(0)} & \cdots & x_{i_1}^{(k)} \\ \vdots & \ddots & \vdots \\ x_{i_k}^{(0)} & \cdots & x_{i_k}^{(k)} \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \lambda_0' \\ \vdots \\ \lambda_k' \end{pmatrix} = \begin{pmatrix} x_{i_1} \\ \vdots \\ x_{i_k} \\ 1 \end{pmatrix}$$

eindeutig durch ein $(\lambda_0', \ldots, \lambda_k') \in \mathbb{R}^{k+1}$ lösbar, das potentiell noch von der Wahl der i_j abhängen könnte. Dies ist jedoch nicht der Fall, da ja nach dem ersten Beweisschritt das Gesamtsystem $\left(\begin{smallmatrix} \mathbf{x}_1^{(0)} & \cdots & \mathbf{x}_1^{(k)} \\ 1 & \cdots & 1 \end{smallmatrix}\right) \boldsymbol{\lambda} = \left(\begin{smallmatrix} \mathbf{x} \\ 1 \end{smallmatrix}\right)$ die eindeutige Lösung $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_k)$ besitzt. Es folgt unabhängig von der Wahl der i_j die Identität $\lambda_j' = \lambda_j$, also z.B. für $i_j = j$

$$\boldsymbol{\lambda} = \underbrace{\begin{pmatrix} x_1^{(0)} & \cdots & x_1^{(k)} \\ \vdots & \ddots & \vdots \\ x_k^{(0)} & \cdots & x_k^{(k)} \\ 1 & \cdots & 1 \end{pmatrix}^{-1}}_{=:\begin{pmatrix} \mathbf{M} & \mathbf{m} \end{pmatrix}} \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 1 \end{pmatrix} = \mathbf{M} \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} + \mathbf{m} = \begin{pmatrix} \mathbf{M} & \mathbf{0} \end{pmatrix} \mathbf{x} + \mathbf{m}.$$

Da $\mathbf{M}$ vollen Rang hat, ist die Abbildung $\mathbf{x} \mapsto \boldsymbol{\lambda}$ bijektiv. Nach Definition der konvexen Hülle ist $\boldsymbol{\lambda} \geq \mathbf{0}$ äquivalent zu $\mathbf{x} \in S$. □

Für spätere Zwecke notieren wir folgendes Lemma:

Lemma 2.4. *Sei $S = \operatorname{conv}\{\mathbf{x}^{(0)}, \dots \mathbf{x}^{(k)}\}$ ein k-Simplex im $\mathbb{R}^n$. Dann ist jede Komponente von $\mathbf{x} \in S$ ein homogenes Polynom (d.h. ohne konstantes Glied) vom Höchstgrad 1 in den $k+1$ Variablen $\boldsymbol{\lambda}$. Umgekehrt ist jede Komponente von $\boldsymbol{\lambda}$ ein Polynom vom Höchstgrad 1 in k der n Variablen $\mathbf{x}$.*

Beweis: Wegen $\mathbf{x} = \sum_{j=0}^{k} \lambda_j \mathbf{x}^{(j)}$ ist jede Komponente von $\mathbf{x}$ ein homogenes Polynom vom Höchstgrad 1 in den $k+1$ Variablen $\lambda_0, \dots, \lambda_k$. Durch Einsetzen der Normierungsbedingung erhält man sogar, dass jede Komponente von $\mathbf{x}$ sich auch als Polynom vom Höchstgrad 1 in k der Variablen $\lambda_0, \dots, \lambda_k$ schreiben lässt, etwa $\lambda_{i_1}, \dots, \lambda_{i_k}$ für $0 \leq i_1 \leq \dots \leq i_k \leq k$.

Aus der Argumentation von Lemma 2.3 entnehmen wir, dass für jede Auswahl von k Zeilenindizes $1 \leq i_1 \leq \dots \leq i_k \leq n$ das quadratische Teilsystem

$$\underbrace{\begin{pmatrix} x_{i_1}^{(0)} & \cdots & x_{i_1}^{(k)} \\ \vdots & \ddots & \vdots \\ x_{i_k}^{(0)} & \cdots & x_{i_k}^{(k)} \\ 1 & \cdots & 1 \end{pmatrix}}_{=:\mathbf{N}} \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_k \end{pmatrix} = \begin{pmatrix} x_{i_1} \\ \vdots \\ x_{i_k} \\ 1 \end{pmatrix}$$

eindeutig lösbar ist. Durchmultiplizieren mit $\mathbf{N}^{-1}$ liefert, dass jedes λ_j Polynom vom Grad 1 in den k Variablen $x_{i_1}, \dots x_{i_k}$ ist. □

Jedes n-Simplex lässt sich affin-linear durch den *Einheitssimplex*

$$S_n := \operatorname{conv}\{\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_n\} \tag{2.4}$$

parametrisieren:

Lemma 2.5. *Für jedes n-dimensionale Simplex $S = \operatorname{conv}\{\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(n)}\}$ existiert genau eine affin-lineare Abbildung*

$$\varphi_S : S_n \to S, \quad \varphi_S(\mathbf{x}) = \mathbf{A}_S \mathbf{x} + \mathbf{b}_S \tag{2.5}$$

mit $\mathbf{A}_S \in \mathbb{R}^{n \times n}$, $\det \mathbf{A}_S \neq 0$, $\mathbf{b}_S \in \mathbb{R}^n$, $\varphi_S(\mathbf{0}) = \mathbf{x}^{(0)}$ und $\varphi_S(\mathbf{e}_j) = \mathbf{x}^{(j)}$, $1 \leq j \leq n$. Es gilt $|\det \mathbf{A}_S| = \frac{|S|}{|S_n|} = n!|S|$ sowie die Abschätzungen

$$\|\mathbf{A}_S\| \leq \frac{h(S)}{\rho(S_n)}, \quad \|\mathbf{A}_S^{-1}\| \leq \frac{h(S_n)}{\rho(S)}, \quad C\rho(S)^n \leq |\det \mathbf{A}_S| \leq Ch(S)^n, \tag{2.6}$$

wobei $C > 0$ nur von der Raumdimension n abhängt.

Beweis: Es ist notwendigerweise $\mathbf{b}_S = \varphi_S(\mathbf{0}) = \mathbf{x}^{(0)}$. Wegen $\varphi_S(\mathbf{e}_i) = \mathbf{x}^{(j)} = \mathbf{A}_S\mathbf{e}_j + \mathbf{x}^{(0)}$ für $1 \le j \le n$ folgt $\mathbf{A}_S = (\mathbf{x}^{(1)} - \mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)} - \mathbf{x}^{(0)})$, es kann offenbar keine andere affin-lineare Abbildung mit den gleichen Abbildungseigenschaften geben. Aufgrund der linearen Unabhängigkeit der $\mathbf{x}^{(j)} - \mathbf{x}^{(0)}$ für $1 \le j \le n$ ist $\mathbf{A}_S$ nichtsingulär.

Sei $\mathbf{x} \in \mathbb{R}^n$ mit $\|\mathbf{x}\| = 1$. Nach Definition von $\rho(S_n)$ existiert ein $\mathbf{x}_0 \in S_n$ mit $\overline{B(\mathbf{x}_0, \frac{\rho(S_n)}{2})} \subset S_n$, also existieren $\mathbf{u}, \mathbf{v} \in S_n$ mit $\mathbf{u} - \mathbf{v} = \rho(S_n)\mathbf{x}$ und folglich

$$\|\mathbf{A}_S\mathbf{x}\| = \frac{1}{\rho(S_n)}\|\mathbf{A}_S\mathbf{u} - \mathbf{A}_S\mathbf{v}\| \le \frac{h(S)}{\rho(S_n)},$$

also die erste Ungleichung in (2.6). Analog existiert $\mathbf{x}_1 \in S$ mit $\overline{B(\mathbf{x}_1, \frac{\rho(S)}{2})} \subset S$, also existieren $\mathbf{y}, \mathbf{z} \in S$ mit $\mathbf{y} - \mathbf{z} = \rho(S)\mathbf{x}$ und damit

$$\|\mathbf{A}_S^{-1}\mathbf{x}\| = \frac{1}{\rho(S)}\|\mathbf{A}_S^{-1}\mathbf{y} - \mathbf{A}_S^{-1}\mathbf{z}\| \le \frac{h(S_n)}{\rho(S)},$$

also die zweite Ungleichung in (2.6). Mit der Transformationsregel rechnet man

$$|S| = \int_S \mathrm{d}\mathbf{x} = \int_{S_n} |\det D\varphi_S(\mathbf{z})|\,\mathrm{d}\mathbf{z} = |\det \mathbf{A}_S||S_n|.$$

Wegen $|S_n| = \frac{1}{n!}$ gilt also $|S| = \frac{1}{n!}|\det \mathbf{A}_S|$. Da das Volumen der n-dimensionalen Kugel mit Radius r gerade $|B(\mathbf{0}, r)| = r^n \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ ist, folgt die untere Abschätzung in (2.6) mit

$$|\det \mathbf{A}_S| = n!|S| \ge n!\big|B(\mathbf{0}, \tfrac{\rho(S)}{2})\big| = \underbrace{\frac{n!\pi^{n/2}}{2^n\Gamma(\frac{n}{2}+1)}}_{=:C}\rho(S)^n,$$

analog gilt

$$|\det \mathbf{A}_S| = n!|S| \le n!\big|B(\mathbf{0}, \tfrac{h(S)}{2})\big| = Ch(S)^n.$$

□

Für die Transformation von Fehlerabschätzungen auf dem Standardsimplex S_n zu solchen über einem beliebigen n-Simplex S benötigen wir folgendes Lemma.

Lemma 2.6. *Sei $S = \mathrm{conv}\{\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathbb{R}^n$ ein n-Simplex mit affin-linearer Parametrisierung φ_S aus (2.5). Dann gelten für $1 \le p \le \infty$, $m \in \mathbb{N}_0$ und $f \in W^m(L_p(S))$ die Abschätzungen*

$$|f \circ \varphi_S|_{W^m(L_p(S_n))} \le C_1 \frac{h(S)^m}{\rho(S_n)^m}\rho(S)^{-n/p}|f|_{W^m(L_p(S))}, \tag{2.7}$$

$$|f|_{W^m(L_p(S))} \le C_2 \frac{h(S_n)^m}{\rho(S)^m}h(S)^{n/p}|f \circ \varphi_S|_{W^m(L_p(S_n))}, \tag{2.8}$$

wobei $C_1, C_2 > 0$ nur von m, n und p abhängen.

Beweis: Sei zunächst $m = 0$. Für $f \in L_p(S)$ gilt

$$\|f\|_{L_p(S)} = \Big(\int_S |f(\mathbf{x})|^p \, \mathrm{d}\mathbf{x} \Big)^{1/p} = \Big(|\det \mathbf{A}_S| \int_{S_n} |f \circ \varphi_S(\mathbf{y})|^p \, \mathrm{d}\mathbf{y} \Big)^{1/p},$$

also mit (2.6) die Abschätzung

$$C^{1/p} \rho(S)^{n/p} \|f \circ \varphi_S\|_{L_p(S_n)} \le \|f\|_{L_p(S)} \le C^{1/p} h(S)^{n/p} \|f \circ \varphi_S\|_{L_p(S_n)}, \tag{2.9}$$

wobei $C > 0$ nur von n abhängt.

Sei dann $m \ge 1$ und $1 \le j \le n$. Es gilt mit der Kettenregel für fast alle $\mathbf{x} \in S_n$

$$\Big| \frac{\partial}{\partial x_j} (f \circ \varphi_S)(\mathbf{x}) \Big| = \big| \nabla f(\varphi_S(\mathbf{x}))(\mathbf{x}^{(j)} - \mathbf{x}^{(0)}) \big| \le \|\mathbf{A}_S\| \Big(\sum_{j=1}^{n} \Big| \frac{\partial f}{\partial x_j} (\varphi_S(\mathbf{x})) \Big|^2 \Big)^{1/2},$$

analog mit Induktion über $m = |\alpha|$ für $f \in W^m(L_p(S))$

$$\begin{aligned} \big| D^\alpha (f \circ \varphi_S)(\mathbf{x}) \big| &\le \|\mathbf{A}_S\|^m \Big(\sum_{|\beta|=m} \big| D^\beta f(\varphi_S(\mathbf{x})) \big|^2 \Big)^{1/2} \\ &\le c(m,n,p) \|\mathbf{A}_S\|^m \Big(\sum_{|\beta|=m} \big| D^\beta f(\varphi_S(\mathbf{x})) \big|^p \Big)^{1/p}. \end{aligned}$$

Integration und Anwendung von (2.9) und (2.6) ergibt daraus

$$\begin{aligned} \big\| D^\alpha (f \circ \varphi_S) \big\|_{L_p(S_n)} &\le c(m,n,p) \|\mathbf{A}_S\|^m \Big(\sum_{|\beta|=m} \big\| (D^\beta f) \circ \varphi_S \big\|_{L_p(S_n)}^p \Big)^{1/p} \\ &\le c'(m,n,p) \frac{h(S)^m}{\rho(S_n)^m} \rho(S)^{-n/p} |f|_{W^m(L_p(S))}, \end{aligned}$$

also ergibt sich (2.7) durch Aufsummation über alle α mit $|\alpha| = m$. Die Abschätzung (2.8) folgt analog. □

Mit Hilfe von Lemma 2.6 lassen sich sogenannte *inverse Abschätzungen* beweisen. Hierbei werden auf einem Simplex S höhere Ableitungen von Polynomen in der Norm abgeschätzt durch niedrigere Ableitungen, auf Kosten negativer Potenzen von $h(S)$.

Satz 2.7. *Sei $\boldsymbol{\alpha} \in \mathbb{N}_0^n$, $\mathbf{0} \le m \le |\boldsymbol{\alpha}|$ und $1 \le p, q \le \infty$. Ist S ein n-Simplex im $\mathbb{R}^n$ und P ein endlich-dimensionaler Raum von Polynomen, dann gilt*

$$\|D^{\boldsymbol{\alpha}} v\|_{L_p(S)} \le C h(S)^{m - |\boldsymbol{\alpha}| - n(\frac{1}{q} - \frac{1}{p})} |v|_{W^m(L_q(S))}, \quad \textit{für alle } v \in P, \tag{2.10}$$

wobei C nur von $\boldsymbol{\alpha}, m, p, q, n, P$ und $\sigma(S)$ abhängt.

Beweis: Seien zunächst $S = S_n$ und $\boldsymbol{\beta} = 0$, also $h(S) = h(S_n)$. Da P endlichdimensional, sind alle Normen äquivalent. Es existiert also $c = c(|\boldsymbol{\alpha}|, p, q, n, P) > 0$ mit

$$\|D^{\boldsymbol{\alpha}} v\|_{L_p(S_n)} \leq \|v\|_{W^{|\boldsymbol{\alpha}|}(L_p(S_n))} \leq c\|v\|_{L_q(S_n)}, \quad \text{für alle } v \in P. \tag{2.11}$$

Da $h(S_n)$ nur von der Raumdimension n abhängt ($h(S_1) = 1$, $h(S_n) = 2^{1/n}$ für $n \geq 2$), entspricht dies der zu zeigenden Abschätzung (2.10). Sei dann $\boldsymbol{\beta} \in \mathbb{N}_0^n$ mit $m = |\boldsymbol{\beta}| > 0$. Der Raum $\tilde{P} := \{D^{\boldsymbol{\beta}} v | v \in P\}$ ist wie P ein Raum von Polynomen. Anwendung von (2.11) auf den Raum $\tilde{P}$ liefert

$$\|D^{\boldsymbol{\alpha}} v\|_{L_p(S_n)} = \|D^{\boldsymbol{\alpha}-\boldsymbol{\beta}} D^{\boldsymbol{\beta}} v\|_{L_p(S_n)} \leq c\|D^{\boldsymbol{\beta}} v\|_{L_q(S_n)} \leq c|v|_{W^m(L_q(S_n))}, \quad \text{für alle } v \in P,$$

wobei $c = c(|\boldsymbol{\alpha}|, m, p, q, n, P)$, das ist wieder die Form (2.10). Aus Lemma 2.5 und der Abschätzung (2.10) auf S_n folgt nun für $v \in P$

$$\begin{aligned}
\|D^{\boldsymbol{\alpha}} v\|_{L_p(S)} &\leq C \frac{h(S_n)^{|\boldsymbol{\alpha}|}}{\rho(S)^{|\boldsymbol{\alpha}|}} h(S)^{n/p} |v \circ \varphi_S|_{W^{|\boldsymbol{\alpha}|}(L_p(S_n))} \\
&\leq C' \frac{h(S_n)^{|\boldsymbol{\alpha}|}}{\rho(S)^{|\boldsymbol{\alpha}|}} h(S)^{n/p} |v \circ \varphi_S|_{W^m(L_q(S_n))} \\
&\leq C'' \frac{h(S_n)^{|\boldsymbol{\alpha}|}}{\rho(S_n)^m} \frac{h(S)^{m+n/p}}{\rho(S)^{|\boldsymbol{\alpha}|+n/q}} |v|_{W^m(L_q(S))}.
\end{aligned}$$

Wegen $\rho(S) = \sigma(S)^{-1} h(S)$ ergibt sich die Behauptung (2.10). □

Zur Definition Finiter Elemente gehört neben einer geeigneten Gebietszerlegung noch die Spezifikation daran angepasster Ansatzsysteme $\{\psi_j\}$. Wir beschränken uns hier auf den Fall stückweise polynomialer Funktionen. Hierzu sei für $k \in \mathbb{N}_0$

$$P^k := \operatorname{span}\left\{ \mathbf{x}^\alpha = \prod_{j=1}^n x_j \,\middle|\, |\boldsymbol{\alpha}| \leq k \right\} \tag{2.12}$$

der Raum aller Polynome in n Variablen vom Höchstgrad k. Wir vermerken, dass $\dim P^k = \binom{n+k}{k}$, insbesondere $\dim P^1 = n + 1$. Weiter sei

$$P_{\mathcal{T}}^k := \left\{ f : \overline{\Omega} \to \mathbb{R} \,\middle|\, f|_T \in P^k, T \in \mathcal{T} \right\} \tag{2.13}$$

der Raum aller auf der Triangulierung $\mathcal{T}$ stückweise vom Höchstgrad k polynomialen Funktionen.

Lemma 2.8. *Sei $\mathcal{T}$ eine konforme Triangulierung von Ω. Dann gilt für $k, m \in \mathbb{N}_0$ und alle $1 \leq p \leq \infty$ die Inklusion $P_{\mathcal{T}}^k \cap C^m(\overline{\Omega}) \subset W^{m+1}(L_p(\Omega))$.*

Beweis: Wir führen den Beweis nur für $m = 0$. Seien also $1 \leq i \leq n$, $1 \leq p \leq \infty$ und $f \in P_{\mathcal{T}}^k \cap C^0(\Omega)$ beliebig. Kandidat für die partielle schwache Ableitung $\frac{\partial}{\partial x_i} f$ ist dann offenbar $g : \Omega \to \mathbb{R}$ mit $g|_T := \frac{\partial}{\partial x_i}(f|_T)$ für alle $T \in \mathcal{T}$. Es folgt $g|_T \in L_p(T)$

für alle $T \in \mathcal{T}$ und mit der Zerlegungseigenschaft also $g \in L_p(\Omega)$. Zu zeigen ist noch die Regel der partiellen Integration (1.31). Für $\varphi \in \mathcal{D}(\Omega)$ folgt mit dem simplexweise angewendeten Satz von Gauß (1.38)

$$\begin{aligned}\int_\Omega f\frac{\partial}{\partial x_i}\varphi\,\mathrm{d}\mathbf{x} = \sum_{T\in\mathcal{T}}\int_T f\frac{\partial}{\partial x_i}\varphi\,\mathrm{d}\mathbf{x} &= \sum_{T\in\mathcal{T}}\Big(\int_{\partial T} f\varphi\mathbf{n}\cdot\mathbf{e}_i\,\mathrm{d}S - \int_T \varphi\frac{\partial}{\partial x_i}(f|_T)\,\mathrm{d}\mathbf{x}\Big)\\ &= -\sum_{T\in\mathcal{T}}\int_T g\varphi\,\mathrm{d}\mathbf{x} = -\int_\Omega g\varphi\,\mathrm{d}\mathbf{x}.\end{aligned}$$

Hierbei verschwinden die Randintegrale über ∂T aus zwei Gründen: bei Integralen über $\partial T \cap \partial\Omega$ benutzt man $\varphi \in \mathcal{D}(\Omega)$, Integrale über innere Randstücke treten doppelt auf mit verschiedenen Vorzeichen (gespiegelte äußere Normalen bei Nachbarsimplizes). □

Die Verbindung zwischen Triangulierung und Ansatzfunktionen geschieht über Interpolationsbedingungen an den Eckpunkten oder den niederdimensionalen Seitenflächen der Triangulierung. Dazu sei für ein Simplex T im Weiteren N_T die Menge der Ecken von T, E_T sei die Menge der $(n-1)$-dimensionalen Seitenflächen (*Facetten*). Für eine Triangulierung $\mathcal{T}$ von Ω sei $N_\mathcal{T} = \bigcup_{T\in\mathcal{T}} N_T$ die Gesamt-Eckenmenge (*Knoten*), $E_\mathcal{T} = \bigcup_{T\in\mathcal{T}} E_T$ sei analog die Facettenmenge. $N_\mathcal{T}$ zerfällt disjunkt in die Menge der *Randknoten* $N_{\mathcal{T},\partial\Omega} := \{\mathbf{x} \in N_\mathcal{T} | \mathbf{x} \in \partial\Omega\}$ und die Menge der *inneren Knoten* $N_{\mathcal{T},\Omega} := \{\mathbf{x} \in N_\mathcal{T} | \mathbf{x} \in \Omega\}$. Analog zerfällt $E_\mathcal{T}$ disjunkt in die Menge der *Randfacetten* $E_{\mathcal{T},\partial\Omega} = \{e \in E_\mathcal{T} | e \subset \partial\Omega\}$ und die Menge der *inneren Facetten* $E_{\mathcal{T},\Omega} = \{e \in E_\mathcal{T} | e \subset \Omega\}$.

Bei der Spezifikation geeigneter Finiter Elemente beschränken wir uns im Folgenden auf den Fall stückweise linearer, stetiger Ansatzfunktionen, die durch Interpolationsbedingungen in den Knoten $N_\mathcal{T}$ gegeben sind.

Satz 2.9 (Courant). *(i) Sei $S = \mathrm{conv}\{\mathbf{x}^{(0)},\dots,\mathbf{x}^{(n)}\}$ ein n-Simplex und $\mathbf{y} = (y_j)_{j=0}^n \in \mathbb{R}^{n+1}$. Dann gibt es genau ein $p \in P^1$ mit $p(\mathbf{x}^{(j)}) = y_j$ für alle $0 \le j \le n$, es ist $p(\mathbf{x}) = \sum_{j=0}^n \lambda_j y_j$ mit den baryzentrischen Koordinaten $\boldsymbol{\lambda} = (\lambda_j)_{0\le j\le n}$ von $\mathbf{x} \in S$.*

(ii) Sei $\mathcal{T}$ eine Triangulierung von Ω mit den Knoten $N_\mathcal{T} = \{\mathbf{z}^{(1)},\dots,\mathbf{z}^{(m)}\}$ und sei $\mathbf{y} = (y_j)_{j=1}^m \in \mathbb{R}^m$ beliebig. Dann gibt es genau ein $q \in P^1_\mathcal{T} \cap C^0(\Omega)$ mit $q(\mathbf{z}^{(j)}) = y_j$ für alle $1 \le j \le m$. Für $\mathbf{y} = \mathbf{e}_k$ heißt das dazugehörige q simpliziales Lagrange-Element *zum Knoten $\mathbf{z}^{(k)}$.*

Beweis: Zu (i): Für die $n+1$ Koeffizienten des gesuchten Polynoms $p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|\le 1} c_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}}$ gelten die $n+1$ linearen Gleichungen $p(\mathbf{x}^{(j)}) = y_j$, $0 \le j \le n$. Es reicht daher, eine lineare, injektive Lösungsabbildung $L : \mathbf{y} \mapsto (c_{\boldsymbol{\alpha}})$ anzugeben. Hierzu sei die Abbildung $\varphi : \mathbf{y} \mapsto \varphi(\mathbf{y})$ in den Raum aller homogenen Polynome vom Höchstgrad 1 gegeben durch $\varphi(\mathbf{y})(\boldsymbol{\lambda}) := \sum_{j=0}^n y_j\lambda_j$. Dann ist φ offenbar linear und injektiv (und aus Dimensionsgründen bijektiv), da die Teilvariablen λ_j als linear unabhängige Monome agieren. Ferner gilt $\varphi(\mathbf{y})(\mathbf{e}_j) = y_j$. Sei jetzt $\mathbf{x} = \sum_{j=0}^n \lambda_j\mathbf{x}^{(j)} \in \mathbb{R}^n$. Nach Lemma 2.4 ist jedes λ_j ein Polynom vom Höchstgrad 1 in $\mathbf{x}$. Wir setzen $L(\mathbf{y})(\mathbf{x}) := p(\mathbf{x}) := \varphi(\mathbf{y})(\boldsymbol{\lambda}(\mathbf{x}))$, dies ist folglich als Verkettung ebenfalls ein Polynom vom Höchstgrad 1 in $\mathbf{x}$. Da die

Koeffizienten von $\varphi(\mathbf{y})$ linear von $\mathbf{y} \in \mathbb{R}^{k+1}$ abhängen und die Koeffizienten von p eine Linearkombination der Koeffizienten von $\varphi(\mathbf{y})$ sind, ist $L : \mathbf{y} \mapsto L(\mathbf{y})$ linear. Die Interpolationsbedingungen für p sind wegen $p(\mathbf{x}^{(j)}) = \varphi(\mathbf{y})(\mathbf{e}_j) = y_j$ erfüllt, $0 \le j \le n$. Zu zeigen bleibt noch die Injektivität von L. Sei also $p(\mathbf{x}) = 0$ für alle $\mathbf{x} \in \mathbb{R}^n$. Es folgt $0 = p(\mathbf{x}^{(j)}) = y_j$ für alle $0 \le j \le n$, also $\mathbf{y} = 0$.

Zu (ii): Wegen (i) existiert zu jedem n-Simplex $T \in \mathcal{T}$ ein eindeutiges $q|_T \in P^1$ mit $q(\mathbf{z}^{(j)}) = y_j$ für alle $\mathbf{z}^{(j)} \in N_T$. Es reicht zu zeigen, dass q über die niederdimensionalen Seitenflächen der Simplizes $T \in \mathcal{T}$ hinweg stetig ist. Seien also $S, T \in \mathcal{T}$ mit einem $(n-k)$-dimensionalen Seitensimplex $G := S \cap T$. Da auf G k der baryzentrischen Koordinaten verschwinden (die zu den an G nicht beteiligten Ecken gehören), ist $q|_G$ ein homogenes Polynom in $n+1-k$ Variablen. Mit (i) ist es durch seine Werte an den $n+1-k$ Eckpunkten von G bereits eindeutig bestimmt, also ist q stetig bei G. □

Lagrange-Elemente q zu inneren Knoten $\mathbf{z}^{(j)} \in N_{\mathcal{T},\Omega}$ sind offenbar wegen $q|_{\partial\Omega} = 0$ enthalten in $W_0^1(L_p(\Omega))$ für alle $1 \le p < \infty$. Sie können also als Galerkin-Ansatzsysteme zur Diskretisierung von Variationsproblemen mit homogenen Dirichlet-Randbedingungen verwendet werden.

2.2 Lokale Fehlerschätzer vom Residuum-Typ

Sei u die Lösung der Variationsgleichung (1.40) und $u_{\mathcal{T}}$ deren Galerkin-Approximation (1.46) bzgl. linearer Lagrange-Elemente zur Triangulierung $\mathcal{T}$. Es ist unser Ziel, zumindest für die schwache Formulierung der Poisson-Gleichung (1.41) *lokale* Fehlerschätzer η_e, $e \in E_{\mathcal{T},\Omega}$, herzuleiten sowie deren Verlässlichkeit und Effizienz nachzuweisen. Die Darstellung erfolgt dabei gemäß [15, 11].

Satz 2.10. *Sei $\mathcal{T}$ quasi-uniform mit $\sigma(T) \le \sigma_0$ für alle $T \in \mathcal{T}$. Dann existiert zu jeder inneren Facette $e \in E_{\mathcal{T},\Omega}$ eine berechenbare Zahl $\eta_e \ge 0$, so dass für gewisse, nur von σ_0 und n abhängige Konstanten $C_1, C_2 > 0$ gilt:*

$$|u - u_{\mathcal{T}}|_{H^1(\Omega)} \le C_1 \Big(\sum_{e \in E_{\mathcal{T},\Omega}} \eta_e^2 \Big)^{1/2}, \tag{2.14}$$

$$\eta_e \le C_2 \Big(|u - u_{\mathcal{T}}|^2_{H^1(\omega_e)} + \sum_{T \subset \omega_e} h(T)^2 \|f - f_T\|^2_{L_2(T)} \Big)^{1/2}, \quad \textit{für alle } e \in E_{\mathcal{T},\Omega}. \tag{2.15}$$

Hierbei ist $f_T := \frac{1}{|T|} \int_T f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ der Mittelwert von f auf $T \in \mathcal{T}$, $\omega_e := \bigcup_{e \subset T \in \mathcal{T}} T$ ist die (zweielementige) Menge aller über die Facette $e \in E_{\mathcal{T},\Omega}$ benachbarten Simplizes aus $\mathcal{T}$.

Eine von mehreren Möglichkeiten, solche lokalen Fehlerschätzer η_e herzuleiten, basiert auf dem aus der Einsetzprobe von $u_{\mathcal{T}}$ in (1.40) entstehenden *Residuum* $R(u_{\mathcal{T}}) \in X'$,

$$R(u_{\mathcal{T}})(v) := a(u - u_{\mathcal{T}}, v) = F(v) - a(u_{\mathcal{T}}, v), \quad \text{für alle } v \in X. \tag{2.16}$$

Wegen (1.46) ist offenbar $R(u_\mathcal{T})(v) = 0$ für alle $v \in X_h = P^1_\mathcal{T} \cap C^0(\Omega)$, für allgemeines $v \in X$ ergeben sich nichttriviale Werte. Wir beobachten zunächst, dass $\|R(u_\mathcal{T})\|_{X'}$ im Fall einer symmetrischen Bilinearform a ein verlässlicher und effizienter Schätzer für den $\|\cdot\|_X$-Fehler von $u_\mathcal{T}$ ist. Denn aus der Stetigkeit der Bilinearform a folgt die Effizienz

$$\|R(u_\mathcal{T})\|_{X'} = \sup_{\|v\|_X=1} a(u - u_\mathcal{T}, v) \le C_a \|u - u_\mathcal{T}\|_X. \tag{2.17}$$

Da a symmetrisch, ist $a(\cdot,\cdot)$ ein Skalarprodukt auf X mit der Cauchy-Schwarzschen Ungleichung $a(v,w) \le \sqrt{a(v,v)}\sqrt{a(w,w)}$. Hieraus folgt direkt $a(v,v) = \sup_{0\neq w\in X} \frac{a(v,w)^2}{a(w,w)}$ und somit die Verlässlichkeit des Fehlerschätzers:

$$C_e \|u - u_\mathcal{T}\|_X^2 \le a(u - u_\mathcal{T}, u - u_\mathcal{T}) = \sup_{0\neq v\in X} \frac{a(u - u_\mathcal{T}, v)^2}{a(v,v)} \le \frac{1}{C_e} \|R(u_\mathcal{T})\|_{X'}^2. \tag{2.18}$$

Da zur Berechnung von $\|R(u_\mathcal{T})\|_{X'}$ ein Supremum gebildet werden müsste, ist dieser Fehlerschätzer so noch nicht implementierbar. Durch geeignete Abschätzungen lässt sich aber eine berechenbare Größe ableiten. Wir verfolgen die dafür notwendigen Umformungen exemplarisch für den Fall der Poisson-Gleichung. Sei also $X = H^1_0(\Omega)$ und $a(v,w) = \int_\Omega \nabla v \nabla w \,\mathrm{d}\mathbf{x}$, die Argumentation für allgemeinere elliptische Differentialgleichungen verläuft dann analog. Das Funktional $F \in H^{-1}(\Omega) := X'$ sei gegeben durch Integration $F(v) := \int_\Omega f v \,\mathrm{d}\mathbf{x}$ gegen die Funktion $f \in L_2(\Omega)$. Mit der Greenschen Formel und $u_\mathcal{T} \in P^1_\mathcal{T}$, also punktweise $\Delta u_\mathcal{T} = 0$ auf $T \in \mathcal{T}$, vereinfacht sich für $v \in H^1_0(\Omega)$ das Residuum zu

$$R(u_\mathcal{T})(v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} - \sum_{T\in\mathcal{T}} \int_T \nabla u_\mathcal{T} \nabla v \,\mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} - \sum_{T\in\mathcal{T}} \int_{\partial T} v \nabla u_\mathcal{T} \cdot \mathbf{n} \,\mathrm{d}S. \tag{2.19}$$

Durch die eindeutige Numerierung der Simplizes $T \in \mathcal{T}$ kann man nun jeder inneren Facette $e \in E_{\mathcal{T},\Omega}$ einen eindeutig bestimmten Normaleneinheitsvektor $\mathbf{n}_e$ zuordnen, z.B. als äußeren Normaleneinheitsvektor zu demjenigen Nachbarsimplex $T \in \omega_e$ mit der kleineren Nummer. Für eine skalar- oder auch vektorwertige, auf der Triangulierung $\mathcal{T}$ stückweise stetige Funktion g sei dann ihr *Sprung* über e erklärt durch

$$[\![g]\!]_e(\mathbf{x}) := \lim_{t\to 0+} \big(g(\mathbf{x} + t\mathbf{n}_e) - g(\mathbf{x} - t\mathbf{n}_e)\big), \quad \text{für alle } \mathbf{x} \in e. \tag{2.20}$$

Auf den Randsegmenten $e \in E_{\mathcal{T},\partial\Omega}$ setzen wir $[\![g]\!]_e := 0$. Dann folgt aus (2.19)

$$R(u_\mathcal{T})(v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} - \sum_{e\in E_{\mathcal{T},\Omega}} \int_e v [\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e \,\mathrm{d}S, \quad \text{für alle } v \in H^1_0(\Omega). \tag{2.21}$$

Zu einer ersten Abschätzung der Residualnorm dient folgendes Lemma.

Lemma 2.11. *Für alle $v \in H^1_0(\Omega)$ gilt*

$$|R(u_\mathcal{T})(v)| \le \inf_{w\in X_h} \Big(\sum_{T\in\mathcal{T}} \|f\|_{L_2(T)} \|v - w\|_{L_2(T)} + \sum_{e\in E_{\mathcal{T},\Omega}} \big\|[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e\big\|_{L_2(e)} \|v - w\|_{L_2(e)} \Big). \tag{2.22}$$

Beweis: Für beliebiges $w \in X_h$ gilt $R(u_\mathcal{T})(w) = 0$, also mit (2.21)

$$R(u_\mathcal{T})(v) = R(u_\mathcal{T})(v) - R(u_\mathcal{T})(w) = \int_\Omega f(v-w)\,\mathrm{d}\mathbf{x} - \sum_{e\in E_{\mathcal{T},\Omega}} \int_e (v-w)[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e \,\mathrm{d}S,$$

woraus sich nach Norm- und Infimumbildung die Abschätzung (2.22) ergibt. □

Durch die Wahl einer ganz bestimmten Approximation $w \in X_h$ von $v \in H^1_0(\Omega)$ lassen sich nun die Normen über $v - w$ durch sogenannte Interpolationsabschätzungen kontrollieren. Da $v \in H^1_0(\Omega)$ allerdings nicht notwendigerweise stetig ist (für $n \geq 2$ gilt $H^1(\Omega) \not\hookrightarrow C^0(\Omega)$), kann dessen Approximation nicht alleine durch punktweise Interpolation entstehen, sondern es muss ein Glättungsschritt vorausgehen. Üblicherweise verwendet man hier zur Approximation gewisse Interpolationsoperatoren vom *Clément*-Typ. Wir verwenden hierfür folgenden generischen Existenzsatz (siehe etwa [2]):

Satz 2.12. *Sei $\mathcal{T}$ quasi-uniform mit $\sigma(T) \leq \sigma_0$ für alle $T \in \mathcal{T}$. Dann existiert ein linearer Operator $I_\mathcal{T} : H^1_0(\Omega) \to X_h$ und Konstanten $C_i = C_i(n, \sigma_0) > 0$, $i \in \{1, 2\}$, mit*

$$\sum_{T\in\mathcal{T}} h(T)^{-2}\|v - I_\mathcal{T} v\|^2_{L_2(T)} \leq C_1 |v|^2_{H^1(\Omega)}, \quad \textit{für alle } v \in H^1_0(\Omega), \tag{2.23}$$

$$\sum_{e\in E_{\mathcal{T},\Omega}} h(e)^{-1}\|v - I_\mathcal{T} v\|^2_{L_2(e)} \leq C_2 |v|^2_{H^1(\Omega)}, \quad \textit{für alle } v \in H^1_0(\Omega). \tag{2.24}$$

Ohne Beweis geben wir zwei Möglichkeiten an, einen solchen Interpolationsoperator zu spezifizieren. In beiden Fällen ist $I_\mathcal{T} v = \sum_{\mathbf{z}^{(k)}\in N_{\mathcal{T},\Omega}} c_k q_k$ mit den Lagrange-Elementen q_k zu inneren Knoten $\mathbf{z}^{(k)}$ und geeigneten Koeffizienten c_k:

- c_k kann durch $c_k = \pi_{\mathbf{z}^{(k)}}(v)(\mathbf{z}^{(k)})$ entstehen, wobei $\pi_\mathbf{x}(v) \in P^1$ gegeben ist durch eine lokale L_2-Projektion $\int_{\omega_\mathbf{x}} \pi_\mathbf{x}(v) w \,\mathrm{d}\mathbf{x} = \int_{\omega_\mathbf{x}} v w \,\mathrm{d}\mathbf{x}$ für alle $w \in P^1$. Dies ist der von Clément in [3] ursprünglich eingeführte Operator.
- c_k kann durch Mittelung $c_k = \frac{1}{\omega_{\mathbf{z}^{(k)}}} \int_{\omega_{\mathbf{z}^{(k)}}} v \,\mathrm{d}\mathbf{x}$ entstehen. Hierbei ist $\omega_\mathbf{x} := \bigcup_{\mathbf{x}\in T\in\mathcal{T}} T$ die Vereinigung aller an $\mathbf{x} \in \overline{\Omega}$ beteiligten Simplizes. Zum Nachweis der Eigenschaften aus Satz 2.12 sei auf [2] verwiesen.

Durch Anwenden eines solchen Clément-Interpolationsoperators ergibt sich aus (2.22), (2.23) und (2.24) für alle $v \in H^1_0(\Omega)$

$$\begin{aligned}
|R(u_\mathcal{T})(v)| &\leq \sum_{T\in\mathcal{T}} \|f\|_{L_2(T)} \|v - I_\mathcal{T} v\|_{L_2(T)} + \sum_{e\in E_{\mathcal{T},\Omega}} \big\|[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e\big\|_{L_2(e)} \|v - I_\mathcal{T} v\|_{L_2(e)} \\
&\leq \Big(\sum_{T\in\mathcal{T}} h(T)^2 \|f\|^2_{L_2(T)}\Big)^{1/2} \Big(\sum_{T\in\mathcal{T}} h(T)^{-2} \|v - I_\mathcal{T} v\|^2_{L_2(T)}\Big)^{1/2} \\
&\quad + \Big(\sum_{e\in E_{\mathcal{T},\Omega}} h(e) \big\|[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e\big\|^2_{L_2(e)}\Big)^{1/2} \Big(\sum_{e\in E_{\mathcal{T},\Omega}} h(e)^{-1} \|v - I_\mathcal{T} v\|^2_{L_2(e)}\Big)^{1/2} \\
&\leq C\|v\|_{H^1(\Omega)} \Big(\sum_{T\in\mathcal{T}} h(T)^2 \|f\|^2_{L_2(T)} + \sum_{e\in E_{\mathcal{T},\Omega}} h(e) \big\|[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e\big\|^2_{L_2(e)}\Big)^{1/2},
\end{aligned}$$

wobei $C > 0$ nach Satz 2.12 nur von n und σ_0 abhängt. Um jeder inneren Facette $e \in E_{\mathcal{T},\Omega}$ einen Fehlerschätzer η_e zuzuordnen, muss noch die erste Summe transformiert werden. Da jedes Simplex $T \in \mathcal{T}$ in mindestens einer der Mengen ω_e auftritt, folgt nach Division durch $\|v\|_{H^1(\Omega)}$ und Supremumbildung

$$\|R(u_\mathcal{T})\|_{H^{-1}(\Omega)} \leq C\Big(\sum_{e \in E_{\mathcal{T},\Omega}} \underbrace{\big(h(\omega_e)^2 \|f\|^2_{L_2(\omega_e)} + h(e) \big\| [\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e \big\|^2_{L_2(e)} \big)}_{=:\eta_e^2} \Big)^{1/2}, \quad (2.25)$$

also mit (2.18) die erste Abschätzung (2.14) aus Satz 2.10. Jedes η_e ist offensichtlich berechenbar, denn es treten nur die rechte Seite f der Differentialgleichung sowie von der Galerkin-Lösung $u_\mathcal{T}$ abhängige Terme auf.

Um die zweite Abschätzung (2.15) zu zeigen, beschränken wir uns zunächst auf den Fall $n = 2$ und führen hier eine Klasse von Hilfsfunktionen ein.

Lemma 2.13. *Sei $\mathcal{T}$ eine Triangulierung von $\Omega \subset \mathbb{R}^2$ und $T = \operatorname{conv}\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}\} \in \mathcal{T}$. Fener sei die* Dreiecks-Bubble-Funktion *$b_T : \overline{\Omega} \to \mathbb{R}$ in baryzentrischen Koordinaten gegeben durch*

$$b_T(\mathbf{x}) := \begin{cases} 27\lambda_0\lambda_1\lambda_2, & \mathbf{x} = \sum_{j=0}^2 \lambda_j \mathbf{x}^{(j)} \in T, \\ 0, & \text{sonst.} \end{cases} \quad (2.26)$$

Dann gilt $\operatorname{supp} b_T = T$, $0 \leq b_T(\mathbf{x}) \leq 1$ *für alle* $\mathbf{x} \in \overline{\Omega}$ *sowie die Abschätzungen*

$$c_1 h(T)^2 \leq \int_T b_T \, \mathrm{d}\mathbf{x} = \frac{9}{20}|T| \leq c_2 h(T)^2, \quad (2.27)$$

wobei $c_1 > 0$ von T unabhängig ist und $c_2 = c_2(\sigma(T)) > 0$. Weiter ist

$$|b_T|_{H^1(T)} \leq c_3 h(T)^{-1} \|b_T\|_{L_2(T)} \quad (2.28)$$

mit einem $c_3 = c_3(\sigma(T)) > 0$.

Beweis: Siehe Übungsaufgabe 8. □

Betrachte nun eine solche Dreiecks-Bubble-Funktion b_T, $T \in \mathcal{T}$, insbesondere die mit einer Konstanten skalierte Version $v_T := f_T b_T$. Zunächst rechnen wir mit (2.27)

$$\int_T f_T v_T \, \mathrm{d}\mathbf{x} = |f_T|^2 \int_T b_T \, \mathrm{d}\mathbf{x} = \frac{9}{20}|T||f_T|^2 = \frac{9}{20}\|f_T\|^2_{L_2(T)}. \quad (2.29)$$

Andererseits gilt wegen $\operatorname{supp} v_T = T$ und $v_T \in H^1_0(\Omega)$ die Identität

$$\begin{aligned} \int_T f_T v_T \, \mathrm{d}\mathbf{x} &= \int_T f v_T \, \mathrm{d}\mathbf{x} + \int_T (f - f_T) v_T \, \mathrm{d}\mathbf{x} \\ &= \int_\Omega f v_T \, \mathrm{d}\mathbf{x} + \int_T (f - f_T) v_T \, \mathrm{d}\mathbf{x} \\ &= \int_\Omega \nabla u \nabla v_T \, \mathrm{d}\mathbf{x} + \int_T (f - f_T) v_T \, \mathrm{d}\mathbf{x} \\ &= \int_T \nabla u \nabla v_T \, \mathrm{d}\mathbf{x} + \int_T (f - f_T) v_T \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Da $u_{\mathcal{T}}|_T \in P^1$, folgt aus $v_T \in H_0^1(T)$ und der Greenschen Formel $\int_T \nabla u_{\mathcal{T}} \nabla v_T \,\mathrm{d}\mathbf{x} = 0$, also mit (2.28) die Abschätzung

$$\begin{aligned}\int_T f_T v_T \,\mathrm{d}\mathbf{x} &= \int_T \nabla(u-u_{\mathcal{T}})\nabla v_T \,\mathrm{d}\mathbf{x} + \int_T (f-f_T) v_T \,\mathrm{d}\mathbf{x} \\ &\le |u-u_{\mathcal{T}}|_{H^1(T)} |v_T|_{H^1(T)} + \|f-f_T\|_{L_2(T)} \|v_T\|_{L_2(T)} \\ &= |f_T|\big(|u-u_{\mathcal{T}}|_{H^1(T)} |b_T|_{H^1(T)} + \|f-f_T\|_{L_2(T)} \|b_T\|_{L_2(T)}\big) \\ &\le |T|^{-1/2} \|f_T\|_{L_2(T)} \|b_T\|_{L_2(T)} \big(c_3 h(T)^{-1} |u-u_{\mathcal{T}}|_{H^1(T)} + \|f-f_T\|_{L_2(T)}\big).\end{aligned}$$

Wegen $0 \le b_T \le 1$ und (2.26) ist $\|b_T\|_{L_2(T)}^2 \le \int_T b_T \,\mathrm{d}\mathbf{x} = \frac{9}{20}|T|$, also folgt

$$\int_T f_T v_T \,\mathrm{d}\mathbf{x} \le C \|f_T\|_{L_2(T)} \big(h(T)^{-1} |u-u_{\mathcal{T}}|_{H^1(T)} + \|f-f_T\|_{L_2(T)}\big)$$

mit nur von $\sigma(T)$ abhängiger Konstanten $C > 0$, woraus sich mit Anfügen der Identität (2.29) und Division durch $\|f_T\|_{L_2(T)}$

$$\|f_T\|_{L_2(T)} \le C' \big(h(T)^{-1} |u-u_{\mathcal{T}}|_{H^1(T)} + \|f-f_T\|_{L_2(T)}\big) \tag{2.30}$$

ergibt, $C' = C'(\sigma(T)) > 0$.

Eine zu (2.30) analoge Abschätzung für die Sprungnormen $\|[\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e\|_{L_2(e)}$ kann mit einer zweiten Klasse von Hilfsfunktionen hergeleitet werden.

Lemma 2.14. *Sei $\mathcal{T}$ eine Triangulierung von $\Omega \subset \mathbb{R}^2$ und sei $e = \mathrm{conv}\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}\} \in E_{\mathcal{T},\Omega}$ eine innere Kante. Die beiden Nachbardreiecke $T_1, T_2 \subset \omega_e$ seien durch Hinzunahme der Ecken $\mathbf{x}^{(2)}$ bzw. $\mathbf{x}^{(3)}$ aufgespannt. Dann gelten für die in baryzentrischen Koordinaten gegebenen* Kanten-Bubble-Funktion $b_e : \overline{\Omega} \to \mathbb{R}$

$$b_e(\mathbf{x}) := \begin{cases} 4\lambda_0\lambda_1, & \mathbf{x} = \sum_{j=0}^2 \lambda_j \mathbf{x}^{(j)} \in T_1, \\ 4\lambda_0\lambda_1, & \mathbf{x} = \lambda_0\mathbf{x}^{(0)} + \lambda_1\mathbf{x}^{(1)} + \lambda_3\mathbf{x}^{(3)} \in T_2, \\ 0, & \textit{sonst} \end{cases} \tag{2.31}$$

die Eigenschaften $\mathrm{supp}\, b_e = \omega_e$, $0 \le b_e(\mathbf{x}) \le 1$ *für alle* $\mathbf{x} \in \overline{\Omega}$ *sowie* $\int_e b_e \,\mathrm{d}S = \frac{2}{3}h(e)$. *Weiter gelten für $\mathcal{T} \ni T \subset \omega_e$ die Abschätzungen*

$$c_4 h(e)^2 \le \int_S b_e \,\mathrm{d}\mathbf{x} = \frac{1}{3}|T| \le c_5 h(e)^2 \tag{2.32}$$

und

$$|b_e|_{H^1(T)} \le c_6 h(e)^{-1} \|b_e\|_{L_2(T)} \tag{2.33}$$

mit höchstens von $\sigma(T)$ abhängigen Konstanten $c_4, c_5, c_6 > 0$.

Beweis: Wir zeigen nur $\int_e b_e \,\mathrm{d}S = \frac{2}{3}h(e)$ und (2.33), siehe Übungsaufgabe 8 für die anderen Behauptungen. Zunächst kann die Kante $e = \mathrm{conv}\{\mathbf{x}^{(0)}, \mathbf{x}^{(1)}\}$ parametrisiert

werden durch $\varphi(t) = (1-t)\mathbf{x}^{(0)} + t\mathbf{x}^{(1)}$. Wegen $D\varphi(t) = \mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ gilt $\det D\varphi(t)^\top D\varphi(t) = \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|^2 = h(e)^2$, also

$$\int_e b_e \,\mathrm{d}S = 4h(e) \int_0^1 t(1-t)\,\mathrm{d}t = \frac{2}{3}h(e).$$

Bezeichne dann die Kanten des Einheitsdreiecks $S_2 = \mathrm{conv}\{(0,0),(1,0),(0,1)\}$ mit e_i, $0 \le i \le 2$ (e_i liege gegenüber $\mathbf{x}^{(i)}$). Dann induzieren die drei Kanten entsprechende Bubble-Funktionen $b_i \in P_2$,

$$b_0(\mathbf{x}) := 4x_1x_2, \quad b_1(\mathbf{x}) := 4x_2(1-x_1-x_2), \quad b_2(\mathbf{x}) := 4x_1(1-x_1-x_2).$$

Ist jetzt $\mathcal{T} \ni T \subset \omega_e$, dann gilt offenbar dann $b_e|_T = b_i \circ \varphi_T^{-1}$ für genau ein $0 \le i \le 2$, mit der affin-linearen Parametrisierung φ_T aus Lemma (2.5). Man rechnet leicht nach (ohne Beweis), dass $|b_i|_{H^1(S_2)} = \frac{2\sqrt{6}}{3}$ und $\|b_i\|_{L_2(S_2)} = \frac{2\sqrt{5}}{15}$ unabhängig von i. Mit Hilfe des Transformationslemmas 2.6 ergibt sich dann (2.33). Denn zunächst folgt aus (2.8) die Abschätzung

$$|b_e|_{H^1(T)} \le Ch(T)\rho(T)^{-1}|b_i|_{H^1(S_2)} = C'\sigma(T),$$

mit von T unabhängigen Konstanten $C, C' > 0$. Hieraus ergibt sich mit (2.7)

$$|b_e|_{H^1(T)} \le C'\|b_i\|_{L_2(S_2)}^{-1}\sigma(T)\|b_i\|_{L_2(S_2)} \le C''\sigma(T)\rho(T)^{-1}\|b_e\|_{L_2(T)}, \tag{2.34}$$

mit von T unabhängigem $C'' > 0$. Wegen $\rho(T)^{-1} = \sigma(T)h(T)^{-1}$ und $h(T)^{-1} \le h(e)^{-1}$ (da $e \subset T$) folgt aus (2.34) sofort (2.33) mit $c_6 = C''\sigma(T)^2$. □

Analog zur Abschätzung von $\|f_T\|_{L_2(T)}$ betrachte nun für $e \in E_{\mathcal{T},\Omega}$ die Hilfsfunktion $v_e := [\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e b_e$. Dann rechnet man zunächst mit (2.32)

$$\int_e [\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e v_e \,\mathrm{d}S = |[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e|^2 \underbrace{\int_e b_e \,\mathrm{d}S}_{=\frac{2}{3}h(e)} = \frac{2}{3}\|[\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e\|_{L_2(e)}^2. \tag{2.35}$$

Andererseits gilt wegen $v_e \in H_0^1(\omega_e)$ und $u_\mathcal{T}|_T \in P^1$ mit der Greenschen Formel

$$\begin{aligned}
\int_e [\![\nabla u_\mathcal{T}]\!]_e \cdot \mathbf{n}_e v_e \,\mathrm{d}S &= \sum_{\mathcal{T} \ni T \subset \omega_e} \int_{\partial T \cap e} v_e \nabla u_\mathcal{T} \cdot \mathbf{n} \,\mathrm{d}S \\
&= \sum_{\mathcal{T} \ni T \subset \omega_e} \int_{\partial T} v_e \nabla u_\mathcal{T} \cdot \mathbf{n} \,\mathrm{d}S \\
&= \sum_{\mathcal{T} \ni T \subset \omega_e} \int_T \nabla u_\mathcal{T} \nabla v_e \,\mathrm{d}\mathbf{x}. \\
&= \int_{\omega_e} \nabla u_\mathcal{T} \nabla v_e \,\mathrm{d}\mathbf{x}.
\end{aligned}$$

Da $\operatorname{supp} v_e = \omega_e$ und $v_e \in H_0^1(\Omega)$, gilt

$$\int_{\omega_e} \nabla u \nabla v_e \,\mathrm{d}\mathbf{x} = \int_{\Omega} \nabla u \nabla v_e \,\mathrm{d}\mathbf{x} = \int_{\Omega} f v_e \,\mathrm{d}\mathbf{x} = \int_{\omega_e} f v_e \,\mathrm{d}\mathbf{x},$$

so dass eine Addition von $0 = \int_{\omega_e} (f v_e - \nabla u \nabla v_e) \,\mathrm{d}\mathbf{x}$ ergibt

$$\begin{aligned}\int_e [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e v_e \,\mathrm{d}S &= \int_{\omega_e} f v_e \,\mathrm{d}\mathbf{x} - \int_{\omega_e} \nabla(u - u_{\mathcal{T}}) \nabla v_e \,\mathrm{d}\mathbf{x} \\ &\leq \sum_{\mathcal{T} \ni T \subset \omega_e} \big(\|f\|_{L_2(T)} \|v_e\|_{L_2(T)} + |u - u_{\mathcal{T}}|_{H^1(T)} |v_e|_{H^1(T)} \big) \\ &= |[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e| \sum_{\mathcal{T} \ni T \subset \omega_e} \big(\|f\|_{L_2(T)} \|b_e\|_{L_2(T)} + |u - u_{\mathcal{T}}|_{H^1(T)} |b_e|_{H^1(T)} \big).\end{aligned}$$

Wegen $0 \leq b_e(\mathbf{x}) \leq 1$ folgt $\|b_e\|_{L_2(T)}^2 \leq \int_e b_e \,\mathrm{d}S \leq c_5 h(e)^2$, also mit (2.33)

$$\begin{aligned}\int_e [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e v_e \,\mathrm{d}S &\leq |[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e| \|b_e\|_{L_2(T)} \sum_{\mathcal{T} \ni T \subset \omega_e} \big(\|f\|_{L_2(T)} + c_6 h(e)^{-1} |u - u_{\mathcal{T}}|_{H^1(T)} \big) \\ &\leq \sqrt{c_5} \|[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e\|_{L_2(e)} \sum_{\mathcal{T} \ni T \subset \omega_e} \big(h(e)^{1/2} \|f\|_{L_2(T)} + c_6 h(e)^{-1/2} |u - u_{\mathcal{T}}|_{H^1(T)} \big).\end{aligned}$$

Mit $\|f\|_{L_2(T)} \leq \|f - f_T\|_{L_2(T)} + \|f_T\|_{L_2(T)}$, Einsetzen von (2.30) und $h(T)^{-1} \leq h(e)^{-1}$ ergibt sich für den ersten Term der inneren Summe

$$\begin{aligned}h(e)^{1/2} \|f\|_{L_2(T)} &\leq h(e)^{1/2} \|f - f_T\|_{L_2(T)} + C' h(e)^{1/2} \big(h(T)^{-1} |u - u_{\mathcal{T}}|_{H^1(T)} + \|f - f_T\|_{L_2(T)} \big) \\ &= (1 + C') h(e)^{1/2} \|f - f_T\|_{L_2(T)} + C' h(e)^{-1/2} |u - u_{\mathcal{T}}|_{H^1(T)},\end{aligned}$$

also insgesamt

$$\int_e [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e v_e \,\mathrm{d}S \leq C |[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e| \sum_{\mathcal{T} \ni T \subset \omega_e} \big(h(e)^{1/2} \|f - f_T\|_{L_2(T)} + h(e)^{-1/2} |u - u_{\mathcal{T}}|_{H^1(T)} \big),$$

wobei $C > 0$ nur von $\max_{\mathcal{T} \ni T \subset \omega_e} \sigma(T)$ abhängt. Durch Anfügen der Identität (2.35) und Division durch $\|[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e\|_{L_2(e)}$ erhält man schließlich die gewünschte Abschätzung

$$\|[\![\nabla u_{\mathcal{T}}]\!] \cdot \mathbf{n}_e\|_{L_2(e)} \leq C' \sum_{\mathcal{T} \ni T \subset \omega_e} \big(h(e)^{1/2} \|f - f_T\|_{L_2(T)} + h(e)^{-1/2} |u - u_{\mathcal{T}}|_{H^1(T)} \big), \quad (2.36)$$

mit $C' = C'(\max_{\mathcal{T} \ni T \subset \omega_e} \sigma(T)) > 0$.

Mit (2.30) und (2.36) kann nun der Beweis der Effizienzabschätzung (2.15) angegangen werden. Auf jeder inneren Kante $e \in E_{\mathcal{T},\Omega}$ erhält man

$$\begin{aligned}\eta_e^2 &= h(\omega_e)^2 \|f\|_{L_2(\omega_e)}^2 + h(e) \big\| [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e \big\|_{L_2(e)}^2 \\ &= h(\omega_e)^2 \sum_{\mathcal{T} \ni T \subset \omega_e} \|f\|_{L_2(T)}^2 + h(e) \big\| [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e \big\|_{L_2(e)}^2 \\ &\leq 2 h(\omega_e)^2 \sum_{\mathcal{T} \ni T \subset \omega_e} \big(\|f_T\|_{L_2(T)}^2 + \|f - f_T\|_{L_2(T)}^2 \big) + h(e) \big\| [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e \big\|_{L_2(e)}^2.\end{aligned}$$

Einsetzen von (2.30) liefert zunächst

$$\eta_e^2 \leq C'' h(\omega_e)^2 \sum_{\mathcal{T} \ni T \subset \omega_e} \big(h(T)^{-2} |u - u_{\mathcal{T}}|^2_{H^1(T)} + \|f - f_T\|^2_{L_2(T)} \big) + h(e) \big\| [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e \big\|^2_{L_2(e)}.$$

Ist nun e eine Kante mit Nachbardreiecken $\mathcal{T} \ni T_1, T_2 \subset \omega_e$, so gilt die Abschätzung $h(T_i) \leq h(\omega_e) \leq h(T_1) + h(T_2)$. Andererseits gilt auf jedem Dreieck $h(e) \geq \rho(T_i) = \sigma(T_i)^{-1} h(T_i)$, also

$$h(\omega_e) \leq (\sigma(T_1) + \sigma(T_2)) h(e) \leq (\sigma(T_1) + \sigma(T_2)) h(T_i), \quad i \in \{1, 2\}.$$

Dann kann der Vorfaktor $h(\omega_e)$ verrechnet werden, wir erhalten

$$\eta_e^2 \leq C''' \sum_{\mathcal{T} \ni T \subset \omega_e} \big(|u - u_{\mathcal{T}}|^2_{H^1(T)} + h(T)^2 \|f - f_T\|^2_{L_2(T)} \big) + h(e) \big\| [\![\nabla u_{\mathcal{T}}]\!]_e \cdot \mathbf{n}_e \big\|^2_{L_2(e)} \quad (2.37)$$

mit einem nur von $\max_{\mathcal{T} \ni T \subset \omega_e} \sigma(T)$ abhängigen $C''' > 0$. Einsetzen der strukturell ähnlichen Abschätzung (2.36) in (2.37) liefert schließlich

$$\begin{aligned} \eta_e^2 \leq C''' & \sum_{\mathcal{T} \ni T \subset \omega_e} \big(|u - u_{\mathcal{T}}|^2_{H^1(T)} + h(T)^2 \|f - f_T\|^2_{L_2(T)} \big) \\ & + C h(e) \sum_{\mathcal{T} \ni T \subset \omega_e} \big(h(e) \|f - f_T\|^2_{L_2(T)} + h(e)^{-1} |u - u_{\mathcal{T}}|^2_{H^1(T)} \big), \end{aligned}$$

woraus sich nach Zusammenfassen der Terme und mittels $h(e) \leq h(T)$ endlich (2.15) ergibt.

Bemerkung 2.15. *Die Abschätzungstechnik über Bubble-Funktionen funktioniert auch in anderen Raumdimensionen als $n = 2$ sowie für allgemeinere Randbedingungen und Differentialoperatoren.*

2.3 Konvergenz adaptiver FEM

Zumindest für den Fall der Poisson-Gleichung mit homogenen Dirichlet-Randbedingungen ist die Existenz verlässlicher und effizienter, lokaler a posteriori-Fehlerschätzer η_e für das Variationsproblem (1.40) nun gesichert. Es stellt sich dann die Frage, wie ein darauf aufsetzendes adaptives Verfahren zu konstruieren ist und unter welchen Bedingungen dieses konvergiert. Bei der Darstellung orientieren wir uns an der Übersichtsarbeit [11].

Im Allgemeinen wird man eine iterative Verfeinerungsstrategie verfolgen, etwa durch einen generischen Algorithmus der Form **FEMSOLVE** (siehe Algorithmus 3). In jedem Iterationsschritt wird, sofern der Fehlerschätzer noch zu groß ist, die Triangulierung $\mathcal{T}_k$ verfeinert. Falls **FEMSOLVE** terminiert, wird so automatisch die Konvergenzbedingung (1.2) gesichert. *Dass* **FEMSOLVE** terminiert, ist allerdings noch zu zeigen. Unklar ist ebenfalls noch, wie die konkrete Wahl der Verfeinerungsstrategie aussieht, d.h. die Entscheidung, welche Facetten $e \in E_{\mathcal{T},\Omega}$ und die dazugehörenden Nachbarsimplizes genau zu unterteilen sind. Um die Terminierungseigenschaft zu sichern, wird man den

Algorithmus 3 FEMSOLVE$[f, \epsilon] \to u_\epsilon$

Wähle Starttriangulierung $\mathcal{T}_0$ von Ω.
$k := 0$
loop
  Berechne Galerkin-Lösung $u_{\mathcal{T}_k}$ von (1.40).
  Berechne η_e für alle $e \in E_{\mathcal{T}_k,\Omega}$.
  if $\sum \eta_e^2 \leq \epsilon^2$ **then**
    $u_\epsilon := u_{\mathcal{T}_k}$
    return
  else
    Verfeinere $\mathcal{T}_k \mapsto \mathcal{T}_{k+1}$.
    $k := k + 1$
  end if
end loop

Verfeinerungsschritt zudem gerne so gestalten wollen, dass eine gewisse Fehlerreduktion pro Iterationsschritt zugesichert wird.

Generell zerfällt der Verfeinerungsschritt bei adaptiven Finite Elemente-Verfahren meist in zwei Schritte:

(i) Markieren zu verfeinernder Simplizes (Markierungsstrategie);

(ii) Verfeinern der markierten Simplizes, inklusive Abschluss (z.B. *green closure*).

Eine Markierungsstrategie liefert für die aktuelle Triangulierung $\mathcal{T}$ inklusive bereits berechneter Fehlerschätzer e genau eine Teilmenge $E \subset E_{\mathcal{T},\Omega}$. Markiert werden dann alle Simplizes $T \in \mathcal{T}$ mit einer Facette in E. Eine typische Markierungsstrategie könnte so aussehen:

$$0 < \theta < 1 \;\Rightarrow\; \text{wähle } E \subset E_{\mathcal{T},\Omega} \text{ mit } \Big(\sum_{e\in E} \eta_e^2\Big)^{1/2} \geq \theta \Big(\sum_{e\in E_{\mathcal{T},\Omega}} \eta_e^2\Big)^{1/2}, \tag{2.38}$$

d.h. es werden hinreichend viele $e \in E_{\mathcal{T},\Omega}$ ausgewählt. Eine konkrete Implementierung der Markierungsstrategie (2.38) könnte so aussehen, dass E iterativ aufgebaut wird, beginnend mit denjenigen $e \in E_{\mathcal{T},\Omega}$ zu großen Werten η_e. Der folgende zentrale Satz dieses Abschnitts zeigt, dass eine auf (2.38) basierende Verfeinerung bis auf einen Oszillationsterm zu einer Fehlerreduktion führt:

Satz 2.16. *Sei $\mathcal{T}$ quasi-uniform mit $\sigma(T) \leq \sigma_0$ für alle $T \in \mathcal{T}$ und sei $\hat{\mathcal{T}}$ aus Markierungsstrategie* (2.38)*. Sei weiter $\mathcal{T}'$ eine Verfeinerung von $\mathcal{T}$, so dass gilt:*

Jedes $T \in \hat{\mathcal{T}}$ und jede seiner Facetten enthält einen inneren Knoten aus $\mathcal{T}'$. (2.39)

Dann existiert ein $0 < \alpha < 1$ mit $\alpha = \alpha(\sigma_0, \theta, C_a, C_e)$, so dass für die Galerkin-Lösungen $u_{\mathcal{T}}$, $u_{\mathcal{T}'}$ gilt

$$|u - u_{\mathcal{T}'}|_{H^1(\Omega)} \leq \alpha |u - u_{\mathcal{T}}|_{H^1(\Omega)} + \operatorname{osc}(f, \mathcal{T}). \tag{2.40}$$

Falls für den Oszillationsterm eine Abschätzung der Form $\operatorname{osc}(f,\mathcal{T}) \le \mu|u-u_{\mathcal{T}}|_{H^1(\Omega)}$ gezeigt werden kann mit einem $0<\mu<1-\alpha$, so würde aus Satz 2.16 offenbar die Reduktion

$$|u-u_{\mathcal{T}'}|_{H^1(\Omega)} \le \beta|u-u_{\mathcal{T}}|_{H^1(\Omega)}, \quad \beta=\alpha+\mu<1, \tag{2.41}$$

folgen.

Bemerkung 2.17. *Die Erzeugung innerer Knoten gemäß* (2.39) *ist eine Anforderung an die konkrete Simplexverfeinerung. Erfüllt wird diese z.B. durch die* newest vertex bisection*, siehe Übungsaufgabe.*

Bevor wir Satz 2.16 beweisen, zeigen wir zunächst, wie man eine Reduktion des Oszillationsterms erzwingen kann.

Lemma 2.18. *Sei $0<\gamma<1$ der Reduktionsfaktor beim Verfeinern von Simplizes, d.h. für einen verfeinerten Simplex $T'\subset T\in\mathcal{T}$ gelte $h(T')\le\gamma h(T)$. Für ein $0<\tilde{\theta}<1$ sei weiter $\hat{\mathcal{T}}\subset\mathcal{T}$ mit*

$$\operatorname{osc}(f,\hat{\mathcal{T}}) \ge \tilde{\theta}\operatorname{osc}(f,\mathcal{T}). \tag{2.42}$$

Ist dann $\mathcal{T}'$ eine Verfeinerung von $\mathcal{T}$, bei der jedes $T\in\hat{\mathcal{T}}$ auch wirklich verfeinert wurde, dann gilt

$$\operatorname{osc}(f,\mathcal{T}') \le \hat{\alpha}\operatorname{osc}(f,\mathcal{T}), \tag{2.43}$$

wobei $0<\hat{\alpha}:=(1-(1-\gamma^2)\hat{\theta}^2)^{1/2}<1$.

Beweis: Sei $T\in\mathcal{T}'$ enthalten in $\hat{T}\in\hat{\mathcal{T}}$. Die Abbildung $L_2(T)\ni f\mapsto f_T$ ist die Orthogonalprojektion auf den Raum der konstanten Funktionen, denn für jede Konstante $g\in\mathbb{R}$ gilt

$$\langle f-f_T,g\rangle_{L_2(T)} = g\int_T f(\mathbf{x})\,\mathrm{d}\mathbf{x} - \frac{1}{|T|}g\int_T \mathrm{d}\mathbf{x}\int_T f(\mathbf{y})\,\mathrm{d}\mathbf{y} = 0.$$

Daher ist $f-f_T$ orthogonal zu konstanten Funktion $f_T-f_{\hat{T}}$, so dass mit dem Satz von Pythagoras folgt

$$\|f-f_{\hat{T}}\|^2_{L_2(T)} = \|f-f_T\|^2_{L_2(T)} + \|f_T-f_{\hat{T}}\|^2_{L_2(T)} \ge \|f-f_T\|^2_{L_2(T)}.$$

Mit $h(T)\le\gamma h(T')$ folgt daraus für alle $\hat{T}\in\hat{\mathcal{T}}$

$$\begin{aligned}\sum_{\mathcal{T}'\ni T\subset\hat{T}} h(T)^2\|f-f_T\|^2_{L_2(T)} &\le \gamma^2 h(\hat{T})^2 \sum_{\mathcal{T}'\ni T\subset\hat{T}} \|f-f_T\|^2_{L_2(T)}\\ &\le \gamma^2 h(\hat{T})^2 \sum_{\mathcal{T}'\ni T\subset\hat{T}} \|f-f_{\hat{T}}\|^2_{L_2(T)}\\ &= \gamma^2 h(\hat{T})^2 \|f-f_{\hat{T}}\|^2_{L_2(\hat{T})}\end{aligned}$$

und daher

$$\begin{aligned}\operatorname{osc}(f,\mathcal{T}')^2 &= \sum_{\hat{T}\in\hat{\mathcal{T}}}\sum_{\mathcal{T}'\ni T\subset\hat{T}} h(T)^2\|f-f_T\|^2_{L_2(T)} + \sum_{T\in\mathcal{T}\setminus\hat{\mathcal{T}}} h(T)^2\|f-f_T\|^2_{L_2(T)}\\ &\le \gamma^2\sum_{\hat{T}\in\hat{\mathcal{T}}} h(\hat{T})^2\|f-f_{\hat{T}}\|^2_{L_2(\hat{T})} + \sum_{T\in\mathcal{T}\setminus\hat{\mathcal{T}}} h(T)^2\|f-f_T\|^2_{L_2(T)}\\ &= (\gamma^2-1)\operatorname{osc}(f,\hat{\mathcal{T}})^2 + \operatorname{osc}(f,\mathcal{T})^2\\ &\le \big(1-(1-\gamma^2)\hat{\theta}^2\big)\operatorname{osc}(f,\mathcal{T})^2.\end{aligned}$$

□

Wir erweitern daher die Markierungsstrategie (2.38) um einen zweiten Teil, um die Sättigungsbedingung (2.42) sicher zu stellen:

$$0<\hat{\theta}<1 \quad\Rightarrow\quad \text{erweitere } E \text{ aus (2.38), so dass } \operatorname{osc}(f,\hat{\mathcal{T}})\ge\hat{\theta}\operatorname{osc}(f,\mathcal{T}). \tag{2.44}$$

Hiermit erhalten wir folgenden Algorithmus:

Algorithmus 4 FEMSOLVE2$[f,\epsilon]\to u_\epsilon$

Wähle Starttriangulierung $\mathcal{T}_0$ von Ω.
$k:=0$
loop
  Berechne Galerkin-Lösung $u_{\mathcal{T}_k}$ von (1.40).
  Berechne η_e für alle $e\in E_{\mathcal{T}_k,\Omega}$.
  if $\sum\eta_e^2\le\epsilon^2$ **then**
    $u_\epsilon := u_{\mathcal{T}_k}$
    return
  else
    Bestimme $E\subset E_{\mathcal{T},\Omega}$ gemäß Markierungsstrategie (2.38).
    Erweitere E gemäß Markierungsstrategie (2.44).
    Verfeinere $\mathcal{T}_k\mapsto\mathcal{T}_{k+1}$ gemäß (2.39).
    $k:=k+1$
  end if
end loop

Für **FEMSOLVE2** lässt sich schließlich Konvergenz zeigen:

Satz 2.19. *Sei $\mathcal{T}$ quasi-uniform mit $\sigma(T)\le\sigma_0$ für alle $T\in\mathcal{T}$. Zu $0<\theta,\hat{\theta}<1$ seien weiter α aus Satz 2.16 und $\hat{\alpha}$ aus Lemma 2.18, dazu $\alpha_0:=\max\{\alpha,\hat{\alpha}\}$. Dann erfüllt die Folge $\{u_{\mathcal{T}_k}\}$ der Galerkin-Lösungen in* **FEMSOLVE2** *zu jedem $\alpha_0<\beta<1$ die Fehlerabschätzung*

$$|u-u_{\mathcal{T}_k}|_{H^1(\Omega)}\le C_0\beta^k \tag{2.45}$$

mit $C_0:=|u-u_{\mathcal{T}_0}|_{H^1(\Omega)}+\frac{1}{\beta-\alpha_0}\operatorname{osc}(f,\mathcal{T}_0)$, d.h. **FEMSOLVE2** *terminiert nach endlich vielen Iterationen.*

Beweis: Aus Lemma 2.18 erhält man zunächst

$$\operatorname{osc}(f,\mathcal{T}_k) \le \hat{\alpha}\operatorname{osc}(f,\mathcal{T}_{k-1}) \le \cdots \le \hat{\alpha}^k \operatorname{osc}(f,\mathcal{T}_0).$$

Mit Satz 2.18 folgt hieraus

$$\begin{aligned} |u-u_{\mathcal{T}_{k+1}}|_{H^1(\Omega)} &\le \alpha|u-u_{\mathcal{T}_k}|_{H^1(\Omega)} + \hat{\operatorname{osc}}(f,\mathcal{T}_k) \\ &\le \alpha|u-u_{\mathcal{T}_k}|_{H^1(\Omega)} + \hat{\alpha}^k \operatorname{osc}(f,\mathcal{T}_0) \end{aligned}$$

und mit Induktion über k

$$|u-u_{\mathcal{T}_{k+1}}|_{H^1(\Omega)} \le \alpha^{k+1}|u-u_{\mathcal{T}_0}|_{H^1(\Omega)} + \operatorname{osc}(f,\mathcal{T}_0)\sum_{j=0}^{k}\alpha^j\hat{\alpha}^{k-j}.$$

Für die letzte Summe gilt wegen $\alpha \le \alpha_0$ und $\hat{\alpha} < \beta$

$$\sum_{j=0}^{k}\alpha^j\hat{\alpha}^{k-j} \le \beta^k\sum_{j=0}^{k}\Big(\frac{\alpha_0}{\beta}\Big)^j \le \beta^k\frac{1}{1-\frac{\alpha_0}{\beta}} = \beta^{k+1}\frac{1}{\beta-\alpha_0},$$

woraus die Behauptung folgt. □

Für den Beweis von Satz 2.16 benötigen wir noch folgendes Lemma.

Lemma 2.20. *Sei $\mathcal{T}$ quasi-uniform mit $\sigma(T) \le \sigma_0$ für alle $T \in \mathcal{T}$. Weiter sei $E \subset E_{\mathcal{T},\Omega}$ gemäß Markierungsstrategien* (2.38) *und* (2.44) *bestimmt worden und $\mathcal{T}'$ sei eine Verfeinerung von $\mathcal{T}$ mit* (2.39)*. Dann existiert ein $C = C(\sigma_0, C_a, C_e)$, so dass mit den Fehlerschätzern η_e aus* (2.25) *gilt*

$$\eta_e^2 \le C\Big(|u_{\mathcal{T}'}-u_{\mathcal{T}}|^2_{H^1(\omega_e)} + \sum_{\mathcal{T}\in T\subset\omega_e} h(T)^2\|f-f_T\|^2_{L_2(T)}\Big), \quad \textit{für alle } e \in E. \tag{2.46}$$

Beweis: Mit Hilfe der inneren Knoten verfeinerter Simplizes und Facetten lassen sich Bubble-Funktionen $b_i \in P^1_{\mathcal{T}'} \cap H^1_0(\omega_e)$ definieren. Der Nachweis von (2.46) verläuft dann vollkommen analog zum Nachweis der Effizienz (2.15) der η_e. Für die genauen Details verweisen wir auf [11]. □

Kommen wir nun zum Beweis von Satz 2.16. Da $\mathcal{T}'$ eine Verfeinerung von $\mathcal{T}$ ist, liegt $u_{\mathcal{T}'} - u_{\mathcal{T}}$ im Ansatzraum $P^1_{\mathcal{T}'} \cap H^1_0(\Omega)$. Folglich gilt $a(u-u_{\mathcal{T}'}, u_{\mathcal{T}'}-u_{\mathcal{T}}) = 0$, so dass man mit dem Satz von Pythagoras folgert

$$|u-u_{\mathcal{T}'}|^2_{H^1(\Omega)} = |u-u_{\mathcal{T}}|^2_{H^1(\Omega)} - |u_{\mathcal{T}'}-u_{\mathcal{T}}|^2_{H^1(\Omega)}. \tag{2.47}$$

Eine Reduktion von $|u-u_{\mathcal{T}'}|_{H^1(\Omega)}$ gegenüber dem vorigen Fehler $|u-u_{\mathcal{T}}|_{H^1(\Omega)}$ kann also erreicht werden, falls $|u_{\mathcal{T}'}-u_{\mathcal{T}}|_{H^1(\Omega)}$ nach unten abgeschätzt werden kann. Dies

geschieht mit Hilfe der Markierungsstrategie (2.38) und Lemma 2.20 wie folgt:

$$\begin{aligned}\theta^2 \sum_{e\in E_{\mathcal{T},\Omega}} \eta_e^2 &\le \sum_{e\in E} \eta_e^2 \\ &\le C \sum_{e\in E} \Big(|u_{\mathcal{T}'} - u_{\mathcal{T}}|^2_{H^1(\omega_e)} + \sum_{\mathcal{T}\in T\subset\omega_e} h(T)^2 \|f - f_T\|^2_{L_2(T)} \Big) \\ &\le (n+1)C \Big(|u_{\mathcal{T}'} - u_{\mathcal{T}}|^2_{H^1(\Omega)} + \sum_{T\in\mathcal{T}} h(T)^2 \|f - f_T\|^2_{L_2(T)} \Big) \\ &= (n+1)C \big(|u_{\mathcal{T}'} - u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f,\mathcal{T})^2 \big),\end{aligned}$$

denn jedes Simplex T hat $n+1$ Facetten. Umstellen nach der gesuchten Größe liefert mit C_1 aus (2.14)

$$|u_{\mathcal{T}'} - u_{\mathcal{T}}|^2_{H^1(\Omega)} \ge \frac{\theta^2}{(n+1)C} \sum_{e\in E_{\mathcal{T},\Omega}} \eta_e^2 - \operatorname{osc}(f,\mathcal{T})^2 \ge \frac{\theta^2 C_1^2}{(n+1)C},$$

also mit (2.47) insgesamt

$$\begin{aligned}|u - u_{\mathcal{T}'}|^2_{H^1(\Omega)} &= |u - u_{\mathcal{T}}|^2_{H^1(\Omega)} - |u_{\mathcal{T}'} - u_{\mathcal{T}}|^2_{H^1(\Omega)} \\ &\le \Big(1 - \frac{\theta^2 C_1^2}{(n+1)C}\Big) |u - u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f,\mathcal{T})^2,\end{aligned}$$

so dass der Beweis von Satz 2.16 mit Hilfe von $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1$ und $\alpha := (1 - \frac{\theta^2 C_1^2}{(n+1)C})^{1/2}$ abgeschlossen ist. □

3 Adaptive Wavelet-Verfahren

3.1 Riesz-Basen und Wavelets

Eine alternative Diskretisierung elliptischer Variationsprobleme lässt sich mit Hilfe sogenannter *Wavelet*-Systeme durchführen. Die Philosophie ist hierbei, nicht unbedingt von vornherein mit endlich-dimensionalen Approximationen zu arbeiten (wie z.B. stückweise polynomiale Funktionen auf einer festen Triangulierung), sondern eine Basis für den *gesamten* Lösungs-Hilbertraum X des Variationsproblems (1.40) zu betrachten.

Definition 3.1. *Eine abzählbare Menge $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X$ heißt* Hilbertraum-Basis *oder* Riesz-Basis *von X, falls* $\operatorname{span}\Psi$ *dicht in X liegt und für Konstanten $c_R, C_R > 0$ die Abschätzung gilt*

$$c_R\|\mathbf{u}\|_{\ell_2(\mathcal{I})} \le \|u\|_X \le C_R\|\mathbf{u}\|_{\ell_2(\mathcal{I})}, \quad \text{für alle } \mathbf{u} = (u_\lambda)_{\lambda\in\mathcal{I}} \in \ell_2(\mathcal{I}). \tag{3.1}$$

Hierbei ist $\|\mathbf{u}\|_{\ell_2(\mathcal{I})} := (\sum_{\lambda\in\mathcal{I}} |u_\lambda|^2)^{1/2}$ die Norm auf dem Hilbertraum $\ell_2(\mathcal{I})$.

Beispiel 3.2. *Jede Orthonormalbasis Ψ von X ist eine Riesz-Basis mit $c_R = C_R = 1$, dies folgt aus der* Parseval-Gleichung

$$\Big\|\sum_{\lambda\in\mathcal{I}} u_\lambda\psi_\lambda\Big\|_X^2 = \sum_{\lambda\in\mathcal{I}} |u_\lambda|^2. \tag{3.2}$$

So ist zum Beispiel das Fourier-System $\{1, \sqrt{2}\sin(k\pi\cdot), \sqrt{2}\cos(k\pi\cdot)\}$ eine Orthonormal- und daher Rieszbasis von $L_2(0,1)$.

Für allgemeinere Hilberträume als $L_2(\Omega)$, insbesondere für die bei der Diskretisierung elliptischer Differentialgleichungen auftretenden Sobolevräume $H^s(\Omega)$, ist die Konstruktion von Orthonormalbasen allerdings schwierig. Daher verzichtet man in der Praxis meist auf die Orthogonalitätsforderung.

Lemma 3.3. *Zu jeder Riesz-Basis $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X$ eines Hilbertraums X gibt es genau ein System $\tilde\Psi = \{\tilde\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X'$ mit der Biorthogonalitätsbedingung*

$$\langle\psi_\mu, \tilde\psi_\lambda\rangle_{X\times X'} = \delta_{\mu,\lambda}, \quad \text{für } \mu,\lambda \in \mathcal{I}. \tag{3.3}$$

$\tilde\Psi$ ist eine Riesz-Basis von X' (die duale Riesz-Basis*) mit Riesz-Konstanten $\tilde c_R = C_R^{-1}$, $\tilde C_R = c_R^{-1}$.*

Beweis: Wir zeigen zunächst, dass jedes $u \in X$ eine eindeutige Reihendarstellung $u = \sum_{\lambda\in\mathcal{I}} u_\lambda \psi_\lambda$ besitzt mit einer Koeffizientenfolge $\mathbf{u} = (u_\lambda)_{\lambda\in\mathcal{I}} \in \ell_2(\mathcal{I})$. Wegen (3.1) ist der Operator $T : \ell_2(\mathcal{I}) \to X$, $T\mathbf{u} = \sum_{\lambda\in\mathcal{I}} u_\lambda\psi_\lambda$, wohldefiniert und stetig mit $\|T\mathbf{u}\|_X \le C_R\|\mathbf{u}\|_{\ell_2(\mathcal{I})}$, ferner injektiv, $\|T\mathbf{u}\|_X \ge c_R\|\mathbf{u}\|_{\ell_2(\mathcal{I})}$. Da $\operatorname{span}\Psi$ dicht in X liegt, gilt dies auch für das Bild von T, also ist T surjektiv und damit bijektiv. Daher ist der Operator $T^{-1} : X \to \ell_2(\mathcal{I})$ nach oben und unten beschränkt, $C_R^{-1}\|u\|_X \le \|T^{-1}u\|_{\ell_2(\mathcal{I})} \le c_R^{-1}\|u\|_X$ für alle $u \in X$. Aufgrund der ersten Ungleichung in (3.1) sind die Koeffizienten u_λ von u eindeutig.

Die linearen Koeffizientenfunktionale $\tilde\psi_\lambda(u) = \langle u, \tilde\psi_\lambda\rangle_{X\times X'} := u_\lambda$ sind beschränkt für alle $\lambda \in \mathcal{I}$ (sogar gleichmäßig in λ) und damit $\tilde\Psi = \{\tilde\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X'$. Die Biorthogonalitätsbedingung (3.3) ist wegen

$$\sum_{\mu\in\mathcal{I}} \delta_{\mu,\lambda}\psi_\mu = \psi_\lambda = \sum_{\mu\in\mathcal{I}} \langle\psi_\lambda, \tilde\psi_\mu\rangle_{X\times X'}\psi_\mu$$

erfüllt. Für ein zweites System $\check\Psi = \{\check\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X'$ mit (3.3) folgt $\langle\psi_\mu, \tilde\psi_\lambda - \check\psi_\lambda\rangle_{X\times X'} = 0$ für alle $\mu, \lambda \in \mathcal{I}$. Für jedes $u = \sum_{\lambda\in\mathcal{I}} u_\lambda\psi_\lambda \in X$ erhält man dann aber aufgrund der Stetigkeit der Funktionale $\tilde\psi_\lambda - \check\psi_\lambda$, dass

$$\langle u, \tilde\psi_\lambda - \check\psi_\lambda\rangle_{X\times X'} = \sum_{\mu\in\mathcal{I}} u_\mu\langle\psi_\mu, \tilde\psi_\lambda - \check\psi_\lambda\rangle_{X\times X'} = 0.$$

Da $u \in X$ beliebig war, folgt $\tilde\psi_\lambda = \check\psi_\lambda$ in X', d.h. $\tilde\Psi$ ist durch (3.3) eindeutig festgelegt.

Zum Nachweis der Riesz-Basis-Eigenschaft von $\tilde\Psi$ zeigen wir zunächst die Darstellung $\langle\cdot, \tilde\psi_\lambda\rangle_{X\times X'} = \langle\cdot, (TT^*)^{-1}\psi_\lambda\rangle_X$, mit dem durch

$$\langle T^*u, \mathbf{v}\rangle_{\ell_2(\mathcal{I})} = \langle u, T\mathbf{v}\rangle_{\ell_2(\mathcal{I})}, \quad \text{für alle } u \in X, \mathbf{v} \in \ell_2(\mathcal{I}),$$

gegebenen adjungierten Operator $T^* : X \to \ell_2(\mathcal{I})$. Denn es ist

$$\begin{aligned} &\langle v, (TT^*)^{-1}\psi_\lambda\rangle_X = \langle v, \tilde\psi_\lambda\rangle_{X\times X'}, \quad \text{für alle } v \in X \\ &\quad\Leftrightarrow\quad \langle \underbrace{T^{-1}v}_{=\mathbf{v}}, \underbrace{T^{-1}\psi_\lambda}_{=\mathbf{e}_\lambda}\rangle_{\ell_2(\mathcal{I})} = v_\lambda, \quad \text{für alle } v = \sum_{\lambda\in\mathcal{I}} v_\lambda\psi_\lambda \in X, \end{aligned}$$

und letzteres ist offenbar richtig.

Für die Dichtheit von $\tilde\Psi$ sei $\tilde u \in X'$ beliebig und $\epsilon > 0$. Dann existiert genau ein $u \in X$ mit $\langle v, \tilde u\rangle_{X\times X'} = \langle v, u\rangle_X$ für alle $v \in X$. Da $\operatorname{span}\Psi$ dicht in X ist und TT^* ein Isomorphismus (als Verkettung von Isomorphismen), ist auch $\operatorname{span}\{(TT^*)^{-1}\psi_\lambda \mid \lambda \in \mathcal{I}\}$ dicht in X. Es existiert daher eine endliche Menge $\Lambda \subset \mathcal{I}$ und Koeffizienten c_λ, $\lambda \in \Lambda$, so dass $\|u - \sum_{\lambda\in\Lambda} c_\lambda(TT^*)^{-1}\psi_\lambda\|_X \le \epsilon$. Folglich ist

$$\begin{aligned} \Big\|\tilde u - \sum_{\lambda\in\Lambda} c_\lambda\tilde\psi_\lambda\Big\|_{X'} &= \sup_{v\in X} \frac{|\langle v, \tilde u - \sum_{\lambda\in\Lambda} c_\lambda\tilde\psi_\lambda\rangle_{X\times X'}|}{\|v\|_X} \\ &= \sup_{v\in X} \frac{|\langle v, u - \sum_{\lambda\in\Lambda} c_\lambda(TT^*)^{-1}\psi_\lambda\rangle_X|}{\|v\|_X} \le \epsilon, \end{aligned}$$

d.h. $\operatorname{span}\tilde{\Psi}$ ist dicht in X'.

Sei dann $\mathbf{c} = (c_\lambda)_{\lambda\in\mathcal{I}} \in \ell_2(\mathcal{I})$ beliebig. Es folgt

$$\Big\| \sum_{\lambda\in\mathcal{I}} c_\lambda \tilde{\psi}_\lambda \Big\|_{X'} = \sup_{v\in X} \frac{|\langle v, \sum_{\lambda\in\mathcal{I}} c_\lambda \tilde{\psi}_\lambda\rangle_{X\times X'}|}{\|v\|_X} = \sup_{\mathbf{v}\in\ell_2(\mathcal{I})} \frac{|\langle \mathbf{v}, \mathbf{c}\rangle_{\ell_2(\mathcal{I})}|}{\|T\mathbf{v}\|_X}.$$

Einsetzen der Ungleichung (3.1) in den Nenner liefert schließlich wie behauptet

$$C_R^{-1}\|\mathbf{c}\|_{\ell_2(\mathcal{I})} \le \Big\| \sum_{\lambda\in\mathcal{I}} c_\lambda \tilde{\psi}_\lambda \Big\|_{X'} \le c_R^{-1}\|\mathbf{c}\|_{\ell_2(\mathcal{I})}.$$

□

Bemerkung 3.4. *Falls X mit seinem Dualraum X' identifiziert wird, kann man die duale Riesz-Basis $\tilde{\Psi}$ auch als Teilmenge von X auffassen, $\tilde{\psi}_\lambda = (TT^*)^{-1}\psi_\lambda$.*

Wir haben in Lemma 3.3 gesehen, dass eine Riesz-Basis Ψ von X notwendigerweise biorthogonal ist zu ihrer dualen Riesz-Basis $\tilde{\Psi}$. Allerdings ist der direkte Nachweis der Riesz-Basis-Eigenschaft im Allgemeinen nicht einfach. Zumindest die Biorthogonalität zweier Systeme $\Psi, \tilde{\Psi} \subset X$ lässt sich aber leicht überprüfen. Den Nachweis der Riesz-Basis-Eigenschaft erleichtert dann die Betrachtung von Funktionensystemen eines bestimmten Typs.

Definition 3.5. *Ein System $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}} \subset X$ heißt* vom Wavelet-Typ*, wenn eine Skalenabbildung $\lambda \mapsto |\lambda| \in \mathbb{N}_0$ existiert. Das Gesamtsystem Ψ zerfällt dadurch disjunkt in die Mengen $\Psi_j := \{\psi_\lambda \mid |\lambda| = j\}$, $j \in \mathbb{N}_0$. Wir bezeichnen ferner die Erzeugendensysteme der aufsteigenden Räume $V_j := \operatorname{clos}_X \operatorname{span}\Psi^j$ mit $\Psi^j := \{\psi_\lambda \mid |\lambda| \le j\}$.*

Der folgende Satz zeigt, dass gewisse Approximations- und Regularitätsannahmen über ein biorthogonales System $\Psi, \tilde{\Psi}$ hinreichen, um eine Riesz-Basis zu erhalten.

Satz 3.6. *Seien $\Psi, \tilde{\Psi} \subset X$ biorthogonal, d.h. es gelte $\langle \psi_\lambda, \tilde{\psi}_\mu\rangle_X = \delta_{\lambda,\mu}$. Weiter seien die folgenden beiden Eigenschaften erfüllt:*

(i) Es gebe Projektoren $Q_j : X \to V_j$, die gleichmäßig beschränkt sind, d.h. es gelte $\|Q_j\| \le C$ für ein $C > 0$ und alle $j \in \mathbb{N}_0$, weiter sei $Q_jQ_k = Q_j$ für $k \ge j$.

(ii) Es gebe einen dicht eingebetteten Hilbertraum $Y \hookrightarrow X$, $C > 0$ und $\delta > 0$, so dass die Jackson-Ungleichungen

$$\inf_{v\in V_j} \|u - v\|_{Y'} \le C2^{-j\delta}\|u\|_X, \quad \inf_{v\in V_j} \|u - v\|_X \le C2^{-j\delta}\|u\|_Y, \quad \text{für alle } u \in Y, \tag{3.4}$$

und die Bernstein-Ungleichungen

$$\|v\|_Y \le C2^{j\delta}\|v\|_X, \quad \|v\|_X \le C2^{j\delta}\|v\|_{Y'}, \quad \text{für alle } v \in V_j, \tag{3.5}$$

gelten, d.h. insbesondere sei $\operatorname{span}\Psi$ dicht in X und $\Psi \subset Y$. Ferner seien die Projektoren Q_j^* auf Y' gleichmäßig beschränkt.

Dann gilt mit $Q_{-1} := 0$, $Q_{-1}^* := 0$ *und einer Konstanten* $C' > 0$

$$\|v\|_X^2 \le C' \sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2, \quad \textit{für alle } v \in X, \tag{3.6}$$

$$\|v\|_X^2 \le C' \sum_{j=0}^{\infty} \|(Q_j^* - Q_{j-1}^*)v\|_X^2, \quad \textit{für alle } v \in X. \tag{3.7}$$

Beweis: Da Y dicht in X, gilt für alle $u \in Y$ und $v \in X$ die verbesserte Cauchy-Schwarz-Ungleichung

$$\langle u, v\rangle_X = \langle u, v\rangle_{Y\times Y'} \le \|u\|_Y \|v\|_{Y'}.$$

Wir können daher für $i, j \in \mathbb{N}_0$ abschätzen

$$\langle (Q_j - Q_{j-1})v, (Q_i - Q_{i-1})v\rangle_X \le \begin{cases} \|(Q_j - Q_{j-1})v\|_{Y'} \|(Q_i - Q_{i-1})v\|_Y, & i \le j \\ \|(Q_j - Q_{j-1})v\|_Y \|(Q_i - Q_{i-1})v\|_{Y'}, & i > j \end{cases}.$$

Mit $\|\cdot\| \in \{\|\cdot\|_{Y'}, \|\cdot\|_X\}$ gilt

$$\|Q_j v - v\| \le \inf_{w\in V_j} \big(\|Q_j v - w\| + \|v - w\|\big) \le \big(1 + \|Q_j\|\big) \inf_{w\in V_j} \|v - w\|,$$

so dass aus $Q_{j-1}(Q_j - Q_{j-1}) = 0$ und $Q_j \to I$ folgt

$$\begin{aligned}
\|v\|_X^2 &= \sum_{i,j=0}^{\infty} \langle (Q_j - Q_{j-1})v, (Q_i - Q_{i-1})v\rangle_X \\
&\le \sum_{i\le j} \|(Q_j - Q_{j-1})v\|_{Y'} \|(Q_i - Q_{i-1})v\|_Y + \sum_{i>j} \|(Q_j - Q_{j-1})v\|_Y \|(Q_i - Q_{i-1})v\|_{Y'} \\
&= \sum_{i\le j} \|(Q_j - Q_{j-1} - Q_{j-1}(Q_j - Q_{j-1}))v\|_{Y'} \|(Q_i - Q_{i-1})v\|_Y \\
&\quad + \sum_{i>j} \|(Q_j - Q_{j-1} - Q_{j-1}(Q_j - Q_{j-1}))v\|_Y \|(Q_i - Q_{i-1})v\|_{Y'} \\
&\le C' \sum_{i\le j} 2^{-(j-i)\delta} \|(Q_j - Q_{j-1})v\|_X \|(Q_i - Q_{i-1})v\|_X \\
&\quad + C' \sum_{i>j} 2^{-(i-j)\delta} \|(Q_j - Q_{j-1})v\|_X \|(Q_i - Q_{i-1})v\|_X \\
&= C' \sum_{i,j=0}^{\infty} 2^{-\delta|i-j|} \|(Q_j - Q_{j-1})v\|_X \|(Q_i - Q_{i-1})v\|_X.
\end{aligned}$$

Diese Summe hat die Form $\langle \mathbf{Mx}, \mathbf{x}\rangle_{\ell_2}$, mit einer Matrix $\mathbf{M} = (m_{i,j})$ mit summierbaren Zeilen und Spalten, $\sum_i |m_{i,j}| \le K$, $\sum_j |m_{i,j}| \le K$, wobei $K = K(\delta)$. Nach dem Schur-Lemma (siehe Satz 3.14) folgt $\|\mathbf{Mx}\|_{\ell_2} \le K\|\mathbf{x}\|_{\ell_2}$, so dass wir (3.6) erhalten,

$$\|v\|_X^2 \le C'' \sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2.$$

Die zweite Abschätzung (3.7) folgt analog durch ein Dualitätsargument; man weist die Gültigkeit von Jackson- und Bernstein-Ungleichungen für die adjungierten Operatoren Q_j^* nach. □

Lemma 3.7. *Unter den Voraussetzungen von Satz 3.6 folgt aus* (3.7) *die Abschätzung*

$$\|v\|_X^2 \geq C'' \sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2, \quad \textit{für alle } v \in X, \tag{3.8}$$

mit einer Konstanten $C'' > 0$, also insgesamt $\|v\|_X^2 \eqsim \sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2$ für alle $v \in X$ (Äquivalenz bis auf Konstanten).

Beweis: Zunächst ist

$$\begin{aligned}\sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2 &= \sum_{j=0}^{\infty} \langle (Q_j^* - Q_{j-1}^*)(Q_j - Q_{j-1})v, v\rangle_X \\ &\leq \Big\| \sum_{j=0}^{\infty} (Q_j^* - Q_{j-1}^*)(Q_j - Q_{j-1})v \Big\|_X \|v\|_X.\end{aligned}$$

Andererseits folgt aus (3.7) wegen $(Q_j^* - Q_{j-1}^*)(Q_k^* - Q_{k-1}^*) = \delta_{j,k}(Q_j^* - Q_{j-1}^*)$ aber

$$\begin{aligned}&\Big\| \sum_{j=0}^{\infty} (Q_j^* - Q_{j-1}^*)(Q_j - Q_{j-1})v \Big\|_X^2 \\ &\leq C' \sum_{k=0}^{\infty} \Big\| (Q_k^* - Q_{k-1}^*) \sum_{j=0}^{\infty} (Q_j^* - Q_{j-1}^*)(Q_j - Q_{j-1})v \Big\|_X^2 \\ &= C' \sum_{k=0}^{\infty} \Big\| (Q_k^* - Q_{k-1}^*)(Q_k - Q_{k-1})v \Big\|_X^2 \\ &\leq CC' \sum_{k=0}^{\infty} \Big\| (Q_k - Q_{k-1})v \Big\|_X^2,\end{aligned}$$

mit $C \geq \|Q_j^*\|$ und C' aus (3.7). Nach Kürzen von $(\sum_{j=0}^{\infty} \|(Q_j - Q_{j-1})v\|_X^2)^{1/2}$ folgt damit die Behauptung (3.8). □

Bemerkung 3.8. *Satz 3.6 und Lemma 3.7 übertragen die "levelweise" Stabilität*

$$c_j \Big(\sum_{|\lambda|=j} |c_\lambda|^2 \Big)^{1/2} \leq \Big\| \sum_{|\lambda|=j} c_\lambda \psi_\lambda \Big\|_X \leq C_j \Big(\sum_{|\lambda|=j} |c_\lambda|^2 \Big)^{1/2} \tag{3.9}$$

der Teilsysteme Ψ_j in den Komplementräumen $W_j := \operatorname{Im}(Q_j - Q_{j-1})$ auf die Stabilität des Gesamtsystems Ψ, sofern die Stabilitätskonstanten $c_j, C_j > 0$ nicht von j abhängen.

Zum Nachweis der Riesz-Basis-Eigenschaft von Ψ in X ist dann nur noch zu zeigen, dass $\operatorname{span}\Psi_j$ *in W_j dicht ist.*

Für die konkrete Anwendung auf die Diskretisierung elliptischer Variationsprobleme ist es besonders wichtig, den Fall von Riesz-Basen für Sobolevräume $H^s(\Omega)$ (oder Teilräume davon) abzudecken. Hierbei sind die Bedingungen aus Satz 3.6 allerdings noch etwas unhandlich. In der Arbeit [6] wurden in mehreren Schritten vereinfachende Bedingungen hergeleitet, die wir hier zusammenfassend angeben (ohne Beweis):

Satz 3.9. *Für ein beschränktes Lipschitzgebiet $\Omega \subset \mathbb{R}^n$ und $s > 0$ sei $H \subset H^s(\Omega)$ ein abgeschlossener Teilraum. Weiter seien folgende Bedingungen erfüllt:*

(i) *$(V_j)_{j\geq 0} \subset H$ sei eine aufsteigende Folge dichter, $L_2(\Omega)$-abgeschlossener Räume mit gleichmäßig $L_2(\Omega)$-beschränkten Projektoren $Q_j : X \to V_j$ und $Q_jQ_k = Q_j$ für $k \geq j$.*

(ii) *Für ein $m > s$ gelte die Jackson-Ungleichung*

$$\inf_{v\in V_j} \|u - v\|_{L_2(\Omega)} \leq 2^{-jm}|u|_{H^m(\Omega)}, \quad \text{für alle } u \in H \cap H^m(\Omega), \tag{3.10}$$

und die Bernstein-Ungleichung

$$\|v\|_{H^m(\Omega)} \leq C2^{jm}\|v\|_{L_2(\Omega)}, \quad \text{für alle } u \in H \cap H^m(\Omega). \tag{3.11}$$

(iii) *Mit den dualen Räumen $\tilde{V}_j := \operatorname{Im}(Q_j^*)$ gelte für ein $\delta > 0$ die Jackson-Ungleichung*

$$\inf_{v\in \tilde{V}_j} \|u - v\|_{L_2(\Omega)} \leq 2^{-j\delta}|u|_{H^\delta(\Omega)}, \quad \text{für alle } u \in H^\delta(\Omega), \tag{3.12}$$

und die Bernstein-Ungleichung

$$\|v\|_{H^\delta(\Omega)} \leq C2^{j\delta}\|v\|_{L_2(\Omega)}, \quad \text{für alle } u \in H^\delta(\Omega). \tag{3.13}$$

Dann gilt

$$\|v\|_{H^s(\Omega)} \eqsim \Big(\sum_{j=0}^{\infty} 2^{2js}\|(Q_j - Q_{j-1})v\|_{L_2(\Omega)}\Big)^{1/2}, \quad \text{für alle } v \in H. \tag{3.14}$$

3.2 Spline-Wavelets: ein Beispiel

Als konkretes Beispiel soll im Folgenden eine Riesz-Basis vom Wavelet-Typ für den Raum $H_0^1(0,1)$ angegeben werden, die aus stückweise linearen, stetigen Splines besteht. Die hinreichenden Bedingungen aus Satz 3.9 werden dabei im Laufe der Konstruktion sukzessive sichergestellt. Einzelheiten lassen sich z.B. der Arbeit [8] entnehmen.

Schritt 1: Konstruktion einer Generatorbasis von V_j

Im Allgemeinen legt man sich zuerst auf die Folge der *primalen* Approximationsräume V_j fest, $j \geq 0$. Hier betrachten wir konkret die Räume stetiger linearer Splines mit Nullrandbedingungen

$$V_j := \left\{ v \in C[0,1] \,\middle|\, v|_{2^{-j}[k,k+1)} \in P_1, 0 \leq k < 2^j, v(0) = v(1) = 0 \right\} \subset H_0^1(0,1). \quad (3.15)$$

Dieser Splineraum mit dyadischen Knoten (d.h. bei $2^{-j}k$) wird offensichtlich aufgespannt von den skalierten Hutfunktionen

$$\varphi_{j,k}(x) := 2^{j/2} N(2^j x - k), \quad k \in \mathcal{I}_j := \{1, \ldots, 2^j - 1\}, \quad N(x) := \max\left\{0, 1 - |x|\right\}, \quad (3.16)$$

die wir auch als *Generatoren* bezeichnen. Fasst man die Generatoren von V_j zu einem Vektor $\Phi_j := \{\varphi_{j,k} \,|\, k \in \mathcal{I}_j\}$ zusammen, so gilt wegen $V_j \subset V_{j+1}$ eine *Verfeinerungsrelation*

$$\Phi_j = \mathbf{M}_{j,0}^\top \Phi_{j+1}, \quad \mathbf{M}_{j,0} = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{1}{2} & & & & & & \\ 1 & & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & & \\ & 1 & & & & & \\ & \frac{1}{2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{2} & & & \\ & & & 1 & & & \\ & & & \frac{1}{2} & \frac{1}{2} & & \\ & & & & 1 & & \\ & & & & \frac{1}{2} & & \end{pmatrix} \in \mathbb{R}^{(2^{j+1}-1)\times(2^j-1)}. \quad (3.17)$$

In der Praxis betrachtet man nur hinreichend feine Räume V_j mit $j \geq j_0 := 3$, denn ab dieser Feinheit sind die Träger aller zu konstruierenden Basisfunktionen im Intervall $[0,1]$ enthalten. Für die theoretischen Überlegungen aus dem vorigen Abschnitt macht die Beschränkung auf ein beliebiges Startlevel j_0 keinen Unterschied.

Wichtige Eigenschaften der Spline-Generatoren $\varphi_{j,k}$ sind:

- Lokalität, $\operatorname{diam} \operatorname{supp} \varphi_{j,k} \sim 2^{-j}$
- lokale Endlichkeit, $\#\{k \in \mathcal{I}_j \,|\, x \in \operatorname{supp} \varphi_{j,k}\} \leq C$ unabhängig von $x \in [0,1]$ und j
- lokale Polynomreproduktion, $\forall p \in P_1 \exists v \in V_j$ mit $v = p$ auf $[2^{-j}, 1 - 2^{-j}]$
- Regularität, $\varphi_{j,k} \in H^s(0,1)$ für $0 < s < m := \frac{3}{2}$

Die ersten drei Eigenschaften sind Grundvoraussetzungen für den Nachweis der Jackson-Ungleichung (3.10), die Regularitätseigenschaft wird für die Bernstein-Ungleichung (3.11) der primalen Räume V_j benötigt.

Schritt 2: Konstruktion einer dualen Generatorbasis von $\tilde{V}_j$

Im zweiten Schritt legt man sich auf die dualen Approximationsräume $\tilde{V}_j \subset L_2(0,1)$ fest und konstruiert Projektoren $Q_j : L_2(0,1) \to V_j$ mit $\mathrm{Im}(Q_j^*) = \tilde{V}_j$. Dies geschieht mit Hilfe sogenannter *dualer Generatoren* $\tilde{\varphi}_{j,k}$, so dass einerseits $\tilde{V}_j = \mathrm{span}\{\tilde{\varphi}_{j,k} \,|\, k \in \mathcal{I}_j\}$, andererseits aber auch die *Biorthogonalitätsbedingung*

$$\langle \varphi_{j,k}, \tilde{\varphi}_{j,k'} \rangle_{L_2(0,1)} = \delta_{k,k'} \tag{3.18}$$

erfüllt ist. Aus (3.18) folgt, dass die Operatoren

$$Q_j f := \sum_{k \in \mathcal{I}_j} \langle f, \tilde{\varphi}_{j,k} \rangle_{L_2(0,1)} \varphi_{j,k} \tag{3.19}$$

genau die Orthogonalprojektoren von $L_2(0,1)$ auf V_j sind, d.h. $\langle f - Q_j f, g \rangle_{L_2(0,1)} = 0$ für alle $g \in L_2(0,1)$. Der adjungierte Operator

$$Q_j^* f = \sum_{k \in \mathcal{I}_j} \langle f, \varphi_{j,k} \rangle_{L_2(\Omega)} \tilde{\varphi}_{j,k} \tag{3.20}$$

projiziert dann offenbar auf $\tilde{V}_j = \mathrm{Im}(Q_j^*)$. Weiter folgt unter der Biorthogonalitätsbedingung bereits die Jackson-Ungleichung (3.10) für $m = 2$ (vgl. Übungsaufgabe zur Haar-Basis) sowie die Bernstein-Ungleichung (3.11) für alle $m < \frac{3}{2}$.

Um analog auch für die dualen Approximationsräume $\tilde{V}_j$ die entsprechenden Jackson- und Bernstein-Ungleichungen (3.12) bzw. (3.13) nachzuweisen, sind folgende weitere Bedingungen an die dualen Generatoren notwendig:

- Lokalität, $\operatorname{diam} \operatorname{supp} \tilde{\varphi}_{j,k} \sim 2^{-j}$
- lokale Endlichkeit, $\#\{k \in \mathcal{I}_j \,|\, x \in \operatorname{supp} \tilde{\varphi}_{j,k}\} \leq C$ unabhängig von $x \in [0,1]$ und j
- Reproduktion von Konstanten, $\forall p \in P_0 \exists v \in \tilde{V}_j$ mit $v = p$ auf $[0,1]$
- Regularität, $\tilde{\varphi}_{j,k} \in H^s(0,1)$ für $0 < s < \delta$

In [5, 8] wurde eine solche duale Generatorbasis konstruiert, die sogar alle Polynome vom Grad 1 auf $[0,1]$ reproduzieren kann und im Raum $H^{0.44}(0,1)$ liegt. Zudem sind

auch die dualen Generatoren ebenfalls verfeinerbar, für $\tilde{\Phi}_j := \{\tilde{\varphi}_{j,k} \,|k \in \mathcal{I}_j\}$ gilt konkret

$$\tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^{\top}\tilde{\Phi}_{j+1}, \quad \tilde{\mathbf{M}}_{j,0} = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{3}{2} & -\frac{1}{2} & & & & & \\ 1 & 0 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & & \\ -\frac{1}{4} & \frac{3}{2} & -\frac{1}{4} & & & & \\ & \frac{1}{2} & \frac{1}{2} & & & & \\ & -\frac{1}{4} & \frac{3}{2} & & & & \\ & & \frac{1}{2} & & & & \\ & & -\frac{1}{4} & \ddots & & & \\ & & & \ddots & -\frac{1}{4} & & \\ & & & & \frac{1}{2} & & \\ & & & & \frac{3}{2} & -\frac{1}{4} & \\ & & & & \frac{1}{2} & \frac{1}{2} & \\ & & & & 0 & 1 & \\ & & & & -\frac{1}{2} & \frac{3}{2} & \end{pmatrix}. \tag{3.21}$$

Die Skala der dualen Approximationsräume $(\tilde{V}_j)_{j\geq j_0}$ ist somit ebenfalls geschachtelt. Man beachte, dass nur die Größe der Verfeinerungsmatrizen $\mathbf{M}_{j,0}$ und $\tilde{\mathbf{M}}_{j,0}$ von der Skala j abhängt, nicht aber deren Einträge. Matrizen mit dieser Eigenschaft nennt man auch *quasi-stationär*.

Durch die Gültigkeit der entsprechenden Jackson- und Bernsteinungleichungen sowie der gleichmäßigen Beschränktheit der Projektoren Q_j ergibt sich mit Satz 3.9 die gesuchte Charakterisierung für $H_0^1(0,1)$,

$$\|v\|_{H^1(0,1)} \eqsim \Big(\sum_{j=j_0}^{\infty} 2^{2js} \big\|(Q_j - Q_{j-1})v\big\|_{L_2(0,1)}^2 \Big)^{1/2}, \quad \text{für alle } v \in H_0^1(0,1). \tag{3.22}$$

3. Schritt: Konstruktion der Wavelet-Basis

Um mit Hilfe von (3.22) eine Riesz-Basis für den Gesamtraum $H_0^1(0,1)$ zu konstruieren, müssen wir nur jeweils eine L_2-stabile Basis $\Psi_j := \{\psi_{j,k} \,|\, k \in \mathcal{J}_j\}$ für die Komplementräume $W_j = \mathrm{Im}(Q_{j+1} - Q_j)$ angeben. Dies geschieht in der Praxis überwiegend so, dass gleichzeitig auch eine duale Basis $\tilde{\Psi}_j := \{\tilde{\psi}_{j,k} \,|\, k \in \mathcal{J}_j\}$ für $\tilde{W}_j = \mathrm{Im}(Q_{j+1}^* - Q_j^*)$ mitkonstruiert wird, d.h. die duale Riesz-Basis besitzt später ebenfalls eine Multiskalenstruktur. In der Praxis kommt zur Waveletkonstruktion auf beschränkten Gebieten häufig die *Methode der stabilen Vervollständigung* aus [1] zum Einsatz. Wir geben hier nur ein Endergebnis dieses Algorithmus an, für die konkrete Situation linearer Splines $\varphi_{j,k}$ und mit den oben gewählten dualen Generatoren $\tilde{\varphi}_{j,k}$. Die primären Wavelets erge-

ben sich wie die Generatoren aus einer quasi-stationären Relation

$$\Psi_j = \mathbf{M}_{j,1}^\top \Phi_{j+1}, \quad \mathbf{M}_{j,1} = \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{5}{8} & -\frac{1}{8} & & & & & \\ -\frac{3}{4} & -\frac{1}{4} & & & & & \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{8} & & & & \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & & & & \\ \frac{1}{8} & -\frac{1}{8} & \frac{3}{4} & & & & \\ & & -\frac{1}{4} & & & & \\ & & -\frac{1}{8} & \ddots & & & \\ & & & \ddots & -\frac{1}{8} & \frac{1}{8} & \\ & & & & -\frac{1}{4} & \frac{1}{4} & \\ & & & & \frac{3}{4} & -\frac{1}{4} & \\ & & & & -\frac{1}{4} & -\frac{3}{4} & \\ & & & & -\frac{1}{8} & \frac{5}{8} & \end{pmatrix}. \tag{3.23}$$

Mit Hilfe der Quasi-Stationarität von $\mathbf{M}_{j,1}$ kann man zeigen, dass die Wavelet-Basis Ψ_j in W_j gleichmäßig L_2-stabil ist,

$$c\Big(\sum_{k\in\mathcal{J}_j} |c_{j,k}|^2\Big)^{1/2} \le \Big\| \sum_{k\in\mathcal{J}_j} c_{j,k}\psi_{j,k}\Big\|_{L_2(0,1)} \le C\Big(\sum_{k\in\mathcal{J}_j} |c_{j,k}|^2\Big)^{1/2}. \tag{3.24}$$

Die Konstanten $c, C > 0$ hängen dabei nicht vom Level j ab. Setzt man (3.24) in (3.22) ein und benutzt ein analoges Stabilitätsresultat für Φ_{j_0} in V_{j_0}, so erhält man die Riesz-basiseigenschaft für $H_0^1(0,1)$,

$$\|v\|^2_{H^1(0,1)} \eqsim \sum_{k\in\mathcal{I}_{j_0}} \big|\langle v, \tilde\varphi_{j_0,k}\rangle_{L_2(0,1)}\big|^2 + \sum_{j=j_0}^{\infty} 2^{2j} \sum_{k\in\mathcal{J}_j} \big|\langle v, \tilde\psi_{j,k}\rangle_{L_2(0,1)}\big|^2, \quad v \in H_0^1(0,1). \tag{3.25}$$

Diese Normäquivalenz ist in ihrer Form typisch für Wavelet-Systeme: eine Glattheitsnorm (hier eine L_2-Sobolevnorm) ist äquivalent zu einer speziell gewichteten Folgennorm der Entwicklungskoeffizienten.

Eine weitere wichtige Eigenschaft von Wavelets, speziell auch ihre geometrische Form, folgt aus der Reproduktion von Polynomen durch die dualen Generatoren $\tilde\varphi_{j,k}$. Wegen $P_1 \subset \tilde V_j$ und der Biorthogonalitätsrelation (3.18) folgt, dass jedes Wavelet $\psi_{j,k}$ zwei *verschwindende Momente* besitzt, d.h. es gilt $\langle \psi_{j,k}, p\rangle_{L_2(0,1)} = 0$ für alle $p \in P_1$.

3.3 Diskretisierung elliptischer Variationsprobleme mit Riesz-Basen

Nach der Einführung von Riesz-Basen, insbesondere vom Wavelet-Typ, wollen wir nun elliptische Variationsprobleme der Form (1.40) diskretisieren. Zunächst überlegen wir uns, dass es sich bei hierbei um eine lineare Operatorgleichung $Au = F$ handelt.

Satz 3.10. *Sei X ein Hilbertraum und sei $a : X \times X \to \mathbb{R}$ eine stetige, elliptische Bilinearform. Dann ist durch $(Av)(w) = a(v, w)$ ein linearer, stetig invertierbarer Operator $A : X \to X'$ festgelegt, es gilt $C_e\|v\|_X \le \|Av\|_{X'} \le C_a\|v\|_X$ für alle $v \in X$.*

Beweis: Es ist offenbar $Av \in X'$ wegen $|(Av)(w)| \le C_a\|v\|_X\|w\|_X$, denn hieraus folgt direkt $\|Av\|_{X'} \le C_a\|v\|_X$. Umgekehrt ist $\|Av\|_{X'} = \sup_{0\neq w\in X} \frac{|a(v,w)|}{\|w\|_X} \ge C_e\|v\|_X$, wenn man im Supremum $w = v$ wählt. □

Mit Hilfe einer Riesz-Basis von X lässt sich die Operatorgleichung $Au = F$ und damit das Variationsproblem äquivalent zu einem unendlich-dimensionalen Gleichungssystem umschreiben.

Satz 3.11. *Sei $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}}$ eine Riesz-Basis des Hilbertraums X, $a : X \times X \to \mathbb{R}$ eine stetige, elliptische Bilinearform und sei $F \in X'$. Dann ist das Variationsproblem* (1.40) *äquivalent zum unendlich-dimensionalen Gleichungssystem $\mathbf{Au} = \mathbf{F}$, wobei*

$$\mathbf{A} = (a(\psi_\mu, \psi_\lambda))_{\lambda,\mu\in\mathcal{I}}, \quad \mathbf{F} = (F(\psi_\lambda))_{\lambda\in\mathcal{I}}. \tag{3.26}$$

Der Operator $\mathbf{A} : \ell_2(\mathcal{I}) \to \ell_2(\mathcal{I})$ ist beschränkt und beschränkt invertierbar mit

$$c_1\|\mathbf{v}\|_{\ell_2(\mathcal{I})} \le \|\mathbf{Av}\|_{\ell_2(\mathcal{I})} \le c_2\|\mathbf{v}\|_{\ell_2(\mathcal{I})}, \quad \textit{für alle } \mathbf{v} \in \ell_2(\mathcal{I}), \tag{3.27}$$

wobei $c_1 = C_e c_R^2$, $c_2 = C_a C_R^2$. Der Vektor $\mathbf{F}$ ist enthalten in $\ell_2(\mathcal{I})$ mit $\|\mathbf{F}\|_{\ell_2(\mathcal{I})} \le C_R\|F\|_{X'}$.

Beweis: Nach Definition einer Riesz-Basis ist $\operatorname{span}\Psi$ dicht in X. Daher ist das Variationsproblem (1.40) äquivalent zu den abzählbar vielen Bedingungsgleichungen

$$a(u, \psi_\lambda) = F(\psi_\lambda), \quad \text{für alle } \lambda \in \mathcal{I}. \tag{3.28}$$

Da jedes $u \in X$ bezüglich der Riesz-Basis Ψ eine eindeutige Darstellung mit einem Koeffizientenvektor $\mathbf{u} \in \ell_2(\mathcal{I})$ besitzt, $u = \sum_{\lambda\in\mathcal{I}} u_\lambda\psi_\lambda$, ist (3.28) aufgrund der Linearität von a äquivalent zu

$$\sum_{\mu\in\mathcal{I}} u_\mu a(\psi_\mu, \psi_\lambda) = F(\psi_\lambda), \quad \text{für alle } \lambda \in \mathcal{I}.$$

Dies ist ein lineares Gleichungssystem der Form $\mathbf{Au} = \mathbf{F}$, mit der Systemmatrix $\mathbf{A}$ und der rechten Seite $\mathbf{F}$ aus (3.26). Es gilt

$$\begin{aligned}
\|\mathbf{Av}\|_{\ell_2(\mathcal{I})} &= \sup_{\mathbf{0}\neq\mathbf{w}\in\ell_2(\mathcal{I})} \frac{\big|\langle\mathbf{Av}, \mathbf{w}\rangle_{\ell_2(\mathcal{I})}\big|}{\|\mathbf{w}\|_{\ell_2(\mathcal{I})}} \\
&= \sup_{\mathbf{0}\neq\mathbf{w}\in\ell_2(\mathcal{I})} \frac{\Big|\sum\limits_{\mu,\lambda} v_\mu w_\lambda a(\psi_\mu, \psi_\lambda)\Big|}{\|\mathbf{w}\|_{\ell_2(\mathcal{I})}} \\
&= \sup_{\mathbf{0}\neq\mathbf{w}\in\ell_2(\mathcal{I})} \frac{\Big|a\Big(\sum\limits_{\mu} v_\mu\psi_\mu, \sum\limits_{\lambda} w_\lambda\psi_\lambda\Big)\Big|}{\|\mathbf{w}\|_{\ell_2(\mathcal{I})}}
\end{aligned}$$

und somit einerseits

$$\|\mathbf{A}\mathbf{v}\|_{\ell_2(\mathcal{I})} \le C_a \sup_{\mathbf{0}\neq\mathbf{w}\in\ell_2(\mathcal{I})} \frac{\Big\|\sum_{\lambda\in\mathcal{I}} w_\lambda\psi_\lambda\Big\|_X}{\|\mathbf{w}\|_{\ell_2(\mathcal{I})}} \Big\|\sum_{\mu\in\mathcal{I}} v_\mu\psi_\mu\Big\|_X \le C_a C_R^2 \|\mathbf{v}\|_{\ell_2(\mathcal{I})},$$

andererseits

$$\|\mathbf{A}\mathbf{v}\|_{\ell_2(\mathcal{I})} \ge \frac{\Big|a\Big(\sum_\mu v_\mu\psi_\mu, \sum_\lambda v_\lambda\psi_\lambda\Big)\Big|}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}} \ge C_e \frac{\Big\|\sum_{\mu\in\mathcal{I}} v_\lambda\psi_\lambda\Big\|_X^2}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}} \ge C_e c_R^2 \|\mathbf{v}\|_{\ell_2(\mathcal{I})}.$$

Für den Vektor $\mathbf{F}$ rechnet man

$$\|\mathbf{F}\|_{\ell_2(\mathcal{I})} = \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\mathcal{I})} \frac{\big|\langle\mathbf{F},\mathbf{v}\rangle_{\ell_2(\mathcal{I})}\big|}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}} = \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\mathcal{I})} \frac{\Big|F\Big(\sum_{\lambda\in\mathcal{I}} v_\lambda\psi_\lambda\Big)\Big|}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}} \le C_R\|F\|_{X'}.$$

□

Aufgrund der Elliptizität der Bilinearform a ist die Systemmatrix $\mathbf{A}$ positiv definit, dies rechnet man wie in (1.50) nach:

$$\langle\mathbf{A}\mathbf{v},\mathbf{v}\rangle = \sum_{\mu,\lambda\in\mathcal{I}} v_\mu v_\lambda a(\psi_\mu,\psi_\lambda) = a\Big(\sum_{\mu\in\mathcal{I}} v_\mu\psi_\mu, \sum_{\lambda\in\mathcal{I}} v_\lambda\psi_\lambda\Big) \ge C_e\Big\|\sum_{\lambda\in\mathcal{I}} v_\lambda\Big\|_X^2 \ge C_e c_R^2 \|\mathbf{v}\|_{\ell_2(\mathcal{I})}^2. \tag{3.29}$$

Ferner ist $\mathbf{A}$ symmetrisch genau dann, wenn die Bilinearform a symmetrisch ist. Um das unendliche System $\mathbf{A}\mathbf{u} = \mathbf{F}$ numerisch zu behandeln, gibt es mehrere grundlegende Ansätze. Neben den in den folgenden Abschnitten behandelten inexakten Iterationsverfahren gibt es noch die Möglichkeit, eine endliche Indexmenge $\Lambda \subset \mathcal{I}$ auszuwählen, die Ausschnitte $\mathbf{A}_\Lambda := (a(\psi_\mu,\psi_\lambda))_{\lambda,\mu\in\Lambda}$ und $\mathbf{F}_\Lambda := (F(\psi_\lambda))_{\lambda\in\Lambda}$ zu betrachten und eine Approximation $\mathbf{u}_\Lambda$ zu berechnen mit $\mathbf{A}_\Lambda\mathbf{u}_\Lambda = \mathbf{F}_\Lambda$. Wie der folgende Satz zeigt, ist bei einer Diskretisierung mit Riesz-Basen im Vorhinein klar, dass die Konditionszahlen $\kappa(\mathbf{A}_\Lambda) = \|\mathbf{A}_\Lambda\|\|\mathbf{A}_\Lambda^{-1}\|$ der Ausschnittsmatrizen $\mathbf{A}_\Lambda$ gleichmäßig beschränkt bleiben. Bei der Methode der Finiten Elemente etwa wäre hierfür noch Zusatzaufwand notwendig, in Form gewisser Vorkonditionierungsstrategien.

Satz 3.12. *Sei $\mathbf{A}$ wie in Satz 3.11 und zu $\Lambda \subset \mathcal{I}$ sei $\mathbf{A}_\Lambda := (a(\psi_\mu,\psi_\lambda))_{\lambda,\mu\in\Lambda}$. Dann gilt $\|\mathbf{A}_\Lambda\| \le \|\mathbf{A}\|$ und $\mathbf{A}_\Lambda$ ist ebenfalls positiv definit. Umgekehrt gilt $\|\mathbf{A}_\Lambda^{-1}\| \le \|\mathbf{A}^{-1}\|$ und somit $\kappa(\mathbf{A}_\Lambda) \le \kappa(\mathbf{A})$.*

Beweis: Sei $\mathbf{P}_\Lambda : \ell_2(\Lambda) \to \ell_2(\mathcal{I})$ der Nullfortsetzungsoperator, d.h. $(\mathbf{P}_\Lambda\mathbf{v})_\lambda = v_\lambda$ für $\lambda \in \Lambda$ und ansonsten $(\mathbf{P}_\Lambda\mathbf{v})_\lambda = 0$. Dann gilt $\|\mathbf{P}_\Lambda\mathbf{v}\|_{\ell_2(\mathcal{I})} = \|\mathbf{v}\|_{\ell_2(\Lambda)}$ für alle $\mathbf{v} \in \ell_2(\Lambda)$. Da $\mathbf{A}$ positiv definit, folgt für $\mathbf{0} \neq \mathbf{v} \in \ell_2(\Lambda)$

$$0 < \langle\mathbf{A}\mathbf{P}_\Lambda\mathbf{v},\mathbf{P}_\Lambda\mathbf{v}\rangle_{\ell_2(\mathcal{I})} = \langle\mathbf{A}_\Lambda\mathbf{v},\mathbf{v}\rangle_{\ell_2(\Lambda)},$$

folglich ist $\mathbf{A}_\Lambda$ positiv definit und somit invertierbar. Es gilt weiter

$$\begin{aligned}
\|\mathbf{A}_\Lambda\| &= \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\Lambda)} \frac{\|\mathbf{A}_\Lambda\mathbf{v}\|_{\ell_2(\Lambda)}}{\|\mathbf{v}\|_{\ell_2(\Lambda)}} \\
&= \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\Lambda)} \sup_{\mathbf{0}\neq\mathbf{w}\in\ell_2(\Lambda)} \frac{\big|\langle\mathbf{A}_\Lambda\mathbf{v},\mathbf{w}\rangle_{\ell_2(\Lambda)}\big|}{\|\mathbf{v}\|_{\ell_2(\Lambda)}\|\mathbf{w}\|_{\ell_2(\Lambda)}} \\
&= \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\Lambda)} \sup_{\substack{\mathbf{0}\neq\mathbf{w}\in\ell_2(\mathcal{I}),\\ w_\lambda=0 \text{ für } \lambda\notin\Lambda}} \frac{\big|\langle\mathbf{A}\mathbf{P}_\Lambda\mathbf{v},\mathbf{w}\rangle_{\ell_2(\mathcal{I})}\big|}{\|\mathbf{P}_\Lambda\mathbf{v}\|_{\ell_2(\mathcal{I})}\|\mathbf{w}\|_{\ell_2(\mathcal{I})}} \\
&\leq \sup_{\mathbf{0}\neq\mathbf{v},\mathbf{w}\in\ell_2(\mathcal{I})} \frac{\big|\langle\mathbf{A}\mathbf{v},\mathbf{w}\rangle_{\ell_2(\mathcal{I})}\big|}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}\|\mathbf{w}\|_{\ell_2(\mathcal{I})}} = \|\mathbf{A}\|.
\end{aligned}$$

Für die Abschätzung von $\|\mathbf{A}_\Lambda^{-1}\|$ überlegen wir uns zunächst allgemein, allgemein, dass für einen positiv definiten Operator $B : Y \to Y$ auf einem Hilbertraum folgende Beziehung gilt:

$$\|B^{-1}\| = \sup_{0\neq v\in Y} \frac{\langle v,v\rangle_Y}{\langle Bv,v\rangle_Y}. \tag{3.30}$$

Denn einerseits ist

$$\|B^{-1}\| = \sup_{0\neq v\in Y} \frac{\|B^{-1}v\|_Y}{\|v\|_Y} = \sup_{0\neq v\in Y} \frac{\|v\|_Y}{\|Bv\|_Y},$$

also mit Cauchy-Schwarz

$$\sup_{0\neq v\in Y} \frac{\langle v,v\rangle_Y}{\langle Bv,v\rangle_Y} \geq \sup_{0\neq v\in Y} \frac{\|v\|_Y}{\|Bv\|_Y} = \|B^{-1}\|$$

und folglich die Abschätzung "$\leq$" in (3.30). Umgekehrt gilt

$$\|Bv\|_Y = \sup_{0\neq w\in Y} \frac{\langle Bv,w\rangle_Y}{\|w\|_Y} \geq \frac{\langle Bv,v\rangle_Y}{\|v\|_Y^2}\|v\|_Y$$

und damit aufgrund der Definitheit

$$\sup_{0\neq v\in Y} \frac{\|v\|_Y}{\|Bv\|_Y} \leq \sup_{0\neq v\in Y} \frac{\|v\|_Y^2}{\langle Bv,v\rangle_Y},$$

also "$\geq$" in (3.30). Für den positiv definiten Operator $\mathbf{A}_\Lambda$ rechnet man dann

$$\begin{aligned}
\|\mathbf{A}_\Lambda^{-1}\|^{-1} &= \inf_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\Lambda)} \frac{\langle\mathbf{A}_\Lambda\mathbf{v},\mathbf{v}\rangle_{\ell_2(\Lambda)}}{\langle\mathbf{v},\mathbf{v}\rangle_{\ell_2(\Lambda)}} \\
&= \inf_{\substack{\mathbf{0}\neq\mathbf{v}\in\ell_2(\mathcal{I}),\\ v_\lambda=0 \text{ für } \lambda\notin\Lambda}} \frac{\langle\mathbf{A}\mathbf{v},\mathbf{v}\rangle_{\ell_2(\mathcal{I})}}{\langle\mathbf{v},\mathbf{v}\rangle_{\ell_2(\mathcal{I})}} \\
&\geq \inf_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\mathcal{I})} \frac{\langle\mathbf{A}\mathbf{v},\mathbf{v}\rangle_{\ell_2(\mathcal{I})}}{\langle\mathbf{v},\mathbf{v}\rangle_{\ell_2(\mathcal{I})}} = \|\mathbf{A}^{-1}\|^{-1}.
\end{aligned}$$

□

3.4 Wavelet-Matrixkompression

Die Diskretisierung elliptischer Variationsprobleme (1.40) mit Hilfe einer Riesz-Basis vom Wavelet-Typ $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}}$ führt, wie oben gesehen, auf ein unendlich-dimensionales Gleichungssystem $\mathbf{Au} = \mathbf{F}$ im Folgenraum $\ell_2(\mathcal{I})$. Es ist unser Ziel, hierfür ein numerisches Verfahren zu entwickeln, etwa eine inexakte Variante der *Richardson-Iteration*

$$\mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} + \alpha(\mathbf{F} - \mathbf{Au}^{(n)}), \quad 0 < \alpha < \frac{2}{\|\mathbf{A}\|}. \tag{3.31}$$

Hier und im Folgenden sei $\|\mathbf{A}\| = \sup_{\mathbf{0}\neq\mathbf{v}\in\ell_2(\mathcal{I})} \frac{\|\mathbf{Av}\|_{\ell_2(\mathcal{I})}}{\|\mathbf{v}\|_{\ell_2(\mathcal{I})}}$. Offenbar ist es für eine konkrete Durchführung von (3.31) notwendig, die biinfinite Matrix $\mathbf{A}$ zumindest approximativ auf Vektoren $\mathbf{v} \in \ell_2(\mathcal{I})$ mit endlich vielen Einträgen anzuwenden. Dies ist in der Tat möglich, da durch die Wavelet-Eigenschaften die Matrixeinträge in $\mathbf{A}$ abseits der Hauptdiagonalen ein gewisses Abfallverhalten zeigen. Im Folgenden werden wir letzteres für den Spezialfall $X = H_0^1(\Omega)$ und $a(v, w) = \int_\Omega \nabla v \nabla w \, \mathrm{d}\mathbf{x}$ genauer analysieren. Derartige Kompressionsresultate lassen sich allerdings auch für Differentialoperatoren verschiedener Ordnung mit allgemeineren Koeffizienten sowie für eine große Klasse von Integraloperatoren nachweisen, siehe etwa [7, 14].

Definition 3.13. *Eine biinfinite Matrix $\mathbf{B} = (b_{\lambda,\mu})_{\lambda,\mu\in\mathcal{I}}$ heißt* s^*-komprimierbar*, falls zu jedem $J \in \mathbb{N}$ eine biinfinite Matrix $\mathbf{B}_J$ existiert, die pro Zeile und Spalte aber nur höchstens 2^J Einträge besitzt, so dass $\sum_{J\in\mathbb{N}} 2^{Js}\|\mathbf{B} - \mathbf{B}_J\| < \infty$ für alle $0 < s < s^*$ gilt.*

Für den Nachweis der Komprimierbarkeit von $\mathbf{A}$ in einem Waveletsystem wird unter anderem das folgende fundamentale Lemma benutzt.

Lemma 3.14 (Schur)**.** *Seien $\mathbf{B} = (b_{\lambda,\mu})_{\lambda,\mu\in\mathcal{I}}$ und $\omega_\lambda > 0$ für alle $\lambda \in \mathcal{I}$. Dann folgt aus*

$$\sup_{\lambda\in\mathcal{I}} \omega_\lambda^{-1} \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu \leq C, \quad \sup_{\mu\in\mathcal{I}} \omega_\mu^{-1} \sum_{\lambda\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\lambda \leq C, \tag{3.32}$$

dass $\|\mathbf{B}\| \leq C$.

Beweis: Sei $\mathbf{v} = (v_\lambda)_{\lambda\in\mathcal{I}} \in \ell_2(\mathcal{I})$. Dann gilt für jedes $\lambda \in \mathcal{I}$ mit der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}
\big|(\mathbf{Bv})_\lambda\big| &\leq \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}||v_\mu| \\
&= \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|^{1/2}\omega_\mu^{1/2}|b_{\lambda,\mu}|^{1/2}\omega_\mu^{-1/2}|v_\mu| \\
&\leq \Big(\sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu\Big)^{1/2}\Big(\sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu^{-1}|v_\mu|^2\Big)^{1/2} \\
&= \Big(\omega_\lambda^{-1}\sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu\Big)^{1/2}\Big(\omega_\lambda \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu^{-1}|v_\mu|^2\Big)^{1/2}.
\end{aligned}$$

Summiert man dies quadratisch über λ auf, erhält man

$$\begin{aligned}\|\mathbf{Bv}\|_{\ell_2(\mathcal{I})}^2 &\leq \sum_{\lambda\in\mathcal{I}} \Big(\underbrace{\omega_\lambda^{-1} \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu}_{\leq C} \Big) \Big(\omega_\lambda \sum_{\mu\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\mu^{-1}|v_\mu|^2 \Big) \\ &\leq C \sum_{\mu\in\mathcal{I}} \underbrace{\omega_\mu^{-1} \Big(\sum_{\lambda\in\mathcal{I}} |b_{\lambda,\mu}|\omega_\lambda \Big)}_{\leq C} |v_\mu|^2 \\ &\leq C^2 \|\mathbf{v}\|_{\ell_2(\mathcal{I})}^2.\end{aligned}$$

□

Wir treffen folgende Grundannahmen über die Waveletbasis Ψ:

- $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}}$ ist eine Riesz-Basis von $L_2(\Omega)$, das durchskalierte (d.h. H^1-normierte) System $\{2^{-|\lambda|}\psi_\lambda\}_{\lambda\in\mathcal{I}}$ ist eine Riesz-Basis von $H_0^1(\Omega)$, daher gilt in leichter Abweichung zum vorigen Abschnitt

$$\mathbf{A} = \big(2^{-(|\mu|+|\lambda|)} a(\psi_\mu, \psi_\lambda)\big)_{\lambda,\mu\in\mathcal{I}}; \tag{3.33}$$

- Ψ ist lokal:

$$\operatorname{diam}\operatorname{supp}\psi_\lambda \lesssim 2^{-|\lambda|}, \qquad \sup_{x\in\Omega, j\geq j_0} \#\big\{|\lambda| = j \,\big|\, B(x,2^{-j}) \cap \operatorname{supp}\psi_\lambda \neq \emptyset\big\} < \infty;$$

- Jedes ψ_λ ist stückweise polynomial der Ordnung $d \in \mathbb{N}$, $d \geq 2$ (vom Grad $d-1$);
- Es gilt $\gamma := \sup\{s | \Psi \subset H^s(\Omega)\} = d - \frac{1}{2}$;
- Für $|\lambda| > j_0$ hat jedes Wavelet ψ_λ genau $\tilde{d} \in \mathbb{N}$ verschwindende Momente, $\tilde{d} \geq d$.

Mit diesen Hilfsmitteln können wir ein im "Levelabstand" exponentielles Abfallverhalten der Matrixeinträge von $\mathbf{A}$ aus (3.33) nachweisen.

Satz 3.15. *Unter obigen Annahmen gilt für die Einträge von* $\mathbf{A}$ *für jedes* $0 < s < \gamma$ *die Abfallbedingung*

$$|\mathbf{A}_{\lambda,\mu}| = 2^{-(|\lambda|+|\mu|)} \big|a(\psi_\mu, \psi_\lambda)\big| \leq C 2^{-||\lambda|-|\mu||(s+n/2-1)}, \quad \lambda,\mu \in \mathcal{I},\ |\lambda|,|\mu| > j_0, \tag{3.34}$$

wobei $C = C(s) > 0$.

<u>*Beweis:*</u> Sei $\lambda \in \mathcal{I}$ mit $|\lambda| > j_0$, $1 \leq k \leq n$ und $p \in P_{\tilde{d}}$. Dann ist $\frac{\partial}{\partial x_k} p \in P_{\tilde{d}-1}$, so dass partielle Integration wegen $\psi_\lambda \in H_0^1(\Omega)$ und den verschwindenden Momenten liefert

$$\int_\Omega \frac{\partial \psi_\lambda}{\partial x_k} p \,\mathrm{d}\mathbf{x} = \int_{\partial\Omega} \psi_\lambda p \mathbf{n}\cdot\mathbf{e}_k \,\mathrm{d}S - \int_\Omega \psi_\lambda \frac{\partial p}{\partial x_k} \,\mathrm{d}\mathbf{x} = 0.$$

Daher hat $\frac{\partial}{\partial x_k}\psi_\lambda$ sogar $\tilde{d}+1$ verschwindende Momente. Sei dann $\mu \in \mathcal{I}$ mit $|\mu| > j_0$ und $|\mu| \geq |\lambda|$. Es folgt für alle $p \in P_{\tilde{d}}$ und für alle $\mu \in \mathcal{I}$, dass

$$\Big| \int_\Omega \frac{\partial \psi_\lambda}{\partial x_k} \frac{\partial \psi_\mu}{\partial x_k} \,\mathrm{d}\mathbf{x} \Big| = \Big| \int_\Omega \frac{\partial \psi_\lambda}{\partial x_k} \Big(\frac{\partial \psi_\mu}{\partial x_k} - p \Big) \,\mathrm{d}\mathbf{x} \Big| \leq \Big\| \frac{\partial \psi_\lambda}{\partial x_k} \Big\|_{L_1(\Omega)} \Big\| \frac{\partial \psi_\mu}{\partial x_k} - p \Big\|_{L_\infty(\operatorname{supp}\psi_\lambda)}.$$

Da $p \in P_{\tilde{d}}$ beliebig ist, können wir rechts auch zum Infimum übergehen. Für die lokale Approximation mit Polynomen P_m auf einer Menge M gilt mit $m \geq s-1$ eine sogenannte *Whitney-Abschätzung*

$$\inf_{p \in P_m} \|f - p\|_{L_p(M)} \leq C \operatorname{diam}(M)^s |f|_{W^s(L_p(M))}, \tag{3.35}$$

vergleiche auch die Herleitung der Jackson-Ungleichung in den Übungen. Anwendung von (3.35) für $M = \operatorname{supp}\psi_\mu$ liefert mit $\|\psi_\lambda\|_{L_p(\Omega)} \eqsim 2^{|\lambda| n(1/2-1/p)}$ und infolgedessen (durch die Skalierung) $|\psi_\lambda|_{W^s(L_p(\Omega))} \eqsim 2^{|\lambda|(s+n(1/2-1/p))}$ für $s < \gamma$

$$\begin{aligned}
\Big| \int_\Omega \frac{\partial \psi_\lambda}{\partial x_k} \frac{\partial \psi_\mu}{\partial x_k} \,\mathrm{d}\mathbf{x} \Big| &\lesssim 2^{|\lambda|(1-n/2)} \inf_{p \in P_{\tilde{d}}} \Big\| \frac{\partial \psi_\mu}{\partial x_k} - p \Big\|_{L_\infty(\operatorname{supp}\psi_\lambda)} \\
&\lesssim 2^{|\lambda|(1-n/2)} (2^{-|\lambda|})^{s-1} |\psi_\mu|_{W^s(L_\infty(\Omega))} \\
&\eqsim 2^{|\lambda|(1-n/2)} (2^{-|\lambda|})^{s-1} 2^{|\mu|(s+n/2)} \\
&= 2^{|\lambda|+|\mu|} 2^{-(|\lambda|-|\mu|)(s+n/2-1)}.
\end{aligned}$$

Ist umgekehrt $|\lambda| \leq |\mu|$, so folgt mit vertauschten Rollen

$$\Big| \int_\Omega \frac{\partial \psi_\lambda}{\partial x_k} \frac{\partial \psi_\mu}{\partial x_k} \,\mathrm{d}\mathbf{x} \Big| \lesssim 2^{|\lambda|+|\mu|} 2^{-(|\mu|-|\lambda|)(s+n/2-1)},$$

also in allen Fällen (3.34) nach Summation über k. □

Es folgt, dass $\mathbf{A}$ s^*-komprimierbar ist für $s^* = \frac{\gamma-1}{n}$:

Satz 3.16. *Sei die Matrix $\mathbf{A}_J$ gegeben durch die Abschneideregel*

$$(\mathbf{A}_J)_{\lambda,\mu} := \begin{cases} (\mathbf{A})_{\lambda,\mu}, & ||\lambda| - |\mu|| \leq J/n \\ 0, & \textit{sonst} \end{cases}. \tag{3.36}$$

Dann hat $\mathbf{A}_J$ bis auf einen konstanten Faktor höchstens 2^J nichttriviale Einträge pro Zeile und Spalte und für alle $s < \frac{\gamma-1}{n}$ gilt die Abschätzung $\|\mathbf{A} - \mathbf{A}_J\| \leq C 2^{-Js}$ mitr $C = C(s) > 0$.

Beweis: Betrachte einen Zeilen-/Spaltenindex $\lambda \in \mathcal{I}$ zur Skala $j = |\lambda|$. Dann ist in der entsprechenden Zeile/Spalte im Levelblock zu $j' \geq j_0$ die Anzahl der nichttrivialen Matrixeinträge von $\mathbf{A}$ höchstens

$$\#\big\{\mu \,\big|\, |\mu| = j', \operatorname{supp}\psi_\mu \cap \operatorname{supp}\psi_\lambda \neq \emptyset\big\} \leq C \begin{cases} 1, & j' < j \\ 2^{(j'-j)n}, & j' \geq j. \end{cases} = C \max\{1, 2^{(j'-j)n}\}. \tag{3.37}$$

Denn für $j' \geq j$ überlegt man sich, dass die Waveletbasis auf einer Skala im Wesentlichen durch Translation gegeben ist. Jeder Würfel der Kantenlänge 2^{-j} wird dann durch die Träger von höchstens $c2^{j'-j}n$ Translaten ψ_μ, $|\mu| = j'$, überlappt. Ist umgekehrt $j' < j$, so ist der Träger von ψ_μ, $|\mu| = j'$, so groß, dass ein Würfel der Kantenlänge 2^{-j} höchstens von den Trägern konstant vieler ψ_μ, $|\mu| = j'$, überlappt wird, wobei die Konstante nicht mehr von der Leveldifferenz $j' - j$ abhängt.

Mit (3.37) und aus der Abschneideregel (3.36) folgt, dass die Anzahl nichttrivialer Einträge in der Zeile λ der ausgedünnten Matrix $\mathbf{A}_J$ beschränkt ist durch

$$\begin{aligned}\sum_{|j'-j|\leq J/n} \max\{1, 2^{(j'-j)n}\} &= \sum_{j'=j-\lfloor J/n\rfloor}^{j-1} 1 + \sum_{j'=j}^{j+\lfloor J/n\rfloor} 2^{(j'-j)n} \\ &= \lfloor J/n\rfloor + \sum_{k=0}^{\lfloor J/n\rfloor} 2^{kn} = \lfloor J/n\rfloor + \frac{2^{(\lfloor J/n\rfloor+1)n}-1}{2^n-1} \lesssim 2^J.\end{aligned}$$

Zum Nachweis von $\|\mathbf{A} - \mathbf{A}_J\| \leq C2^{-Js}$ für alle $0 < s < \frac{\gamma-1}{n}$ benutzen wir das Schur-Lemma 3.14 für die Gewichte $w_\lambda = 2^{-|\lambda|n/2} > 0$. Es reicht also zu zeigen, dass für jedes $0 < s < \frac{\gamma-1}{n}$

$$\begin{cases} \sup_{\mu\in\mathcal{I}} 2^{|\mu|n/2} \sum_{\lambda\in\mathcal{I}} \big|(\mathbf{A} - \mathbf{A}_J)_{\lambda,\mu}\big| 2^{-|\lambda|n/2} \leq C2^{-Js}, \\ \sup_{\lambda\in\mathcal{I}} 2^{|\lambda|n/2} \sum_{\mu\in\mathcal{I}} \big|(\mathbf{A} - \mathbf{A}_J)_{\lambda,\mu}\big| 2^{-|\mu|n/2} \leq C2^{-Js}. \end{cases}$$

Wir zeigen hiervon nur die zweite Ungleichung, die erste folgt analog. Es gilt zunächst für festes $\lambda \in \mathcal{I}$ mit $j = |\lambda|$

$$2^{|\lambda|n/2} \sum_{\mu\in\mathcal{I}} \big|(\mathbf{A} - \mathbf{A}_J)_{\lambda,\mu}\big| 2^{-|\mu|n/2} = 2^{jn/2} \sum_{|j'-j|>J/n} \sum_{|\mu|=j'} 2^{-(j+j')} \big|a(\psi_\mu, \psi_\lambda)\big| 2^{-j'n/2}.$$

Für die rechte Summe machen wir die Fallunterscheidung $j' > j$ und $j' < j$. Sei zunächst $j' > j$ und $0 < s < \frac{\gamma-1}{n}$, also $1 < sn + 1 < \gamma$. Dann folgt aus (3.34) und (3.37)

$$\begin{aligned}&2^{jn/2} \sum_{|j'-j|>J/n} \sum_{|\mu|=j'} 2^{-(j+j')} \big|a(\psi_\mu, \psi_\lambda)\big| 2^{-j'n/2} \\ &\lesssim 2^{jn/2} \sum_{j'=j+\lceil J/n\rceil}^{\infty} 2^{(j'-j)n} 2^{-(j'-j)(s+1/2)n} 2^{-j'n/2} \\ &= \sum_{j'=j+\lceil J/n\rceil}^{\infty} 2^{-(j'-j)sn} = \sum_{k=\lceil J/n\rceil}^{\infty} 2^{-ksn} \lesssim 2^{-Js}.\end{aligned}$$

Die Argumentation für den umgekehrten Fall $j' < j$ verläuft analog, denn aus (3.34)

und (3.37) folgert man für $0 < s < \frac{\gamma-1}{n}$

$$2^{jn/2} \sum_{|j'-j|>J/n} \sum_{|\mu|=j'} 2^{-(j+j')} \big|a(\psi_\mu, \psi_\lambda)\big| 2^{-j'n/2} \lesssim 2^{jn/2} \sum_{j'=j_0}^{j-\lceil J/n \rceil} 2^{-(j-j')(s+1/2)n} 2^{-j'n/2}$$
$$= \sum_{j'=j_0}^{j-\lceil J/n \rceil} 2^{-(j-j')sn}$$
$$= \sum_{k=\lceil J/n \rceil}^{j-j_0} 2^{-ksn} \lesssim 2^{-Js}.$$

□

Basierend auf der s^*-Komprimierbarkeit der Systemmatrix $\mathbf{A}$ kann man einen Algorithmus **APPLY** angeben, der die Matrix-Vektor-Multiplikation $\mathbf{Av}$ für endlich getragene Vektoren $\mathbf{v}$ in endlicher Zeit approximativ ausführt.

Algorithmus 5 APPLY$[\mathbf{A}, \mathbf{v}, \epsilon] \to \mathbf{w}$

Sei $0 < s < s^*$ fest, $\mathbf{A}$ s^*-komprimierbar, $p = (s + \frac{1}{2})^{-1}$.
$N := \#\operatorname{supp} \mathbf{v}$
Sortiere die Einträge von $\mathbf{v}$ betragsweise absteigend.
Berechne $\mathbf{v}_j$ zu den jeweils 2^j betragsgrößten Einträgen, $\mathbf{v}_j := \mathbf{v}$ für $j > \log_2 N$.
Wähle $k \in \mathbb{N}_0$ so groß, dass $2^{-sk}\big(\|\mathbf{A}\| + \sum_{j=0}^{k} 2^{js}\|\mathbf{A} - \mathbf{A}_j\|\big)\|\mathbf{v}\|_{\ell_p(\mathcal{I})} \le \epsilon$.
$\mathbf{w} := \mathbf{A}_k\mathbf{v}_0 + \sum_{j=0}^{k-1} \mathbf{A}_j(\mathbf{v}_{k-j} - \mathbf{v}_{k-j-1})$

Satz 3.17. APPLY *terminiert in endlicher Zeit, es gilt* $\|\mathbf{Av} - \mathbf{w}\|_{\ell_2(\mathcal{I})} \le C\epsilon$ *mit einer Konstante* $C = C(s) > 0$.

Beweis: Da $\mathbf{v}$ ein endlich getragener Vektor ist, können die Einträge in endlicher Zeit betragsweise fallend sortiert werden (Kosten: $\mathcal{O}(N \log N)$). Für das Aufstellen der Vektoren $\mathbf{v}_j$ fallen bei optimaler Implementierung keine Kopierkosten an, da das sortierte Feld wiederverwendet werden kann.

Für den Approximationsfehler gilt zunächst die Umformung

$$\mathbf{Av} - \mathbf{w}_k = \mathbf{Av} - \Big(\mathbf{A}_k\mathbf{v}_0 + \sum_{j=0}^{k-1} \mathbf{A}_j(\mathbf{v}_{k-j} - \mathbf{v}_{k-j-1})\Big)$$
$$= \mathbf{Av} - \Big(\mathbf{A}_k\mathbf{v}_0 + \sum_{j=0}^{k-1} (\mathbf{A}_j - \mathbf{A})(\mathbf{v}_{k-j} - \mathbf{v}_{k-j-1}) + \mathbf{A} \underbrace{\sum_{j=0}^{k-1} (\mathbf{v}_{k-j} - \mathbf{v}_{k-j-1})}_{=\mathbf{v}_k - \mathbf{v}_0}\Big)$$
$$= \mathbf{A}(\mathbf{v} - \mathbf{v}_k) + (\mathbf{A} - \mathbf{A}_k)\mathbf{v}_0 + \sum_{j=0}^{k-1} (\mathbf{A} - \mathbf{A}_j)(\mathbf{v}_{k-j} - \mathbf{v}_{k-j-1}),$$

so dass wir mit der Dreiecksungleichung erhalten

$$\|\mathbf{A}\mathbf{v}-\mathbf{w}_k\|_{\ell_2(\mathcal{I})} \le \|\mathbf{A}\|\|\mathbf{v}-\mathbf{v}_k\|_{\ell_2(\mathcal{I})}+\|\mathbf{A}-\mathbf{A}_k\|\|\mathbf{v}_0\|_{\ell_2(\mathcal{I})}+\sum_{j=0}^{k-1}\|\mathbf{A}-\mathbf{A}_j\|\|\mathbf{v}_{k-j}-\mathbf{v}_{k-j-1}\|_{\ell_2(\mathcal{I})}.$$

Man kann nun zeigen, dass für die besten 2^j-Term-Approximationen $\mathbf{v}_j$ die Abschätzung $\|\mathbf{v}-\mathbf{v}_j\|_{\ell_2(\mathcal{I})} \le C2^{-js}\|\mathbf{v}\|_{\ell_p(\mathcal{I})}$ gilt, wobei $C = C(s) > 0$, siehe [9] für einen Beweis. Hieraus und aus der s^*-Komprimierbarkeit von $\mathbf{A}$ folgt

$$\begin{aligned}
&\|\mathbf{A}\mathbf{v}-\mathbf{w}_k\|_{\ell_2(\mathcal{I})}\\
&\le C\|\mathbf{A}\|2^{-ks}\|\mathbf{v}\|_{\ell_p(\mathcal{I})} + \|\mathbf{A}-\mathbf{A}_k\|\|\mathbf{v}\|_{\ell_p(\mathcal{I})} + 2C\sum_{j=0}^{k-1}\|\mathbf{A}-\mathbf{A}_j\|2^{-(k-j)s}\|\mathbf{v}\|_{\ell_p(\mathcal{I})}\\
&\le C'\Big(\|\mathbf{A}\| + \sum_{j=0}^{k}2^{js}\|\mathbf{A}-\mathbf{A}_j\|\Big)2^{-ks}\|\mathbf{v}\|_{\ell_p(\mathcal{I})},
\end{aligned}$$

mit $C' = C'(s) > 0$.

□

Bemerkung 3.18. *Es existieren Varianten von* **APPLY**, *so dass für s^*-komprimierbare Matrizen $\mathbf{A}$, $0 < s < s^*$ und $p = (s+\frac{1}{2})^{-1}$ folgende Abschätzung gilt:*

$$\#\operatorname{supp}\mathbf{w} \lesssim \epsilon^{-1/s}\|\mathbf{v}\|_{\ell_p(\mathcal{I})}^{1/s}. \tag{3.38}$$

Weiter ist die Anzahl der benötigten Rechenoperationen zur Durchführung von **APPLY** *proportional zu $\#\operatorname{supp}\mathbf{w} + \#\operatorname{supp}\mathbf{v} + 1$.*

3.5 Inexakte Iterationsverfahren

Zur adaptiven numerischen Behandlung des unendlich-dimensionalen Gleichungssystems $\mathbf{A}\mathbf{u} = \mathbf{F}$ wollen wir im Folgenden eine inexakte Variante der Richardson-Iteration (3.31) besprechen. Für die Konvergenzanalyse betrachten wir zunächst das Originalverfahren mit exakten Operatorauswertungen. Dann überträgt sich der Konvergenz-Beweis aus dem endlich-dimensionalen Fall, denn wegen

$$\mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} + \alpha(\mathbf{F}-\mathbf{A}\mathbf{u}^{(n)}) = (\mathbf{I}-\alpha\mathbf{A})\mathbf{u}^{(n)} + \alpha\mathbf{F}$$

folgt die lineare Fehlerentwicklung

$$\mathbf{u}^{(n+1)} - \mathbf{u} = (\mathbf{I}-\alpha\mathbf{A})\mathbf{u}^{(n)} + \alpha\mathbf{F} - \mathbf{u} = (\mathbf{I}-\alpha\mathbf{A})(\mathbf{u}^{(n)}-\mathbf{u}). \tag{3.39}$$

Wegen $0 < \alpha < 2/\|\mathbf{A}\| \leq 2\|\mathbf{A}^{-1}\|$ folgt aus der Symmetrie und positiven Definitheit von $\mathbf{A}$, dass für alle $\mathbf{0} \neq \mathbf{v} \in \ell_2(\mathcal{I})$

$$\frac{\langle (\mathbf{I} - \alpha\mathbf{A})\mathbf{v}, \mathbf{v}\rangle_{\ell_2(\mathcal{I})}}{\langle \mathbf{v}, \mathbf{v}\rangle_{\ell_2(\mathcal{I})}} = 1 - \alpha \underbrace{\frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v}\rangle_{\ell_2(\mathcal{I})}}{\langle \mathbf{v}, \mathbf{v}\rangle_{\ell_2(\mathcal{I})}}}_{\in \left[1/\|\mathbf{A}^{-1}\|, \|\mathbf{A}\|\right]},$$

und damit

$$q := \|\mathbf{I} - \alpha\mathbf{A}\| = \max\left\{\left|1 - \alpha\|\mathbf{A}\|\right|, \left|1 - \alpha/\|\mathbf{A}^{-1}\|\right|\right\} < 1. \tag{3.40}$$

Einsetzen in (3.39) liefert die Kontraktionseigenschaft der Iteration (3.31) auf $\ell_2(\mathcal{I})$. Mit dem Fixpunktsatz von Banach folgert man die Fehlerabschätzungen

$$\|\mathbf{u}^{(n)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} \leq \frac{q}{1-q}\|\mathbf{u}^{(n)} - \mathbf{u}^{(n-1)}\|_{\ell_2(\mathcal{I})} \leq \frac{q^n}{1-q}\|\mathbf{u}^{(1)} - \mathbf{u}^{(0)}\|_{\ell_2(\mathcal{I})}. \tag{3.41}$$

Ist die exakte Iteration nun gestört durch approximative Auswertungen von $\mathbf{F}$ und $\mathbf{A}\mathbf{u}^{(n-1)}$, so kann die Fehlerreduktion um den Faktor $0 < q < 1$ nicht mehr erreicht werden. Eine Analyse der Fehlerfortpflanzung bei der gestörten Iteration

$$\tilde{\mathbf{u}}^{(n)} = \tilde{\mathbf{u}}^{(n-1)} + \alpha(\tilde{\mathbf{F}} - \mathbf{w}) \tag{3.42}$$

mit Approximationen $\tilde{\mathbf{u}}^{(n)} \approx \mathbf{u}^{(n)}$, $\tilde{\mathbf{F}} \approx \mathbf{F}$ und $\mathbf{w} \approx \mathbf{A}\tilde{\mathbf{u}}^{(n-1)}$ ergibt

$$\begin{aligned}\tilde{\mathbf{u}}^{(n)} - \mathbf{u} &= \tilde{\mathbf{u}}^{(n-1)} + \alpha(\tilde{\mathbf{F}} - \mathbf{w}) - \mathbf{u} \\ &= (\mathbf{I} - \alpha\mathbf{A})(\tilde{\mathbf{u}}^{(n-1)} - \mathbf{u}) + \alpha(\mathbf{A}\tilde{\mathbf{u}}^{(n-1)} - \mathbf{w}) + \alpha(\tilde{\mathbf{F}} - \mathbf{F}),\end{aligned}$$

also

$$\|\tilde{\mathbf{u}}^{(n)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} \leq q\|\tilde{\mathbf{u}}^{(n-1)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} + \alpha\big(\|\mathbf{A}\tilde{\mathbf{u}}^{(n-1)} - \mathbf{w}\|_{\ell_2(\mathcal{I})} + \|\tilde{\mathbf{F}} - \mathbf{F}\|_{\ell_2(\mathcal{I})}\big). \tag{3.43}$$

Mit der gestörten Iteration (3.42) lässt sich zumindest noch eine Reduktion des Fehlers um einen Faktor $q < \tilde{q} < 1$ erreichen, sofern die Fehlerkomponenten $\|\mathbf{A}\tilde{\mathbf{u}}^{(n-1)} - \mathbf{w}\|_{\ell_2(\mathcal{I})}$ und $\|\tilde{\mathbf{F}} - \mathbf{F}\|_{\ell_2(\mathcal{I})}$ sich nach oben jeweils durch $\frac{\tilde{q}-q}{2\alpha}\|\tilde{\mathbf{u}}^{(n-1)} - \mathbf{u}\|_{\ell_2(\mathcal{I})}$ abschätzen lassen. In einer praktischen Implementierung könnte man dies wie im Algorithmus **SOLVE** realisieren, siehe Algorithmus 6.

Satz 3.19. SOLVE *terminiert mit* $\|\mathbf{u} - \mathbf{u}_\epsilon\|_{\ell_2(\mathcal{I})} \leq \epsilon$.

Beweis: Wir zeigen mit Induktion über i, dass $\|\mathbf{u}^{(i)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} \leq \epsilon_i$, die Behauptung folgt dann aus $\epsilon_i = \tilde{q}^i\epsilon_0$ und $0 < \tilde{q} < 1$.

Für $i = 0$ rechnet man $\|\mathbf{u}^{(0)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} = \|\mathbf{u}\|_{\ell_2(\mathcal{I})} = \|\mathbf{A}^{-1}\mathbf{F}\|_{\ell_2(\mathcal{I})} \leq \epsilon_0$. Angenommen, die Behauptung gelte für $i-1$. Dann rechnet man mit (3.43)

$$\begin{aligned}\|\mathbf{u}^{(i)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} &\leq q\|\tilde{\mathbf{u}}^{(i-1)} - \mathbf{u}\|_{\ell_2(\mathcal{I})} + \alpha\big(\|\mathbf{A}\mathbf{u}^{(i-1)} - \mathbf{w}\|_{\ell_2(\mathcal{I})} + \|\tilde{\mathbf{F}} - \mathbf{F}\|_{\ell_2(\mathcal{I})}\big) \\ &\leq q\epsilon_{i-1} + \alpha\Big(\frac{\tilde{q}-q}{2\alpha\tilde{q}}\epsilon_i + \frac{\tilde{q}-q}{2\alpha\tilde{q}}\epsilon_i\Big) = \epsilon_i.\end{aligned}$$

□

Algorithmus 6 SOLVE$[\mathbf{F}, \epsilon] \to \mathbf{u}_\epsilon$

wähle $0 < \alpha < 2/\|\mathbf{A}\|$
$q := \|\mathbf{I} - \alpha\mathbf{A}\|$
wähle $q < \tilde{q} < 1$
$\mathbf{u}^{(0)} := \mathbf{0}$
$\epsilon_0 := \|\mathbf{A}^{-1}\|\|\mathbf{F}\|_{\ell_2(\mathcal{I})}$
$i := 0$
while $\epsilon_i > \epsilon$ **do**
  $i := i + 1$
  $\epsilon_i := \tilde{q}\epsilon_{i-1}$
  berechne $\tilde{\mathbf{F}}$ mit $\|\tilde{\mathbf{F}} - \mathbf{F}\|_{\ell_2(\mathcal{I})} \leq \frac{\tilde{q}-q}{2\alpha\tilde{q}}\epsilon_i$
  $\mathbf{w} := \mathbf{APPLY}[\mathbf{A}, \mathbf{u}^{(i-1)}, \frac{\tilde{q}-q}{2\alpha\tilde{q}}\epsilon_i]$
  $\mathbf{u}^{(i)} := \mathbf{u}^{(i-1)} + \alpha(\tilde{\mathbf{F}} - \mathbf{w})$
end while
$\mathbf{u}_\epsilon := \mathbf{u}^{(i)}$

Bemerkung 3.20. *Es existieren Varianten von* **SOLVE***, so dass unter gewissen Annahmen für s^*-komprimierbare Matrizen $\mathbf{A}$, $0 < s < s^*$ und $p = (s + \frac{1}{2})^{-1}$ folgende Abschätzung gilt:*

$$\#\operatorname{supp}\mathbf{u}_\epsilon \lesssim \epsilon^{-1/s}\|\mathbf{u}\|_{\ell_p(\mathcal{I})}^{1/s}. \tag{3.44}$$

Weiter ist die Anzahl der benötigten Rechenoperationen zur Durchführung von **SOLVE** *proportional zu $\#\operatorname{supp}\mathbf{u}_\epsilon + 1$.*

Im Sinne der Definition 1.11 ist $\mathbf{u}_\epsilon$ damit ℓ_2-konvergent gegen $\mathbf{u}$ mit Rate s, sofern $\mathbf{u} \in \ell_p(\mathcal{I})$.

4 Optimalität adaptiver Verfahren

Neben der Konvergenzanalyse adaptiver Verfahren ist natürlich auch die Frage zu klären, inwiefern sich Adaptivität bei einem gegebenen Problem *überhaupt* lohnt. Denn bei der konkreten Umsetzung adaptiver Approximationsmethoden ist immer ein gewisser Aufwand zur Verwaltung der dynamischen Datenstrukturen in Kauf zu nehmen, der sich zum Erreichen einer vorgegebenen Genauigkeit unter Umständen gegenüber einer einfachen (z.B. uniformen, "unterteile gleichmäßig") Verfeinerungsstrategie gar nicht lohnt. Diese Frage kann durch eine genauere Analyse der Konvergenz*ordnung* eines adaptiven Verfahrens beantwortet werden, siehe Definition 1.11. Im Einzelnen ist hierbei zu klären:

- Unter welchen Bedingungen ist ein adaptives Verfahren konvergent mit der Rate $s > 0$?
- Unter welchen Bedingungen konvergieren nichtadaptive Verfahren mit der gleichen Rate? Sind diese Bedingungen stärker als jene für adaptive Verfahren und gibt es relevante Praxisbeispiele, wo diese Bedingungen nicht erfüllt sind?
- Kann ein adaptives Verfahren sogar die Rate der besten N-Term-Approximation erreichen ($\rightarrow$ Optimalität)?

Zur Klärung der Fragen sei X ein Hilbertraum, $u \in X$ sei die zu approximierende Lösung des Variationsproblems (1.40) und $\Psi = \{\psi_\lambda\}_{\lambda\in\mathcal{I}}$ sei eine Riesz-Basis von X.

4.1 Konvergenzraten nichtadaptiver Verfahren

Eine einfache nichtadaptive Approximation von $u \in X$ durch Elemente $u_j \in \operatorname{span}\Psi$ kann man erhalten, wenn die Riesz-Basis Ψ eine Multiskalenstruktur besitzt. Denn dann sind die Mengen $\Psi^j = \{\psi_\lambda \mid |\lambda| \le j\}$ in j aufsteigend und induzieren Galerkin-Projektionen $u_j \in V_j$ mit $a(u_j, v) = F(v)$ für alle $v \in V_j$.

Betrachten wir den Spezialfall, dass es sich bei X um einen abgeschlossenen Teilraum von $H^t(\Omega)$ handelt für $t \ge 0$, also etwa $X = H_0^t(\Omega)$, und dass man mit Linearkombinationen aus V_j lokal alle Polynome vom Grad $r-1$ reproduzieren kann. Dann gilt die Jackson-Ungleichung (3.10) und ihre Verallgemeinerung

$$\inf_{v\in V_j} \|u - v\|_{H^t(\Omega)} \le C 2^{-j(r-t)} |u|_{H^r(\Omega)}. \tag{4.1}$$

Mit Satz 1.29 folgt $\|u - u_j\|_{H^t(\Omega)} \lesssim 2^{-j(r-t)}|u|_{H^r(\Omega)}$. Die Konvergenzrate dieses einfachen Approximationsverfahrens $j \mapsto u_j$ kann ermittelt werden über das Verhältnis zwischen der Anzahl benutzter Freiheitsgrade und der erreichten Genauigkeit. Da ein

Element aus V_j typischerweise $\dim V_j \approx N := 2^{jn}$ aktive Koeffizienten besitzt, ergibt sich wegen $2^{-j(r-t)} = N^{-(r-t)/n}$ eine Approximationsrate von $(r-t)/n$, falls die Regularitätsbedingung $u \in H^r(\Omega)$ erfüllt ist. Mit anderem Exponenten $r = sn + t$ ausgedrückt bedeutet dies, dass die nichtadaptive Galerkin-Approximation $j \mapsto u_j$ mit Rate s in $H^t(\Omega)$ konvergiert, falls $u \in H^{sn+t}(\Omega)$.

Beispiel 4.1. *Sei $\Omega \subset \mathbb{R}^2$, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ und Ψ eine Wavelet-Basis von $L_2(\Omega)$, die lokal Polynome vom Grad 1 reproduzieren kann (lineare Spline-Wavelets). Dann konvergiert die Galerkin-Approximation u_j in der H^1-Norm gegen u mit Rate $\frac{2-1}{2} = \frac{1}{2}$.*

4.2 Konvergenzraten inexakter Iterationsverfahren

Nach Bemerkung 3.20 existieren Varianten des inexakten Iterationsverfahrens **SOLVE**, so dass $\mathbf{u}_\epsilon = \mathbf{SOLVE}[\epsilon]$ gegen $\mathbf{u}$ mit Rate s in $\ell_2(\mathcal{I})$ konvergiert, sofern $\mathbf{u} \in \ell_p(\mathcal{I})$ für $p = (s + \frac{1}{2})^{-1}$. Es stellt sich die Frage, ob man diese hinreichende Bedingung für eine Konvergenzrate s auch durch das Enthaltensein von u in gewissen Funktionenräumen ausdrücken kann.

Falls es sich bei X um einen abgeschlossenen Unterraum von $H^t(\Omega)$ handelt für $t \geq 0$, z.B. $X = H_0^t(\Omega)$, und falls Ψ wie im letzten Abschnitt eine Riesz-Basis vom Wavelet-Typ für $L_2(\Omega)$ ist, so dass das skalierte System $\{2^{-|\lambda|t}\psi_\lambda\}_{\lambda\in\mathcal{I}}$ eine Riesz-Basis für X ist, dann lassen sich diejenigen $u \in X$ mit Entwicklungskoeffizienten in $\ell_p(\mathcal{I})$ genau charakterisieren (siehe etwa [4]):

Satz 4.2. *Seien gewisse technische Voraussetzungen an das Wavelet-System Ψ erfüllt. Dann gilt für jedes $u = \sum_{\lambda\in\mathcal{I}} u_\lambda 2^{-|\lambda|t}\psi_\lambda$ mit Konvergenz der Reihe in $H^t(\Omega)$ die Normäquivalenz*

$$\|u\|_{B_p^{sn+t}(L_p(\Omega))} \eqsim \|\mathbf{u}\|_{\ell_p(\mathcal{I})}, \quad p = (s + \tfrac{1}{2})^{-1}. \tag{4.2}$$

Hierbei ist $B_p^{sn+t}(L_p(\Omega))$ ein Besovraum.

Besovräume sind Funktionenräume mit Glattheit in $L_p(\Omega)$. Der Vollständigkeit halber geben wir eine Definition an:

Definition 4.3. *Zu $h \in \mathbb{R}^n$ und $r \in \mathbb{N}_0$ sei Δ_h^r die r-te Vorwärtsdifferenz*

$$\Delta_h^0 f := f, \quad \Delta_h^1 f := f(\cdot + h) - f, \quad \Delta_h^{k+1} := \Delta_h^1 \Delta_h^k, \tag{4.3}$$

definiert für Funktionen f auf den Mengen

$$\Omega_h := \big\{x \in \Omega : x + th \in \Omega,\ t \in [0,1]\big\}, \quad h \in \mathbb{R}^n. \tag{4.4}$$

Sei weiter der r-te L_p-Glattheitsmodul gegeben durch

$$\omega_r(f,t)_{L_p(\Omega)} := \sup_{\|h\|\leq t} \|\Delta_h^r f\|_{L_p(\Omega_{rh})}, \quad t > 0. \tag{4.5}$$

Für $s > 0$, $r := \lfloor s \rfloor + 1$ und $0 < p, q \leq \infty$ ist dann der Besov-Raum $B_q^s(L_p(\Omega))$ definiert durch

$$B_q^s\big(L_p(\Omega)\big) := \big\{f \in L_p(\Omega) : |f|_{B_q^s(L_p(\Omega))} < \infty\big\}, \tag{4.6}$$

wobei

$$|f|_{B_q^s(L_p(\Omega))} := \begin{cases} \Big(\int_0^\infty \big(t^{-s}\omega_r(f,t)_{L_p(\Omega)}\big)^q \,\mathrm{d}t/t \Big)^{1/q} & , \ 0 < q < \infty \\ \sup\limits_{t\geq 0} t^{-s}\omega_r(f,t)_{L_p(\Omega)} & , \ q = \infty \end{cases}, \tag{4.7}$$

und $\|\cdot\|_{B_q^s(L_p(\Omega))} := \|\cdot\|_{L_p(\Omega)} + |\cdot|_{B_q^s(L_p(\Omega))}$.

Der Raum $B_q^s(L_p(\Omega))$ ist ein vollständiger, im allgemeinen quasi-normierter Raum. Von besonderer Bedeutung sind für unsere Zwecke gewisse Einbettungsresultate. Diese macht man sich häufig grafisch klar anhand eines s-$\frac{1}{p}$-Diagramms, siehe Abbildung 4.1.

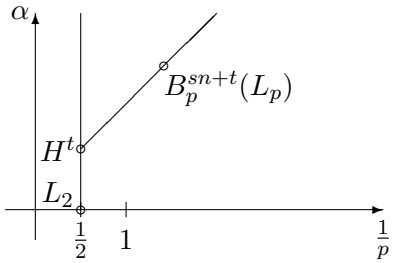

Abbildung 4.1: α-$\frac{1}{p}$-Diagramm, Einbettungslinien linearer und nichtlinearer Approximation in $H^t(\Omega)$

Jeder Punkt im Diagramm 4.1 gehört zu einem Funktionenraum der Glattheit α in L_p über Ω. Auf der Abszisse $\alpha = 0$ befinden sich die L_p-Räume, auf der Vertikalen bei $\frac{1}{p}$ liegen die L_p-Sobolevräume $W^s(L_p(\Omega))$. Ignoriert man für einen Moment den Zusatzparameter q, kann man auch die Besov-Räume $B_q^s(L_p(\Omega))$ in das Diagramm eintragen.

Die Beziehungen zwischen verschiedenen Räumen in Diagramm 4.1 kann man durch Einbettungslinien verdeutlichen. Offensichtlich gilt $H^{sn+t}(\Omega) \hookrightarrow H^t(\Omega)$ für alle $s, t \geq 0$, dies entspricht der vertikalen Einbettungslinie (von oben nach unten). Aufgrund der Beschränktheit von Ω gilt $H^s(\Omega) \hookrightarrow B_q^s(L_p(\Omega))$ für alle $p, q < 2$, dies entspricht der horizontalen Einbettung (von links nach rechts). Ferner gibt es noch die Sobolev-Einbettungslinie diagonal von rechts oben nach links unten, mit Steigung n. Die Einbettung $B_p^{sn+t}(L_p(\Omega)) \hookrightarrow H^t(\Omega)$, $p = (s + \frac{1}{2})^{-1}$, ist genau von diesem Typ, siehe die diagonale Linie in Abbildung 4.1.

Für das Enthaltensein von $\mathbf{u}$ in $\ell_p(\mathcal{I})$ und demnach die Gültigkeit einer Konvergenzrate s bei der ℓ_2-Approximation von $\mathbf{u}$ durch $\mathbf{u}_\epsilon$ ist also äquivalent das Enthaltensein von u in $B_p^{sn+t}(L_p(\Omega))$, mit $p = (s+\frac{1}{2})^{-1}$. Diese Bedingung ist wegen $H^{sn+t}(\Omega) \hookrightarrow B_p^{sn+t}(L_p(\Omega))$ aber wesentlich schwächer als zum Erreichen der gleichen Konvergenzrate durch das einfache nichtadaptive Verfahren aus Abschnitt 4.1. Man kann daher erwarten, dass unter gleichen Regularitätsbedingungen adaptive Verfahren eine höhere Konvergenzrate aufweisen als nichtadaptive. Dies ist für praxisrelevante Problemklassen in der Tat gegeben:

Beispiel 4.4. *Bei der Lösung elliptischer Variationsprobleme auf nichtglatten, nichtkonvexen Gebieten (z.B. polygonale bzw. polyhedrale Gebiete in* $\mathbb{R}^2$ *und* $\mathbb{R}^3$*) treten generisch sogenannte* Gebietssingularitäten *als additive Lösungsbestandteile auf. Diese haben häufig eine signifikant höhere Regularität in der Skala der Besov-Räume* $B_p^{sn+t}(L_p(\Omega))$

als in der Skala der Sobolevräume $H^{sn+t}(\Omega)$ (im Sinne des Laufbereichs von s). Bei der Lösung der Poisson-Gleichung mit rechter Seite $f \in C^\infty(\Omega)$ auf dem L-förmigen Gebiet im $\mathbb{R}^2$ etwa gilt nur $u \in H^{2s+1}(\Omega)$ für $s < \frac{1}{3}$, aber es gilt $u \in B_p^{2s+1}(L_p(\Omega))$, $p = (s + \frac{1}{2})^{-1}$, für $s < \frac{1}{2}$.

Algorithmenverzeichnis

Literaturverzeichnis

[1] Carnicer, J. M., W. Dahmen und J. M. Peña: *Local Decomposition of Refinable Spaces and Wavelets.* Appl. Comput. Harmon. Anal., 3:127–153, 1996.

[2] Carstensen, C. und S. A. Funken: *Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods.* East-West J. Numer. Math., 8(3):153–175, 2000.

[3] Clément, P.: *Approximation by finite element functions using local regularization.* RAIRO Sér. Rouge Anal. Numér., R-2:77–84, 1975.

[4] Cohen, A.: *Numerical analysis of wavelet methods*, Band 32 der Reihe *Studies in Mathematics and its Applications.* North-Holland, Amsterdam, 2003.

[5] Cohen, A., I. Daubechies und J.-C. Feauveau: *Biorthogonal bases of compactly supported wavelets.* Commun. Pure Appl. Math., 45:485–560, 1992.

[6] Dahmen, W.: *Stability of Multiscale Transformations.* J. Fourier Anal. Appl., 4:341–362, 1996.

[7] Dahmen, W.: *Wavelet and multiscale methods for operator equations.* Acta Numerica, 6:55–228, 1997.

[8] Dahmen, W., A. Kunoth und K. Urban: *Biorthogonal spline-wavelets on the interval — Stability and moment conditions.* Appl. Comput. Harmon. Anal., 6:132–196, 1999.

[9] DeVore, R.: *Nonlinear approximation.* Acta Numerica, 7:51–150, 1998.

[10] Evans, L.C.: *Partial differential equations*, Band 19 der Reihe *Graduate Studies in Mathematics.* American Mathematical Society (AMS), Providence, RI, 1998.

[11] Morin, P., R. H. Nochetto und K. G. Siebert: *Convergence of adaptive finite element methods.* SIAM Rev., 44(4):631–658, 2002.

[12] Schwab, C.: *p- and hp-finite element methods. Theory and applications in solid and fluid mechanics.* Clarendon Press, Oxford, 1998.

[13] Stein, E.M.: *Singular Integrals and Differentiability Properties of Functions.* Princeton University Press, Princeton, New Jersey, 1970.

[14] Stevenson, R.: *On the compressibility of operators in wavelet coordinates.* SIAM J. Math. Anal., 35(5):1110–1132, 2004.

[15] Verfürth, R.: *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Wiley-Teubner, Chichester, UK, 1996.