Deterministische und stochastische Rundungsfehleranalysen von schnellen trigonometrischen Algorithmen in Gleitkomma- bzw. Festkomma-Arithmetik

Dissertation

zur

Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Dipl. Math. Katja Ihsberner, geb. am 05.09.1979 in Rostock

aus Rostock

Rostock, 31. Januar 2011

urn:nbn:de:gbv:28-diss2011-0075-8

Dekan: Prof. Dr. Christoph Schick

1. Gutachter: Prof. Dr. Manfred Tasche (Universität Rostock)

2. Gutachter: Prof. Dr. Gerlind Plonka-Hoch (Universität Göttingen)

Tag der öffentlichen Verteidigung: 21. April 2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

Eine der Hauptaufgaben der numerischen Mathematik ist die Entwicklung von schnellen und numerisch stabilen Algorithmen. Zu den bekanntesten schnellen und numerisch stabilen Algorithmen gehört die schnelle Fourier-Transformation (FFT). Diese Methode zur Berechnung der diskreten Fourier-Transformation (DFT) wurde von Cooley und Tukey [12] im Jahre 1965 vorgeschlagen und beruht auf der Teile-und-Herrsche-Strategie. Bei einer Transformationslänge $n = 2^t$ reduziert die FFT die Anzahl der arithmetischen Operationen auf $\mathcal{O}(n \log_2 n)$. Eine naive Implementierung der DFT benötigt dagegen $\mathcal{O}(n^2)$ Operationen. Der nach Strang [52] wichtigste Algorithmus des 20. Jahrhunderts ist bereits 1805 von Gauss verwendet worden, wie aus seinen gesammelten Werken [22] hervorgeht.
Heutzutage ist die Fourieranalysis ein hoch entwickelter Zweig der Mathematik mit vielfältigen Anwendungen [31]. Bei zahlreichen Aufgabenstellungen sind jedoch reelle Datenvektoren gegeben. Um die komplexe Arithmetik bei der DFT zu umgehen, werden beispielsweise die *diskreten Kosinus-Transformationen* (DCT) und die *diskreten Sinus-Transformationen* (DST) angewandt. Eine herausragende Rolle spielen dabei die DCT, deren Eigenschaften sich für die Datenkompression als besonders geeignet herausgestellt haben [2]. Aufgrund ihrer Linearität, leichten Invertierbarkeit und Nähe zur DFT sind die DCT und DST ebenso in der digitalen Signalverarbeitung wichtige Werkzeuge beim Entwurf von effizienten Filterbanken [15, S. 203f und S. 559f].
Für die DCT und DST lassen sich reelle, schnelle und numerisch stabile Algorithmen angeben. Diese beruhen ebenfalls auf der Teile-und-Herrsche-Strategie. Wie bei der DFT gibt es für die DCT und DST verschiedene Möglichkeiten einer schnellen Realisierung. Im Vordergrund stehen hier Algorithmen, welche auf Matrixfaktorisierungen

$$A \;=\; \prod_{m=1}^{\nu} A^{(m)} \;:=\; A^{(\nu)} \dots A^{(2)} A^{(1)}$$

der entsprechenden orthogonalen Transformationsmatrix A basieren, deren Faktoren $A^{(m)}$ *dünnbesetzt* und orthogonal sind. Dabei bedeutet hier die Dünnbesetztheit von $A^{(m)}$, dass in jeder Zeile und in jeder Spalte maximal zwei Elemente ungleich Null stehen.
Neben der Reduzierung der Anzahl der Rechenoperationen haben die Dünnbesetztheit und die Orthogonalität einen positiven Einfluss auf die Genauigkeit der Rechenergebnisse. Darüber hinaus lassen sich orthogonale Matrizen denkbar einfach invertieren, so dass Transponieren der entsprechenden Faktorisierung sofort einen Algorithmus für die inverse Transformation liefert. Ist die Transformationsmatrix A sowohl orthogonal als auch symmetrisch, wie es bei der Sinus- und Kosinusmatrix vom Typ IV der Fall ist, dann sind Hin- und Rücktransformation mit demselben Algorithmus möglich. Aufgrund dieser Eigenschaft ist die DCT-IV innerhalb der digitalen Signalverarbeitung besonders attraktiv [15, 57].
Die Effizienz eines schnellen Algorithmus lässt sich im Vergleich zu seinen Stabilitätseigenschaften wesentlich einfacher vorhersagen [24, S. 438]. Letztere erfordert eine sorgfältige Analyse. Erste systematische Untersuchungen zum Rundungsfehler-Verhalten eines Computers gehen auf die wegweisenden Arbeiten von J. von Neumann und Goldstine [41] (bei der Festkomma-Arithmetik) sowie von Wilkinson [66, 67] (bei der Gleitkomma-Arithmetik) zurück. Seitdem ist vor allem die Rundungsfehleranalyse in der Gleitkomma-Arithmetik weiterentwickelt worden. Dies liegt insbesondere daran, dass alle gängigen Computer die Gleitkomma-Arithmetik verwenden, wie wir dem Standardwerk [24] von Higham entnehmen können. Untersuchungen von Rundungsfehlern schneller Algorithmen in Festkomma-Arithmetik sind dagegen kaum durchgeführt worden.
Aufgrund der rasanten Entwicklung digitaler Technologien innerhalb der letzten Jahre, deren Auswirkungen in nahezu allen Lebensbereichen zu spüren sind, ist der Bedarf an digitalen Signalprozessoren enorm gestiegen. Die digitale Signalverarbeitung spielt derzeit eine Schlüsselrolle für die drahtlose Kommunikation, Audio- und Videoverarbeitung und bei industriellen Steuerungsaufgaben. Um die Effizienz zu steigern, werden daher viele digitale Signalprozessoren anwendungsspezifisch konzipiert.

Nicht selten wird dabei auf die im Vergleich zur konventionellen Gleitkomma-Arithmetik *einfachere* und bezüglich der Hardware *weniger kostenintensive* Festkomma-Arithmetik zurückgegriffen [15, S. 669]. Somit wächst zunehmend das Interesse an deterministischen bzw. stochastischen Rundungsfehleranalysen in Festkomma-Arithmetik. Bei der deterministischen Rundungsfehleranalyse wird der Rundungsfehler im ungünstigsten Fall (worst-case) bei jedem Schritt nach oben abgeschätzt. Die auf diese Weise ermittelten oberen Schranken sind im Allgemeinen jedoch viel zu pessimistisch. Daher wird im Rahmen einer stochastischen Rundungsfehleranalyse zusätzlich das durchschnittliche Verhalten (average-case) des Rundungsfehlers abgeschätzt. Entsprechende Untersuchungen existieren bisher ausschließlich für die FFT [9, 34, 64, 65]. Lediglich bei Yun und Lee [72, 73] finden sich Erwartungswert und Varianz der Komponenten des bei ausgewählten schnellen DCT-Algorithmen entstehenden Fehlervektors, wobei dort die stark vereinfachenden Annahmen verwendet werden, dass sämtliche Matrix-Einträge exakt und die Eingangsdaten unabhängig identisch verteilt sind. Daher hat sich diese Arbeit zum Ziel gesetzt, diese Lücke für die in [43, 50, 54] vorgestellten schnellen DCT- und DST-Algorithmen zu schließen.

In dieser Arbeit wird eine umfassende und einheitliche Stabilitätsanalyse sowohl in Festkomma- als auch in Gleitkomma-Arithmetik für eine Klasse von schnellen DCT- und DST-Algorithmen durchgeführt, welche auf Faktorisierungen in dünnbesetzte orthogonale Matrizen beruhen. Dabei betrachten wir nur Transformationslängen $n = 2^t$ $(t \in \mathbb{N})$. Insbesondere verwenden sämtliche Untersuchungen, dass sich alle beteiligten Matrixfaktoren durch Permutationen sowie durch zeilen- und spaltenweise Vorzeichenskalierungen auf eine einheitliche Blockdiagonalgestalt zurückführen lassen. Aus diesem Grund sind einheitliche Ergebnisse wie in den Sätzen 3.9 und 3.16 bzw. wie in den Sätzen 4.8 – 4.10 bei Analysen der numerischen Stabilität überhaupt erst möglich. Dabei hängen die entsprechenden Stabilitätskonstanten (gegebenenfalls näherungsweise) linear von der Anzahl der beteiligten Matrixfaktoren ab. Auf diese Weise lassen sich die in Kapitel 1 vorgestellten schnellen Algorithmen 1.15 – 1.22 bezüglich ihrer numerischen Stabilität in drei Gruppen einteilen.
Neben Untersuchungen für den ungünstigsten Fall wird sowohl in Gleitkomma- als auch in Festkomma-Arithmetik eine stochastische Rundungsfehleranalyse durchgeführt, welche jeweils das Verhalten des durchschnittlichen relativen bzw. absoluten Rundungsfehlers abschätzt. Im Fall der Gleitkomma-Arithmetik basiert die stochastische Modellierung des relativen Rundungsfehlers auf einer Weiterentwicklung der von Zeuner [75] ursprünglich für die Multiplikation von zwei komplexen Zahlen entwickelten Idee. Neu ist hierbei, dass auf die beispielsweise in der Bildverarbeitung selten gegebene Unkorreliertheit der Eingangsdaten verzichtet werden kann und nur wenige Informationen über die Verteilung aller beteiligten Größen bekannt sein müssen. Als weitere Verallgemeinerung werden zusätzlich die Matrizeneinträge unter den sinnvollen Annahmen (3.61) und (3.61a) bzw. (3.61b) als Zufallsgrößen modelliert.
Stochastische Rundungsfehleranalysen in Festkomma-Arithmetik tauchen erstmals in [41, S. 1029ff] im Zusammenhang mit linearen Gleichungssystemen auf. Dabei werden absolute Rundungsfehler als gleichverteilte Zufallsvariablen modelliert. In ähnlicher Weise verfährt auch Henrici [23, S. 50ff] bei der Untersuchung numerischer Lösungsverfahren für gewöhnliche Differentialgleichungen. Im Gegensatz dazu benötigt das in Kapitel 4 neu vorgestellte Modell 4.16 lediglich, dass die auftretenden absoluten Fehler zentrierte Zufallsgrößen mit endlichem zweiten Moment sind. An die Eingangsdaten werden – bis auf die zur Vermeidung von Überlauf notwendige Beschränkung der euklidischen Norm gemäß Modellannahme (A0) – keine speziellen Forderungen bezüglich ihrer Verteilung gestellt, was eine bisher nicht gegebene Allgemeinheit bietet.

Die Arbeit ist in 5 Kapitel gegliedert:

In **Kapitel 1** wird eine spezielle Klasse von schnellen Algorithmen zu diskreten Kosinus- und Sinustransformationen vom Typ II – IV vorgestellt. Bei einer solchen Transformation wird für einen Vektor $\mathbf{x}$ fester Länge $n \in \mathbb{N}$ jeweils eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ mit einer quadratischen vollbesetzten, jedoch orthogonalen Matrix ausgeführt. Die den jeweiligen Transformationen zugrunde liegenden Kosinus- und Sinusmatrizen der Größe $n \times n$ werden in Abschnitt 1.1 definiert sowie einige wichtige Beziehungen zwischen ihnen nachgewiesen. Insbesondere interessiert uns der Fall $n = 2^t$ $(t \in \mathbb{N})$. Zu jeder dieser vollbesetzten Kosinus- und Sinusmatrizen werden in Abschnitt 1.2 – angefangen bei den aus [43, Lemma 2.2 und 2.4] bereits bekannten Ergebnissen in Lemma 1.6 – nacheinander Matrix-

faktorisierungen angegeben, welche als Faktoren zwei dünnbesetzte orthogonale Matrizen und eine Blockdiagonalmatrix besitzen. Darüber hinaus enthält die jeweilige Blockdiagonalmatrix als Blöcke wiederum Kosinus- bzw. Sinusmatrizen der Größe $\frac{n}{2} \times \frac{n}{2}$ – jedoch nicht notwendigerweise vom gleichen Typ. Auf diese Weise lassen sich induktiv Faktorisierungen herleiten, die nur noch aus dünnbesetzten orthogonalen Matrizen bestehen, wobei sich jeder Faktor durch Permutationen und zeilen- bzw. spaltenweise Vorzeichenskalierungen in eine Blockdiagonalmatrix mit Blöcken der Gestalt

$$Q_2(\varphi) \;:=\; \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix} \qquad \left(\varphi \in \left[0, \tfrac{\pi}{4}\right]\right)$$

überführen lässt. Die Multiplikation eines Vektors $\mathbf{x} \in \mathbb{R}^2$ mit einer Matrix $Q_2(\varphi)$ bedeutet geometrisch genau eine Drehung um den Winkel $-\varphi$, so dass es sich bei $Q_2(\varphi)$ wegen $Q_2(\varphi)^{-1} = Q_2(-\varphi) = Q_2(\varphi)^{\mathrm{T}}$ um eine orthogonale Matrix handelt.
In den Aussagen 1.11 – 1.13 werden ausgehend von den Ergebnissen von [54] bzw. [50] weitere Faktorisierungen angegeben, deren dünnbesetzte orthogonale Faktoren sich in gleicher Weise auf die oben beschriebenen Blockdiagonalmatrizen überführen lassen. Aus den in Abschnitt 1.2 hergeleiteten bzw. zitierten Faktorisierungen werden in Abschnitt 1.3 schnelle Algorithmen entworfen. Dabei bestechen insbesondere die *rekursiven Algorithmen* 1.19 – 1.22 durch ihre einfache Struktur.

Kapitel 2 bildet den Hauptteil dieser Arbeit. In Abschnitt 2.1 wird zunächst ein Modell zur Gleitkomma-Arithmetik (2.1) in Erinnerung gebracht, welches beispielsweise in [24] zu finden ist, jedoch schon von Wilkinson [67] verwendet worden ist. Innerhalb dieser Arithmetik werden vom Wilkinson-Modell (2.6) ausgehend jeweils Abschätzungen für den entstehenden relativen Rundungsfehler bei elementaren Operationen (Addition, Subtraktion und Multiplikation) und ebenso beim Skalarprodukt von Vektoren sowie bei der Matrix-Vektor-Multiplikation hergeleitet. Insbesondere interessieren dabei die Ergebnisse für Matrix-Vektor-Multiplikationen mit den Drehmatrizen $Q_2(\varphi)$. Unter Verwendung der in [43, Lemma 5.1] angegebenen Ungleichung (2.19) erhalten wir in Satz 2.12 das Hauptergebnis von Abschnitt 2.1. In Abschnitt 2.2 untersuchen wir die Möglichkeiten für ein auf (2.6) abgestimmtes stochastisches Modell, welches unter zusätzlichen Annahmen Vorhersagen für die euklidische Norm des Vektors der auftretenden relativen Rundungsfehler liefern soll. Dabei erweist sich die im Unterabschnitt 2.2.1 zunächst geforderte Annahme der stochastischen Unabhängigkeit an die Eingangsdaten als ungeeignet, da sie im Allgemeinen bereits nach einer Matrix-Vektor-Multiplikation nicht mehr gewährleistet werden kann. Unter Hinzunahme der speziellen Blockdiagonalgestalt der auftretenden Matrizen gelingt in Unterabschnitt 2.2.2 die Übertragung der Idee von Zeuner [75], welcher ein stochastisches Modell für die Multiplikation komplexer Zahlen entwickelt hat, auf den Fall von Drehmatrizen (vgl. Lemma 2.20). Nach geringfügiger Modifikation liefert dieses durch Satz 2.22 den Grundbaustein für alle nachfolgenden Ergebnisse innerhalb des stochastischen Modells zur Gleitkomma-Arithmetik. Neu ist hierbei, dass die Eingangsdaten korreliert sein können. Unter Verwendung der sinnvollen Annahmen (2.37) und (2.37a) bzw. (2.37b) sind die Hauptergebnisse aus Abschnitt 2.2 in Satz 2.26 zusammengefasst. Abschnitt 2.3 betrachtet das bereits auf J. von Neumann und Goldstine [41] zurückgehende Modell zur Festkomma-Arithmetik, welches die Vorzeichenbetragsdarstellung (2.49) (sign-magnitude representation) verwendet. Diese ist der Zweierkomplementdarstellung (2.54) aus den in Bemerkung 2.28 genannten Gründen überlegen. Im Vergleich zur Gleitkomma-Arithmetik ist die Addition innerhalb der Festkomma-Arithmetik fehlerfrei ausführbar, solange kein Überlauf auftritt. Ausgehend von (2.53) werden analog zu Abschnitt 2.1 obere Schranken für den entstehenden absoluten Rundungsfehler bei elementaren Operationen, Skalarprodukt und Matrix-Vektor-Multiplikation hergeleitet, wobei letztere nicht nur in einfacher Genauigkeit sondern auch – wie von [41] vorgeschlagen – in doppelter Genauigkeit untersucht werden. Unter Berücksichtigung von (2.62) und (2.63) erhalten wir mit Satz 2.31 die wichtigste Aussage von Abschnitt 2.3. Schließlich untersuchen wir in Abschnitt 2.4 ein stochastisches Modell zur Festkomma-Arithmetik, welches die in Abschnitt 2.3 verwendeten Annahmen berücksichtigt. Ebenso wie bei der Gleitkomma-Arithmetik kann auf die Unkorreliertheit der Eingangsdaten verzichtet werden. Die in dieser allgemeinen Form völlig neuen Hauptergebnisse aus Abschnitt 2.4, welche die besondere Gestalt der betrachteten Matrizen essentiell verwendet, finden sich in den Sätzen 2.37, 2.38 und 2.39.

In **Kapitel 3** untersuchen wir zunächst allgemein alle Situationen für die Gleitkomma-Arithmetik, bei der Rundungsfehler auftreten können. Eingangsfehler der Daten werden zusätzlich berücksichtigt. Anschließend werden für die Algorithmen 1.15 – 1.22 gemäß Definitionen 3.1 und 3.21 Stabilitätskon-

stanten k_n bzw. $\breve{k}_n$ für den ungünstigsten bzw. für den durchschnittlich auftretenden Fall hergeleitet. Für die deterministische Rundungsfehleranalyse in Abschnitt 3.1 sind die auf den neuen Ungleichungen (3.21) bzw. (3.40a) und (3.40b) basierenden Hauptergebnisse in den Sätzen 3.12 und 3.18 zusammengefasst. Letzterer berücksichtigt, dass einige der Matrixfaktoren höchstens skalierte Butterfly-Matrizen enthalten, so dass hier pro Block zwei Rundungsfehler erzeugende Multiplikationen eingespart werden können. Im Fall, dass alle Matrix-Einträge tabelliert sind und somit für das ε aus Satz 2.1 jeweils $|\varepsilon| \leq \frac{u}{2}$ gilt, unterbieten die Konstanten (3.27) bzw. (3.46) jeweils die in [54, Theorem 8.3] ermittelte Schranke (3.25). Für die auf den Annahmen aus Modell 3.22 basierende stochastische Rundungsfehleranalyse in Abschnitt 3.2 sind die Hauptergebnisse in Satz 3.23 zusammengefasst. Im Vergleich zu [54, (M3)] verzichtet Modell 3.22 darauf, die Eingangsdaten als unkorreliert oder stochastisch unabhängig anzunehmen. Auf diese Weise ist das Anwendungsspektrum erheblich größer.

In **Kapitel 4** werden die Algorithmen 1.15 – 1.22 auf ihre numerische Stabilität in Festkomma-Arithmetik gemäß Definitionen 4.6 bzw. 4.15 untersucht. Abschnitt 4.1 betrachtet zunächst Abschätzungen für den ungünstigsten Fall sowohl in einfacher als auch in doppelter Genauigkeit. Die in Satz 4.5 zusammengefassten Resultate aus Unterabschnitt 4.1.1 für eine spezielle Blockdiagonalmatrix und einen Eingangsvektor $\mathbf{x} \in [-1, 1]^n$ mit $\|\mathbf{x}\|_2 \leq 1$ bilden die Grundlage zur Herleitung der Hauptergebnisse in Unterabschnitt 4.1.2, welche in den Sätzen 4.8 – 4.10 zusammengefasst sind. Die zur Vermeidung von Überlauf notwendige Vorskalierung findet dabei Berücksichtigung. Unterabschnitt 4.1.3 liefert in Gestalt der Sätze 4.12 – 4.14 schließlich die Stabilitätskonstanten für die Algorithmen 1.15 – 1.22 bei verschiedenen Skalierungsarten. Basierend auf Modell 4.16 befasst sich Abschnitt 4.2 analog zur Gleitkomma-Arithmetik mit einer stochastischen Rundungsfehleranalyse. Als Hauptergebnis sind die entsprechenden Stabilitätskonstanten in Satz 4.17 zusammengefasst.

Schließlich werden in **Kapitel 5** die theoretischen Schranken aus den Sätzen 3.9, 3.16, 3.23, 4.12 und 4.17 auf ihre Qualität überprüft, indem die Algorithmen aus Abschnitt 1.3 für verschiedene Transformationslängen $n = 2^t$ und zufällig erzeugte Testvektoren ausgeführt werden. Dazu sind die Algorithmen 1.15 – 1.22 jeweils in MATLAB implementiert worden. Für die Simulationen in Festkomma-Arithmetik hat insbesondere Methode 5.1 bei der jeweiligen Implementation der Algorithmen Berücksichtigung gefunden.

Als Vorbereitung zu den stochastischen Modellierungen wird im Anhang A.1 ein kurzer Überblick der Maß- und Integrationstheorie sowie der Wahrscheinlichkeitstheorie gegeben. Um die Lesbarkeit von Kapitel 2 zu erleichtern, werden einige längere Beweise im Anhang A.2 nachgeholt. Schließlich stützen die im Anhang A.3 durchgeführten Tests die im Abschnitt 2.2 verwendeten Modellannahmen für den bei einfachen arithmetischen Operationen auftretenden Rundungsfehler.

Die wichtigsten neuen Ergebnisse lassen sich wie folgt zusammenfassen:

(1) Die teilweise bekannten Matrixfaktorisierungen für Kosinus- und Sinusmatrizen vom Typ II – IV werden systematisch ausgearbeitet und führen zu einer Klasse von einfachen, rekursiven Algorithmen für die DCT und DST vom Typ II – IV, falls die Transformationslänge n eine Zweierpotenz ist.

(2) Es werden deterministische und stochastische Modelle zur Gleit- und Festkomma-Arithmetik bereitgestellt. Für Matrix-Vektor-Multiplikationen erfolgt eine deterministische und stochastische Rundungsfehleranalyse. Besitzen die Matrizen spezielle Blockdiagonalgestalt, so bleiben die Ergebnisse auch im Fall korrelierter Eingangsdaten gültig.

(3) Die entwickelten Methoden der deterministischen und stochastischen Rundungsfehleranalysen werden auf die schnellen DCT- und DST-Algorithmen 1.15 – 1.22 angewandt. Für die numerische Stabilität dieser Algorithmen werden erstmals bzw. verbesserte Konstanten angegeben. Die numerischen Testrechnungen illustrieren die theoretischen Ergebnisse und ermöglichen einen sehr guten Vergleich des Rundungsfehler-Verhaltens in beiden Rechnerarithmetiken.

Danksagungen

An dieser Stelle möchte ich mich ganz herzlich bei Prof. Dr. sc. nat. Manfred Tasche für die ausgezeichnete Betreuung und die vielen wertvollen Hinweise bedanken, durch welche diese Arbeit in vielerlei Hinsicht Verbesserung erfahren hat. Insbesondere bedanke ich mich für das entgegengebrachte Vertrauen, die Geduld sowie das überdurchschnittliche Engagement und die zahlreichen Anregungen.

Ebenso herzlich möchte ich mich bei Prof. Dr. rer. nat. habil. Rybakowski dafür bedanken, dass mir durch eine Stelle als wissenschaftliche Mitarbeiterin an seinem Lehrstuhl die Möglichkeit zur Promotion an der Universität Rostock gegeben worden ist. Der herzliche Umgang, das freundliche Entgegenkommen und der Raum zur Selbständigkeit hat mich gerne an seinem Lehrstuhl arbeiten lassen.

Allen Mitarbeiterinnen und Mitarbeitern des Instituts für Mathematik danke ich ebenfalls für die überaus angenehme Arbeitsatmosphäre, genauso wie der Arbeitsgruppe Rechentechnik für die Unterstützung bezüglich Hard- und Software.

Weitere Dankesworte möchte ich meiner Gemeinde und insbesondere der Warnemünder Kantorei widmen, die mir in den letzten 5 Jahren viel Freude durch das gemeinsame Musizieren ermöglicht hat. Die dort gelebte Herzlichkeit und christliche Nächstenliebe haben mich seitdem durch alle Arten von Schwierigkeiten getragen.

Schließlich bedanke ich mich bei meinen Eltern sowie bei meinem Lebensgefährten für ihre Unterstützung und ihr Vertrauen. Ihre aufmunternden Worte und liebevolle Motivation haben in großem Maße zum Gelingen meiner Dissertation beigetragen.

Bezeichnungen

Wir verwenden in dieser Arbeit die folgenden Standardbezeichnungen. Weitere Bezeichnungen werden im Text erklärt.

$\emptyset$	leere Menge
$\mathbb{F}_q$	Menge der Festkomma-Zahlen in Zweierkomplementdarstellung (siehe S. 47)
$\mathbb{G} := \mathbb{G}_{\text{norm}}(\beta, \tau, \gamma_{\min}, \gamma_{\max})$	Menge der normierten Gleitkomma-Zahlen (siehe S. 23)
$\mathbb{G}^n$	Menge der Vektoren $(g_k)_{k=0}^{n-1}$ mit $g_k \in \mathbb{G}$ $(k = 0, \ldots, n-1)$
$\mathbb{G}^{n\times n}$	Menge der Matrizen $(g_{jk})_{j,k=0}^{n-1}$ mit $g_{jk} \in \mathbb{G}$ $(j, k = 0, \ldots, n-1)$
$\mathbb{G}(\beta, \tau, \gamma_{\min}, \gamma_{\max})$	Menge der Gleitkomma-Zahlen (siehe S. 22)
$\mathbb{M}_q$	Menge der Festkomma-Zahlen in Vorzeichen-Betragsdarstellung (siehe S. 46)
$\mathbb{M}_q^n$	Menge der Vektoren $(m_k)_{k=0}^{n-1}$ mit $m_k \in \mathbb{M}_q$ $(k = 0, \ldots, n-1)$
$\mathbb{M}_q^{n\times n}$	Menge der Matrizen $(m_{jk})_{j,k=0}^{n-1}$ mit $m_{jk} \in \mathbb{M}_q$ $(j, k = 0, \ldots, n-1)$
$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{N}_0 := \mathbb{N} \cup \{0\}$	Menge der nichtnegativen ganzen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$	Menge der nichtnegativen reellen Zahlen
$\mathbb{R}^n$	n-dimensionaler Vektorraum über $\mathbb{R}$
$\mathbb{R}^{n\times n}$	Menge der Matrizen $(a_{jk})_{j,k=0}^{n-1}$ mit $a_{jk} \in \mathbb{R}$ $(j, k = 0, \ldots, n-1)$
$\mathbb{Z}$	Menge der ganzen Zahlen
$[a, b] := \{x \in \mathbb{R} \ : \ a \leq 0 \leq b\}$	abgeschlossenes Intervall mit Randpunkten $a, b \in \mathbb{R}$
$[a, b[:= \{x \in \mathbb{R} \ : \ a \leq 0 < b\}$	rechtsseitig halboffenes Intervall mit Randpunkten $a, b \in \mathbb{R}$
$]a, b] := \{x \in \mathbb{R} \ : \ a < 0 \leq b\}$	linksseitig halboffenes Intervall mit Randpunkten $a, b \in \mathbb{R}$
$]a, b[:= \{x \in \mathbb{R} \ : \ a < 0 < b\}$	offenes Intervall mit Randpunkten $a, b \in \mathbb{R}$
$\text{fit} : [-1, 1] \to \mathbb{F}_q$	Rundungsoperator der Festkomma-Arithmetik $\mathbb{F}_q$ (siehe S. 47)
$\text{fix} : [-1, 1] \to \mathbb{M}_q$	Rundungsoperator der Festkomma-Arithmetik $\mathbb{M}_q$ (siehe S. 46)
$\text{fl} : R_{\mathbb{G}} \to \mathbb{G}$	Rundungsoperator der Gleitkomma-Arithmetik $\mathbb{G}$ (siehe S. 23 bzw. 23)
e	Elementarfunktion (siehe S. 103)

$m_1 \times m_2$	Pseudo-Multiplikation in Festkomma-Arithmetik (siehe S. 47)
n	natürliche Zahl
$n_s := \frac{n}{2^s}$	Komplementärteiler zu 2^s $(s = 0, \ldots, t-1)$ für $n = 2^t$
$t_{nk} := \cos\left(\frac{(2k+1)\pi}{2n}\right)$	äquidistante Stützstellen für Chebyshev-Polynome
u	Rundungseinheit (siehe S. 24 bzw. 46)
$v_j(x) := \dfrac{\cos\left(\frac{(2j+1)\arccos(x)}{2}\right)}{\cos\left(\frac{\arccos(x)}{2}\right)}$	Chebyshev-Polynom dritter Art
$\lfloor x \rfloor := \max\{z \in \mathbb{Z} \ : \ z \leq x\}$	floor-Funktion bzw. ganzer Teil einer reellen Zahl
$\lceil x \rceil := \min\{z \in \mathbb{Z} \ : \ x \leq z\}$	ceil-Funktion (siehe S. 25)
$\langle z \rangle_n$	nichtnegativer Rest von $z \in \mathbb{Z}$ modulo n (siehe S. 9)
$\delta(j) := \begin{cases} 1 & \text{für } j = 0, \\ 0 & \text{für } j \in \mathbb{Z} \setminus \{0\} \end{cases}$	Kronecker-Symbol
ρ	Anzahl der Nichtnullelemente pro Zeile (siehe S. 27)
$\sigma, \sigma_\odot \in [0,1]$	Parameter für die Varianz bei Festkomma-Multiplikation (siehe S. 93)
$\sigma_\bullet$ für $\bullet \in \{+, -, \times\}$	Parameter für die Varianz bei Gleitkomma-Operationen (siehe S. 30)
ς_n	Ersatzgröße für den relativen Fehler (siehe S. 95)
$\tau_j(x) := \cos(j \arccos(x))$	Chebyshev-Polynom erster Art
$\mathbf{0} := (0)_{k=0}^{n-1} \in \mathbb{R}^n$	Nullvektor der Länge n
$\mathbf{1} := (1)_{k=0}^{n-1} \in \mathbb{R}^n$	Einsvektor der Länge n
$\mathbf{c}_n := \left(\cos\left(\frac{(2k+1)\pi}{8n}\right)\right)_{k=0}^{n-1}$	Vektor spezieller Drehfaktoren
$\mathbf{s}_n := \left(\sin\left(\frac{(2k+1)\pi}{8n}\right)\right)_{k=0}^{n-1}$	Vektor spezieller Drehfaktoren
$\mathbf{x} := (x_k)_{k=0}^{n-1} \in \mathbb{R}^n$	Spaltenvektor der Länge n
$\hat{\mathbf{x}}$	Approximation eines Vektors $\mathbf{x} \in R_{\mathbb{G}}^n$ in der Gleitkomma-Menge $\mathbb{G}^n$ gemäß $\hat{\mathbf{x}} := \mathrm{fl}(\mathbf{x})$ (siehe S. 23) oder Approximation eines Vektors $\mathbf{x} \in [-1,1]^n$ in der Festkomma-Menge $\mathbb{M}_q^n$ gemäß $\hat{\mathbf{x}} := \mathrm{fix}(\mathbf{x})$ (siehe S. 47)
$\mathbf{x} \circ \mathbf{y} := (x_k y_k)_{k=0}^{n-1} \in \mathbb{R}^n$	Hadamard-Produkt zweier Vektoren der Länge n
$\mathbf{x}^{\mathrm{T}}$	Zeilenvektor der Länge n
$\mathbf{x}^{\mathrm{T}}\mathbf{y} := \sum_{k=0}^{n-1} x_k y_k$	Skalarprodukt zweier Vektoren der Länge n
$\hat{\mathbf{x}}^{\mathrm{T}} \times \hat{\mathbf{y}} := \sum_{i=0}^{n-1} \hat{x}_i \times \hat{y}_i$	inneres Pseudo-Produkt zweier Vektoren $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{M}_q^n$ (siehe S. 47)
$\hat{\mathbf{x}}^{\mathrm{T}} \odot \hat{\mathbf{y}}$	doppelt genaues inneres Pseudo-Produkt (siehe S. 48)
$\lvert\mathbf{x}\rvert := (\lvert x_k \rvert)_{k=0}^{n-1} \in \mathbb{R}^n$	Vektor der Beträge
$\mathbf{x} \leq \mathbf{y}$	Halbordnung der Vektoren mit $x_k \leq y_k$ für alle $k = 0, \ldots, n-1$

$\|\mathbf{x}\|_1 := \sum_{k=0}^{n-1} \|x_k\|$	Betragssummennorm des Vektors $\mathbf{x} \in \mathbb{R}^n$
$\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}} \;=\; \sqrt{\sum_{k=0}^{n-1} x_k^2}$	euklidische Norm des Vektors $\mathbf{x} \in \mathbb{R}^n$
$\|\mathbf{x}\|_\infty := \max_{k=0,\ldots,n-1} \|x_k\|$	Maximumnorm des Vektors $\mathbf{x} \in \mathbb{R}^n$
$\boldsymbol{\beta}_s, \tilde{\boldsymbol{\beta}}_s, \breve{\boldsymbol{\beta}}_s$	spezielle Binärvektoren der Länge 2^s (siehe S. 16 und 19)
$\boldsymbol{\gamma}_s, \tilde{\boldsymbol{\gamma}}_s, \breve{\boldsymbol{\gamma}}_s$	spezielle Binärvektoren der Länge 2^s (siehe S. 17 und 20)
$\eth_{\mathbf{x}}$	Dirac-Maß (siehe S. 110)
$\varpi_n(\mathbf{x})$	Abkürzung definiert in (4.34)
$A := (a_{jk})_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$	Matrix der Größe $n \times n$
$\hat{A}$	Approximation einer Matrix $A \in \mathbb{R}^{n\times n}$ in der Menge $\mathbb{G}^{n\times n}$ gemäß $\hat{A} := \mathrm{fl}(A)$ (siehe S. 23) oder Approximation einer Matrix $A \in [-1,1]^{n\times n}$ in der Menge $\mathbb{M}_q^{n\times n}$ gemäß $\hat{A} := \mathrm{fix}(A)$ (siehe S. 47)
$A \circ B := (a_{jk}b_{jk})_{j,k=0}^{n-1}$	Hadamard-Produkt von Matrizen $A = (a_{jk})_{j,k=0}^{n-1}$ und $B = (b_{jk})_{j,k=0}^{n-1}$
$\hat{A} \times \hat{\mathbf{x}}$	einfach genaues Matrix-Vektor-Pseudo-Produkt (siehe S. 48)
$\hat{A} \odot \hat{\mathbf{x}}$	doppelt genaues Matrix-Vektor-Pseudo-Produkt (siehe S. 48)
$A \oplus B := \begin{pmatrix} A & \\ & B \end{pmatrix}$	direkte Summe von $A = (a_{jk})_{j,k=0}^{n-1}$ und $B = (a_{lr})_{l,r=0}^{m-1}$
$A \otimes B$	Kronecker-Produkt (siehe S. 9)
$A_n(0)$	Einheitsmatrix bei Definition von $A_n(\boldsymbol{\beta}_s)$ (siehe S. 17)
$A_n(1)$	modifizierte Additionsmatrix (siehe S. 6)
$A_n(\boldsymbol{\beta}_s)$	(permutierte) dünnbesetzte Blockdiagonalmatrix (siehe S. 17)
$A_n(\breve{\boldsymbol{\beta}}_s)$	(permutierte) dünnbesetzte Blockdiagonalmatrix (siehe S. 20)
$A_n(\boldsymbol{\gamma}_s)$	(permutierte) dünnbesetzte Blockdiagonalmatrix (siehe S. 18)
$A_n(\breve{\boldsymbol{\gamma}}_s)$	(permutierte) dünnbesetzte Blockdiagonalmatrix (siehe S. 21)
$\breve{A}_n(1)$	modifizierte Additionsmatrix (siehe S. 7)
$C(F) := \{y \in \mathbb{R} \;:\; F \text{ stetig in } y\}$	Menge der Stetigkeitspunkte einer Funktion $F : \mathbb{R} \to \mathbb{R}$ (siehe S. 119)
$B_n^{(s)}$	modifizierte Butterfly-Matrix (siehe S. 9)
$\tilde{B}_n^{(s)}$	modifizierte Butterfly-Matrix (siehe S. 10)
$C_n^{\mathrm{II}}, C_n^{\mathrm{III}}, C_n^{\mathrm{IV}}$	Kosinusmatrizen vom Typ II - IV (siehe S. 1)
$C_n(\boldsymbol{\beta}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 17)
$C_n(\boldsymbol{\gamma}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 18)

$D_n^{(s)}$	spezielle Drehmatrix (siehe S. 10)
$\tilde{D}_n^{(s)}$	spezielle Drehmatrix (siehe S. 10)
F_n	Fourier-Matrix (siehe S. 1)
$I_n := (\delta(j-k))_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$	Einheitsmatrix
$J_n := (\delta(j+k-n+1))_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$	Gegenidentitätsmatrix
L_4	spezielle Drehmatrix (siehe S. 10)
$O_n := (0)_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$	Nullmatrix
P_n	2-Schritt-Permutationsmatrix (siehe S. 6)
$P_n(s) := \bigoplus_{i=1}^{2^s} P_{n_s}^{\mathrm{T}}$	transponierte verallgemeinerte 2-Schritt-Permutationsmatrix
$P_n^{(s)}$	spezielle Permutationsmatrix (siehe S. 9)
$\tilde{P}_n^{(s)}$	spezielle Permutationsmatrix (siehe S. 10)
$Q_2(\varphi)$	Drehmatrix (siehe S. 7)
$R_{\mathbb{G}}$	Definitionsbereich der Funktion fl (siehe S. 23)
$S_n^{\mathrm{II}}, S_n^{\mathrm{III}}, S_n^{\mathrm{IV}}$	Sinusmatrizen vom Typ II – IV (siehe S. 1)
$S_n(\breve{\boldsymbol{\beta}}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 20)
$S_n(\breve{\boldsymbol{\gamma}}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 21)
$T_n(0)$	verallgemeinerte Butterfly-Matrix (siehe S. 6)
$T_n(1)$	kreuzförmige Drehmatrix (siehe S. 6)
$T_n(\boldsymbol{\beta}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 17)
$T_n(\breve{\boldsymbol{\beta}}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 20)
$T_n(\boldsymbol{\gamma}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 18)
$T_n(\breve{\boldsymbol{\gamma}}_s)$	dünnbesetzte Blockdiagonalmatrix (siehe S. 21)
$\breve{T}_n(1)$	kreuzförmige Drehmatrix (siehe S. 7)
U_ν	modifizierte Bit-Umkehrmatrix (siehe S. 9)
$\Sigma_n := \operatorname{diag}\left((-1)^k\right)_{k=0}^{n-1} \in \mathbb{R}^{n\times n}$	spezielle Diagonalmatrix (siehe S. 2)
$\lvert A\rvert := (\lvert a_{j,k}\rvert)_{j,k=0}^{n-1}$	Matrix der Beträge
$\lVert A\rVert_2 := \sqrt{\max_{\lambda\in\sigma(A^T A)} \lvert\lambda\rvert}$	Spektralnorm einer Matrix $A = (a_{jk})_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$
$\lVert A\rVert_F := \sqrt{\sum_{j,k=0}^{n-1} a_{jk}^2}$	Frobeniusnorm einer Matrix $A = (a_{jk})_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$ (siehe S. 60)
$\sigma(A)$	Spektrum (Menge der Eigenwerte) einer Matrix $A \in \mathbb{R}^{n\times n}$
$\operatorname{tr}(A) := \sum_{k=0}^{n-1} a_{kk}$	Spur der Matrix $A = (a_{jk})_{j,k=0}^{n-1} \in \mathbb{R}^{n\times n}$

$\bigoplus_{k=1}^{n} A_k := A_1 \oplus A_2 \oplus \ldots \oplus A_n$	direkte Summe von n Matrizen
$\prod_{k=1}^{n} A_k := A_n A_{n-1} \cdot \ldots \cdot A_2 A_1$	Matrizenprodukt (von rechts nach links)
$\Delta\mathbf{x}$	Rückwärtsfehler (siehe S. 57)
$\mathfrak{S}$	σ-Algebra (in Ω)
Ω	beliebige Menge
$(\Omega, \mathfrak{S})$	Messraum (siehe S. 102)
$(\Omega, \mathfrak{S}, \mu)$	Maßraum (siehe S. 102)
$\mathcal{P}(\Omega)$	Potenzmenge von Ω
$\mathcal{T}_n(\beta_1, \beta_2, \beta_3, \beta_4)$	spezielle Tridiagonalmatrix (siehe S. 2)
$X_n \xrightarrow{\mathcal{L}} X$	X_n verteilungskonvergent gegen X (siehe S. 119)

1 Diskrete trigonometrische Transformationen

In diesem Kapitel definieren wir jeweils verschiedene Kosinus- und Sinus-Matrizen, welche in gewissem Sinn als reellwertige Versionen der Fourier-Matrix

$$F_n := \frac{1}{\sqrt{n}} \left(\mathrm{e}^{-\frac{2\pi \mathrm{i} jk}{n}} \right)_{j,k=0}^{n-1} \in \mathbb{C}^{n \times n} \tag{1.1}$$

aufgefasst werden können. Die entsprechenden diskreten Kosinus- und Sinus-Transformationen (DCT, DST), welche wir zusammenfassend als diskrete trigonometrische Transformationen bezeichnen, werden in Abschnitt 1.3 erklärt. Für die hier vorgestellten Transformationen gibt es schnelle Algorithmen, deren Unterschied in der Struktur derjenigen Matrizen besteht, in welche die jeweils zugrunde liegende Transformationsmatrix faktorisiert wird. An dieser Stelle werden nur Faktorisierungen in dünnbesetzte orthogonale Matrizen untersucht. Dabei heißt hier eine Matrix *dünnbesetzt*, wenn sie in jeder Zeile und in jeder Spalte nicht mehr als zwei Nichtnulleinträge besitzt.

1.1 Trigonometrische Matrizen

Diskrete Kosinus-Transformationen treten erstmals bei Ahmed, Natarajan und Rao [1] auf, welche einen Zusammenhang zwischen den Eigenvektoren spezieller Toeplitz-Matrizen und den Spalten der später als Kosinus-Matrix vom Typ II bezeichneten Matrix feststellen [1, 11]. Zehn Jahre darauf finden sich in [63] insgesamt 16 verschiedene trigonometrische Matrizen, welche Rao und Yip [46] als Sinus- und Kosinus-Matrizen vom Typ I – VIII klassifizieren. Wir beschränken uns hier jedoch auf Matrizen vom Typ II – IV, da sie in der Signal- und Bildverarbeitung eine herausragende Rolle spielen [2, 15, 57]. Es ist zu beachten, dass in [52] und z.T. in [45] die Rollen von Typ II und III vertauscht sind.

Definition 1.1. *Sei $n \geq 2$ eine natürliche Zahl. Die Kosinus-Matrizen vom Typ* II–IV *werden durch*

$$C_n^{\mathrm{II}} := \sqrt{\frac{2}{n}} \left(\varepsilon_n(j) \cos \frac{j(2k+1)\pi}{2n} \right)_{j,k=0}^{n-1}, \tag{1.2}$$

$$C_n^{\mathrm{III}} := \left(C_n^{\mathrm{II}} \right)^{\mathrm{T}}, \tag{1.3}$$

$$C_n^{\mathrm{IV}} := \sqrt{\frac{2}{n}} \left(\cos \frac{(2j+1)(2k+1)\pi}{4n} \right)_{j,k=0}^{n-1} \tag{1.4}$$

und die Sinus-Matrizen vom Typ II–IV *werden durch*

$$S_n^{\mathrm{II}} := \sqrt{\frac{2}{n}} \left(\varepsilon_n(j+1) \sin \frac{(j+1)(2k+1)\pi}{2n} \right)_{j,k=0}^{n-1}, \tag{1.5}$$

$$S_n^{\mathrm{III}} := \left(S_n^{\mathrm{II}} \right)^{\mathrm{T}}, \tag{1.6}$$

$$S_n^{\mathrm{IV}} := \sqrt{\frac{2}{n}} \left(\sin \frac{(2j+1)(2k+1)\pi}{4n} \right)_{j,k=0}^{n-1} \tag{1.7}$$

mit

$$\varepsilon_n(j) := \begin{cases} \frac{\sqrt{2}}{2} & j \in \{0, n\}, \\ 1 & j \in \{1, \ldots, n-1\} \end{cases}$$

definiert.

Über die Größe $\varepsilon_n(j)$ erfolgt bei den trigonometrischen Matrizen vom Typ II und III jeweils eine Skalierung der ersten bzw. letzten Zeile bzw. Spalte, welche bewirkt, dass einerseits alle Zeilen und Spalten dieselbe euklidische Norm besitzen und andererseits alle Zeilen bzw. alle Spalten jeder Matrix in (1.2)–(1.7) jeweils paarweise orthogonal zueinander sind (vgl. Satz 1.4). Der Faktor $\sqrt{\frac{2}{n}}$ normiert zusätzlich alle Zeilen- und Spaltenvektoren auf die Länge 1.

Bemerkung 1.2. In [52, 45] wird der Zusammenhang von diskreten trigonometrischen Matrizen (in der unskalierten Variante) mit diskreten Randwertaufgaben hergestellt. Insbesondere zeigen Strang [52] und Püschel [45], dass es sich bei (1.2) – (1.7) (gegebenenfalls in unskalierter Form) um Eigenmatrizen spezieller tridiagonaler Matrizen der Größe $n \times n$ handelt, welche in der Gestalt

$$\mathcal{T}_n(\beta_1,\beta_2,\beta_3,\beta_4) \;=\; \frac{1}{2}\begin{pmatrix} \beta_1 & \beta_2 & & & & & \\ 1 & 0 & 1 & & & & \\ & 1 & 0 & 1 & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 1 & 0 & 1 & \\ & & & & \beta_3 & \beta_4 \end{pmatrix} \tag{1.8}$$

gewählt werden können. Betrachten wir nun die aus der Differentialgleichung $u'' = 0$ resultierende diskrete Randwertaufgabe

$$\begin{cases} a_k &= \frac{1}{2}(a_{k-1} + a_{k+1}), \qquad 0 \le k \le n-1\ , \\ a_{-1} &= \zeta\ , \\ a_n &= \eta \end{cases} \tag{1.9}$$

mit Konstanten $\zeta, \eta \in \mathbb{R}$ und beachten, dass die Größen β_1, β_2 durch (1.9) eindeutig festgelegt sind, so kann die Randwertaufgabe (1.9) mit $\mathbf{a} = (a_0, \ldots, a_{n-1})^{\mathrm{T}}$ in der Gestalt

$$\mathbf{a} \;=\; \mathcal{T}_n(\beta_1,\beta_2,\beta_3,\beta_4)\cdot\mathbf{a}$$

formuliert werden. Dabei entsprechen beispielsweise $\eta = 0$ bzw. $\eta = -a_{n-1}$ jeweils einer diskreten Variante der Dirichlet-Randbedingung und $\eta = a_{n-2}$ bzw. $\eta = a_{n-1}$ einer diskreten Variante der Neumann-Randbedingung [52, 45]. In Tabelle 1.1 sind die Beziehungen für die hier betrachteten Matrizen (1.2) – (1.7) zusammengefasst, welche sich aus Table 5.2 und Table 5.3 in [45] ableiten lassen. Es ist zu beachten, dass die Matrizen $\mathcal{T}_n(0,2,1,0)$ und $\mathcal{T}_n(0,1,2,0)$ nicht symmetrisch sind und daher in [45] bewusst unskalierte und somit insbesondere nicht-orthogonale Varianten der trigonometrischen Matrizen betrachtet werden. Für die orthogonalen Varianten (1.2) und (1.5) können wir alternativ jedoch die symmetrisierten Tridiagonalmatrizen

$$\operatorname{diag}\left(\tfrac{1}{\sqrt{2}},1,\ldots,1\right)\mathcal{T}_n(0,2,1,0)\operatorname{diag}(\sqrt{2},1,\ldots,1)\ ,\quad \operatorname{diag}\left(1,\ldots,1,\tfrac{1}{\sqrt{2}}\right)\mathcal{T}_n(0,1,2,0)\operatorname{diag}(1,\ldots,1,\sqrt{2})$$

wählen, für welche C_n^{II} bzw. S_n^{II} Eigenmatrizen sind (vgl. auch [52] sowie [50, Satz 2.3]). □

ζ	η	β_1	β_2	β_3	β_4	Eigenmatrix von $\mathcal{T}_n(\beta_1,\beta_2,\beta_3,\beta_4)$
a_1	0	0	2	1	0	$\operatorname{diag}(\sqrt{2},1,\ldots,1)C_n^{\mathrm{II}}$
a_0	a_{n-1}	1	1	1	1	C_n^{III}
a_0	$-a_{n-1}$	1	1	1	-1	C_n^{IV}
0	a_{n-2}	0	1	2	0	$\operatorname{diag}(1,\ldots,1,\sqrt{2})S_n^{\mathrm{II}}$
$-a_0$	$-a_{n-1}$	-1	1	1	-1	S_n^{III}
$-a_0$	a_{n-1}	-1	1	1	1	S_n^{IV}

Tabelle 1.1: Beziehung zwischen den Tridiagonal-Matrizen $\mathcal{T}_n(\beta_1,\beta_2,\beta_3,\beta_4)$ und den trigonometrischen Matrizen (1.2)–(1.7) in Anlehnung an Table 5.2, 5.3 in [45].

Mittels der hier *Vorzeichenskalierungsmatrix* genannten speziellen Diagonalmatrix

$$\Sigma_n := \operatorname{diag}((-1)^k)_{k=0}^{n-1}$$

und der mit J_n bezeichneten Matrix der Größe $n \times n$, welche einen Vektor $\mathbf{x} := (x_k)_{k=0}^{n-1} \in \mathbb{R}^n$ auf

$$J_n\mathbf{x} := (x_{n-1-k})_{k=0}^{n-1}$$

abbildet, lassen sich die in (1.2)–(1.7) definierten Kosinus- und Sinus-Matrizen zum Teil ineinander überführen. Aufgrund des folgenden Lemmas genügt es daher, die Matrizen (1.2)–(1.4) zu betrachten.

Lemma 1.3 (vgl. [50], Lemma 2.2)**.** *Sei $n \geq 2$ eine natürliche Zahl. Dann genügen die trigonometrischen Matrizen den folgenden Beziehungen*

$$J_n C_n^{\mathrm{II}} \;=\; S_n^{\mathrm{II}} \Sigma_n, \qquad C_n^{\mathrm{III}} J_n \;=\; \Sigma_n S_n^{\mathrm{III}}, \qquad C_n^{\mathrm{IV}} J_n \;=\; \Sigma_n S_n^{\mathrm{IV}}.$$

Beweis: Wir halten zunächst fest, dass aus dem Additionstheorem

$$\cos(x \pm y) \;=\; \cos x \cos y \mp \sin x \sin y \tag{1.10}$$

für $x, y \in \mathbb{R}$ sofort die Beziehung

$$\cos(j\pi \pm x) \;=\; (-1)^j \cos x \tag{1.11}$$

für $j \in \mathbb{Z}$ folgt. Als weitere Vorbemerkung sei erwähnt, dass die Multiplikation mit J_n von links die Reihenfolge der Zeilen (bzw. von rechts analog die Reihenfolge der Spalten) umkehrt, wodurch lediglich die entsprechenden Indizes gemäß der Abbildung $j \longleftrightarrow n-1-j$ (bzw. $k \longleftrightarrow n-1-k$) vertauscht werden. Wegen $\varepsilon_n(0) = \varepsilon_n(n) = \frac{\sqrt{2}}{2}$ und $\varepsilon_n(j) = 1$ für $j \in \{1, \ldots, n-1\}$ folgt demnach

$$\begin{aligned}
J_n C_n^{\mathrm{II}} &\overset{(1.2)}{=} J_n \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j) \cos \tfrac{j\ (2k+1)\pi}{2n} \right)_{j,k=0}^{n-1} \\
&\overset{j \longleftrightarrow n-1-j}{=} \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(n-1-j) \cos \tfrac{(n-1-j)\ (2k+1)\pi}{2n} \right)_{j,k=0}^{n-1} \\
&\overset{\varepsilon_n(0)=\varepsilon_n(n)}{=} \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j+1) \cos \tfrac{(n-(j+1))\ (2k+1)\pi}{2n} \right)_{j,k=0}^{n-1} \\
&= \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j+1) \cos \left(k\pi - \left(\tfrac{(j+1)\ (2k+1)\pi}{2n} - \tfrac{\pi}{2} \right) \right) \right)_{j,k=0}^{n-1} \\
&\overset{(1.11)}{=} \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j+1) \cos \left(\tfrac{(j+1)\ (2k+1)\pi}{2n} - \tfrac{\pi}{2} \right) (-1)^k \right)_{j,k=0}^{n-1} \\
&\overset{\cos(x-\frac{\pi}{2})=\sin x}{=} \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j+1) \sin \tfrac{(j+1)\ (2k+1)\pi}{2n}\ (-1)^k \right)_{j,k=0}^{n-1} \\
&\overset{\Sigma_n=\mathrm{diag}((-1)^k)_{k=0}^{n-1}}{=} \sqrt{\tfrac{2}{n}} \left(\varepsilon_n(j+1) \sin \tfrac{(j+1)\ (2k+1)\pi}{2n} \right)_{j,k=0}^{n-1} \Sigma_n \\
&\overset{(1.5)}{=} S_n^{\mathrm{II}} \Sigma_n.
\end{aligned}$$

Unter Beachtung von $J_n^{\mathrm{T}} = J_n$, $\Sigma_n^{\mathrm{T}} = \Sigma_n$ sowie der Definitionen von C_n^{III} und S_n^{III} ergibt sich daraus sofort auch $C_n^{\mathrm{III}} J_n \;=\; \Sigma_n S_n^{\mathrm{III}}$. Die letzte der drei Beziehungen lässt sich nun analog zu der ersten herleiten. Es gilt

$$\begin{aligned}
C_n^{\mathrm{IV}} J_n &\overset{(1.4)}{=} \sqrt{\tfrac{2}{n}} \left(\cos \tfrac{(2j+1)\ (2k+1)\pi}{4n} \right)_{j,k=0}^{n-1} J_n \\
&\overset{k \longleftrightarrow n-1-k}{=} \sqrt{\tfrac{2}{n}} \left(\cos \tfrac{(2j+1)\ (2(n-1-k)+1)\pi}{4n} \right)_{j,k=0}^{n-1} \\
&\overset{2(n-1-k)+1=2n-(2k+1)}{=} \sqrt{\tfrac{2}{n}} \left(\cos \tfrac{2n(2j+1)-(2j+1)(2k+1))\pi}{4n} \right)_{j,k=0}^{n-1} \\
&= \sqrt{\tfrac{2}{n}} \left(\cos \left(j\pi - \left(\tfrac{(2j+1)(2k+1)\pi}{4n} - \tfrac{\pi}{2} \right) \right) \right)_{j,k=0}^{n-1} \\
&\overset{(1.11)}{=} \sqrt{\tfrac{2}{n}} \left((-1)^j \cos \left(\tfrac{(2j+1)(2k+1)\pi}{4n} - \tfrac{\pi}{2} \right) \right)_{j,k=0}^{n-1} \\
&\overset{\cos(x-\frac{\pi}{2})=\sin x}{=} \sqrt{\tfrac{2}{n}} \left((-1)^j \sin \tfrac{(2j+1)(2k+1)\pi}{4n} \right)_{j,k=0}^{n-1} \\
&\overset{\Sigma_n=\mathrm{diag}((-1)^j)_{j=0}^{n-1}}{=} \Sigma_n \sqrt{\tfrac{2}{n}} \left(\sin \tfrac{(2j+1)\ (2k+1)\pi}{4n} \right)_{j,k=0}^{n-1} \\
&\overset{(1.7)}{=} \Sigma_n S_n^{\mathrm{IV}},
\end{aligned}$$

womit der Beweis beendet ist. ■

Eine wesentliche Eigenschaft der Matrizen (1.2)–(1.7) ist, dass wir sehr schnell ihre Inversen angeben können. Insbesondere sind dafür keine Rechenoperationen notwendig, was für spätere Fehlerbetrachtungen sehr interessant ist.

Satz 1.4 (vgl. auch [46] oder [50]). *Es gelten die Aussagen:*

(i) *Die Matrizen* (1.2) – (1.7) *sind orthogonal.*

(ii) *Es sind* $\left(C_n^{\mathrm{II}}\right)^{-1} = C_n^{\mathrm{III}}$ *und* $\left(S_n^{\mathrm{II}}\right)^{-1} = S_n^{\mathrm{III}}$. *Desweiteren sind die Matrizen* C_n^{IV} *und* S_n^{IV} *zu sich selbst invers.*

Beweis: (i) Aus dem für beliebige Winkel $\alpha, \beta \in \mathbb{R}$ gültigen Additionstheorem

$$\sin(\alpha)\cos(\beta) \;=\; \frac{1}{2}\left(\sin(\alpha-\beta)+\sin(\alpha+\beta)\right)$$

erhalten wir – da der Sinus eine ungerade Funktion ist – zunächst

$$\sin(x)\sum_{k=0}^{n-1}\cos((2k+1)x) \;=\; \frac{1}{2}\sum_{k=0}^{n-1}\left(\sin(2k+2)x-\sin(2kx)\right) \;=\; \frac{1}{2}\sin(2nx)$$

für alle $x \in \mathbb{R}$. Auf dem Intervall $[0,\pi[$ folgt die hilfreiche Summenformel

$$\sum_{k=0}^{n-1}\cos((2k+1)x) \;=\; \begin{cases} n & \text{für } x = 0\ , \\ \frac{\sin(2nx)}{2\sin(x)} & \text{für } x \in]0,\pi[\ , \end{cases} \tag{1.12}$$

die insbesondere bei $x = \frac{m\pi}{2n}$ für $|m| \in \{1,\dots,2n-1\}$ verschwindet. Mit dem wiederum für beliebige Winkel $\alpha, \beta \in \mathbb{R}$ gültigen Additionstheorem

$$\cos(\alpha)\cos(\beta) \;=\; \frac{1}{2}\big(\cos(\alpha+\beta)+\cos(\alpha-\beta)\big)\ , \tag{1.13}$$

welches sich sofort aus (1.10) herleiten lässt, ergibt sich nun wegen

$$\begin{aligned}
C_n^{\mathrm{II}}\left(C_n^{\mathrm{II}}\right)^{\mathrm{T}} \;&=\; \sqrt{\frac{2}{n}}\Big(\varepsilon_n(j)\cos\frac{j(2k+1)\pi}{2n}\Big)_{j,k=0}^{n-1}\cdot\sqrt{\frac{2}{n}}\Big(\varepsilon_n(l)\cos\frac{l(2k+1)\pi}{2n}\Big)_{k,l=0}^{n-1} \\
&=\; \frac{2}{n}\Big(\varepsilon_n(j)\varepsilon_n(l)\sum_{k=0}^{n-1}\cos\frac{j(2k+1)\pi}{2n}\cdot\cos\frac{l(2k+1)\pi}{2n}\Big)_{j,l=0}^{n-1} \\
&\overset{(1.13)}{=}\; \frac{1}{n}\left(\varepsilon_n(j)\varepsilon_n(l)\left(\sum_{k=0}^{n-1}\cos\frac{(j+l)(2k+1)\pi}{2n}+\sum_{k=0}^{n-1}\cos\frac{(j-l)(2k+1)\pi}{2n}\right)\right)_{j,l=0}^{n-1} \\
&\overset{(1.12)}{=}\; \Big(\big(\varepsilon_n(j)\big)^2\big(\delta(j+l)+\delta(j-l)\big)\Big)_{j,l=0}^{n-1} \;=\; \Big(\delta(j-l)\Big)_{j,l=0}^{n-1} \;=\; I_n
\end{aligned}$$

die Orthogonalität der Matrizen C_n^{II} und $C_n^{\mathrm{III}} = \left(C_n^{\mathrm{II}}\right)^{\mathrm{T}}$. Dabei haben wir verwendet, dass hier jeweils $j+l \in \{0,\dots,2n-1\}$ und $|j-l| \in \{0,\dots,2n-1\}$ erfüllt sind.
Aufgrund der aus der Definition ersichtlichen Beziehungen $\Sigma_n\Sigma_n^{\mathrm{T}} = \Sigma_n^2 = I_n$ und $J_nJ_n^{\mathrm{T}} = J_n^2 = I_n$ liefert Lemma 1.3 nun ebenso

$$S_n^{\mathrm{II}}\left(S_n^{\mathrm{II}}\right)^{\mathrm{T}} \;=\; J_nC_n^{\mathrm{II}}\Sigma_n\left(J_nC_n^{\mathrm{II}}\Sigma_n\right)^{\mathrm{T}} \;=\; J_nC_n^{\mathrm{II}}\Sigma_n\Sigma_nC_n^{\mathrm{III}}J_n \;=\; J_nC_n^{\mathrm{II}}C_n^{\mathrm{III}}J_n \;=\; J_n^2 \;=\; I_n$$

und demzufolge die Orthogonalität der Matrizen S_n^{II} und $S_n^{\mathrm{III}} = \left(S_n^{\mathrm{II}}\right)^{\mathrm{T}}$.
Analog folgt nun aus

$$\begin{aligned}
C_n^{\mathrm{IV}}\left(C_n^{\mathrm{IV}}\right)^{\mathrm{T}} \;&=\; \frac{2}{n}\left(\cos\frac{(2j+1)(2k+1)\pi}{4n}\right)_{j,k=0}^{n-1}\left(\cos\frac{(2k+1)(2l+1)\pi}{4n}\right)_{k,l=0}^{n-1} \\
&=\; \frac{2}{n}\left(\sum_{k=0}^{n-1}\cos\frac{(2j+1)(2k+1)\pi}{4n}\cos\frac{(2k+1)(2l+1)\pi}{4n}\right)_{j,l=0}^{n-1} \\
&\overset{(1.13)}{=}\; \frac{1}{n}\left(\sum_{k=0}^{n-1}\left(\cos\frac{(2k+1)(j-l)\pi}{2n}+\cos\frac{(2k+1)(j+l+1)\pi}{2n}\right)\right)_{j,l=0}^{n-1} \\
&\overset{(1.12)}{=}\; \Big(\delta(j-l)+\delta(j+l+1)\Big)_{j,l=0}^{n-1} \;=\; \Big(\delta(j-l)\Big)_{j,l=0}^{n-1} \;=\; I_n
\end{aligned}$$

die Orthogonalität von C_n^{IV} und mit Lemma 1.3 weiter

$$S_n^{\mathrm{IV}} \left(S_n^{\mathrm{IV}}\right)^{\mathrm{T}} \;=\; \Sigma_n C_n^{\mathrm{IV}} J_n \left(\Sigma_n C_n^{\mathrm{IV}} J_n\right)^{\mathrm{T}} \;=\; \Sigma_n C_n^{\mathrm{IV}} J_n J_n C_n^{\mathrm{IV}} \Sigma_n \;=\; \Sigma_n C_n^{\mathrm{IV}} C_n^{\mathrm{IV}} \Sigma_n \;=\; \Sigma_n^2 \;=\; I_n \;,$$

also auch die Orthogonalität von S_n^{IV}.
(ii) Mit der in (i) gezeigten Orthogonalität und der Definition 1.1 erhalten wir

$$\left(C_n^{\mathrm{II}}\right)^{-1} \;=\; \left(C_n^{\mathrm{II}}\right)^{\mathrm{T}} \;=\; C_n^{\mathrm{III}} \;, \qquad \left(S_n^{\mathrm{II}}\right)^{-1} \;=\; \left(S_n^{\mathrm{II}}\right)^{\mathrm{T}} \;=\; S_n^{\mathrm{III}} \;.$$

Aufgrund der bereits aus der Definition 1.1 erkennbaren Symmetrien $C_n^{\mathrm{IV}} = \left(C_n^{\mathrm{IV}}\right)^{\mathrm{T}}$ und $S_n^{\mathrm{IV}} = \left(S_n^{\mathrm{IV}}\right)^{\mathrm{T}}$ liefert die in (i) nachgewiesene Orthogonalität von C_n^{IV} und S_n^{IV} nun auch die Gültigkeit des letzten Teils der Behauptung. ■

Im Beweis der vorangegangenen Folgerung haben wir verwendet, dass das Produkt orthogonaler Matrizen wiederum eine orthogonale Matrix liefert. Daher wird die Orthogonalität der Matrizen (1.2) – (1.7) in Abschnitt 1.3 durch eine jeweilige Faktorisierung in orthogonale Matrixfaktoren erneut bestätigt.

Bemerkung 1.5. In [1, 50] wird der Zusammenhang von Kosinus-Matrizen und Chebyshev-Polynomen nachgewiesen. Werten wir die *Chebyshev-Polynome erster Art*, definiert durch $\tau_j(x) := \cos(j \arccos x)$, an den äquidistanten Stützstellen $t_{nk} := \cos \frac{(2k+1)\pi}{2n}$ aus, so gilt

$$\tau_j(t_{nk}) \;=\; \cos\left(j \arccos\left(\cos \frac{(2k+1)\pi}{2n}\right)\right) \;=\; \cos \frac{j(2k+1)\pi}{2n}$$

und daher

$$C_n^{\mathrm{II}} \;=\; \sqrt{\frac{2}{n}}\Big(\varepsilon_n(j)\tau_j(t_{nk})\Big)_{j,k=0}^{n-1} \;.$$

Um nun $C_n^{\mathrm{II}}\big(C_n^{\mathrm{II}}\big)^{\mathrm{T}} = I_n$ zu zeigen, wird lediglich die in [50, Satz 1.6] angegebene Orthogonalitätsbeziehung

$$\sum_{k=0}^{n-1} \tau_j(t_{nk})\tau_l(t_{nk}) \;=\; \frac{n}{2}\, \delta(j-l)\, (\varepsilon_n(j))^{-2}$$

im ersten Teil des Beweises von Satz 1.4 (i) hergeleitet und angewandt.

Ebenso erhalten wir mit der für alle $x \in]-\pi, \pi[$ gültigen Beziehung $0 < \cos \frac{x}{2} = \sqrt{\frac{1+\cos(x)}{2}}$, eingesetzt an den oben definierten Stützstellen t_{nk}, für die Matrix C_n^{IV} die Darstellung

$$\begin{aligned} C_n^{\mathrm{IV}} \;&=\; \sqrt{\frac{2}{n}}\left(\cos \frac{(2j+1)\arccos t_{nk}}{2}\right)_{j,k=0}^{n-1} \;=\; \sqrt{\frac{2}{n}}\left(\frac{\sqrt{\frac{1+t_{nk}}{2}}}{\cos \frac{\arccos t_{nk}}{2}} \cos \frac{(2j+1)\arccos t_{nk}}{2}\right)_{j,k=0}^{n-1} \\ &=\; \sqrt{\frac{1}{n}}\left(\sqrt{1+t_{nk}}\; \frac{\cos \frac{(2j+1)\arccos t_{nk}}{2}}{\cos \frac{\arccos t_{nk}}{2}}\right)_{j,k=0}^{n-1} \;=\; \sqrt{\frac{1}{n}}\Big(\sqrt{1+t_{nk}}\; v_j(t_{nk})\Big)_{j,k=0}^{n-1} \end{aligned}$$

mit den *Chebyshev-Polynomen dritter Art*, definiert durch $v_j(x) := \dfrac{\cos \frac{(2j+1)\arccos x}{2}}{\cos \frac{\arccos x}{2}}$. Demnach ist die Orthogonalität von C_n^{IV} zu der Beziehung

$$\sum_{k=0}^{n-1} (1+t_{nk})\; v_j(t_{nk}) v_l(t_{nk}) \;=\; n\delta(l-j)$$

äquivalent, welche ebenso in [50, Satz 1.6] angegeben wird, jedoch fälschlicherweise mit einem zusätzlichen Faktor $\frac{1}{2}$. □

Aufgrund der Orthogonalität ist die Matrix-Vektor-Multiplikation mit jeder der Matrizen (1.2) – (1.7) gut konditioniert. Insbesondere besitzen orthogonale Matrizen bestmögliche Kondition bezüglich der Spektralnorm. Da es sich jeweils um vollbesetzte Matrizen handelt, sind wir – um die arithmetischen Kosten zu reduzieren – an Faktorisierungen in weniger vollbesetzte Matrizen interessiert. Um die Kondition nicht zu verschlechtern, favorisieren wir dabei orthogonale Matrixfaktoren.

1.2 Orthogonale Faktorisierungen von Sinus- und Kosinusmatrizen

In diesem Abschnitt betrachten wir für gerades $n \in \mathbb{N}$ verschiedene Faktorisierungen der in Abschnitt 1.1 eingeführten Kosinus- und Sinusmatrizen. Ergänzend zu den Matrizen J_n und Σ_n, welche bereits in Abschnitt 1.1 eingeführt worden sind, bezeichnen wir mit P_n die 2-*Schritt-Permutationsmatrix* der Ordnung n, welche einen Vektor $\mathbf{x} = (x_k)_{k=0}^{n-1}$ auf

$$P_n\mathbf{x} \;:=\; (x_0, x_2, \ldots, x_{n-2}, x_1, x_3, \ldots, x_{n-1})^{\mathrm{T}} \tag{1.14}$$

abbildet. Mit I_n wird wie üblich die Einheitsmatrix der Größe $n \times n$ bezeichnet. Zunächst geben wir die in [43] hergeleiteten Beziehungen zwischen den Matrizen C_n^{II} und C_n^{IV} an.

Lemma 1.6 (vgl. [43], Lemma 2.2 (*i*) und Lemma 2.4)**.** *Sei $n \in \mathbb{N}$ gerade mit $n \geq 4$ gegeben und $n_1 := \frac{n}{2}$. Dann gelten die Faktorisierungen*

$$C_n^{\mathrm{II}} = P_n^{\mathrm{T}} \begin{pmatrix} C_{n_1}^{\mathrm{II}} & \\ & C_{n_1}^{\mathrm{IV}} \end{pmatrix} T_n(0), \qquad C_n^{\mathrm{IV}} = P_n^{\mathrm{T}} A_n(1) \begin{pmatrix} C_{n_1}^{\mathrm{II}} & \\ & C_{n_1}^{\mathrm{II}} \end{pmatrix} T_n(1) \tag{1.15}$$

mit der verallgemeinerten Butterfly-Matrix

$$T_n(0) := \tfrac{1}{\sqrt{2}} \begin{pmatrix} I_{n_1} & J_{n_1} \\ I_{n_1} & -J_{n_1} \end{pmatrix}, \tag{1.16}$$

der modifizierten Additionsmatrix

$$A_n(1) := \tfrac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2} & & & \\ & I_{n_1-1} & I_{n_1-1} & \\ & I_{n_1-1} & -I_{n_1-1} & \\ & & & -\sqrt{2} \end{pmatrix} \begin{pmatrix} I_{n_1} & \\ & \Sigma_{n_1} J_{n_1} \end{pmatrix} \tag{1.17}$$

und der kreuzförmigen Drehmatrix

$$T_n(1) := \begin{pmatrix} I_{n_1} & \\ & \Sigma_{n_1} \end{pmatrix} \begin{pmatrix} \operatorname{diag} \mathbf{c}_{n_1} & (\operatorname{diag} \mathbf{s}_{n_1}) J_{n_1} \\ -J_{n_1}(\operatorname{diag} \mathbf{s}_{n_1}) & \operatorname{diag}(J_{n_1}\mathbf{c}_{n_1}) \end{pmatrix} \tag{1.18}$$

mit den die entsprechenden Drehfaktoren enthaltenden Vektoren

$$\mathbf{c}_{n_1} := \left(\cos \tfrac{(2k+1)\pi}{8n_1}\right)_{k=0}^{n_1-1}, \quad \mathbf{s}_{n_1} := \left(\sin \tfrac{(2k+1)\pi}{8n_1}\right)_{k=0}^{n_1-1}.$$

Werden die Definitionen von $T_n(0)$ und $T_n(1)$ auch für $n = 2$ verwendet, so stimmt $T_2(0)$ genau mit C_2^{II} überein. Weiterhin gilt $C_2^{\mathrm{IV}} = \Sigma_2 T_2(1)$. Aufgrund der Symmetrie der Matrix C_n^{IV} und mittels (1.3) lassen sich durch Transponieren von (1.15) sofort ähnliche Beziehungen für die Matrizen C_n^{III} und C_n^{IV} herleiten.

Folgerung 1.7. *Sei $n \in \mathbb{N}$ gerade mit $n \geq 4$ gegeben und $n_1 := \frac{n}{2}$. Mit den Bezeichungen aus Lemma* 1.6 *gelten dann ebenso die Faktorisierungen*

$$C_n^{\mathrm{III}} = T_n(0)^{\mathrm{T}} \begin{pmatrix} C_{n_1}^{\mathrm{III}} & \\ & C_{n_1}^{\mathrm{IV}} \end{pmatrix} P_n, \qquad C_n^{\mathrm{IV}} = T_n(1)^{\mathrm{T}} \begin{pmatrix} C_{n_1}^{\mathrm{III}} & \\ & C_{n_1}^{\mathrm{III}} \end{pmatrix} A_n(1)^{\mathrm{T}} P_n\,. \tag{1.19}$$

Aufgrund ihrer Orthogonalität und ihrer Symmetrie sind die Matrizen J_n und Σ_n darüber hinaus zu sich selbst invers. Demnach liefert Lemma 1.3 sofort die Darstellungen $S_n^{\mathrm{III}} = \Sigma_n C_n^{\mathrm{III}} J_n$ sowie $S_n^{\mathrm{IV}} = \Sigma_n C_n^{\mathrm{IV}} J_n$. Mit der Identität $P_n J_n = J_n P_n$ ergibt sich

$$\begin{aligned} \begin{pmatrix} J_{n_1} & \\ & J_{n_1} \end{pmatrix} A_n(1)^{\mathrm{T}} P_n J_n \;&=\; \tfrac{1}{\sqrt{2}} \begin{pmatrix} J_{n_1} & \\ & \Sigma_{n_1} \end{pmatrix} \begin{pmatrix} \sqrt{2} & & & \\ & I_{n_1-1} & I_{n_1-1} & \\ & I_{n_1-1} & -I_{n_1-1} & \\ & & & -\sqrt{2} \end{pmatrix} J_n P_n \\ &=\; \tfrac{1}{\sqrt{2}} \begin{pmatrix} J_{n_1} & \\ & \Sigma_{n_1} \end{pmatrix} \begin{pmatrix} & & & \sqrt{2} \\ & J_{n_1-1} & J_{n_1-1} & \\ & -J_{n_1-1} & J_{n_1-1} & \\ -\sqrt{2} & & & \end{pmatrix} P_n\,. \end{aligned}$$

Beachten wir darüber hinaus die Gültigkeit von

$$\begin{aligned}\Sigma_n J_n \mathbf{x} &= \Sigma_n \left(x_{n-1-k}\right)_{k=0}^{n-1} = \left((-1)^k x_{n-1-k}\right)_{k=0}^{n-1} = (-1)^{n-1}\left((-1)^{n-1-k} x_{n-1-k}\right)_{k=0}^{n-1} \\ &= (-1)^{n-1} J_n \left((-1)^k x_k\right)_{k=0}^{n-1} = (-1)^{n-1} J_n \Sigma_n \mathbf{x}\end{aligned}$$

und $\Sigma_n(\operatorname{diag}\mathbf{x}) = (\operatorname{diag}\mathbf{x})\Sigma_n$ für einen beliebigen Vektor $\mathbf{x} \in \mathbb{R}^n$, so folgen

$$\begin{aligned}\Sigma_n T_n(0)^{\mathrm{T}} \begin{pmatrix}\Sigma_{n_1} & \\ & \Sigma_{n_1}\end{pmatrix} &= \frac{1}{\sqrt{2}} \begin{pmatrix}\Sigma_{n_1} & \\ & (-1)^{n_1}\Sigma_{n_1}\end{pmatrix} \begin{pmatrix} I_{n_1} & I_{n_1} \\ J_{n_1} & -J_{n_1}\end{pmatrix} \begin{pmatrix}\Sigma_{n_1} & \\ & \Sigma_{n_1}\end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} I_{n_1} & I_{n_1} \\ (-1)^{n_1}\Sigma_{n_1} J_{n_1} \Sigma_{n_1} & (-1)^{n_1-1}\Sigma_{n_1} J_{n_1} \Sigma_{n_1}\end{pmatrix} = T_n(0)^{\mathrm{T}} \begin{pmatrix} & I_{n_1} \\ I_{n_1} & \end{pmatrix}\end{aligned}$$

sowie

$$\begin{aligned}\Sigma_n T_n(1)^{\mathrm{T}} \begin{pmatrix}\Sigma_{n_1} & \\ & \Sigma_{n_1}\end{pmatrix} &= \begin{pmatrix}\Sigma_{n_1} & \\ & (-1)^{n_1}\Sigma_{n_1}\end{pmatrix} \begin{pmatrix} \operatorname{diag}\mathbf{c}_{n_1} & -(\operatorname{diag}\mathbf{s}_{n_1}) J_{n_1} \\ J_{n_1}(\operatorname{diag}\mathbf{s}_{n_1}) & \operatorname{diag}(J_{n_1}\mathbf{c}_{n_1})\end{pmatrix} \begin{pmatrix}\Sigma_{n_1} & \\ & I_{n_1}\end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{n_1}\operatorname{diag}\mathbf{c}_{n_1} & -\Sigma_{n_1}(\operatorname{diag}\mathbf{s}_{n_1}) J_{n_1} \\ (-1)^{n_1}\Sigma_{n_1} J_{n_1}(\operatorname{diag}\mathbf{s}_{n_1}) & (-1)^{n_1}\Sigma_{n_1}\operatorname{diag}(J_{n_1}\mathbf{c}_{n_1})\end{pmatrix} \begin{pmatrix}\Sigma_{n_1} & \\ & I_{n_1}\end{pmatrix} \\ &= \begin{pmatrix} \operatorname{diag}\mathbf{c}_{n_1} & (\operatorname{diag}\mathbf{s}_{n_1}) J_{n_1} \\ -J_{n_1}(\operatorname{diag}\mathbf{s}_{n_1}) & \operatorname{diag}(J_{n_1}\mathbf{c}_{n_1})\end{pmatrix} \begin{pmatrix} I_{n_1} & \\ & (-1)^{n_1}\Sigma_{n_1}\end{pmatrix}.\end{aligned}$$

Insgesamt erhalten wir somit die nachstehenden beiden Faktorisierungpaare, wobei das zweite Paar wieder mittels Transponieren aus dem ersten Paar erhalten wird.

Folgerung 1.8. *Sei $n \in \mathbb{N}$ gerade mit $n \geq 4$ gegeben und $n_1 := \frac{n}{2}$. Weiter seien $\mathbf{c}_{n_1}, \mathbf{s}_{n_1}$ und $T_n(0)$ wie in Lemma* 1.6 *definiert. Dann gelten*

$$S_n^{\mathrm{III}} = T_n(0)^{\mathrm{T}} \begin{pmatrix} S_{n_1}^{\mathrm{IV}} & \\ & S_{n_1}^{\mathrm{III}}\end{pmatrix} P_n, \qquad S_n^{\mathrm{IV}} = \breve{T}_n(1) \begin{pmatrix} S_{n_1}^{\mathrm{III}} & \\ & S_{n_1}^{\mathrm{III}}\end{pmatrix} \breve{A}_n(1) P_n \tag{1.20}$$

sowie

$$S_n^{\mathrm{II}} = P_n^{\mathrm{T}} \begin{pmatrix} S_{n_1}^{\mathrm{IV}} & \\ & S_{n_1}^{\mathrm{II}}\end{pmatrix} T_n(0)\,, \qquad S_n^{\mathrm{IV}} = P_n^{\mathrm{T}} \breve{A}_n(1)^{\mathrm{T}} \begin{pmatrix} S_{n_1}^{\mathrm{II}} & \\ & S_{n_1}^{\mathrm{II}}\end{pmatrix} \breve{T}_n(1)^{\mathrm{T}} \tag{1.21}$$

mit der modifizierten Additionsmatrix

$$\breve{A}_n(1) := \tfrac{1}{\sqrt{2}} \begin{pmatrix} J_{n_1} & \\ & \Sigma_{n_1}\end{pmatrix} \begin{pmatrix} & & & \sqrt{2} \\ & & J_{n_1-1} & J_{n_1-1} \\ & & -J_{n_1-1} & J_{n_1-1} \\ -\sqrt{2} & & & \end{pmatrix} \tag{1.22}$$

und der kreuzförmigen Drehmatrix

$$\breve{T}_n(1) = \begin{pmatrix} \operatorname{diag}\mathbf{c}_{n_1} & (\operatorname{diag}\mathbf{s}_{n_1}) J_{n_1} \\ -J_{n_1}(\operatorname{diag}\mathbf{s}_{n_1}) & \operatorname{diag}(J_{n_1}\mathbf{c}_{n_1})\end{pmatrix} \begin{pmatrix} I_{n_1} & \\ & (-1)^{n_1}\Sigma_{n_1}\end{pmatrix}. \tag{1.23}$$

Offensichtlich lassen sich die Faktorisierungen aus (1.15) für $n = 2^t$ rekursiv ineinander einsetzen, bis der mittlere Matrixfaktor zu einer Blockdiagonalmatrix wird, deren Blöcke entweder C_2^{II} oder C_2^{IV} sind. Auf diese Weise lassen sich sowohl C_n^{II} als auch C_n^{IV} jeweils durch ein Produkt von dünnbesetzten orthogonalen Matrizen darstellen. Diese Darstellungen bilden die Grundlage für entsprechend schnelle Algorithmen, wie sie im nächsten Abschnitt eingeführt werden. Analog lassen sich auch aus den Faktorisierungspaaren (1.19) – (1.21) Produktdarstellungen von C_n^{III}, C_n^{IV} sowie $S_n^{\mathrm{II}}, S_n^{\mathrm{III}}$ und S_n^{IV} finden, welche nur dünnbesetzte orthogonale Faktoren beinhalten. Überdies lassen sich diese Faktoren mittels geeigneter Permutationen und durch spalten- bzw. zeilenweise Vorzeichenskalierungen als direkte Summen von Drehmatrizen

$$Q_2(\varphi) := \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi)\end{pmatrix} \qquad \left(\varphi \in \left]0, \tfrac{\pi}{4}\right]\right) \tag{1.24}$$

ausdrücken. Dabei bezeichnet die *direkte Summe* zweier quadratischer Matrizen A und B von nicht notwendigerweise gleicher Größe die Blockdiagonalmatrix

$$A \oplus B \;:=\; \begin{pmatrix} A & \\ & B \end{pmatrix} \;.$$

Im nachstehenden Lemma finden sich nun die entsprechenden Darstellungen der an (1.15) beteiligten Matrizen, welche erneut ihre Orthogonalität aufzeigen und darüber hinaus die Einschränkung unserer Betrachtungen auf Drehmatrizen der Gestalt (1.24) rechtfertigen.

Lemma 1.9. *Es sei $n = 2^t$ mit einer natürlichen Zahl $t \geq 2$ gegeben. Dann existieren für die in Lemma* 1.6 *definierten Matrizen die Faktorisierungen*

$$\left.\begin{aligned}
T_n(0) &= P_n \Sigma_n \left(\bigoplus_{i=1}^{n_1} Q_2(\tfrac{\pi}{4}) \right) P_n^{\mathrm{T}} \; (I_{n_1} \oplus J_{n_1}), \\
T_n(1) &= (I_{n_1} \oplus J_{n_1} \Sigma_{n_1}) P_n \left(\bigoplus_{i=1}^{n_1} Q_2\left(\tfrac{(2i-1)\pi}{4n}\right) \right) P_n^{\mathrm{T}} \; (I_{n_1} \oplus J_{n_1}), \\
A_n(1) &= (1 \oplus P_{n-2} \oplus 1) \left(1 \oplus \left(\bigoplus_{i=1}^{n_1-1} Q_2(\tfrac{\pi}{4}) \right) \oplus (-1) \right) \left(1 \oplus P_{n-2}^{\mathrm{T}} \oplus 1\right) (I_{n_1} \oplus \Sigma_{n_1} J_{n_1}) \,.
\end{aligned}\right\} \tag{1.25}$$

Weiterhin gelten

$$C_2^{\mathrm{II}} \;=\; T_2(0) \;=\; \Sigma_2 Q_2(\tfrac{\pi}{4}), \qquad C_2^{\mathrm{IV}} \;=\; \Sigma_2 T_2(1) \;=\; \Sigma_2 Q_2(\tfrac{\pi}{8}). \tag{1.26}$$

Der **Beweis** besteht aus einfachem Nachrechnen und wird daher an dieser Stelle weggelassen. ■

Zu (1.25) ähnliche Faktorisierungen sind bereits in [42] zu finden, wobei dort $Q_2(\varphi)$ mit $R_2(\varphi)$ bezeichnet und $(I_{n_1} \oplus J_{n_1})P_n$ zu Q_n zusammengefasst wird.

Da offensichtlich $Q_2(\varphi)^{\mathrm{T}} = \Sigma_2 Q_2(\varphi) \Sigma_2$ erfüllt ist, lassen sich aus (1.25) in ähnlicher Weise Faktorsierungen für $T_n(0)^{\mathrm{T}}$, $T_n(1)^{\mathrm{T}}$ und $A_n(1)^{\mathrm{T}}$ sowie $\breve{T}_n(1)$ und $\breve{A}_n(1)$ angeben. Mit

$$\breve{A}_n(1) \;=\; (J_{n_1} \oplus J_{n_1})\, A_n(1)^{\mathrm{T}} J_n$$

und

$$\breve{T}_n(1) \;=\; \Sigma_n T_n(1)^{\mathrm{T}} \, (\Sigma_{n_1} \oplus \Sigma_{n_1})$$

erhalten wir insgesamt

Folgerung 1.10. *Es sei $n = 2^t$ mit einer natürlichen Zahl $t \geq 2$ gegeben. Dann existieren für die in den Folgerungen* 1.7 *und* 1.8 *verwendeten Matrizen die Faktorisierungen*

$$\left.\begin{aligned}
T_n(0)^{\mathrm{T}} &= (I_{n_1} \oplus J_{n_1}) P_n \Sigma_n \left(\bigoplus_{i=1}^{n_1} Q_2(\tfrac{\pi}{4}) \right) P_n^{\mathrm{T}} \;, \\
T_n(1)^{\mathrm{T}} &= (I_{n_1} \oplus J_{n_1}) P_n \Sigma_n \left(\bigoplus_{i=1}^{n_1} Q_2\left(\tfrac{(2i-1)\pi}{4n}\right) \right) \Sigma_n P_n^{\mathrm{T}} (I_{n_1} \oplus \Sigma_{n_1} J_{n_1}), \\
\breve{T}_n(1) &= (\Sigma_{n_1} \oplus \Sigma_{n_1} J_{n_1}) P_n \Sigma_n \left(\bigoplus_{i=1}^{n_1} Q_2\left(\tfrac{(2i-1)\pi}{4n}\right) \right) \Sigma_n P_n^{\mathrm{T}} (\Sigma_{n_1} \oplus J_{n_1})
\end{aligned}\right\} \tag{1.27}$$

sowie

$$\left.\begin{aligned}
A_n(1)^{\mathrm{T}} &= (I_{n_1} \oplus J_{n_1} \Sigma_{n_1}) \, (1 \oplus P_{n-2} \oplus 1) \\
&\qquad \times \left(1 \oplus \Sigma_{n-2} \left(\bigoplus_{i=1}^{n_1-1} Q_2(\tfrac{\pi}{4}) \right) \Sigma_{n-2} \oplus (-1) \right) \left(1 \oplus P_{n-2}^{\mathrm{T}} \oplus 1\right), \\
\breve{A}_n(1) &= (J_{n_1} \oplus \Sigma_{n_1}) \, (1 \oplus P_{n-2} \oplus 1) \\
&\qquad \times \left(1 \oplus \Sigma_{n-2} \left(\bigoplus_{i=1}^{n_1-1} Q_2(\tfrac{\pi}{4}) \right) \Sigma_{n-2} \oplus (-1) \right) \left(1 \oplus P_{n-2}^{\mathrm{T}} \oplus 1\right) J_n \;.
\end{aligned}\right\} \tag{1.28}$$

Weiterhin gelten

$$S_2^{\mathrm{II}} \;=\; T_2(0) \;=\; \Sigma_2 Q_2(\tfrac{\pi}{4}), \qquad S_2^{\mathrm{IV}} \;=\; Q_2(\tfrac{\pi}{8}) J_2. \tag{1.29}$$

In [54] bzw. [50] finden sich für die Matrizen C_n^{IV} und C_n^{III} nach geringfügiger Modifizierung jeweils eine vollständige, nichtrekursive Faktorisierung in dünnbesetzte, orthogonale Matrizen. Um diese angeben zu können, bedarf es noch der Einführung einiger Bezeichnungen. Das Produkt quadratischer Matrizen $A_k \in \mathbb{R}^{n\times n},\ k = 1, \ldots, n$, von rechts nach links kürzen wir mit

$$\prod_{k=1}^{n} A_k := A_n A_{n-1} \cdot \ldots \cdot A_2 A_1 \tag{1.30}$$

ab. Darüber hinaus bezeichnet das *Kronecker-Produkt* zweier Matrizen $A = (a_{ij})_{i,j=0}^{k-1,l-1} \in \mathbb{R}^{k\times l}$ und $B \in \mathbb{R}^{m\times n}$ die Blockmatrix

$$A \otimes B \;:=\; (a_{ij}B)_{i,j=0}^{k-1,l-1} \in \mathbb{R}^{km\times ln}\ . \tag{1.31}$$

Ähnlich wie bei der Faktorisierung der Fourier-Matrix benötigen wir für $\nu = 2^\tau$ ($\tau \geq 2$) eine modifizierte Bitumkehr-Matrix U_ν, die sich über die Rekursion

$$U_4 := (I_2 \oplus J_2)P_4\ , \qquad U_\nu := (I_{\frac{\nu}{2}} \oplus J_{\frac{\nu}{2}})P_\nu\big(U_{\frac{\nu}{2}} \otimes I_2\big)$$

definieren lässt. Die in der Produktdarstellung der Fourier-Matrix auftretende Bitumkehr-Matrix

$$\prod_{s=0}^{t-2} (I_{2^s} \otimes P_{n_s})$$

ist eine spezielle Permutationsmatrix der Größe $2^n \times 2^n$, welche ihren Namen der Eigenschaft verdankt, dass für jede Komponente des zu permutierenden Vektors der alte und der neue Index in der Binärdarstellung durch Vertauschung der Bit-Reihenfolge ineinander übergehen (vgl. [58], S. 36, [50], S. 26). Besitzt also ein Element b der Menge $\{0, \ldots, 2^t - 1\}$ die Binärdarstellung $(b_{t-1}, \ldots, b_0)_2$ mit $b_k \in \{0, 1\}, k = 0, \ldots, t-1$, so wird b durch die Bitumkehr-Permutation auf dasjenige Element der Menge $\{0, \ldots, 2^t - 1\}$ abgebildet, welches die Binärdarstellung $(b_0, \ldots, b_{t-1})_2$ besitzt. Analog ist die modifizierte Bitumkehr-Permutation dadurch gegeben, dass b auf dasjenige Element der Menge $\{0, \ldots, 2^t - 1\}$ abgebildet wird, welches die Binärdarstellung

$$(\langle b_0 + b_1\rangle_2, \langle b_1 + b_2\rangle_2, \ldots, \langle b_{t-2} + b_{t-1}\rangle_2, b_{t-1})_2$$

besitzt. Dabei bezeichnet $\langle z\rangle_n$ den nichtnegativen Rest von $z \in \mathbb{Z}$ modulo n. Die modifizierte Bitumkehr-Matrix U_ν erfüllt nun genau die Eigenschaft, dass für jede Komponente des zu permutierenden Vektors der neue Index aus dem alten durch Ausführen der modifizierten Bitumkehr-Permutation hervorgeht (vgl. [50], S. 29). Desweiteren tritt U_n bei der folgenden Faktorisierung von C_n^{IV} an der entsprechenden Position auf, an der die Bitumkehr-Matrix innerhalb der Faktorisierung der Fourier-Matrix steht.

Lemma 1.11 (vgl. [54], Example 8.10 und [50], Algorithmus 2.21)**.** *Es sei $n = 2^t$ mit einer natürlichen Zahl $t \geq 3$ gegeben. Weiter sei $n_s := 2^{t-s}$ für $s = 0, \ldots, t$. Dann gilt*

$$C_n^{\mathrm{IV}} \;=\; U_n \left(\prod_{s=1}^{t-1} \left(D_n^{(s)} B_n^{(s)} P_n^{(s)} \right) \right) D_n^{(0)} P_n^{(0)} \tag{1.32}$$

mit den Permutationsmatrizen

$$\begin{aligned} P_n^{(0)} \;&:=\; ((I_{n_1} \oplus J_{n_1})P_n)^{\mathrm{T}}\ , \\ P_n^{(s)} \;&:=\; I_{2^{s-1}} \otimes (I_{n_{s+1}} \oplus J_{n_{s+1}}) \otimes I_2\ , \qquad s = 1, \ldots, t-1, \end{aligned}$$

den modifizierten Butterfly-Matrizen

$$B_n^{(s)} := I_{2^{s-1}} \otimes T_2(0) \otimes I_{n_s}\ , \qquad s = 1, \ldots, t-1,$$

und den Drehmatrizen

$$\begin{aligned} D_n^{(0)} &:= \bigoplus_{j=0}^{n_1-1} \Sigma_2^{j+1} Q_2\left(\frac{(2j+1)\pi}{4n}\right) = \bigoplus_{j=0}^{n_1-1} \begin{pmatrix} \cos\left(\frac{(2j+1)\pi}{4n}\right) & \sin\left(\frac{(2j+1)\pi}{4n}\right) \\ (-1)^j \sin\left(\frac{(2j+1)\pi}{4n}\right) & -(-1)^j \cos\left(\frac{(2j+1)\pi}{4n}\right) \end{pmatrix}, \\ D_n^{(s)} &:= I_{2^{s-1}} \otimes \left(I_{n_s} \oplus \bigoplus_{j=0}^{\frac{n_s}{2}-1} \Sigma_2^{j+1} Q_2\left(\frac{(2j+1)\pi}{2n_s}\right) \Sigma_2^j \right) \\ &= I_{2^{s-1}} \otimes \left(I_{n_s} \oplus \bigoplus_{j=0}^{\frac{n_s}{2}-1} \begin{pmatrix} \cos\left(\frac{(2j+1)\pi}{2n_s}\right) & (-1)^j \sin\left(\frac{(2j+1)\pi}{2n_s}\right) \\ (-1)^j \sin\left(\frac{(2j+1)\pi}{2n_s}\right) & -\cos\left(\frac{(2j+1)\pi}{2n_s}\right) \end{pmatrix} \right), \qquad s = 1, \ldots, t-1. \end{aligned}$$

Mit $L_4 := (I_2 \oplus T_2(0))T_4(0)(I_2 \oplus J_2)$ *gilt weiterhin*

$$C_4^{\mathrm{IV}} = U_4 L_4 D_4^{(0)} U_4^{\mathrm{T}} . \tag{1.33}$$

Wiederum sind hier alle beteiligten Matrizen sowohl orthogonal als auch dünnbesetzt. Dies ist für die Permutationsmatrizen $P_n^{(s)}$, $s = 0, \ldots, t$, offensichtlich. Da die Kronecker-Multiplikation mit einer Einheitsmatrix von links lediglich eine Blockdiagonalmatrix mit identischen Blöcken liefert, folgt hieraus auch die Orthogonalität bzw. Dünnbesetztheit der Drehmatrizen $D_n^{(s)}$, $s = 0, \ldots, t-1$, aus den Eigenschaften ihrer Blöcke. Mit (1.26) und beispielweise nach [54, Theorem 8.1] gilt schließlich

$$B_n^{(s)} = I_{2^{s-1}} \otimes \Sigma_2 Q\left(\tfrac{\pi}{4}\right) \otimes I_{n_s} = \left(I_{2^{s-1}} \otimes P_{2^{t-s+1}}\right)\left(I_{n_1} \otimes \Sigma_2 Q\left(\tfrac{\pi}{4}\right)\right)\left(I_{2^{s-1}} \otimes P_{2^{t-s+1}}^{\mathrm{T}}\right) \tag{1.34}$$

für die Butterfly-Matrizen $B_n^{(s)}$, $s = 1, \ldots, t-1$, so dass hier mit derselben Argumentation die Orthogonalität und Dünnbesetztheit folgt. Zusätzlich haben wir in (1.34) und in Lemma 1.11 für alle beteiligten Matrizen Faktorisierungen in lediglich ebene Drehungen enthaltende Blockdiagonalmatrizen sowie in Permutations- und Vorzeichenskalierungsmatrizen vorliegen.
Analog sind auch alle an der folgenden Darstellung von C_n^{III} beteiligten Matrizen faktorisierbar. Demzufolge können wir auch hier unsere Betrachtungen auf ebene Drehungen beschränken.

Lemma 1.12 (vgl. [54], Example 8.11 und [50], Algorithmus 2.23)**.** *Es sei* $n = 2^t$ *mit einer natürlichen Zahl* $t \geq 3$ *gegeben. Weiter sei* $n_s := 2^{t-s}$ *für* $s = 0, \ldots, t$*. Mit den Bezeichnungen aus Lemma* 1.11 *gilt dann*

$$C_n^{\mathrm{III}} = \left(\prod_{s=1}^{t} \left(T_{2^s}(0)^{\mathrm{T}} \oplus I_{n-2^s}\right)\right) \tilde{P}_n^{(t-1)} \left(\prod_{s=1}^{t-2} \left(\tilde{D}_n^{(s)} \tilde{B}_n^{(s)} \tilde{P}_n^{(s)}\right)\right) \tilde{D}_n^{(0)} \tilde{P}_n^{(0)} \tag{1.35}$$

mit den Permutationsmatrizen

$$\begin{aligned} \tilde{P}_n^{(0)} &:= \left(I_4 \oplus \bigoplus_{s=1}^{t-2} P_{2^{s+1}}^{\mathrm{T}} (I_{2^s} \oplus J_{2^s})\right) \prod_{s=0}^{t-2} (P_{n_s} \oplus I_{n-n_s}), \qquad \tilde{P}_n^{(t-1)} := I_4 \oplus \bigoplus_{s=2}^{t-1} U_{2^s} , \\ \tilde{P}_n^{(s)} &:= I_{n_s} \oplus \left(I_{2^s-1} \otimes (I_{n_{s+2}} \oplus J_{n_{s+2}}) \otimes I_2\right) , \qquad s = 1, \ldots, t-2, \end{aligned}$$

den modifizierten Butterfly-Matrizen

$$\tilde{B}_n^{(s)} := I_{n_s} \oplus \left(I_{2^s-1} \otimes T_2(0) \otimes I_{n_{s+1}}\right) , \qquad s = 1, \ldots, t-1,$$

und den Drehmatrizen

$$\begin{aligned} \tilde{D}_n^{(0)} &:= I_2 \oplus \bigoplus_{s=1}^{t-1} D_{2^s}^{(0)} = I_2 \oplus \bigoplus_{s=1}^{t-1} \left(\bigoplus_{j=0}^{2^{s-1}-1} \Sigma_2^{j+1} Q\left(\frac{(2j+1)\pi}{2^{s+2}}\right) \right), \\ \tilde{D}_n^{(s)} &:= I_{n_s} \oplus \left(I_{2^s-1} \otimes \left(I_{n_{s+1}} \oplus \bigoplus_{j=0}^{n_{s+2}-1} \Sigma_2^{j+1} Q\left(\frac{(2j+1)\pi}{n_s}\right) \Sigma_2^j \right) \right) , \qquad s = 1, \ldots, t-2. \end{aligned}$$

Weiterhin gilt

$$C_4^{\mathrm{III}} = T_4(0)^{\mathrm{T}} \tilde{D}_4^{(0)} P_4 . \tag{1.36}$$

Zu bemerken ist, dass die Matrizen $\tilde{B}_n^{(s)}$ in [54] nur als Vielfache von orthogonalen Matrizen definiert werden, während sie hier orthogonal sind.
Offensichtlich ergeben sich wiederum mittels Transponieren weitere Darstellungen für die Matrizen C_n^{II} und C_n^{IV}.

Folgerung 1.13. *Mit den Bezeichnungen aus Lemma* 1.11 *und* 1.12 *gelten die Faktorisierungen*

$$C_n^{\mathrm{IV}} = P_n^{(0)^{\mathrm{T}}} D_n^{(0)^{\mathrm{T}}} \left(\prod_{s=1}^{t-1} \left(P_n^{(t-s)^{\mathrm{T}}} B_n^{(t-s)^{\mathrm{T}}} D_n^{(t-s)^{\mathrm{T}}} \right) \right) U_n^{\mathrm{T}} , \tag{1.37}$$

$$C_n^{\mathrm{II}} = \tilde{P}_n^{(0)^{\mathrm{T}}} \tilde{D}_n^{(0)^{\mathrm{T}}} \left(\prod_{s=2}^{t-1} \left(\tilde{P}_n^{(t-s)^{\mathrm{T}}} \tilde{B}_n^{(t-s)^{\mathrm{T}}} \tilde{D}_n^{(t-s)^{\mathrm{T}}} \right) \right) \tilde{P}_n^{(t-1)^{\mathrm{T}}} \left(\prod_{s=0}^{t-1} (T_{n_s}(0) \oplus I_{n-n_s}) \right). \tag{1.38}$$

Weiterhin gelten

$$C_4^{\mathrm{IV}} = U_4 D_4^{(0)^{\mathrm{T}}} L_4^{\mathrm{T}} U_4 , \qquad C_4^{\mathrm{II}} = P_4^{\mathrm{T}} \tilde{D}_4^{(0)^{\mathrm{T}}} T_4(0) . \tag{1.39}$$

1.3 Schnelle DCT- und DST-Algorithmen

Analog zur diskreten Fourier-Transformation der Länge $n \in \mathbb{N}$, die mit Hilfe der Fourier-Matrix (1.1) als lineare Abbildung von $\mathbb{C}^n$ in sich erklärt wird, definieren wir nun die *diskreten trigonometrischen Transformationen.*

Definition 1.14. *Unter einer* diskreten Kosinus-Transformation vom Typ $*$ ($* \in \{\mathrm{II}, \mathrm{III}, \mathrm{IV}\}$) *der Länge* $n \in \mathbb{N}$ (DCT-$*(n)$) *verstehen wir diejenige lineare Abbildung von* $\mathbb{R}^n$ *in sich, die jedem Vektor* $\mathbf{x} := (x_j)_{j=0}^{n-1} \in \mathbb{R}^n$ *den Vektor* $\mathbf{y} := (y_j)_{j=0}^{n-1}$ *mit*

$$\mathbf{y} := C_n^* \mathbf{x}$$

zuordnet. Analog definieren wir eine diskrete Sinus-Transformation vom Typ $*$ ($* \in \{\mathrm{II}, \mathrm{III}, \mathrm{IV}\}$) *der Länge* $n \in \mathbb{N}$ (DST-$*(n)$) *als diejenige lineare Abbildung von* $\mathbb{R}^n$ *in sich, die jedem Vektor* $\mathbf{x} := (x_j)_{j=0}^{n-1} \in \mathbb{R}^n$ *den Vektor* $\mathbf{y} := (y_j)_{j=0}^{n-1}$ *mit*

$$\mathbf{y} := S_n^* \mathbf{x}$$

zuordnet.

Aufgrund der Vollbesetztheit der Matrizen (1.2) – (1.7) würde eine naive Implementierung der entsprechenden Transformationen einen Rechenaufwand von n^2 Multiplikationen und $n(n-1)$ Additionen pro transformierten Vektor der Länge n erfordern. Ähnlich wie bei der DFT ist es jedoch möglich, mit Hilfe von Teile-und-Herrsche-Strategien den Rechenaufwand zu reduzieren. Dies lässt sich über entsprechende Matrixfaktorisierungen darstellen, wobei die beteiligten Matrizen möglichst wenige Nichtnulleinträge besitzen sollten. Lemmata 1.11 und 1.12 liefern für $n = 2^t$ bereits zwei derartige Faktorisierungen. Aus (1.32) und (1.33) lässt sich ein iterativer Algorithmus zur Berechnung der DCT-IV ableiten.

Algorithmus 1.15 ($\cos -\mathrm{IV}(\mathbf{x}, n)$)**.**
Eingabe: $n = 2^t$ ($t \geq 1$), $\mathbf{x} \in \mathbb{R}^n$.

- *Falls* $n = 2$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{IV}} \mathbf{x}$.
- *Falls* $n = 4$ *, berechne* $\mathbf{y} \leftarrow U_4 L_4 D_4^{(0)} U_4^{\mathrm{T}} \mathbf{x}$.
- *Falls* $n \geq 8$ *, setze* $n_1 := \frac{n}{2}$.
  - *Permutiere und berechne*

$$\begin{aligned} \mathbf{v}' &\leftarrow P_n^{(0)} \mathbf{x}, \\ \mathbf{x}^{(1)} &\leftarrow D_n^{(0)} \mathbf{v}'. \end{aligned}$$

– *Für* $s = 1, \ldots, t-1$ *permutiere und berechne*

$$\begin{aligned} \mathbf{v}' &\leftarrow P_n^{(s)}\, \mathbf{x}^{(s)}, \\ \mathbf{v}'' &\leftarrow B_n^{(s)}\, \mathbf{v}', \\ \mathbf{x}^{(s+1)} &\leftarrow D_n^{(s)}\, \mathbf{v}''. \end{aligned}$$

– *Permutiere*

$$\mathbf{y} \leftarrow U_n\, \mathbf{x}^{(t)}.$$

Ausgabe: $\mathbf{y} = C_n^{\mathrm{IV}}\mathbf{x}$.

Aufgrund der Orthogonalität und Symmetrie der Matrix C_n^{IV} liefert eine zweimalige Anwendung auf einen Vektor $\mathbf{x}$ von Algorithmus 1.15 im Idealfall wiederum den Ausgangsvektor $\mathbf{x}$. Auf diese Weise kann die Korrektheit der Implementierung leicht überprüft werden.
Weiterhin erhalten wir aus (1.35) und (1.36) ebenso einen iterativen Algorithmus zur Berechnung der DCT-III.

Algorithmus 1.16 $(\cos -\mathrm{III}(\mathbf{x}, n))$**.**
Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$.

- *Falls* $n = 2$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 4$ *, berechne* $\mathbf{y} \leftarrow T_4^{\mathrm{T}}(0)\tilde{D}_4^{(0)}P_4\mathbf{x}$.
- *Falls* $n \geq 8$ *, setze* $n_1 := \frac{n}{2}$.
  – *Permutiere und berechne*

$$\begin{aligned} \mathbf{v}' &\leftarrow \tilde{P}_n^{(0)}\, \mathbf{x}, \\ \mathbf{x}^{(1)} &\leftarrow \tilde{D}_n^{(0)}\, \mathbf{v}'. \end{aligned}$$

  – *Für* $s = 1, \ldots, t-2$ *permutiere und berechne*

$$\begin{aligned} \mathbf{v}' &\leftarrow \tilde{P}_n^{(s)}\, \mathbf{x}^{(s)}, \\ \mathbf{v}'' &\leftarrow \tilde{B}_n^{(s)}\, \mathbf{v}', \\ \mathbf{x}^{(s+1)} &\leftarrow \tilde{D}_n^{(s)}\, \mathbf{v}''. \end{aligned}$$

  – *Permutiere*

$$\mathbf{x}^{(t)} \leftarrow \tilde{P}_n^{(t-1)}\, \mathbf{x}^{(t-1)}.$$

  – *Für* $s = 1, \ldots, t$ *berechne*

$$\mathbf{x}^{(t+s)} \leftarrow \left(T_{2^s}(0)^{\mathrm{T}} \oplus I_{n-2^s}\right) \mathbf{x}^{(t+s-1)}.$$

  – *Setze* $\mathbf{y} \leftarrow \mathbf{x}^{(2t)}$.

Ausgabe: $\mathbf{y} = C_n^{\mathrm{III}}\mathbf{x}$.

Bei der Implementierung kann hierbei ausgenutzt werden, dass innerhalb der letzten Schleife im Schritt s lediglich die ersten 2^s Komponenten verändert werden. Aufgrund der Orthogonalität aller beteiligten Matrizen kann für die inverse Transformation DCT-II ein entsprechender iterativer Algorithmus aus (1.38) und (1.39) abgelesen werden.

Algorithmus 1.17 $(\cos -\mathrm{II}(\mathbf{x}, n))$**.**
Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$.

- *Falls* $n = 2$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 4$ *, berechne* $\mathbf{y} \leftarrow P_4^{\mathrm{T}}\tilde{D}_4^{(0)^{\mathrm{T}}}T_4(0)\mathbf{x}$.
- *Falls* $n \geq 8$ *, setze* $n_1 := \frac{n}{2}$ *und* $\mathbf{x}^{(0)} := \mathbf{x}$.

– *Für* $s = 0, \ldots, t-1$ *berechne*

$$\mathbf{x}^{(s+1)} \leftarrow (T_{n_s}(0) \oplus I_{n-n_s})\, \mathbf{x}^{(s)}.$$

– *Permutiere*

$$\mathbf{x}^{(t+1)} \leftarrow \tilde{P}_n^{(t-1)^{\mathrm{T}}}\, \mathbf{x}^{(t)}.$$

– *Für* $s = 2, \ldots, t-1$ *berechne und permutiere*

$$\begin{aligned} \mathbf{v}' &\leftarrow \tilde{D}_n^{(t-s)^{\mathrm{T}}}\, \mathbf{x}^{(t+s-1)}, \\ \mathbf{v}'' &\leftarrow \tilde{B}_n^{(t-s)^{\mathrm{T}}}\, \mathbf{v}', \\ \mathbf{x}^{(t+s)} &\leftarrow \tilde{P}_n^{(t-s)^{\mathrm{T}}}\, \mathbf{v}''. \end{aligned}$$

– *Permutiere und berechne*

$$\begin{aligned} \mathbf{v}' &\leftarrow \tilde{D}_n^{(0)^{\mathrm{T}}}\, \mathbf{x}^{(2t-1)}, \\ \mathbf{y} &\leftarrow \tilde{P}_n^{(0)^{\mathrm{T}}}\, \mathbf{v}'. \end{aligned}$$

Ausgabe: $\mathbf{y} = C_n^{\mathrm{II}}\mathbf{x}$.

Ein weiterer iterativer Algorithmus zur Berechnung der DCT-IV lässt sich aus (1.37) und (1.39) ableiten.

Algorithmus 1.18 $(\cos -\mathrm{IVT}(\mathbf{x}, n))$**.**
Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$.

- *Falls* $n = 2$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{IV}}\mathbf{x}$.
- *Falls* $n = 4$ *, berechne* $\mathbf{y} \leftarrow U_4 D_4^{(0)^{\mathrm{T}}} L_4^{\mathrm{T}} U_4 \mathbf{x}$.
- *Falls* $n \geq 8$ *, setze* $n_1 := \frac{n}{2}$.
  – *Permutiere*

$$\mathbf{x}^{(1)} \leftarrow U_n^{\mathrm{T}}\, \mathbf{x}.$$

  – *Für* $s = 1, \ldots, t-1$ *berechne und permutiere*

$$\begin{aligned} \mathbf{v}' &\leftarrow D_n^{(t-s)^{\mathrm{T}}}\, \mathbf{x}^{(s)}, \\ \mathbf{v}'' &\leftarrow B_n^{(t-s)^{\mathrm{T}}}\, \mathbf{v}', \\ \mathbf{x}^{(s+1)} &\leftarrow P_n^{(t-s)^{\mathrm{T}}}\, \mathbf{v}''. \end{aligned}$$

  – *Berechne und permutiere*

$$\begin{aligned} \mathbf{v}' &\leftarrow D_n^{(0)^{\mathrm{T}}}\, \mathbf{x}^{(t)}, \\ \mathbf{y} &\leftarrow P_n^{(0)^{\mathrm{T}}}\, \mathbf{v}'. \end{aligned}$$

Ausgabe: $\mathbf{y} = C_n^{\mathrm{IV}}\mathbf{x}$.

Die Algorithmen 1.15 – 1.18 enthalten jeweils mindestens eine Schleife, welche eine vom Parameter n abhängige endliche Anzahl durchlaufen wird. Neben einer solchen „iterativen“ Implementierung eines Verfahrens besteht in manchen Fällen auch die Möglichkeit, einen Algorithmus rekursiv zu entwerfen. Dabei nennen wir einen Algorithmus rekursiv, wenn er sich in einem Durchlauf mindestens einmal (ggf. mit anderen Parametern) selbst aufruft. Insbesondere eignen sich die in Lemma 1.6 und in den Folgerungen 1.7 und 1.8 angegebenen Faktorisierungen für die Konstruktion von rekursiven Algorithmen für die entsprechende DCT bzw. DST. Aus dem Faktorisierungspaar (1.15) erhalten wir für $n = 2^t$ einen ersten schnellen rekursiven Algorithmus, welcher je nach Wahl des Parameters $k \in \{0, 1\}$ die DCT-II bzw. DCT-IV für einen Vektor der Länge $n = 2^t$ ausführt. Dieser gehört offensichtlich zur Klasse der Zweierpotenz-Algorithmen, zu denen beispielsweise der Cooley-Tukey- sowie der Gentleman-Sande-Algorithmus für die DFT zählen.

Algorithmus 1.19 ($\cos -\mathrm{II/IV}(\mathbf{x}, n, k)$)**.**

Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$, $k \in \{0, 1\}$.

- *Falls* $n = 2$ *und* $k = 0$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 2$ *und* $k = 1$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{IV}}\mathbf{x}$.
- *Falls* $n \geq 4$ *, setze* $n_1 := \frac{n}{2}$.
  - *Für* $k = 0$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow T_n(0)\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \cos -\mathrm{II/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \cos -\mathrm{II/IV}((u_j)_{j=n_1}^{n-1}, n_1, 1),\\ \mathbf{y} &\leftarrow P_n^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

  - *Für* $k = 1$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow T_n(1)\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \cos -\mathrm{II/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \cos -\mathrm{II/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow P_n^{\mathrm{T}} A_n(1)((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

Ausgabe: $\mathbf{y} = \begin{cases} C_n^{\mathrm{II}}\mathbf{x} & \text{für } k = 0,\\ C_n^{\mathrm{IV}}\mathbf{x} & \text{für } k = 1.\end{cases}$

Weitere Zweierpotenz-Algorithmen für DCT und DST vom Typ II–IV lassen sich analog angeben. Aufgrund der bereits erwähnten Orthogonalität erhalten wir aus dem Faktorisierungspaar (1.19) sofort einen Algorithmus für die entsprechende Umkehrtransformation, die im Fall $k = 0$ der DCT-III bzw. im Fall $k = 1$ wiederum der DCT-IV entspricht.

Algorithmus 1.20 ($\cos -\mathrm{III/IV}(\mathbf{x}, n, k)$)**.**

Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$, $k \in \{0, 1\}$.

- *Falls* $n = 2$ *und* $k = 0$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 2$ *und* $k = 1$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{IV}}\mathbf{x}$.
- *Falls* $n \geq 4$ *, setze* $n_1 := \frac{n}{2}$.
  - *Für* $k = 0$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow P_n\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \cos -\mathrm{III/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \cos -\mathrm{III/IV}((u_j)_{j=n_1}^{n-1}, n_1, 1),\\ \mathbf{y} &\leftarrow (T_n(0))^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

  - *Für* $k = 1$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow A_n(1)^{\mathrm{T}} P_n\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \cos -\mathrm{III/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \cos -\mathrm{III/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow T_n(1)^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

Ausgabe: $\mathbf{y} = \begin{cases} C_n^{\mathrm{III}}\mathbf{x} & \text{für } k = 0,\\ C_n^{\mathrm{IV}}\mathbf{x} & \text{für } k = 1.\end{cases}$

Wir verwenden hier außerdem, dass die Matrix C_2^{III} mit C_2^{II} übereinstimmt. Aufgrund der Symmetrie von C_n^{IV} sind uns somit gleich zwei rekursive Algorithmen zur Ausführung der DCT-IV gegeben.
Analog lassen sich für die DST vom Typ II–IV die folgenden zwei rekursiven Algorithmen aus den beiden Faktorisierungspaaren (1.20) und (1.21) herleiten. Wir verwenden dabei, dass die Matrizen S_2^{II} und S_2^{III} ebenso mit C_2^{II} zusammenfallen.

Algorithmus 1.21 $(\sin -\mathrm{III/IV}(\mathbf{x}, n, k))$**.**

Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$, $k \in \{0, 1\}$.

- *Falls* $n = 2$ *und* $k = 0$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 2$ *und* $k = 1$ *, berechne* $\mathbf{y} \leftarrow S_2^{\mathrm{IV}}\mathbf{x}$.
- *Falls* $n \geq 4$ *, setze* $n_1 := \frac{n}{2}$.
  - *Für* $k = 0$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow P_n\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \sin -\mathrm{III/IV}((u_j)_{j=0}^{n_1-1}, n_1, 1),\\ \mathbf{v}'' &\leftarrow \sin -\mathrm{III/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow (T_n(0))^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

  - *Für* $k = 1$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow \breve{A}_n(1)P_n\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \sin -\mathrm{III/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \sin -\mathrm{III/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow \breve{T}_n(1)((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

Ausgabe: $\mathbf{y} = \begin{cases} S_n^{\mathrm{III}}\mathbf{x} & \text{für } k = 0,\\ S_n^{\mathrm{IV}}\mathbf{x} & \text{für } k = 1.\end{cases}$

Algorithmus 1.22 $(\sin -\mathrm{II/IV}(\mathbf{x}, n, k))$**.**

Eingabe: $n = 2^t$ $(t \geq 1)$, $\mathbf{x} \in \mathbb{R}^n$, $k \in \{0, 1\}$.

- *Falls* $n = 2$ *und* $k = 0$ *, berechne* $\mathbf{y} \leftarrow C_2^{\mathrm{II}}\mathbf{x}$.
- *Falls* $n = 2$ *und* $k = 1$ *, berechne* $\mathbf{y} \leftarrow S_2^{\mathrm{IV}}\mathbf{x}$.
- *Falls* $n \geq 4$ *, setze* $n_1 := \frac{n}{2}$.
  - *Für* $k = 0$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow (T_n(0))\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \sin -\mathrm{II/IV}((u_j)_{j=0}^{n_1-1}, n_1, 1),\\ \mathbf{v}'' &\leftarrow \sin -\mathrm{II/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow P_n^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

  - *Für* $k = 1$ *berechne*

$$\begin{aligned}(u_j)_{j=0}^{n-1} &\leftarrow \breve{T}_n(1)^{\mathrm{T}}\ \mathbf{x},\\ \mathbf{v}' &\leftarrow \sin -\mathrm{II/IV}((u_j)_{j=0}^{n_1-1}, n_1, 0),\\ \mathbf{v}'' &\leftarrow \sin -\mathrm{II/IV}((u_j)_{j=n_1}^{n-1}, n_1, 0),\\ \mathbf{y} &\leftarrow P_n^{\mathrm{T}}\breve{A}_n(1)^{\mathrm{T}}((\mathbf{v}')^{\mathrm{T}}, (\mathbf{v}'')^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

$$\textit{Ausgabe: } \mathbf{y} \;=\; \begin{cases} S_n^{\mathrm{II}}\mathbf{x} & \textit{für } k=0, \\ S_n^{\mathrm{IV}}\mathbf{x} & \textit{für } k=1. \end{cases}$$

Die jeweiligen Multiplikationen mit den entsprechend dünnbesetzten Matrizen können aufgrund ihrer speziellen Struktur als unabhängige Prozeduren implementiert werden, so dass keine Matrizen abgespeichert werden müssen. Auf diese Weise können die Algorithmen 1.19 – 1.22 auch noch für große Vektorlängen $n = 2^t$ $(t \gg 1)$ realisiert werden.
Die Algorithmen 1.19 – 1.22 sind aufgrund ihrer rekursiven Gestalt, welche aus der Teile-und-Herrsche-Strategie resultiert, sehr schnell und bequem zu implementieren. Zur Untersuchung ihres Rundungsfehlerverhaltens ist jedoch eine iterative Darstellung der Gestalt

$$\prod_{j=1}^{\nu} A^{(j)} \;:=\; A^{(\nu)} \dots A^{(2)} A^{(1)}$$

mit Matrizen $A^{(j)} \in \mathbb{R}^{n\times n}$, $j = 1, \dots, \nu$, wesentlich geeigneter ([24], §24.1). Insbesondere stellt sich dabei heraus, dass die an den jeweiligen Faktorisierungen beteiligten Matrizen ebenfalls dünnbesetzt, orthogonal und von sehr einfacher Gestalt sind.
Der Übersichtlichkeit halber führen wir wiederum einige Abkürzungen ein. Für $n := 2^t$ $(t \geq 3)$ definieren wir $n_s := 2^{t-s}$ $(s = 0, \dots, t-1)$. Weiterhin bezeichne

$$P_n(s) := \bigoplus_{i=1}^{2^s} P_{n_s}^{\mathrm{T}} \quad (s = 0, \dots, t-2) \tag{1.40}$$

die Transponierte der *verallgemeinerten 2-Schritt-Permutationsmatrix der Ordnung n*, wobei P_{n_s} die in (1.14) angegebene Permutationsmatrix ist. In Anlehnung an [43] (vgl. auch [58], 115 ff) verwenden wir die Binärvektoren

$$\boldsymbol{\beta}_s = (\beta_s(1), \dots, \beta_s(2^s)) \in \{0,1\}^{2^s} \qquad (s \in \{0, \dots, t-1\}), \tag{1.41}$$

welche der Rekursion

$$\boldsymbol{\beta}_0 := (0), \quad \boldsymbol{\beta}_{s+1} := (\boldsymbol{\beta}_s,\ \tilde{\boldsymbol{\beta}}_s) \qquad (s = 0, \dots, t-2) \tag{1.42}$$

genügen. Dabei bezeichnet $\tilde{\boldsymbol{\beta}}_s := \left(\tilde{\beta}_s(1), \dots, \tilde{\beta}_s(2^s)\right)$ denjenigen Binärvektor, welcher bis auf den letzten Eintrag mit $\boldsymbol{\beta}_s$ übereinstimmt, d.h. welcher durch

$$\tilde{\beta}_s(i) := \beta_s(i), \qquad i = 1, \dots, 2^s - 1, \qquad \tilde{\beta}_s(2^s) := 1 - \beta_s(2^s)$$

gegeben ist. In Abbildung 1.1 ist der Zusammenhang der Binärvektoren $\boldsymbol{\beta}_s$ $(s = 0, 1, 2, 3)$ schematisch veranschaulicht, wobei sich die jeweiligen Komponenten zeilenweise von links nach rechts ablesen lassen.

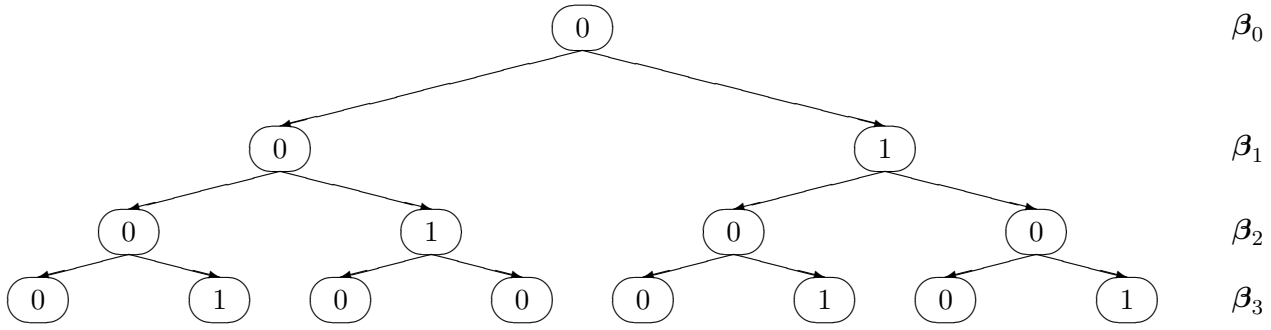


Abbildung 1.1: Diagramm zur Herleitung der Binärvektoren $\boldsymbol{\beta}_s$ $(s = 0, 1, 2, 3)$.

Zu jedem dieser Binärvektoren $\boldsymbol{\beta}_s$ definieren wir nun die dünnbesetzten orthogonalen Matrizen

$$\left.\begin{array}{rcll} A_n(\boldsymbol{\beta}_s) & := & P_n(s)\,\big(A_{n_s}(\beta_s(1)) \oplus \ldots \oplus A_{n_s}(\beta_s(2^s))\big) & (s=0,\ldots,t-2), \\ T_n(\boldsymbol{\beta}_s) & := & T_{n_s}(\beta_s(1)) \oplus \ldots \oplus T_{n_s}(\beta_s(2^s)) & (s=0,\ldots,t-2), \\ C_n(\boldsymbol{\beta}_s) & := & C_{n_s}(\beta_s(1)) \oplus \ldots \oplus C_{n_s}(\beta_s(2^s)) & (s=0,\ldots,t-1) \end{array}\right\} \tag{1.43}$$

mit $A_{n_s}(0) := I_{n_s}$, $C_{n_s}(0) := C_{n_s}^{\mathrm{II}}$ und $C_{n_s}(1) := C_{n_s}^{\mathrm{IV}}$. Die blockweise Anwendung der rekursiven Faktorisierungen (1.15) und (1.19) führt nun zu folgendem Zusammenhang.

Lemma 1.23. *Für die in* (1.43) *definierten Matrizen gilt*

$$C_n(\boldsymbol{\beta}_s) = \; A_n(\boldsymbol{\beta}_s)\; C_n(\boldsymbol{\beta}_{s+1})\; T_n(\boldsymbol{\beta}_s) \qquad (s=0,\ldots,t-2). \tag{1.44}$$

In Abbildung 1.2 ist schematisch dargestellt, wie sich die Matrix C_n^{II} durch Anwenden von (1.44) für $s=0,1,2$ auf eine Faktorisierung zurückführen lässt, welche die direkte Summe von 8 Matrizen der Größe $n_3 \times n_3$ enthält.

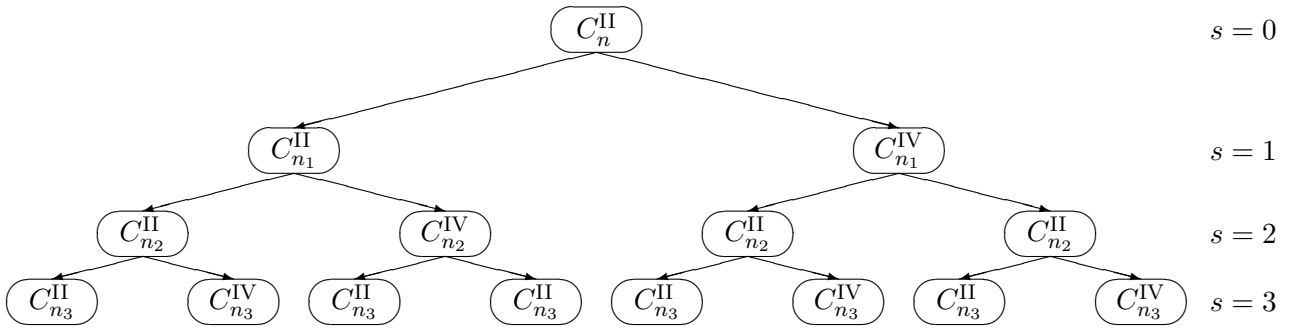


Abbildung 1.2: Faktorisierungsschema von C_n^{II} bis zum Level $s=3$.

Es ist zu beachten, dass (1.44) mit Gleichung (4.4) in [43] übereinstimmt, da die Permutationsmatrizen $P_n(s)$ hier bereits in den gemäß (1.43) definierten Matrizen $A_n(\boldsymbol{\beta}_s)$ enthalten sind. Definieren wir weiterhin $T_n(\boldsymbol{\beta}_{t-1}) := C_n(\boldsymbol{\beta}_{t-1})$, so führt die wiederholte Anwendung von (1.44) bzw. der transponierten Gleichung offenbar auf die gewünschten iterativen Darstellungen von $C_n^{\mathrm{II}} = C_n(\boldsymbol{\beta}_0)$ und C_n^{III}.

Satz 1.24 (vgl. [43], Theorem 4.2)**.** *Die rekursiven Algorithmen* 1.19 *und* 1.20 *basieren für* $k=0$ *auf den lediglich dünnbesetzte orthogonale Matrizen enthaltenden iterativen Faktorisierungen*

$$\left.\begin{array}{rcl} C_n^{\mathrm{II}} & = & A_n(\boldsymbol{\beta}_0)\;\ldots\;A_n(\boldsymbol{\beta}_{t-2})\;T_n(\boldsymbol{\beta}_{t-1})\;T_n(\boldsymbol{\beta}_{t-2})\;\ldots\;T_n(\boldsymbol{\beta}_0)\;, \\ C_n^{\mathrm{III}} & = & T_n(\boldsymbol{\beta}_0)^{\mathrm{T}}\;\ldots\;T_n(\boldsymbol{\beta}_{t-2})^{\mathrm{T}}\;T_n(\boldsymbol{\beta}_{t-1})^{\mathrm{T}}\;A_n(\boldsymbol{\beta}_{t-2})^{\mathrm{T}}\;\ldots\;A_n(\boldsymbol{\beta}_0)^{\mathrm{T}}\;. \end{array}\right\} \tag{1.45}$$

Analog zu $\boldsymbol{\beta}_s$ betrachten wir nun die Binärvektoren

$$\boldsymbol{\gamma}_s = (\gamma_s(1),\ldots,\gamma_s(2^s)) \in \{0,1\}^{2^s} \qquad (s \in \{0,\ldots,t-1\}),$$

welche der Rekursion

$$\boldsymbol{\gamma}_0 := (1), \quad \boldsymbol{\gamma}_{s+1} := (\tilde{\boldsymbol{\gamma}}_s,\; \tilde{\boldsymbol{\gamma}}_s) \qquad (s=0,\ldots,t-2) \tag{1.46}$$

genügen, wobei $\tilde{\boldsymbol{\gamma}}_s := (\tilde{\gamma}_s(1),\ldots,\tilde{\gamma}_s(2^s))$ wiederum durch

$$\tilde{\gamma}_s(i) := \gamma_s(i), \qquad i=1,\ldots,2^s-1, \qquad \tilde{\gamma}_s(2^s) := 1-\gamma_s(2^s)$$

gegeben ist. Analog zur Abbildung 1.1 ist der Zusammenhang der Binärvektoren $\boldsymbol{\gamma}_s$ $(s=0,1,2,3)$ in Abbildung 1.3 schematisch veranschaulicht. In den einzelnen Zeilen lassen sich wiederum die jeweiligen

Komponenten von links nach rechts ablesen lassen. Offenbar ergibt sich das Diagramm in Abbildung 1.3 als Teildiagramm aus Abbildung 1.1.

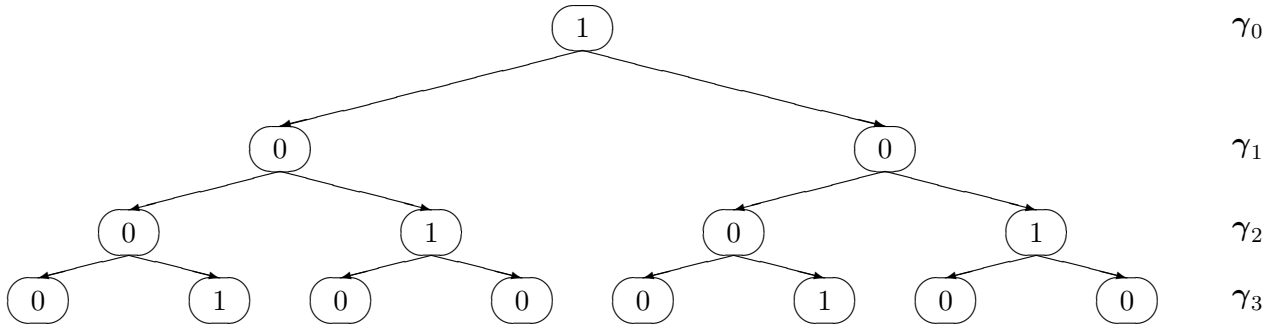


Abbildung 1.3: Diagramm zur Herleitung der Binärvektoren $\boldsymbol{\gamma}_s$ $(s = 0, 1, 2, 3)$.

Aus der blockweisen Anwendung der rekursiven Faktorisierungen (1.15) und (1.19) ergibt sich nun analog zu (1.44) die folgenden Darstellung.

Lemma 1.25. *Für die analog zu* (1.43) *definierten Matrizen* $A_n(\boldsymbol{\gamma}_s)$, $T_n(\boldsymbol{\gamma}_s)$ *und* $C_n(\boldsymbol{\gamma}_s)$ *gilt*

$$C_n(\boldsymbol{\gamma}_s) = A_n(\boldsymbol{\gamma}_s)\, C_n(\boldsymbol{\gamma}_{s+1})\, T_n(\boldsymbol{\gamma}_s) \qquad (s = 0, \ldots, t-2). \tag{1.47}$$

Der aus der Anwendung von (1.47) für $s = 0, 1, 2$ folgende Zusammenhang zwischen C_n^{IV} und den Matrizen $C_{n_3}^{\mathrm{IV}}$ und $C_{n_3}^{\mathrm{IV}}$ ist in Abbildung 1.4 schematisch veranschaulicht.

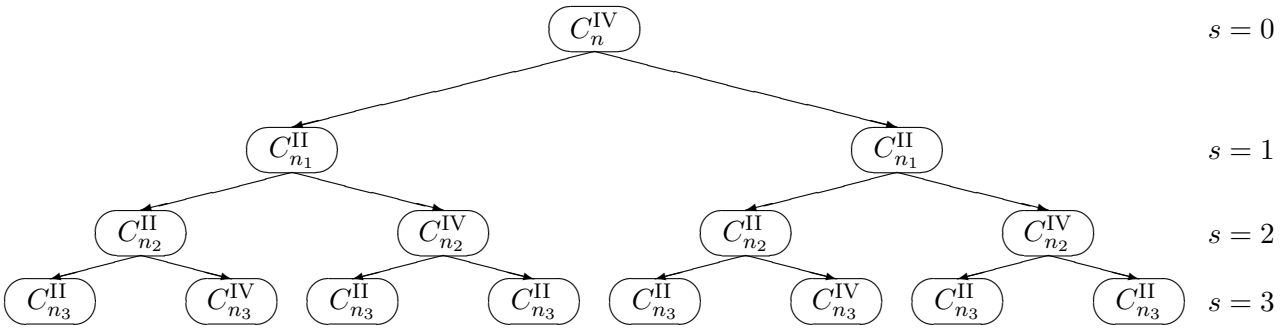


Abbildung 1.4: Faktorisierungsschema von C_n^{IV} bis zum Level $s = 3$.

Wie zuvor liefert nun die wiederholte Anwendung von (1.47) bzw. der transponierten Gleichung die gewünschten iterativen Darstellungen von $C_n^{\mathrm{IV}} = C_n(\boldsymbol{\gamma}_0)$.

Satz 1.26 (vgl. [43], Theorem 4.4)**.** *Die rekursiven Algorithmen* 1.19 *und* 1.20 *basieren für* $k = 1$ *auf den lediglich dünnbesetzte orthogonale Matrizen enthaltenden iterativen Faktorisierungen*

$$\left.\begin{aligned} C_n^{\mathrm{IV}} &= A_n(\boldsymbol{\gamma}_0)\, \ldots\, A_n(\boldsymbol{\gamma}_{t-2})\, T_n(\boldsymbol{\gamma}_{t-1})\, T_n(\boldsymbol{\gamma}_{t-2})\, \ldots\, T_n(\boldsymbol{\gamma}_0)\,, \\ C_n^{\mathrm{IV}} &= T_n(\boldsymbol{\gamma}_0)^{\mathrm{T}}\, \ldots\, T_n(\boldsymbol{\gamma}_{t-2})^{\mathrm{T}}\, T_n(\boldsymbol{\gamma}_{t-1})^{\mathrm{T}}\, A_n(\boldsymbol{\gamma}_{t-2})^{\mathrm{T}}\, \ldots\, A_n(\boldsymbol{\gamma}_0)^{\mathrm{T}}\,. \end{aligned}\right\} \tag{1.48}$$

Da die in den Matrizen $A_n(\boldsymbol{\beta}_s)$ und $A_n(\boldsymbol{\gamma}_s)$ vorkommenden Matrizen $A_{n_s}(0)$ Einheitsmatrizen der Größe $n_s \times n_s$ sind, ist die Anzahl der Einsen innerhalb der Binärvektoren $\boldsymbol{\beta}_s$ und $\boldsymbol{\gamma}_s$ für die Analyse des Rundungsfehlers von Interesse.

Folgerung 1.27 (vgl. [43], Lemma 4.1 und Lemma 4.3)**.** *Die Anzahl der in den Binärvektoren* $\boldsymbol{\beta}_s$ *und* $\boldsymbol{\gamma}_s$ *enthaltenen Einsen ergibt sich aus den Beziehungen*

$$\|\boldsymbol{\beta}_s\|_1 = \frac{2^s - (-1)^s}{3}\,, \qquad \|\boldsymbol{\gamma}_s\|_1 = \frac{2^s + 2(-1)^s}{3}\,. \tag{1.49}$$

Beweis: Mit (1.42) und (1.46) folgen für $s \geq 0$ die Rekursionsgleichungen

$$\|\boldsymbol{\beta}_{s+1}\|_1 \;=\; 2\|\boldsymbol{\beta}_s\|_1 + (-1)^s \;, \qquad\qquad \|\boldsymbol{\gamma}_{s+1}\|_1 \;=\; 2\|\boldsymbol{\gamma}_s\|_1 - 2(-1)^s \;.$$

Wiederholtes Anwenden dieser Rekursionsgleichungen führt mit $\|\boldsymbol{\beta}_0\|_1 = 0$ und $\|\boldsymbol{\gamma}_0\|_1 = 1$ zu

$$\|\boldsymbol{\beta}_s\|_1 \;=\; \sum_{k=0}^{s-1} 2^k (-1)^{s-1-k} \;=\; (-1)^{s-1}\frac{(-2)^s - 1}{-2-1} \;=\; \frac{2^s - (-1)^s}{3}$$

und

$$\|\boldsymbol{\gamma}_s\|_1 \;=\; 2^s - 2\sum_{k=0}^{s-1} 2^k (-1)^{s-1-k} \;=\; 2^s - 2(-1)^{s-1}\frac{(-2)^s - 1}{-2-1} \;=\; \frac{2^s + 2(-1)^s}{3} \;. \qquad \blacksquare$$

In ähnlicher Weise lassen sich vollständige Matrixfaktorisierungen zu den rekursiven Algorithmen 1.21 und 1.22 herleiten. Betrachten wir die in (1.20) und (1.21) angegebenen Zusammenhänge, so ergeben sich für die Matrix S_n^{II} nacheinander die in Abbildung 1.7 schematisch dargestellten Faktorisierungen.

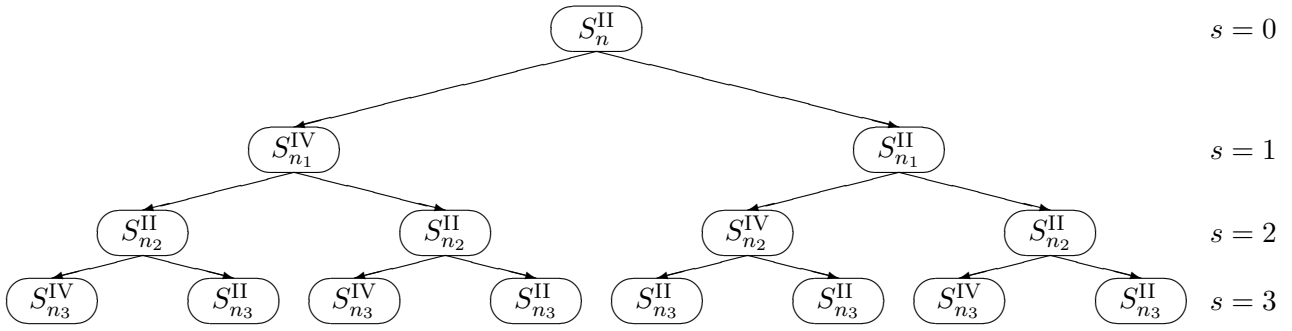


Abbildung 1.5: Faktorisierungsschema von S_n^{II} bis zum Level $s = 3$.

Definieren wir nun die Binärvektoren

$$\breve{\boldsymbol{\beta}}_s \;=\; \Big(\breve{\beta}_s(1), \ldots, \breve{\beta}_s(2^s)\Big) \in \{0,1\}^{2^s} \qquad (s \in \{0, \ldots, t-1\}) \tag{1.50}$$

durch

$$\breve{\beta}_s(i) \;:=\; \beta_s(2^s + 1 - i) \;, \qquad i = 1, \ldots, 2^s, \tag{1.51}$$

so ergibt sich beispielsweise für die Binärvektoren $\breve{\boldsymbol{\beta}}_s$ $(s = 0, 1, 2, 3)$ der in Abbildung 1.6 schematisch dargestellte Zusammenhang. Dabei lassen sich die jeweiligen Komponenten von $\breve{\boldsymbol{\beta}}_s$ wiederum zeilenweise von links nach rechts ablesen.

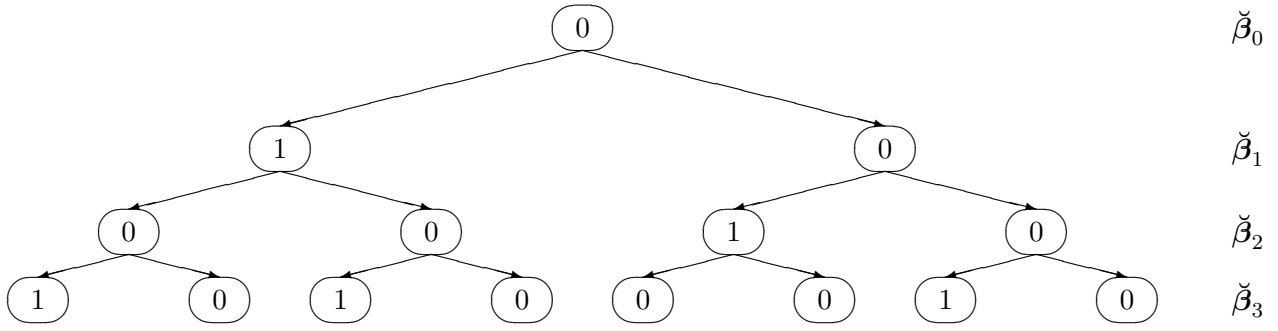


Abbildung 1.6: Diagramm zur Herleitung der Binärvektoren $\breve{\boldsymbol{\beta}}_s$ $(s = 0, 1, 2, 3)$.

Offenbar entsteht das Diagramm in Abbildung 1.6 durch Spiegelung des in Abbildung 1.1 enthaltenen Diagramms an der Vertikalen. Aus (1.51) geht weiter hervor, dass $\breve{\boldsymbol{\beta}}_s$ wegen

$$\|\breve{\boldsymbol{\beta}}_s\|_1 \;=\; \|\boldsymbol{\beta}_s\|_1 \tag{1.52}$$

genauso viele Einsen wie $\boldsymbol{\beta}_s$ enthält. Definieren wir nun wie in (1.43) zu jedem dieser Binärvektoren $\breve{\boldsymbol{\beta}}_s$ die dünnbesetzten orthogonalen Matrizen

$$\left.\begin{array}{lcll} \breve{A}_n(\breve{\boldsymbol{\beta}}_s) & := & P_n(s)\,\left(\breve{A}_{n_s}(\breve{\beta}_s(1))^{\mathrm{T}} \oplus \ldots \oplus \breve{A}_{n_s}(\breve{\beta}_s(2^s))^{\mathrm{T}}\right) & (s=0,\ldots,t-2), \\ \breve{T}_n(\breve{\boldsymbol{\beta}}_s) & := & \breve{T}_{n_s}(\breve{\beta}_s(1))^{\mathrm{T}} \oplus \ldots \oplus \breve{T}_{n_s}(\breve{\beta}_s(2^s))^{\mathrm{T}} & (s=0,\ldots,t-2), \\ S_n(\breve{\boldsymbol{\beta}}_s) & := & S_{n_s}(\breve{\beta}_s(1)) \oplus \ldots \oplus S_{n_s}(\breve{\beta}_s(2^s)) & (s=0,\ldots,t-1) \end{array}\right\} \tag{1.53}$$

mit $\breve{A}_{n_s}(0) := I_{n_s}$, $\breve{T}_{n_s}(0) := T_{n_s}(0)^{\mathrm{T}}$, $S_{n_s}(0) := S^{\mathrm{II}}_{n_s}$ und $S_{n_s}(1) := S^{\mathrm{IV}}_{n_s}$, führt die blockweise Anwendung der rekursiven Faktorisierungen (1.20) und (1.21) zu folgendem Zusammenhang.

Lemma 1.28. *Für die in* (1.53) *definierten Matrizen gilt*

$$S_n(\breve{\boldsymbol{\beta}}_s) \;=\; \breve{A}_n(\breve{\boldsymbol{\beta}}_s)\; S_n(\breve{\boldsymbol{\beta}}_{s+1})\; \breve{T}_n(\breve{\boldsymbol{\beta}}_s) \qquad (s=0,\ldots,t-2). \tag{1.54}$$

Definieren wir weiterhin $\breve{T}_n(\breve{\boldsymbol{\beta}}_{t-1}) := S_n(\boldsymbol{\beta}_{t-1})$, so führt die wiederholte Anwendung von (1.54) bzw. der transponierten Gleichung offenbar auf die gewünschten iterativen Darstellungen von $S^{\mathrm{II}}_n = S_n(\boldsymbol{\beta}_0)$ und S^{III}_n.

Satz 1.29. *Die rekursiven Algorithmen* 1.21 *und* 1.22 *basieren für* $k=0$ *auf den lediglich dünnbesetzte orthogonale Matrizen enthaltenden iterativen Faktorisierungen*

$$\left.\begin{array}{lcl} S^{\mathrm{II}}_n & = & \breve{A}_n(\breve{\boldsymbol{\beta}}_0)\;\ldots\;\breve{A}_n(\breve{\boldsymbol{\beta}}_{t-2})\;\breve{T}_n(\breve{\boldsymbol{\beta}}_{t-1})\;\breve{T}_n(\breve{\boldsymbol{\beta}}_{t-2})\;\ldots\;\breve{T}_n(\breve{\boldsymbol{\beta}}_0)\;, \\ S^{\mathrm{III}}_n & = & \breve{T}_n(\breve{\boldsymbol{\beta}}_0)^{\mathrm{T}}\;\ldots\;\breve{T}_n(\breve{\boldsymbol{\beta}}_{t-2})^{\mathrm{T}}\;\breve{T}_n(\breve{\boldsymbol{\beta}}_{t-1})^{\mathrm{T}}\;\breve{A}_n(\breve{\boldsymbol{\beta}}_{t-2})^{\mathrm{T}}\;\ldots\;\breve{A}_n(\breve{\boldsymbol{\beta}}_0)^{\mathrm{T}}\;. \end{array}\right\} \tag{1.55}$$

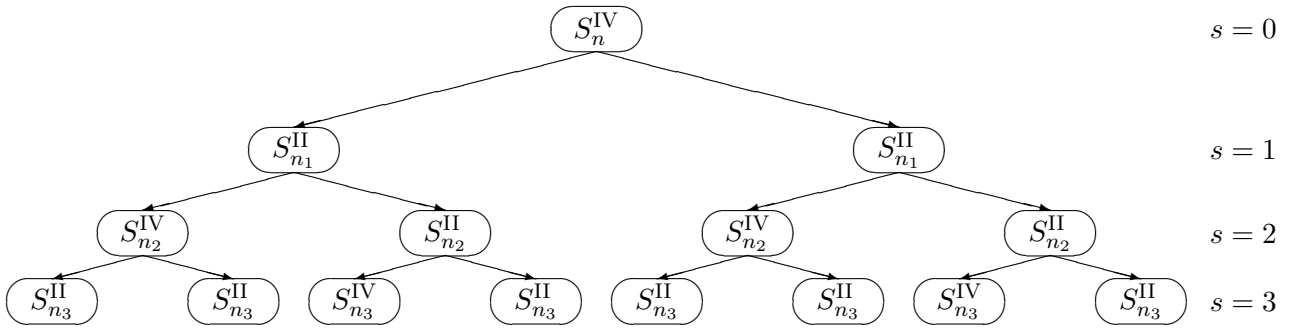


Abbildung 1.7: Faktorisierungsschema von S^{IV}_n bis zum Level $s=3$.

Betrachten wir nun wiederum die in (1.20) und (1.21) angegebenen Zusammenhänge, so ergeben sich für die Matrix S^{IV}_n nacheinander die in Abbildung 1.7 schematisch dargestellten Faktorisierungen.

Aufgrund der Symmetrie der in den Abbildungen 1.4 und 1.7 dargestellten Schemata ergibt sich nach blockweiser Anwendung der rekursiven Faktorisierungen (1.20) und (1.21) mit den Binärvektoren

$$\breve{\boldsymbol{\gamma}}_s \;=\; (\breve{\gamma}_s(1),\ldots,\breve{\gamma}_s(2^s)) \in \{0,1\}^{2^s} \qquad (s\in\{0,\ldots,t-1\})\;, \tag{1.56}$$

welche analog zu (1.51) durch

$$\breve{\gamma}_s(i) \;:=\; \gamma_s(2^s+1-i)\;, \qquad i=1,\ldots,2^s, \tag{1.57}$$

definiert und für $s=0,1,2,3$ in Abbildung 1.8 wiederum zeilenweise von links nach rechts abzulesen sind, die folgende Beziehung.

Lemma 1.30. *Für die analog zu* (1.53) *definierten Matrizen* $\breve{A}_n(\breve{\boldsymbol{\gamma}}_s)$, $\breve{T}_n(\breve{\boldsymbol{\gamma}}_s)$ *und* $S_n(\breve{\boldsymbol{\gamma}}_s)$ *gilt*

$$S_n(\breve{\boldsymbol{\gamma}}_s) \;=\; \breve{A}_n(\breve{\boldsymbol{\gamma}}_s)\; S_n(\breve{\boldsymbol{\gamma}}_{s+1})\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_s) \qquad (s = 0, \ldots, t-2). \tag{1.58}$$

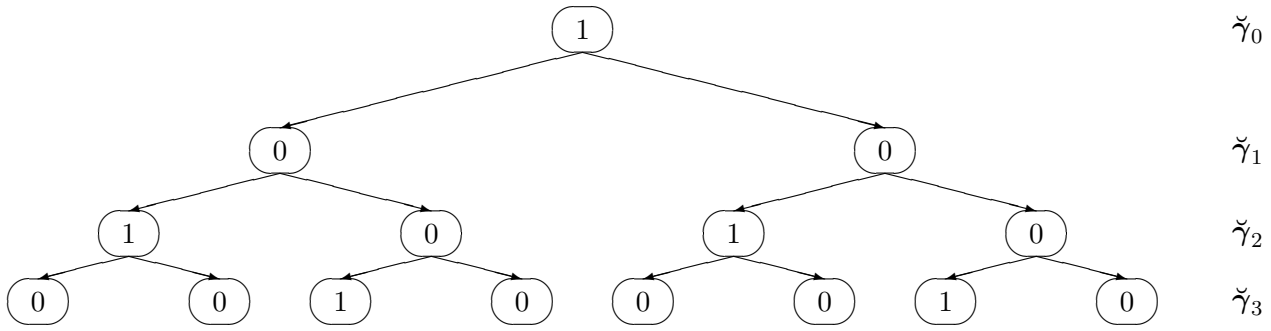


Abbildung 1.8: Diagramm zur Herleitung der Binärvektoren $\breve{\boldsymbol{\gamma}}_s$ $(s = 0, 1, 2, 3)$.

Offenbar ergibt sich das Schema in Abbildung 1.8 aus dem Schema in Abbildung 1.3 genau durch Spiegelung an der Vertikalen. Wegen (1.57) gilt außerdem

$$\|\breve{\boldsymbol{\gamma}}_s\|_1 \;=\; \|\boldsymbol{\gamma}_s\|_1 \;. \tag{1.59}$$

Wie zuvor liefert nun die wiederholte Anwendung von (1.58) bzw. der transponierten Gleichung die gewünschten iterativen Darstellungen von $S_n^{\mathrm{IV}} = S_n(\breve{\boldsymbol{\gamma}}_0)$.

Satz 1.31. *Die rekursiven Algorithmen* 1.21 *und* 1.22 *basieren für* $k = 1$ *auf den lediglich dünnbesetzte orthogonale Matrizen enthaltenden iterativen Faktorisierungen*

$$\left.\begin{aligned} S_n^{\mathrm{IV}} &= \breve{A}_n(\breve{\boldsymbol{\gamma}}_0)\; \ldots\; \breve{A}_n(\breve{\boldsymbol{\gamma}}_{t-2})\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_{t-1})\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_{t-2})\; \ldots\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_0)\;, \\ S_n^{\mathrm{IV}} &= \breve{T}_n(\breve{\boldsymbol{\gamma}}_0)^{\mathrm{T}}\; \ldots\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_{t-2})^{\mathrm{T}}\; \breve{T}_n(\breve{\boldsymbol{\gamma}}_{t-1})^{\mathrm{T}}\; \breve{A}_n(\breve{\boldsymbol{\gamma}}_{t-2})^{\mathrm{T}}\; \ldots\; \breve{A}_n(\breve{\boldsymbol{\gamma}}_0)^{\mathrm{T}}\;. \end{aligned}\right\} \tag{1.60}$$

Basierend auf den Faktorisierungen aus (1.32), (1.35) (1.37) und (1.38) sowie auf den Faktorisierungen aus (1.45), (1.48), (1.55) und (1.60) werden in den Kapiteln 3 und 4 die hier angegebenen schnellen Algorithmen 1.15 – 1.22 auf ihr Rundungsfehlerverhalten sowohl in Gleitkomma- als auch in Festkomma-Arithmetik hin untersucht. Zuvor gehen wir vorbereitend in Kapitel 2 auf diese beiden Rechner-Arithmetiken ein.

Bemerkung 1.32 (Zusammenfassung von Kapitel 1)**.** Zunächst beschäftigt sich Abschnitt 1.1 mit den in Definition 1.1 eingeführten Kosinus- und Sinus-Matrizen (1.2) – (1.7) vom Typ II – IV, deren Orthogonalität in Satz 1.4 nachgewiesen wird. Darüber hinaus wird der Zusammenhang zu diskreten Randwertaufgaben und Chebyshev-Polynomen erster und dritter Art dargelegt.
Abschnitt 1.2 umfasst einerseits in Lemma 1.6 bis Folgerung 1.8 rekursive Faktorisierungen (vgl. [43, Lemma 2.2, 2.4]) aller Matrizen aus Definition 1.1 und andererseits in Lemma 1.11 bis Folgerung 1.13 iterative Faktorisierungen (vgl. [50, 54]) der Matrizen (1.2) – (1.4).
Neben der Einführung diskreter trigonometrischer Transformationen in Definition 1.14 enthält Abschnitt 1.3 die schnellen Algorithmen 1.15 – 1.22. Desweiteren liefern die Sätze 1.24, 1.26, 1.29 und 1.31 die für Rundungsfehleranalysen günstigen iterativen Matrixfaktorisierungen, welche die Wirkung der Algorithmen 1.19 – 1.22 am besten beschreiben.
Insbesondere enthalten die Faktorisierungen (1.32), (1.35) (1.37) und (1.38) einerseits sowie die Faktorisierungen (1.45), (1.48), (1.55) und (1.60) nach (1.25) – (1.29) andererseits nur dünnbesetzte orthogonale Matrizen, welche – gegebenenfalls nach Permutationen und/oder Vorzeichenskalierungen – Blockdiagonalgestalt mit Blöcken der Form (1.24) besitzen, wobei an dieser Stelle $\varphi = 0$ ebenfalls zugelassen ist. □

2 Modellierung von Rechner-Arithmetiken

Entsprechend den Anforderungen kommen in Prozessoren verschiedene Arithmetiken vor. Ist ein Prozessor darauf ausgelegt, viele verschiedene Aufgaben bewältigen zu können, wird häufig Gleitkomma-Arithmetik (engl. floating point arithmetic) verwendet. Diese ist zwar leichter zu handhaben, jedoch – im Vergleich zu anderen Arithmetiken – teurer in der Herstellung. Wird dagegen ein Chip dahingehend konzipiert, nur innerhalb eines sehr begrenzten Aufgabenspektrums zu arbeiten, finden dort auch andere Arithmetiken Anwendung [15, S. 668ff]. Beispiele dafür sind Festkomma-Arithmetik (engl. fixed point arithmetic) oder auch Ganzzahl-Arithmetik (engl. integer arithmetic).
In Abschnitt 2.1 wird zunächst ein Modell zur Gleitkomma-Arithmetik (2.1) in Erinnerung gebracht, welches beispielsweise in [24] zu finden ist, jedoch schon von Wilkinson [67] verwendet worden ist. Innerhalb dieser Arithmetik werden vom Wilkinson-Modell (2.6) ausgehend jeweils Abschätzungen für den entstehenden relativen Rundungsfehler bei elementaren Operationen (Addition, Subtraktion und Multiplikation) und ebenso beim Skalarprodukt von Vektoren sowie bei der Matrix-Vektor-Multiplikation hergeleitet. Insbesondere interessieren dabei die Ergebnisse für Matrix-Vektor-Multiplikationen mit den Repräsentanten der Drehmatrizen $Q_2(\varphi)$.
In Abschnitt 2.2 untersuchen wir die Möglichkeiten für ein auf (2.6) abgestimmtes stochastisches Modell, welches unter zusätzlichen Annahmen Vorhersagen für die euklidische Norm des Vektors der auftretenden relativen Rundungsfehler liefern soll. Dabei erweist sich die im Unterabschnitt 2.2.1 zunächst geforderte Annahme der stochastischen Unabhängigkeit an die Eingangsdaten als ungeeignet, da sie im Allgemeinen bereits nach einer Matrix-Vektor-Multiplikation nicht mehr gewährleistet werden kann. Unter Hinzunahme der speziellen Blockdiagonalgestalt der auftretenden Matrizen gelingt in Unterabschnitt 2.2.2 die Übertragung der Idee von Zeuner [75], welcher ein stochastisches Modell für die Multiplikation komplexer Zahlen entwickelt hat, auf den Fall von Drehmatrizen (vgl. Lemma 2.20). Nach geringfügiger Modifikation bildet dieses durch Satz 2.22 den Grundbaustein für alle nachfolgenden Ergebnisse innerhalb des stochastischen Modells zur Gleitkomma-Arithmetik.
In Abschnitt 2.3 wird das bereits auf J. von Neumann und H.H. Goldstine [41] zurückgehende Modell zur Festkomma-Arithmetik betrachtet, welches die Vorzeichenbetragsdarstellung (2.49) (engl. sign-magnitude representation) verwendet. Diese ist der Zweierkomplementdarstellung (2.54) aus den in Bemerkung 2.28 genannten Gründen überlegen. Im Vergleich zur Gleitkomma-Arithmetik ist die Addition innerhalb der Festkomma-Arithmetik fehlerfrei ausführbar, solange kein Überlauf auftritt. Analog zu Abschnitt 2.1 werden obere Schranken für den entstehenden absoluten Rundungsfehler bei elementaren Operationen, Skalarprodukt und Matrix-Vektor-Multiplikation hergeleitet. Entsprechend der Empfehlung von J. von Neumann und H.H. Goldstine [41] betrachten wir hierbei Skalarprodukte nicht nur in einfacher Genauigkeit, sondern auch in doppelter Genauigkeit.
Schließlich untersuchen wir in Abschnitt 2.4 ein stochastisches Modell zur Festkomma-Arithmetik, welches die in Abschnitt 2.3 verwendeten Annahmen berücksichtigt.

2.1 Gleitkomma-Arithmetik nach Wilkinson

In Anlehnung an [24] (vgl. auch [7]) definieren wir eine Menge von Gleitkomma-Zahlen als Teilmenge von $\mathbb{R}$, welche sich durch

$$\mathbb{G}(\beta,\tau,\gamma_{\min},\gamma_{\max}) := \left\{ \pm\mu\beta^{\gamma-\tau} \;\middle|\; \mu,\gamma\in\mathbb{Z} \text{ mit } 0\le\mu\le\beta^{\tau}-1 \text{ und } \gamma_{\min}\le\gamma\le\gamma_{\max} \right\} \tag{2.1}$$

mit ganzzahligen Parametern $\gamma_{\min}\le\gamma_{\max}$, $\beta\ge 2$ und $\tau>0$ beschreiben lässt. Dabei bezeichnet μ die *Mantisse*, β die *Basis*, τ die *Genauigkeit* und γ den *Exponenten*. Offensichtlich lässt sich jedes $\mu\in\mathbb{Z}$ mit $0\le\mu\le\beta^{\tau}-1$ auch in der Form

$$\mu \;=\; \sum_{k=1}^{\tau}\mu_k\beta^{\tau-k} \quad \text{mit} \quad \mu_k\in\{0,\ldots,\beta-1\} \tag{2.2}$$

darstellen. Zu beachten ist, dass aufgrund der Exponentendarstellung keine äquidistante Verteilung der Elemente vorliegt. Insbesondere steigt der Abstand direkt benachbarter Elemente mit zunehmendem Betrag der Elemente an. Als Beispiel ist in Abbildung 2.1 die Menge $\mathbb{G}(2,3,-1,3)$ veranschaulicht.

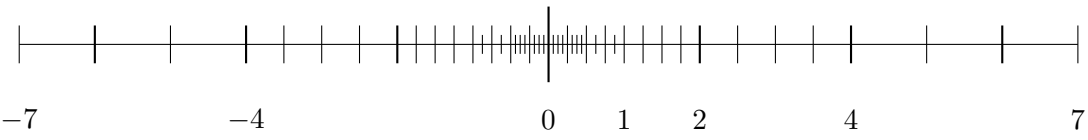


Abbildung 2.1: Gleitkomma-Zahlenmenge $\mathbb{G}(2,3,-1,3)$.

In (2.1) ist die Darstellung nicht eindeutig. Daher wird die Menge der Gleitkomma-Zahlen üblicherweise durch eine Normierungsbedingung (d.h. $\mu_1 = 1$ in (2.2)) verkleinert. Die resultierende Menge

$$\mathbb{G}_{\text{norm}}(\beta,\tau,\gamma_{\min},\gamma_{\max}) := \left\{ \pm\mu\beta^{\gamma-\tau} \;\middle|\; \mu,\gamma \in \mathbb{Z} \text{ mit } \beta^{\tau-1} \le \mu \le \beta^{\tau}-1 \text{ und } \gamma_{\min} \le \gamma \le \gamma_{\max} \right\} \cup \{0\}$$

mit eindeutiger Darstellung ihrer nicht äquidistant verteilten Elemente besitzt jedoch wegen der für alle $g \in \mathbb{G}_{\text{norm}}(\beta,\tau,\gamma_{\min},\gamma_{\max})\backslash\{0\}$ gültigen Ungleichungskette $\beta^{\gamma_{\min}-1} \le |g| \le \beta^{\gamma_{\max}}(1-\beta^{-\tau})$ die bekannte Lücke um den Nullpunkt (vgl. [7]). Zur Illustration ist in Abbildung 2.2 die Menge $\mathbb{G}_{\text{norm}}(2,3,-1,3)$ graphisch dargestellt.

Abbildung 2.2: Gleitkomma-Zahlenmenge $\mathbb{G}_{\text{norm}}(2,3,-1,3)$.

Wie aus den Abbildungen 2.1 und 2.2 ersichtlich, sind die eingeführten Mengen $\mathbb{G}(\beta,\tau,\gamma_{\min},\gamma_{\max})$ und $\mathbb{G}_{\text{norm}}(\beta,\tau,\gamma_{\min},\gamma_{\max})$ endlich, so dass überabzählbar viele reelle Zahlen lediglich näherungsweise durch ein jeweils geeignetes Element aus $\mathbb{G}(\beta,\tau,\gamma_{\min},\gamma_{\max})$ bzw. $\mathbb{G}_{\text{norm}}(\beta,\tau,\gamma_{\min},\gamma_{\max})$ dargestellt werden können. Es sei im Folgenden

$$\mathbb{G} := \mathbb{G}_{\text{norm}}(\beta,\tau,\gamma_{\min},\gamma_{\max}) \tag{2.3}$$

fest gewählt und durch

$$R_{\mathbb{G}} \;:=\; \{x \in \mathbb{R} \mid \beta^{\gamma_{\min}-1} \le |x| \le \beta^{\gamma_{\max}}(1-\beta^{-\tau})\} \cup \{0\}$$

der *Bereich* von $\mathbb{G}$ definiert (vgl. [7], S. 452). Für ein $x \in R_{\mathbb{G}}$ bezeichne $\text{fl}(x)$ die beste Approximation in $\mathbb{G}$. Falls zwei Elemente $g_1, g_2 \in \mathbb{G}$ mit $x - g_1 = g_2 - x = \min\limits_{g\in\mathbb{G}} |g - x|$ existieren, kann beispielsweise durch

$$\text{fl}(x) := \begin{cases} g_1\ , & \text{falls } \max(|g_1|,|g_2|) = |g_1|\ , \\ g_2\ , & \text{falls } \max(|g_1|,|g_2|) = |g_2| \end{cases} \tag{2.4a}$$

(Runden weg von der Null) oder alternativ auch

$$\text{fl}(x) := \begin{cases} g_1\ , & \text{falls } g_1 = \pm\sum\limits_{k=1}^{\tau} \mu_k\beta^{\gamma-k} \text{ mit } \mu_\tau \text{ gerade}, \\ g_2\ , & \text{falls } g_2 = \pm\sum\limits_{k=1}^{\tau} \mu_k\beta^{\gamma-k} \text{ mit } \mu_\tau \text{ gerade} \end{cases} \tag{2.4b}$$

(Runden auf gerades letztes Bit) eine eindeutige Abbildung $\text{fl} : R_{\mathbb{G}} \to \mathbb{G}$ definiert werden. Diese Abbildung setzen wir nun auch kanonisch auf Vektoren $\mathbf{x} = (x_k)_{k=0}^{n-1} \in R_{\mathbb{G}}^n$ und ebenso auf Matrizen $A = (a_{j,k})_{j,k=0}^{n-1} \in R_{\mathbb{G}}^{n\times n}$ fort, indem fl komponentenweise angewandt wird, d.h., wir definieren

$$\text{fl}(\mathbf{x}) \;:=\; (\text{fl}(x_k))_{k=0}^{n-1}\ , \qquad \text{fl}(A) \;:=\; (\text{fl}(a_{j,k}))_{j,k=0}^{n-1}\ .$$

Offenbar folgt für $x \in \mathbb{G}$ sowohl $\mathrm{fl}(x) = x$ als auch $\mathrm{fl}(-x) = -x$ (vgl. [7, S. 453]) und aus $x \geq y$ stets $\mathrm{fl}(x) \geq \mathrm{fl}(y)$. Mit der *Rundungseinheit*

$$u := \frac{\beta^{1-\tau}}{2} \tag{2.5}$$

gilt weiterhin

Satz 2.1 (vgl. [24], Theorem 2.2)**.** *Für jedes $x \in R_{\mathbb{G}}$ existiert ein $\varepsilon \in \mathbb{R}$ mit $|\varepsilon| < u$ und $\mathrm{fl}(x) = x(1+\varepsilon)$.*

In den gebräuchlichen Standards ist $\beta = 2$ (vgl. Tabelle 2.1), so dass die Rundungseinheit dort $u = 2^{-\tau}$ ergibt. Die kleinstmögliche Basis erhält aufgrund des in [24] beschriebenen „wobbling"-Effekts den Vorzug. Weiterhin wird in den Standards die Rundungsart (2.4b) verwendet (vgl. [24]).

Bemerkung 2.2. Anstelle (2.4b) könnte auch auf ungerades letztes Bit gerundet werden. Für $\beta = 2$ ist (2.4b) jedoch zu bevorzugen (vgl. [24, S. 54]). Als weitere Möglichkeit zur Definition von $\mathrm{fl}(x)$ wird beispielsweise im Standard IBM/370 das Abschneiden (engl. *chopping*) verwendet (vgl. [24, S. 54]), bei dem $x \in R_{\mathbb{G}}$ die nächstgelegene Zahl $y \in \mathbb{G}$ zugeordnet wird, für welche $|y| \leq |x|$ erfüllt ist. □

Arithmetik	β	τ	$\gamma_{\min}$	$\gamma_{\max}$	$u = \frac{\beta^{1-\tau}}{2}$
IEEE single	2	24	-125	128	$2^{-24} \approx 5.96 \times 10^{-8}$
IEEE double	2	53	-1021	1024	$2^{-53} \approx 1.11 \times 10^{-16}$
IEEE extended (typisch)	2	64	-16381	16384	$2^{-64} \approx 5.42 \times 10^{-20}$

Tabelle 2.1: Parametergrößen für geläufige Gleitkomma-Arithmetiken (vgl. [24, Table 2.1]).

Für die deterministische Rundungsfehleranalyse von Algorithmen benötigen wir Annahmen über die Genauigkeit arithmetischer Operationen in $\mathbb{G}$. Nach dem Wilkinson-Modell (vgl. [24, 54]) werden die arithmetischen Operationen $\bullet \in \{+, -, \times\}$ für beliebige $g_1, g_2 \in \mathbb{G}$ mit $g_1 \bullet g_2 \in R_{\mathbb{G}}$ durch

$$\boxed{\mathrm{fl}(g_1 \bullet g_2) \;=\; (g_1 \bullet g_2)(1 + \varepsilon^{\bullet}) \;, \qquad |\varepsilon^{\bullet}| \leq u} \tag{2.6}$$

modelliert. Dabei bezeichnet $\varepsilon^{\bullet}$ den bei der Operation $\bullet \in \{+, -, \times\}$ auftretenden *relativen Rundungsfehler*. Durch die Voraussetzung $g_1 \bullet g_2 \in R_{\mathbb{G}}$ wird der Fall von Über- oder Unterlauf ausgeschlossen.

Bemerkung 2.3. Zur Behandlung von Unterlauf kann das Modell (2.6) für beliebige $g_1, g_2 \in \mathbb{G}$ zu

$$\boxed{\mathrm{fl}(g_1 \bullet g_2) \;=\; (g_1 \bullet g_2)(1 + \varepsilon^{\bullet}) \;+\; \eta \;, \qquad |\varepsilon^{\bullet}| \leq u} \tag{2.7}$$

mit $|\eta| \leq \lambda := \beta^{\gamma_{\min}-1}$ bzw. mit $|\eta| \leq \lambda u$ erweitert werden, wobei $\eta \cdot \varepsilon^{\bullet} = 0$ gelte (vgl. [24, S. 56]). Wir beschränken uns hier jedoch auf das Modell (2.6). □

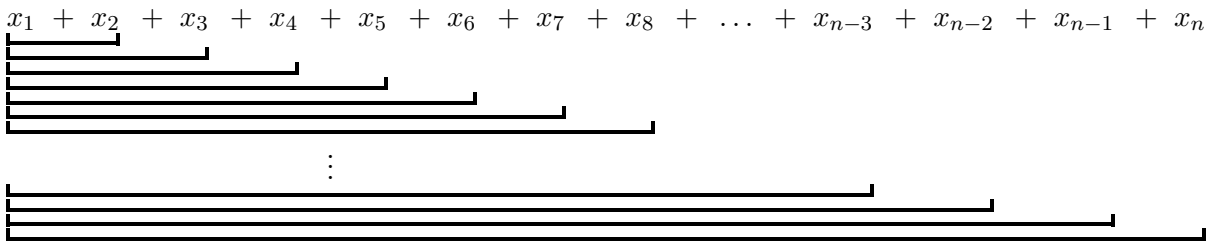


Abbildung 2.3: Schematische Darstellung für die Ausführung der naiven, d.h. sequentiellen Addition von n Zahlen. Das Element x_1 ist offenbar an der Bildung von $n-1$ Summen beteiligt.

Aus der Modellannahme (2.6) folgt, dass die Assoziativität nicht erfüllt ist (vgl. [24, S. 54]). Werden Operationen in verschiedenen Reihenfolgen ausgeführt, treten gegebenenfalls voneinander abweichende Rundungsfehler auf. Insbesondere ergeben sich unterschiedliche Fehlerabschätzungen, wenn n Zahlen einerseits sequentiell (vgl. Abb. 2.3) und andererseits kaskadenförmig (vgl. Abb. 2.4) aufsummiert werden.

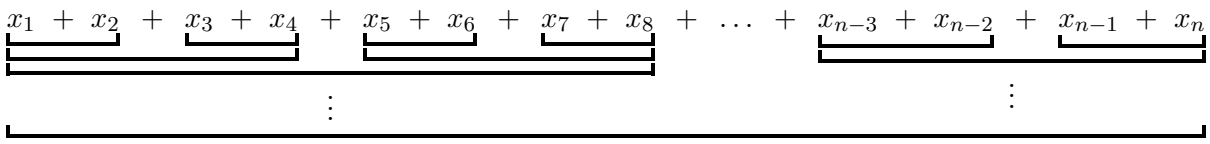


Abbildung 2.4: Schematische Darstellung für die Ausführung der Kaskaden-Summation von n Zahlen. Jedes Element ist an der Bildung von höchstens $\lceil \log_2 n \rceil$ Summen beteiligt. Dabei ist $\lceil x \rceil := \min\{z \in \mathbb{Z} \,:\, x \leq z\}$ für ein $x \in \mathbb{R}$.

Falls n eine Zweierpotenz ist, wird der Unterschied des Rundungsfehlerverhaltens bei sequentieller und Kaskaden-Summation besonders deutlich. Jedoch verbessern sich die Abschätzungen auch für ein beliebiges $n \in \mathbb{N}$, wie an folgendem Beispiel ersichtlich ist.

Beispiel 2.4. Gegeben seien $g_1, \ldots, g_6 \in \mathbb{G}$, deren Summe berechnet werden soll. Bei sequentieller Addition liefert die Modellannahme (2.6) nacheinander die Beziehungen

$$\begin{aligned} \mathrm{fl}\left(\sum_{k=1}^{6}{}^{\mathrm{seq}}\, g_k\right) &:= \mathrm{fl}\left(\mathrm{fl}\left(\mathrm{fl}\left(\mathrm{fl}\left(\mathrm{fl}\left(g_1+g_2\right)+g_3\right)+g_4\right)+g_5\right)+g_6\right) \\ &= \Big(\Big(\Big(\Big(\big(g_1+g_2\big)\left(1+\varepsilon_1^+\right)+g_3\Big)\left(1+\varepsilon_2^+\right)+g_4\Big)\left(1+\varepsilon_3^+\right)+g_5\Big)\left(1+\varepsilon_4^+\right)+g_6\Big)\left(1+\varepsilon_5^+\right) \end{aligned}$$

mit $|\varepsilon_j^+| \leq u$, $j = 1, \ldots, 5$, welches wir zu

$$\mathrm{fl}\left(\sum_{k=1}^{6}{}^{\mathrm{seq}}\, g_k\right) = \sum_{k=1}^{6} g_k + \sum_{j=1}^{5} \varepsilon_j^+ \sum_{k=1}^{j+1} g_k \prod_{l=j+1}^{5} \left(1+\varepsilon_l^+\right)$$

zusammenfassen können. Verwenden wir, dass

$$(1+\mathcal{O}(u))(1+\mathcal{O}(u)) = 1+\mathcal{O}(u) \tag{2.8}$$

gilt, lässt sich das Produkt jeweils durch

$$\left|\prod_{l=j+1}^{5}\left(1+\varepsilon_l^+\right)\right| \leq 1+\mathcal{O}(u)$$

abschätzen. Somit folgt mit Dreiecksungleichung für den entstehenden Rundungsfehler

$$\left|\mathrm{fl}\left(\sum_{k=1}^{6}{}^{\mathrm{seq}}\, g_k\right)-\sum_{k=1}^{6} g_k\right| \leq \sum_{j=1}^{5} u\left|\sum_{k=1}^{j+1} g_k\right|(1+\mathcal{O}(u)) \leq 5\sum_{k=1}^{6}|g_k|\left(u+\mathcal{O}(u^2)\right) .$$

Führen wir die Summation dagegen kaskadenförmig aus, d.h. fassen wir – solange es möglich ist – stets zwei benachbarte Summanden zusammen, ergibt sich mit der Modellannahme (2.6) nacheinander

$$\begin{aligned} \mathrm{fl}\left(\sum_{k=1}^{6}{}^{\mathrm{cas}}\, g_k\right) &:= \mathrm{fl}\left(\mathrm{fl}\left(\mathrm{fl}\left(g_1+g_2\right)+\mathrm{fl}\left(g_3+g_4\right)\right)+\mathrm{fl}\left(g_5+g_6\right)\right) \\ &= \Big(\big(\big(g_1+g_2\big)\left(1+\varepsilon_1^+\right)+\left(g_3+g_4\right)\left(1+\varepsilon_2^+\right)\big)\left(1+\varepsilon_4^+\right)+\left(g_5+g_6\right)\left(1+\varepsilon_3^+\right)\Big)\left(1+\varepsilon_5^+\right) \end{aligned}$$

mit $|\varepsilon_j^+| \leq u$, $j = 1, \ldots, 5$. Ausmultiplizieren ergibt für den Term auf der rechten Seite

$$\sum_{k=1}^{6} g_k + \varepsilon_5^+ \sum_{k=1}^{6} g_k + \left(\sum_{j=1}^{3} \varepsilon_j^+ \left(g_{2j-1}+g_{2j}\right) + \varepsilon_4^+ \sum_{k=1}^{4} g_k + \varepsilon_4^+ \sum_{j=1}^{2} \varepsilon_j^+ \left(g_{2j-1}+g_{2j}\right)\right)\left(1+\varepsilon_5^+\right) .$$

Wiederum mit den Rechenregeln für das Landau-Symbol $\mathcal{O}$ folgt nun für den Rundungsfehler die Abschätzung

$$\left| \mathrm{fl}\left(\sum_{k=1}^{6}{}^{\mathrm{cas}}\, g_k \right) - \sum_{k=1}^{6} g_k \right| \;\leq\; 3 \sum_{k=1}^{6} |g_k| \left(u \;+\; \mathcal{O}(u^2) \right) \;,$$

was offensichtlich eine Verbesserung im Vergleich zur sequentiellen Addition darstellt. □

Bemerkung 2.5. (i) Die Rundungsart (2.4b) ist in dem Sinne stabil, dass

$$\mathrm{fl}\Big(\big(((x+y)-y)+y \big) - y \Big) \;=\; \mathrm{fl}\big((x+y)-y \big)$$

erfüllt wird. Bei der Rundungsart (2.4a) kann dagegen wiederholte Subtraktion und Addition der gleichen Zahl y eine ansteigende Folge liefern, was als *Drift* bezeichnet wird (vgl. [24, S. 54]).

(ii) Eine wichtige Eigenschaft der (korrekt implementierten) Gleitkomma-Arithmetik ist die Monotonie, d.h. für $a, b, c, d \in \mathbb{G}$ und $\bullet \in \{+, -, \times\}$ folgt aus der Ungleichung $a \bullet b \leq c \bullet d$ stets auch die Gültigkeit von $\mathrm{fl}(a \bullet b) \leq \mathrm{fl}(c \bullet d)$, solange kein Überlauf auftritt (vgl. [24, S. 56]).

(iii) In [24, Lemma 3.1] wird im Fall $nu < 1$ und $|\delta_k| \leq u$ für $k = 1, \ldots, n$ zusätzlich die Abschätzung

$$\prod_{k=1}^{n} (1 + \delta_k) \;\leq\; 1 + \frac{nu}{1 - nu}$$

zur Verfügung gestellt, welche aus der für $1 \leq k \leq n$ gültigen Ungleichung $\binom{n}{k} u^k \leq (nu)^k$ und der geometrischen Summenformel resultiert. Auf diese Weise kann die $\mathcal{O}(u^2)$-Schreibweise gegebenenfalls umgangen werden. □

Die Addition von n Zahlen kann ebenso als Ausführung des Skalarproduktes im $\mathbb{R}^n$ mit dem Einsvektor verstanden werden. Allgemein gilt:

Folgerung 2.6. *Gegeben seien der Vektor $\mathbf{g} = (g_k)_{k=0}^{n-1} \in \mathbb{G}^n$ mit $\|\mathbf{g}\|_1 \in R_{\mathbb{G}}$ und der Einsvektor $\mathbf{1} = (1)_{k=0}^{n-1} \in \mathbb{G}^n$. Dann erfüllt der Rundungsfehler für das Skalarprodukt $\mathbf{g}^{\mathrm{T}}\mathbf{1}$ die Fehlerabschätzungen*

$$\left| \mathrm{fl}\left(\mathbf{g}^{\mathrm{T}}\mathbf{1} \right) - \mathbf{g}^{\mathrm{T}}\mathbf{1} \right| \;\leq\; (n-1)\, \|\mathbf{g}\|_1 \, \left(u \;+\; \mathcal{O}(u^2) \right) \;\leq\; n(n-1)\, \|\mathbf{g}\|_\infty \, \left(u \;+\; \mathcal{O}(u^2) \right) \tag{2.9a}$$

im Fall von sequentieller Summation und

$$\left| \mathrm{fl}\left(\mathbf{g}^{\mathrm{T}}\mathbf{1} \right) - \mathbf{g}^{\mathrm{T}}\mathbf{1} \right| \;\leq\; \lceil \log_2 n \rceil \, \|\mathbf{g}\|_1 \, \left(u \;+\; \mathcal{O}(u^2) \right) \;\leq\; n \lceil \log_2 n \rceil \, \|\mathbf{g}\|_\infty \, \left(u \;+\; \mathcal{O}(u^2) \right) \tag{2.9b}$$

im Fall von Kaskaden-Summation.

Beweis: Offenbar gilt zunächst für einen beliebigen Vektor $\mathbf{g} = (g_k)_{k=0}^{n-1} \in \mathbb{G}^n$ stets

$$\sum_{k=0}^{n-1} |g_k| \;=\; \|\mathbf{g}\|_1 \;\leq\; n\, \|\mathbf{g}\|_\infty \;.$$

Bei der sequentiellen Addition ist eine beliebige Komponente g_k maximal an $n-1$ Additionen beteiligt (vgl. Abb. 2.3). Somit treten innerhalb des Rundungsfehlers, welcher analog zu Beispiel 2.4 angegeben werden kann, maximal $n-1$ Summanden auf, welche $|g_k|$ enthalten und durch einen in u linearen Term abgeschätzt werden können. Bei der Kaskaden-Summation hingegen ist eine beliebige Komponente g_k maximal an $\lceil \log_2 n \rceil$ Additionen beteiligt (vgl. Abb. 2.4). Damit folgen die Behauptungen. ■

Analog ergeben sich Abschätzungen für das Skalarprodukt zweier beliebiger Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{G}^n$, wenn wir zunächst alle Produkte $\mathrm{fl}(x_k y_k),\ k = 0, \ldots, n-1$, ausführen, in einem Vektor

$$\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times} \right) := \left(x_k y_k (1 + \varepsilon_k^{\times}) \right)_{k=0}^{n-1}$$

zusammenfassen und anschließend aufsummieren. Mit Folgerung 2.6 ergibt sich:

Folgerung 2.7. *Erfüllen* $\mathbf{x} = (x_k)_{k=0}^{n-1} \in \mathbb{G}^n$ *und* $\mathbf{y} = (y_k)_{k=0}^{n-1} \in \mathbb{G}^n$ *die Bedingung*

$$\|\mathbf{x} \circ \mathbf{y}\|_1 \;=\; \sum_{k=0}^{n-1} |x_k y_k| \in R_{\mathbb{G}} \;, \tag{2.10}$$

so ergibt sich für den Rundungsfehler bei Ausführung des Skalarproduktes

$$\left|\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \mathbf{x}^{\mathrm{T}}\mathbf{y}\right| \;\leq\; n\,\|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \;\leq\; n^2\,\|\mathbf{x} \circ \mathbf{y}\|_\infty \left(u \;+\; \mathcal{O}(u^2)\right) \tag{2.11a}$$

im Fall von sequentieller Summation und

$$\left.\begin{aligned} \left|\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \mathbf{x}^{\mathrm{T}}\mathbf{y}\right| \;&\leq\; (\lceil \log_2 n \rceil + 1)\;\|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \\ &\leq\; n\,(\lceil \log_2 n \rceil + 1)\;\|\mathbf{x} \circ \mathbf{y}\|_\infty \left(u \;+\; \mathcal{O}(u^2)\right) \end{aligned}\right\} \tag{2.11b}$$

im Fall von Kaskaden-Summation.

Beweis: Wegen $\mathbf{x}^{\mathrm{T}}\mathbf{y} = (\mathbf{x} \circ \mathbf{y})^{\mathrm{T}}\mathbf{1}$ und

$$\begin{aligned} \mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \mathbf{x}^{\mathrm{T}}\mathbf{y} \;&=\; \left[\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \left(\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\right)^{\mathrm{T}}\mathbf{1}\right] + \left[\left(\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\right) - (\mathbf{x} \circ \mathbf{y})\right]^{\mathrm{T}}\mathbf{1} \\ &=\; \left[\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \left(\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\right)^{\mathrm{T}}\mathbf{1}\right] + \left(\mathbf{x} \circ \mathbf{y} \circ \boldsymbol{\varepsilon}^{\times}\right)^{\mathrm{T}}\mathbf{1} \end{aligned}$$

ergeben sich nach Folgerung 2.6 für den ersten Term bei sequentieller Summation

$$\begin{aligned} \left|\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \left(\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\right)^{\mathrm{T}}\mathbf{1}\right| \;&\leq\; (n-1)\;\|\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \\ &\leq\; (n-1)\;\|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \end{aligned}$$

und bei Kaskaden-Summation analog

$$\left|\mathrm{fl}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}\right) - \left(\mathbf{x} \circ \mathbf{y} \circ \left(\mathbf{1} + \boldsymbol{\varepsilon}^{\times}\right)\right)^{\mathrm{T}}\mathbf{1}\right| \;\leq\; \lceil \log_2 n \rceil\;\|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right).$$

Da sich $\left(\mathbf{x} \circ \mathbf{y} \circ \boldsymbol{\varepsilon}^{\times}\right)^{\mathrm{T}}\mathbf{1}$ betragsmäßig durch $\|\mathbf{x} \circ \mathbf{y}\|_1\, u$ abschätzen lässt, folgt die Behauptung. ■

Die Ergebnisse aus den Folgerungen 2.6 und 2.7 berücksichtigen noch nicht die Besetzungsstruktur der Vektoren $\mathbf{x}$ und $\mathbf{y}$. Beispielsweise verbessern sich die Abschätzungen, wenn einige Komponenten von $\mathbf{y}$ gleich Null sind, gleichgültig, welche Werte $\mathbf{x}$ in diesen Komponenten annimmt. Ist $\rho \in \mathbb{N}$ die Anzahl der Nichtnullelemente von $\mathbf{y}$, dann können die Faktoren n und n^2 in (2.11a) entsprechend durch ρ und ρ^2 ersetzt werden. Analoges gilt für $\log_2(n)$ und n in (2.11b).
Weiterhin kann der Ausdruck $\left(\mathbf{x} \circ \mathbf{y} \circ \boldsymbol{\varepsilon}^{\times}\right)^{\mathrm{T}}\mathbf{1}$, in welchen die bei den einzelnen Multiplikationen auftretenden relativen Rundungsfehler $\varepsilon_k^{\times},\ k = 0, \ldots, n-1$, eingehen, schärfer abgeschätzt werden. Wird berücksichtigt, dass bei der Multiplikation mit $g \in \{0, \pm 1\}$ kein zusätzlicher Rundungsfehler auftritt, können die zugehörigen $\varepsilon_k^{\times}$ gleich Null gesetzt werden. Führen wir für Matrizen $A = (a_{j,k})_{j,k=0}^{n-1} \in \mathbb{R}^{n \times n}$ und Vektoren $\mathbf{x} = (x_k)_{k=0}^{n-1} \in \mathbb{R}^n$ wie in [54] die Bezeichnungen

$$|A| \;:=\; (|a_{j,k}|)_{j,k=0}^{n-1}\;, \qquad |\mathbf{x}| \;:=\; (|x_k|)_{k=0}^{n-1} \tag{2.12}$$

ein, ergeben sich entsprechende Aussagen für eine Matrix-Vektor-Multiplikation.

Lemma 2.8. *Gegeben seien eine Matrix* $A = (a_{j,k})_{j,k=0}^{n-1} \in \mathbb{G}^{n \times n}$ *und ein Vektor* $\mathbf{x} = (x_k)_{k=0}^{n-1} \in \mathbb{G}^n$. *Weiterhin sei die Matrix* $B = (b_{j,k})_{j,k=0}^{n-1} \in \{0,1\}^{n \times n}$ *durch*

$$b_{j,k} \;:=\; \begin{cases} 1\;, & \text{falls } a_{j,k} \notin \{0, \pm 1\}\;, \\ 0\;, & \text{falls } a_{j,k} \in \{0, \pm 1\} \end{cases}$$

definiert. Enthält jede Zeile $\mathbf{y}^{\mathrm{T}}$ *von* A *höchstens* ρ *Einträge ungleich Null und erfüllt jede Zeile* $\mathbf{y}^{\mathrm{T}}$ *von* A *die Bedingung* (2.10)*, so gelten unter der Modellannahme* (2.6) *für den bei einer Matrix-Vektor-Multiplikation* $A\mathbf{x}$ *auftretenden Rundungsfehler* $\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}$ *die komponentenweisen Abschätzungen*

$$|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}| \;\leq\; \left((\rho - 1)|A| + B \circ |A|\right)|\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right) \tag{2.13a}$$

im Fall von sequentieller Summation und

$$|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}| \;\leq\; \left(\lceil \log_2 \rho \rceil |A| + B \circ |A|\right)|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right) \tag{2.13b}$$

im Fall von Kaskaden-Summation, wobei $B \circ |A| := (b_{j,k}\,|a_{j,k}|)_{j,k=0}^{n-1}$ *bezeichnet.*

Beweis: Sei $\mathbf{a}_j^{\mathrm{T}}$ die j-te Zeile von A und $\mathbf{b}_j^{\mathrm{T}}$ entsprechend die j-te Zeile von B. Gemäß Folgerung 2.7 und anschließender Bemerkung ergibt sich dann für die j-te Komponente des Fehlervektors

$$\begin{aligned}\left|\mathrm{fl}\left(\mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right) - \mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right| &\leq (\rho-1)\,\|\mathbf{a}_j \circ \mathbf{x}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \;+\; \left(\mathbf{a}_j \circ \mathbf{x} \circ \boldsymbol{\varepsilon}^{\times}\right)^{\mathrm{T}} \mathbf{1} \\ &\leq \left((\rho-1)\,|\mathbf{a}_j|^{\mathrm{T}} \;+\; (\mathbf{b}_j \circ |\mathbf{a}_j|)^{\mathrm{T}}\right)|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right)\end{aligned}$$

im Fall von sequentieller Summation und

$$\begin{aligned}\left|\mathrm{fl}\left(\mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right) - \mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right| &\leq \lceil \log_2 \rho \rceil\, \|\mathbf{a}_j \circ \mathbf{x}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) \;+\; \left(\mathbf{a}_j \circ \mathbf{x} \circ \boldsymbol{\varepsilon}^{\times}\right)^{\mathrm{T}} \mathbf{1} \\ &\leq \left(\lceil \log_2 \rho \rceil\; |\mathbf{a}_j|^{\mathrm{T}} \;+\; (\mathbf{b}_j \circ |\mathbf{a}_j|)^{\mathrm{T}}\right)|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right)\end{aligned}$$

im Fall von Kaskaden-Summation. Mit den Identitäten

$$|A||\mathbf{x}| = \left(|\mathbf{a}_j|^{\mathrm{T}}\,|\mathbf{x}|\right)_{j=0}^{n-1}$$

und

$$(B \circ |A|)\,|\mathbf{x}| = \left((\mathbf{b}_j \circ |\mathbf{a}_j|)^{\mathrm{T}}\,|\mathbf{x}|\right)_{j=0}^{n-1}$$

sowie

$$|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}| = \left(\left|\mathrm{fl}\left(\mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right) - \mathbf{a}_j^{\mathrm{T}}\mathbf{x}\right|\right)_{j=0}^{n-1}$$

folgt nun die Behauptung. ■

Ähnliche Ergebnisse sind in verallgemeinerter Form für komplexwertige Gleitkomma-Arithmetik in [54, Lemma 8.4] zu finden, wobei dort wegen $A \in \mathbb{G}^{n\times n}$ und der reellwertigen Arithmetik für unseren Fall $\eta = 0$ sowie $\mu_{\mathbb{C}} = 1$ gewählt werden kann.
Offenbar liefern die Abschätzungen (2.13a) und (2.13b) für Permutationsmatrizen und für die in Abschnitt 1.1 definierte Vorzeichenskalierungsmatrix Σ_n wegen $\rho = 1$ und $B = O_n$ wie gewünscht keinen zusätzlichen Rundungsfehler. Es ist zu beachten, dass sich diese Matrizen tatsächlich schon in $\mathbb{G}^{n\times n}$ befinden. Für eine beliebige Drehmatrix (1.24) gilt im Allgemeinen jedoch $Q_2(\varphi) \notin \mathbb{G}^{2\times 2}$, so dass unter der Modellannahme (2.6) nur die Matrix

$$\hat{Q}_2(\varphi) \;=\; \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \;:=\; \begin{pmatrix} \cos(\varphi)(1+\varepsilon_1) & \sin(\varphi)(1+\varepsilon_2) \\ -\sin(\varphi)(1+\varepsilon_2) & \cos(\varphi)(1+\varepsilon_1) \end{pmatrix} \;=\; Q_2(\varphi) + \begin{pmatrix} \varepsilon_1 & \varepsilon_2 \\ \varepsilon_2 & \varepsilon_1 \end{pmatrix} \circ Q_2(\varphi)$$

mit $|\varepsilon_1|, |\varepsilon_2| \leq u$ zur Verfügung steht. Wenden wir nun (2.6) auf die Matrix-Vektor-Multiplikation $\hat{Q}_2(\varphi)\mathbf{x}$ mit einem $\mathbf{x} \in \mathbb{G}^2$ an, erhalten wir

$$\begin{aligned}\mathrm{fl}\left(\mathrm{fl}(cx_0) + \mathrm{fl}(sx_1)\right) &= \left(cx_0(1+\varepsilon_1^{\times}) + sx_1(1+\varepsilon_2^{\times})\right)(1+\varepsilon_1^{+}) \\ &= (cx_0 + sx_1)\,(1+\varepsilon_1^{+}) + \left(cx_0\varepsilon_1^{\times} + sx_1\varepsilon_2^{\times}\right)\left(1+\varepsilon_1^{+}\right)\end{aligned}$$

mit $|\varepsilon_1^{\times}|, |\varepsilon_2^{\times}|, |\varepsilon_1^{+}| \leq u$ als Ergebnis für die erste Komponente und analog

$$\begin{aligned}\mathrm{fl}\left(-\mathrm{fl}(sx_0) + \mathrm{fl}(cx_1)\right) &= \left(-sx_0(1+\varepsilon_3^{\times}) + cx_1(1+\varepsilon_4^{\times})\right)(1+\varepsilon_2^{+}) \\ &= (-sx_0 + cx_1)\,(1+\varepsilon_2^{+}) + \left(-sx_0\varepsilon_3^{\times} + cx_1\varepsilon_4^{\times}\right)\left(1+\varepsilon_2^{+}\right)\end{aligned}$$

mit $|\varepsilon_3^{\times}|, |\varepsilon_4^{\times}|, |\varepsilon_2^{+}| \leq u$ als Ergebnis für die zweite Komponente. Somit ergibt sich für den Rundungsfehler die komponentenweise Abschätzung

$$\left|\mathrm{fl}\left(\hat{Q}_2(\varphi)\mathbf{x}\right) - \hat{Q}_2(\varphi)\mathbf{x}\right| \;\leq\; \left|\hat{Q}_2(\varphi)\mathbf{x}\right| u \;+\; \left|\hat{Q}_2(\varphi)\right|\,|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right)\;. \tag{2.14}$$

Da wir hierbei lediglich verwendet haben, dass in jeder Zeile genau eine Addition auftritt, lässt sich die Aussage auf Matrizen mit höchstens zwei Nichtnulleinträgen pro Zeile verallgemeinern.

Folgerung 2.9. *Gegeben seien eine Matrix* $A = (a_{j,k})_{j,k=0}^{n-1} \in \mathbb{G}^{n\times n}$ *und ein Vektor* $\mathbf{x} = (x_k)_{k=0}^{n-1} \in \mathbb{G}^n$, *so dass* $\mathbf{a}_j^{\mathrm{T}}$ *für* $j = 0, \ldots, n-1$ *maximal zwei Einträge ungleich Null besitzt und* $|\mathbf{a}_j|^{\mathrm{T}}|\mathbf{x}| \in R_{\mathbb{G}}$ *für alle* $j = 0, \ldots, n-1$ *erfüllt bleibt. Dann gilt*

$$|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}| \;\leq\; |A\mathbf{x}|u \;+\; |A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right). \tag{2.15}$$

Insbesondere ist Folgerung 2.9 auf Blockdiagonalmatrizen

$$A := \bigoplus_{k=0}^{l-1} B_k$$

mit Blöcken $B_k \in \mathbb{G}^{2\times 2}$ anwendbar. Nehmen wir zusätzlich an, dass die Blöcke die Gestalt

$$B_k := \begin{pmatrix} g_{1k} & g_{2k} \\ -g_{2k} & g_{1k} \end{pmatrix} \tag{2.16}$$

mit $g_{1k}^2 + g_{2k}^2 > 0$ besitzen, so sind sie wegen $B_k^{\mathrm{T}} B_k = \left(g_{1k}^2 + g_{2k}^2\right) I_2$, $k = 0, \ldots, l-1$, fast orthogonal – d.h. orthogonal bis auf einen positiven Faktor – und somit regulär. Demnach ist auch A regulär mit Spektralnorm

$$\|A\|_2 \;=\; \max_{k=0,\ldots,l-1} \|B_k\|_2 \;=\; \max_{k=0,\ldots,l-1} \sqrt{g_{1k}^2 + g_{2k}^2}\,. \tag{2.17}$$

Weiterhin besitzen die symmetrischen Matrizen $|B_k|$, $k = 0, \ldots, l-1$, die Eigenwerte $\lambda_{1,2}^{(k)} = |g_{1k}| \pm |g_{2k}|$ zu den Eigenvektoren $\mathbf{v}_{1,2}^{(k)} = (1, \pm 1)^{\mathrm{T}}$, so dass sich für die Spektralnorm von $|A|$ die Abschätzung

$$\big\||A|\big\|_2 \;=\; \max_{k=0,\ldots,l-1} \big\||B_k|\big\|_2 \;\leq\; \sqrt{2}\,\|A\|_2 \tag{2.18}$$

ergibt. Dies ist beispielsweise mittels Cauchy-Schwarz-Ungleichung ersichtlich, wenn sie auf die Vektoren $(1,1)^{\mathrm{T}}$ und $(|g_{1k}|, |g_{2k}|)^{\mathrm{T}}$ angewandt wird. Zusammen mit Folgerung 2.9 erhalten wir:

Lemma 2.10. *Sei* $n \in \mathbb{N}$ *gerade,* $A \in \mathbb{G}^{n\times n}$ *eine Blockdiagonalmatrix mit fast orthogonalen Blöcken wie in* (2.16) *definiert und* $\mathbf{x} \in \mathbb{G}^n$. *Im Fall* $|A||\mathbf{x}| \in R_{\mathbb{G}}^n$ *gilt dann*

$$\big\|\,\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}\big\|_2 \;\leq\; \left(1 + \sqrt{2}\right)\|A\|_2\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right).$$

Beweis: Zunächst halten wir fest, dass sich für beliebige Vektoren $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ mit $|\mathbf{a}| \leq |\mathbf{b}|$ wegen der Nichtnegativität zunächst auch

$$|a_i|^2 \;\leq\; |b_i|^2 \qquad (i = 0, \ldots, n-1)$$

gilt und sich aufgrund der Rechenregeln für Ungleichungen sowie der Monotonie der Wurzelfunktion dann ebenso $\|\mathbf{a}\|_2 \leq \|\mathbf{b}\|_2$ ergibt. Aus (2.15) folgt dann mittels Dreiecksungleichung

$$\begin{aligned} \big\|\,|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}|\,\big\|_2 &\leq \big\||A\mathbf{x}|u \;+\; |A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right)\big\|_2 \\ &\leq \big\||A\mathbf{x}|\|_2 u \;+\; \big\||A||\mathbf{x}|\big\|_2\left(u \;+\; \mathcal{O}(u^2)\right) \end{aligned}$$

Verwenden wir noch die Gültigkeit von $\big\||\mathbf{x}|\big\|_2 = \|\mathbf{x}\|_2$ und $\|A\mathbf{x}\|_2 \leq \|A\|_2\|\mathbf{x}\|_2$ sowie die Abschätzung (2.18), ergibt sich die Behauptung. ■

Bemerkung 2.11. In [43, Lemma 5.1] wird für beliebige $a, b, c, d \in \mathbb{R}$ die Ungleichung

$$(|ac| + |bd| + |ac - bd|)^2 + (|ad| + |bc| + |ad + bc|)^2 \;\leq\; \frac{16}{3}(a^2 + b^2)(c^2 + d^2) \tag{2.19}$$

gezeigt, welche beispielsweise für $a = c = \frac{1}{\sqrt{2}}$ und $b = d = 1$ scharf ist. Demzufolge ist die Konstante $\frac{16}{3}$ für die allgemeine Ungleichung (2.19) bestmöglich. □

Aufgrund der Äquivalenz der offensichtlich erfüllten Ungleichung $\frac{7}{3} < 2\sqrt{2}$ zu $\frac{16}{3} < (1 + \sqrt{2})^2$, lässt sich die Abschätzung in Lemma 2.10 noch etwas verbessern.

Satz 2.12. *Sei $n \in \mathbb{N}$ gerade, $A \in \mathbb{G}^{n\times n}$ eine Blockdiagonalmatrix mit fast orthogonalen Blöcken wie in* (2.16) *definiert und $\mathbf{x} \in \mathbb{G}^n$. Im Fall $|A||\mathbf{x}| \in R^n_{\mathbb{G}}$ gilt dann*

$$\left\| \mathrm{fl}(A\mathbf{x}) - A\mathbf{x} \right\|_2 \;\leq\; \tfrac{4}{3}\sqrt{3}\,\|A\|_2\;\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right).$$

Beweis: Analog wie im Beweis von Lemma 2.10 erhalten wir aus (2.15) zunächst die Ungleichung

$$\left\| \left|\mathrm{fl}(A\mathbf{x}) - A\mathbf{x}\right| \right\|_2^2 \;\leq\; \left\| |A\mathbf{x}| \;+\; |A||\mathbf{x}| \right\|_2^2 \left(u \;+\; \mathcal{O}(u^2)\right)^2 .$$

Aufgrund der Blockstruktur von A und der besonderen Form (2.16) der einzelnen Blöcke können wir jeweils Ungleichung (2.19) anwenden und erhalten

$$\begin{aligned}
\left\| |A\mathbf{x}| \;+\; |A||\mathbf{x}| \right\|_2^2 \;&=\; \sum_{k=0}^{\frac{n}{2}-1} \left(|g_{1k}x_{2k+1}| + |g_{2k}x_{2k}| + |g_{1k}x_{2k+1} + g_{2k}x_{2k}|\right)^2 \\
&\qquad +\; \sum_{k=0}^{\frac{n}{2}-1} \left(|g_{1k}x_{2k}| + |g_{2k}x_{2k+1}| + |g_{1k}x_{2k} - g_{2k}x_{2k+1}|\right)^2 \\
&\leq\; \frac{16}{3} \sum_{k=0}^{\frac{n}{2}-1} \|B_k\|_2^2 \left(x_{2k}^2 + x_{2k+1}^2\right) .
\end{aligned}$$

Wegen (2.17) folgt insbesondere $\|B_k\|_2 \leq \|A\|_2$ für $k = 0, \ldots, \frac{n}{2} - 1$, so dass sich

$$\left\| |A\mathbf{x}| \;+\; |A||\mathbf{x}| \right\|_2^2 \;\leq\; \frac{16}{3}\|A\|_2^2\|\mathbf{x}\|_2^2$$

und nach anschließendem Wurzelziehen die Behauptung ergibt. ■

Nun haben wir alle Werkzeuge bereitgestellt, um in Abschnitt 3.1 eine Rundungsfehleranalyse für die in Abschnitt 1.3 hergeleiteten Algorithmen durchführen zu können.

2.2 Stochastisches Modell für Gleitkomma-Arithmetik

Während in Abschnitt 2.1 für den relativen Rundungsfehler obere Schranken unter Vernachlässigung von Termen der Ordnung $\mathcal{O}(u^2)$ hergeleitet werden, beschäftigen wir uns nun mit der Vorhersage des Fehlerverhaltens. Dazu nehmen wir jede Eingangsgröße X als reelle Zufallsvariable mit bekanntem Erwartungswert $\mathbb{E}(X)$ und endlichem zweiten Moment $\mathbb{E}(X^2) < \infty$ an. Entsprechend [75, 54] fassen wir im Folgenden den durch $\mathrm{fl}(X \bullet Y) = (X \bullet Y)(1 + \varepsilon^\bullet)$ für zufällige Eingangsdaten X, Y und für jede Operation $\bullet \in \{+, -, \times\}$ gegebenen relativen Rundungsfehler $\varepsilon^\bullet$ (vgl. (2.6)) als eine reelle Zufallsvariable mit Erwartungswert

$$\mathbb{E}\left(\varepsilon^\bullet\right) \;=\; \mu_\bullet u \tag{2.20}$$

und Varianz

$$\mathbb{V}\left(\varepsilon^\bullet\right) \;=\; \sigma_\bullet^2 u^2 \tag{2.21}$$

auf, wobei die Konstanten $\mu_\bullet \in \mathbb{R}, \sigma_\bullet \geq 0$ von den Operationen $\bullet \in \{+, -, \times\}$ und der Verteilung der Eingangsdaten abhängen. Desweiteren bezeichnet $\delta^\bullet := (X \bullet Y)\varepsilon^\bullet$ die Zufallsvariable des entsprechenden absoluten Rundungsfehlers.

Bemerkung 2.13. Rundungsfehler zu festen Eingangsdaten sind im Allgemeinen nicht zufällig, sondern hängen im Wesentlichen nur von der jeweiligen Implementation der Gleitkomma-Arithmetik ab. Dennoch können wir die Zufälligkeit von $\varepsilon^\bullet$ mit der Zufälligkeit der Eingangsdaten begründen [75]. Der Einfachheit halber werden wir im Folgenden analog [54, 75] stets annehmen, dass $\varepsilon^\bullet$ sowohl von X als auch von Y stochastisch unabhängig ist. Genau genommen ist dies zwar nur für den trivialen Fall erfüllt, jedoch wird die Annahme der Unabhängigkeit durch die Tatsache motiviert, dass die Größe des relativen Fehlers $\varepsilon^\bullet$ nicht mit der Größe der Operanden in Zusammenhang steht und dass die letzten Kommastellen, welche den größten Einfluss auf den Rundungsfehler ausüben, vom Rest der Mantisse näherungsweise unabhängig sind [75]. Im Anhang A.3 finden sich überdies experimentelle Ergebnisse, welche die in den folgenden Unterabschnitten 2.2.1 und 2.2.2 verwendeten Modellannahmen ebenfalls als näherungsweise erfüllt bestätigen. □

2.2.1 Modell für eine allgemeine Matrix

Ähnlich wie in Abschnitt 2.1 betrachten wir wiederum alle relevanten Szenarien, angefangen bei der Addition von n Zahlen, welche als Skalarprodukt im $\mathbb{R}^n$ mit dem Einsvektor verstanden werden kann.

Lemma 2.14. *Zu $n \in \mathbb{N},\ n \geq 2$, seien gegeben der Einsvektor $\mathbf{1} = (1)_{k=0}^{n-1} \in \mathbb{G}^n$ und der Zufallsvektor $\mathbf{G} = (G_k)_{k=0}^{n-1}$ mit unkorrelierten Einträgen aus $\mathbb{G}^n$, deren Erwartungswerte Null und deren zweite Momente endlich sind. Angenommen, alle auftretenden relativen Rundungsfehler $\varepsilon_j^\bullet$ seien untereinander und von den G_k stochastisch unabhängig. Dann besitzt der Rundungsfehler $\Delta := \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}$ für das Skalarprodukt $\mathbf{G}^{\mathrm{T}}\mathbf{1}$ unabhängig von der Summationsweise den Erwartungswert Null. Die Varianz ist*

$$\begin{aligned}\mathbb{V}(\Delta) \;=\; \mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; &\left((n-1)\mathbb{V}(G_0) + \sum_{j=1}^{n-1}(n-j)\mathbb{V}(G_j)\right)\sigma_+^2\left(u^2 + \mathcal{O}(u^3)\right) \\ &+\left((n-1)^2\mathbb{V}(G_0) + \sum_{k=1}^{n-1}(n-k)^2\mathbb{V}(G_k)\right)\mu_+^2\left(u^2 + \mathcal{O}(u^3)\right)\end{aligned}$$

im Fall von sequentieller Summation und

$$\mathbb{V}(\Delta) = \begin{cases} \left(\left(2\sigma_+^2 + 4\mu_+^2\right)\sum_{k=0}^{1}\mathbb{V}(G_k) \;+\; \left(\sigma_+^2 + \mu_+^2\right)\mathbb{V}(G_2)\right)\left(u^2 + \mathcal{O}(u^3)\right), & n = 3 \\ \left(\left(3\sigma_+^2 + 9\mu_+^2\right)\sum_{k=0}^{3}\mathbb{V}(G_k) + \left(\sigma_+^2 + \mu_+^2\right)\mathbb{V}(G_4)\right)\left(u^2 + \mathcal{O}(u^3)\right), & n = 5 \\ \left(\left(3\sigma_+^2 + 9\mu_+^2\right)\sum_{k=0}^{3}\mathbb{V}(G_k) + \left(2\sigma_+^2 + 4\mu_+^2\right)\sum_{k=4}^{5}\mathbb{V}(G_k)\right)\left(u^2 + \mathcal{O}(u^3)\right), & n = 6 \\ \left(\left(3\sigma_+^2 + 9\mu_+^2\right)\sum_{k=0}^{3}\mathbb{V}(G_k) + \left(2\sigma_+^2 + 4\mu_+^2\right)\sum_{k=4}^{5}\mathbb{V}(G_k) + \left(\sigma_+^2 + \mu_+^2\right)\mathbb{V}(G_6)\right)\left(u^2 + \mathcal{O}(u^3)\right), & n = 7 \end{cases}$$

bzw.

$$\mathbb{V}(\Delta) \;=\; \left(\log_2(n)\cdot\sigma_+^2 + \left(\log_2(n)\right)^2\cdot\mu_+^2\right)\sum_{k=0}^{n-1}\mathbb{V}(G_k)\left(u^2 + \mathcal{O}(u^3)\right)$$

für allgemeine Zweierpotenz $n = 2^t$ $(t \in \mathbb{N})$ im Fall von Kaskaden-Summation.

Der **Beweis** ist in Anhang A.2 zu finden. ■

Unter zusätzlichen Annahmen lassen sich die Ergebnisse aus Lemma 2.14 noch weiter vereinfachen.

Folgerung 2.15. *Sei $n \geq 2$ eine Zweierpotenz. Mit den Voraussetzungen und Bezeichnungen aus Lemma 2.14 gelten die folgenden Aussagen:*

(i) *Ist zusätzlich $\mathbb{V}(G_k) \;=\; \varrho^2$ für alle $k \in \{0, 1, \ldots, n-1\}$, dann ergibt sich*

$$\mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; \left(\frac{(n-1)(n+2)}{2}\sigma_+^2 \;+\; \frac{(n-1)(2n^2+5n-6)}{6}\mu_+^2\right)\varrho^2\left(u^2 + \mathcal{O}(u^3)\right)$$

im Fall von sequentieller Summation und

$$\mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; n\left(\log_2(n)\cdot\sigma_+^2 + \left(\log_2(n)\right)^2\cdot\mu_+^2\right)\varrho^2\left(u^2 + \mathcal{O}(u^3)\right)$$

im Fall von Kaskaden-Summation.

(ii) *Gilt zusätzlich $\mu_+ = 0$, dann folgt*

$$\mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;= \begin{cases} \left((n-1)\mathbb{V}(G_0) + \sum_{j=1}^{n-1}(n-j)\mathbb{V}(G_j)\right)\sigma_+^2\left(u^2 + \mathcal{O}(u^3)\right) & \text{bei seq. Sum.,} \\ \log_2(n)\cdot\sigma_+^2\sum_{k=0}^{n-1}\mathbb{V}(G_k)\left(u^2 + \mathcal{O}(u^3)\right) & \text{bei Kas.-Sum.} \end{cases}$$

(iii) *Ist sowohl* $\mathbb{V}(G_k) = \varrho^2$ *für alle* $k \in \{0, 1, \ldots, n-1\}$ *als auch* $\mu_+ = 0$*, dann ergibt sich*

$$\mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) = \begin{cases} \frac{(n-1)(n+2)}{2} \cdot \varrho^2\sigma_+^2\left(u^2 + \mathcal{O}(u^3)\right) & \text{bei sequentieller Summation,} \\ n\ \log_2(n) \cdot \varrho^2\sigma_+^2\left(u^2 + \mathcal{O}(u^3)\right) & \text{bei Kaskaden-Summation.} \end{cases}$$

Beweis: Alle Aussagen folgen aus den Formeln in Lemma 2.14, wobei gegebenenfalls einfache Summenformeln angewandt werden müssen. ■

Im nächsten Schritt untersuchen wir das Fehlerverhalten bei einem Skalarprodukt.

Lemma 2.16. *Zu* $n \in \mathbb{N},\ n \geq 2$*, seien zwei unabhängige Zufallsvektoren* $\mathbf{X} = (X_k)_{k=0}^{n-1} \in \mathbb{G}^n$ *und* $\mathbf{G} = (G_k)_{k=0}^{n-1} \in \mathbb{G}^n$ *mit unkorrelierten Einträgen gegeben, deren zweite Momente endlich sind. Weiterhin gelte* $\mathbb{E}(X_k) = 0,\ k = 0, \ldots, n-1$*. Angenommen, alle auftretenden relativen Rundungsfehler* $\varepsilon_j^\bullet$ *seien untereinander und von den* G_k *und* X_k *stochastisch unabhängig. Dann besitzt der Rundungsfehler* $\Delta := \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{X}$ *für das Skalarprodukt* $\mathbf{G}^{\mathrm{T}}\mathbf{X}$ *unabhängig von der Summationsweise den Erwartungswert Null. Die Varianz ist*

$$\begin{aligned} \mathbb{V}(\Delta) &= \sum_{k=0}^{n-1} \mathbb{E}\left((G_k)^2\right) \mathbb{V}(X_k)\left(\sigma_\times^2 + \mu_\times^2\right) u^2 \\ &\quad + \left((n-1)\ \mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0) + \sum_{k=1}^{n-1}(n-k)\ \mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\right)\left(\sigma_+^2 + 2\mu_\times\mu_+\right)\left(u^2 + \mathcal{O}(u^3)\right) \\ &\quad + \left((n-1)^2\ \mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0) + \sum_{k=1}^{n-1}(n-k)^2\ \mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\right)\mu_+^2\left(u^2 + \mathcal{O}(u^3)\right) \end{aligned}$$

im Fall von sequentieller Summation und

$$\mathbb{V}(\Delta) = \left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2\right) \sum_{r=0}^{n-1} \mathbb{E}\left(G_r^2\right) \mathbb{V}(X_r)\left(u^2 + \mathcal{O}(u^3)\right)$$

für eine allgemeine Zweierpotenz $n = 2^t$ $(n \in \mathbb{N})$ *im Fall von Kaskaden-Summation.*

Der **Beweis** ist in Anhang A.2 zu finden. ■

Unter zusätzlichen Annahmen lassen sich die Ergebnisse aus Lemma 2.16 noch weiter vereinfachen.

Folgerung 2.17. *Sei* $n \geq 2$ *eine Zweierpotenz. Mit den Voraussetzungen und Bezeichnungen aus Lemma* 2.16 *gelten die folgenden Aussagen:*

(i) *Ist zusätzlich* $\mu_+ = 0$*, dann ergibt sich*

$$\mathbb{V}(\Delta) = \left(\sum_{k=0}^{n-1}\left((n-k)\sigma_+^2 + \sigma_\times^2 + \mu_\times^2\right)\mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k) - \sigma_+^2\mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0)\right)\left(u^2 + \mathcal{O}(u^3)\right)$$

im Fall der sequentiellen Summation und

$$\mathbb{V}(\Delta) = \left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \mu_\times^2\right) \sum_{r=0}^{n-1} \mathbb{E}\left(G_r^2\right) \mathbb{V}(X_r)\left(u^2 + \mathcal{O}(u^3)\right)$$

im Fall der Kaskaden-Summation.

(ii) *Gilt zusätzlich* $\mu_+ = \mu_\times = 0$*, dann ergibt sich sinngemäß die Behauptung aus* [75, Lemma 3.4].

(iii) *Falls* $\mathbb{V}(X_k) = \varrho^2$ *für alle* $k \in \{0, 1, \ldots, n-1\}$*, so gilt bei Kaskaden-Summation*

$$\mathbb{V}(\Delta) = \varrho^2\left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2\right)\mathbb{E}\left(\|\mathbf{G}\|_2^2\right)\left(u^2 + \mathcal{O}(u^3)\right)\ .$$

Beweis: Die Aussagen (i) und (ii) folgen sofort aus Lemma 2.16. Ebenso folgt Aussage (iii), wobei hier noch die Linearität des Erwartungswertes verwendet wird. ■

Ist neben den Voraussetzungen von Lemma 2.16 bzw. Folgerung 2.17 zusätzlich bekannt, dass r ($0 < r < n-1$) der Zufallsgrößen G_k , $k = 0, \ldots, n$, konstant Null sind und demnach auch Varianz Null besitzen, kann Lemma 2.16 bzw. Folgerung 2.17 nach entsprechender Umordnung der Komponenten bereits mit $\tilde{n} = n - r$ angewandt werden.
Da sich eine Matrix-Vektor-Multiplikation aus mehreren voneinander unabhängigen Skalarprodukten zusammensetzt, erhalten wir aus Lemma 2.16 bzw. Folgerung 2.17 sofort auch entsprechende Aussagen über die Komponenten des Vektors der dabei auftretenden absoluten Fehler. Fordern wir zusätzlich, dass die Komponenten des Vektors, welcher mit der Matrix multipliziert werden soll, jeweils Erwartungswert Null besitzen, so gelangen wir zu nachstehenden Ergebnissen.

Satz 2.18. *Zu einer Zweierpotenz $n \in \mathbb{N}$, $n \geq 2$, sei $\mathbf{X} = (X_k)_{k=0}^{n-1} \in \mathbb{G}^n$ ein Zufallsvektor mit unkorrelierten Komponenten, deren zweite Momente endlich sind. Weiterhin sei eine Zufallsmatrix $G := (G_{jk})_{j,k=0}^{n-1} \in \mathbb{G}^{n\times n}$ gegeben, wobei die Vektoren $(G_{jk})_{k=0}^{n-1}$, $j = 0, \ldots, n-1$, jeweils unabhängig von $\mathbf{X}$ seien und unkorrelierte Komponenten besitzen, deren zweite Momente endlich sind. Zusätzlich gelte $\mathbb{E}(X_k) = 0$, $k = 0, \ldots, n-1$. Angenommen, alle auftretenden relativen Rundungsfehler $\varepsilon_m^\bullet$ seien untereinander und von den G_{jk} und X_k stochastisch unabhängig. Dann besitzt der Vektor der absoluten Rundungsfehler*

$$\boldsymbol{\Delta} \;:=\; \mathrm{fl}\,(G\mathbf{X}) - G\mathbf{X} \tag{2.22}$$

der Matrix-Vektor-Multiplikation $G\mathbf{X}$ unabhängig von der bei den jeweils auftretenden Skalarprodukten verwendeten Summationsweise in jeder Komponente Erwartungswert Null und es gilt

$$\mathrm{tr}\,(\mathrm{Cov}(\boldsymbol{\Delta})) \;=\; \left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2\right) \sum_{j,k=0}^{n-1} \mathbb{E}\left(G_{jk}^2\right) \mathbb{V}(X_k) \left(u^2 + \mathcal{O}(u^3)\right)$$

im Fall von Kaskaden-Summation. Falls zusätzlich noch $\mathbb{V}(X_k) = \varrho^2$ für alle $k \in \{0, 1, \ldots, n-1\}$ erfüllt ist, dann ist

$$\mathrm{tr}\,(\mathrm{Cov}(\boldsymbol{\Delta})) \;=\; \varrho^2 \left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2\right) \mathbb{E}\left(\|G\|_F^2\right) \left(u^2 + \mathcal{O}(u^3)\right) \;.$$

Dabei bezeichnet $\mathrm{tr}(A)$ die Spur und $\|A\|_F$ die Frobeniusnorm einer Matrix $A \in \mathbb{R}^{n\times n}$.
Beweis: Mit $\mathbf{G}^{(j)} := (G_{jk})_{k=0}^{n-1}$ besitzt der Fehlervektor (2.22) die Komponenten

$$\Delta_j \;:=\; \mathrm{fl}\left(\mathbf{G^{(j)}}^{\mathrm{T}}\mathbf{X}\right) - \mathbf{G^{(j)}}^{\mathrm{T}}\mathbf{X} \;,$$

die nach Lemma 2.16 Erwartungswert Null und im Fall der Kaskaden-Summation Varianz

$$\mathbb{V}(\Delta_j) \;=\; \left(\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2\right) \sum_{r=0}^{n-1} \mathbb{E}\left(G_{jr}^2\right) \mathbb{V}(X_r) \left(u^2 + \mathcal{O}(u^3)\right) \tag{2.23}$$

besitzen. Wegen

$$\mathrm{tr}\,(\mathrm{Cov}(\boldsymbol{\Delta})) \;=\; \sum_{j=0}^{n-1} \mathbb{V}(\Delta_j)$$

ergibt sich damit der erste Teil der Behauptung. Mit der Linearität des Erwartungswertes und demnach

$$\sum_{j,k=0}^{n-1} \mathbb{E}\left(G_{jr}^2\right) \;=\; \mathbb{E}\left(\sum_{j,k=0}^{n-1} G_{jr}^2\right) \;=\; \mathbb{E}\left(\|G\|_F^2\right)$$

erhalten wir für $\mathbb{V}(X_k) = \varrho^2$, $k = 0, \ldots, n-1$, aus dem ersten Teil der Behauptung dann auch den zweiten Teil. ■

Die Spur der Kovarianzmatrix des Fehlervektors $\boldsymbol{\Delta}$ ist wegen $|\,\mathrm{Cov}(\Delta_j, \Delta_k)|^2 \leq \mathbb{V}(\Delta_j)\mathbb{V}(\Delta_k)$ ein geeignetes Maß, da sie nur in dem Fall verschwindet, wenn fast sicher $\boldsymbol{\Delta} = \mathbb{E}(\boldsymbol{\Delta})$, also in der Situation von Satz 2.18 alle Komponenten von $\boldsymbol{\Delta}$ fast sicher Null sind.

Bemerkung 2.19. Besitzt die Zufallsmatrix $G := (G_{jk})_{j,k=0}^{n-1} \in \mathbb{G}^{n\times n}$ aus Folgerung 2.17 als Erwartungswert näherungsweise eine der vollbesetzten Matrizen aus Definition 1.1 und ist die Varianz jeder einzelnen Komponente kleiner oder gleich u^2, so ergibt sich mit der Orthogonalität von (1.2)–(1.7) und aufgrund der angenommenen Unkorreliertheit $\mathbb{E}\left(\|G\|_F^2\right) \leq n + n^2u^2$. Die zweite Aussage aus Satz 2.18 liest sich dann als

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \leq \breve{k}_n^2 \mathbb{E}\left(\|X\|_2^2\right)\left(u^2 + \mathcal{O}(u^3)\right) \tag{2.24}$$

mit

$$\breve{k}_n = \sqrt{\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2}\,, \tag{2.25}$$

wobei sowohl $\mathbb{E}(\boldsymbol{\Delta}) = \mathbf{0}$ als auch $\mathbb{V}(X_k) = \varrho^2$ und $\mathbb{E}(X_k) = 0,\ k = 0,\ldots,n-1$, berücksichtigt worden ist. Jedoch sind in diesem Fall die Annahmen an $\mathbf{X}$ derart einschränkend, dass ein Anwendungsbezug kaum möglich scheint. Lassen wir die Forderung fallen, dass $\mathbb{V}(X_k) = \varrho^2$ für alle $k = 0,\ldots,n-1$ gilt, so liefert Aufsummieren der n Gleichungen (2.23) mit der aus Definition 1.1 stammenden Beziehung

$$\max_{r=0,\ldots,n-1} \mathbb{E}(G_{jr}^2) \leq \sqrt{\frac{2}{n}} + u^2$$

für die Abschätzung (2.24) die von n abhängige Konstante

$$\breve{k}_n = \sqrt{\sigma_\times^2 + \log_2(n)\sigma_+^2 + \left(\mu_\times + \log_2(n)\mu_+\right)^2}\,\sqrt[4]{2n} \tag{2.26}$$

In Tabelle 2.2 sind explizite Werte der $\breve{k}_n$ für ausgewählte n mit den in Anhang A.3 geschätzten Größen $\sigma_\times = \sigma_+ = 0.425$ und $\mu_\times = \mu_+ = 0$ angegeben. □

n	8	16	32	64	128	256	512	1024	2048
$\breve{k}_n$ aus (2.25)	0.8500	0.9503	1.0410	1.1244	1.2021	1.2750	1.3440	1.4096	1.4722
$\breve{k}_n$ aus (2.26)	1.7000	2.2603	2.9445	3.7822	4.8083	6.0650	7.6026	9.4824	11.7779

Tabelle 2.2: Konstanten $\breve{k}_n$ für die Abschätzung (2.24), falls einerseits die Voraussetzungen aus Satz 2.18 erfüllt sind und andererseits die Matrix der Erwartungswerte $\mathbb{E}(G)$ näherungsweise orthogonal ist.

In Lemma 2.14 bis Folgerung 2.17 haben wir unter bestimmten Modellannahmen jeweils in verschiedenen Situationen den Erwartungswert sowie die Varianz des absoluten Rundungsfehlers ermittelt, welcher bei der Addition von unkorrelierten Zufallsgrößen bzw. beim Skalarprodukt zweier voneinander unabhängiger Zufallsvektoren mit unkorrelierten Komponenten auftritt. Letzterer Fall lässt nun auch Rückschlüsse auf den Vektor der bei einer Matrix- Vektor-Multiplikation auftretenden absoluten Fehler zu. Aufgrund der starken Forderung nach stochastischer Unabhängigkeit an die Eingangsvektoren sind diese Ergebnisse im Allgemeinen jedoch nicht auf iterierte Matrix-Vektor-Multiplikationen anwendbar. Um die Voraussetzungen abschwächen zu können, muss die Struktur der beteiligten Matrizen herangezogen werden.

2.2.2 Modell für spezielle Blockdiagonalmatrizen

Im Folgenden seien alle Zufallsgrößen X nicht trivial, d.h., insbesondere trete das Ereignis $|X| > 0$ mit positiver Wahrscheinlichkeit ein. Desweiteren wandeln wir unser Modell bezüglich der relativen Fehler geringfügig ab, um u-Terme höherer Ordnung vernachlässigen zu können, und beziehen an einigen Stellen die Symmetrie der Verteilung der Eingangsdaten mit ein. Dies wird beispielsweise in [75] während der Untersuchung des relativen und absoluten Fehlers bei der Multiplikation zweier komplexer Zufallsvariablen getan. Da wir

$$(a_1 - \mathrm{i}b_1)\cdot(a_2 + \mathrm{i}b_2) = (a_1a_2 + b_1b_2) + \mathrm{i}(a_1b_2 - b_1a_2)$$

für reelle Zahlen a_1, a_2, b_1, b_2 auch als Matrix-Vektor-Multiplikation

$$\begin{pmatrix} a_1 & b_1 \\ -b_1 & a_1 \end{pmatrix}\begin{pmatrix} a_2 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1b_2 \\ a_1b_2 - b_1a_2 \end{pmatrix}$$

auffassen können, wenden wir die entsprechenden Ergebnisse auf den uns interessierenden Fall an.

Lemma 2.20 (vgl. [75], Proposition 2.3)**.** *Sei* $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ *mit identisch und symmetrisch zum Nullpunkt verteilten Zufallsvariablen* X_1*,* X_2 *aus* $\mathbb{G}$ *gegeben. Weiterhin sei die Matrix*

$$A \;:=\; \begin{pmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{pmatrix} \tag{2.27}$$

mit Zufallsvariablen G_1 *und* G_2 *aus* $\mathbb{G}$ *definiert. Angenommen, es gelte*

$$\mathrm{fl}(A\mathbf{X}) \;=\; \begin{pmatrix} (G_1X_1 + G_2X_2)(1+\varepsilon_1^+) + (G_1X_1)\varepsilon_1^\times + (G_2X_2)\varepsilon_2^\times \\ (G_1X_2 - G_2X_1)(1+\varepsilon_2^+) + (G_1X_2)\varepsilon_3^\times - (G_2X_1)\varepsilon_4^\times \end{pmatrix}$$

mit paarweise unabhängigen Zufallsgrößen $X_1, X_2, \varepsilon_1^\times, \varepsilon_2^\times, \varepsilon_3^\times, \varepsilon_4^\times, \varepsilon_1^+, \varepsilon_2^+$*, welche*

$$\left.\begin{array}{rclrcll} \mathbb{E}(\varepsilon_j^\times) & = & \mu_\times u\,, & \mathbb{V}(\varepsilon_j^\times) & = & \sigma_\times^2 u^2 & (j=1,2,3,4), \\ \mathbb{E}(\varepsilon_j^+) & = & \mu_+ u\,, & \mathbb{V}(\varepsilon_j^+) & = & \sigma_+^2 u^2 & (j=1,2) \end{array}\right\} \tag{2.28}$$

erfüllen. Weiter seien G_1, G_2 *von den übrigen Zufallsgrößen unabhängig und der Vektor* $\mathfrak{E} = (\mathfrak{E}_1, \mathfrak{E}_2)^{\mathrm{T}}$ *durch die Gleichung*

$$\mathrm{fl}(A\mathbf{X}) \;=\; A\mathbf{X} \;+\; \|A\mathbf{X}\|_2\, \mathfrak{E} \tag{2.29}$$

definiert. Dann gelten die folgenden Aussagen:

(i) *Die Verteilung von* $\mathfrak{E}$ *ist symmetrisch. Insbesondere gilt* $\mathbb{E}(\mathfrak{E}_1) = \mathbb{E}(\mathfrak{E}_2) = 0$*.*

(ii) *Die Komponenten* $\mathfrak{E}_1, \mathfrak{E}_2$ *von* $\mathfrak{E}$ *sind untereinander und mit* $\|A\mathbf{X}\|_2$ *unkorreliert.*

(iii) *Die Komponenten* $\mathfrak{E}_1, \mathfrak{E}_2$ *von* $\mathfrak{E}$ *besitzen dieselbe Varianz und es gilt*

$$\mathbb{E}(\|\mathfrak{E}\|_2^2) \;=\; \left(\sigma_\times^2 + \sigma_+^2 + (\mu_\times + \mu_+)^2\right) u^2\,.$$

Beweis: (i) Zunächst einmal halten wir fest, dass sich für die Norm

$$\|A\mathbf{X}\|_2 \;=\; \sqrt{(G_1X_1 + G_2X_2)^2 + (G_1X_2 - G_2X_1)^2} \;=\; \sqrt{G_1^2 + G_2^2}\sqrt{X_1^2 + X_2^2} \tag{2.30}$$

ergibt. Damit folgen für die Komponenten $\mathfrak{E}_1, \mathfrak{E}_2$ von $\mathfrak{E}$ die Darstellungen

$$\begin{aligned} \mathfrak{E}_1 &= \frac{G_1X_1}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_1^+ + \varepsilon_1^\times\right) + \frac{G_2X_2}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_1^+ + \varepsilon_2^\times\right)\,, \\ \mathfrak{E}_2 &= \frac{G_1X_2}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_2^+ + \varepsilon_3^\times\right) - \frac{G_2X_1}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_2^+ + \varepsilon_4^\times\right)\,. \end{aligned}$$

Ersetzen wir (X_1, X_2) durch $(-X_1, -X_2)$, so ändern sich die Verteilungen von $\mathfrak{E}_1$ und $\mathfrak{E}_2$ aufgrund der Symmetrie der Verteilungen von X_1 und X_2 nicht, womit die Behauptung folgt.
(ii) Aufgrund der Unabhängigkeit, wegen $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$ und zusammen mit Aussage (i) erhalten wir

$$\begin{aligned} \mathrm{Cov}\left(\mathfrak{E}_1, \|A\mathbf{X}\|_2\right) &= \mathbb{E}\left((G_1X_1 + G_2X_2)\varepsilon_1^+ + G_1X_1\varepsilon_1^\times + G_2X_2\varepsilon_2^\times\right) - \mathbb{E}(\mathfrak{E}_1)\mathbb{E}\left(\|A\mathbf{X}\|_2\right) \\ &= \left(\mathbb{E}(G_1)\mathbb{E}(X_1) + \mathbb{E}(G_2)\mathbb{E}(X_2)\right)\mathbb{E}(\varepsilon_1^+) + \mathbb{E}(G_1)\mathbb{E}(X_1)\mathbb{E}(\varepsilon_1^\times) + \mathbb{E}(G_2)\mathbb{E}(X_2)\mathbb{E}(\varepsilon_2^\times) \\ &= 0 \end{aligned}$$

und ebenso

$$\mathrm{Cov}\left(\mathfrak{E}_2, \|A\mathbf{X}\|_2\right) \;=\; \mathbb{E}\left((G_1X_2 - G_2X_1)\varepsilon_2^+ + G_1X_2\varepsilon_3^\times - G_2X_1\varepsilon_4^\times\right) \;=\; 0\,.$$

Wiederum wegen Aussage (i) und der Unabhängigkeit ergibt sich weiterhin

$$\begin{aligned}
&\mathrm{Cov}\,(\mathfrak{E}_1,\mathfrak{E}_2)\\
&= \mathbb{E}\left(\frac{\left((G_1X_1+G_2X_2)\varepsilon_1^+ + G_1X_1\varepsilon_1^\times + G_2X_2\varepsilon_2^\times\right)\left((G_1X_2-G_2X_1)\varepsilon_2^+ + G_1X_2\varepsilon_3^\times - G_2X_1\varepsilon_4^\times\right)}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)\\
&= T_1\mathbb{E}(\varepsilon_1^+)\mathbb{E}(\varepsilon_2^+) + T_2\mathbb{E}(\varepsilon_1^\times)\mathbb{E}(\varepsilon_3^\times) - T_3\mathbb{E}(\varepsilon_1^\times)\mathbb{E}(\varepsilon_4^\times) + T_4\mathbb{E}(\varepsilon_2^\times)\mathbb{E}(\varepsilon_3^\times) - T_5\mathbb{E}(\varepsilon_2^\times)\mathbb{E}(\varepsilon_4^\times)\\
&\quad +(T_2+T_4)\mathbb{E}(\varepsilon_1^+)\mathbb{E}(\varepsilon_3^\times) - (T_3+T_5)\mathbb{E}(\varepsilon_1^+)\mathbb{E}(\varepsilon_4^\times) + (T_2-T_3)\mathbb{E}(\varepsilon_1^\times)\mathbb{E}(\varepsilon_2^+) + (T_4-T_5)\mathbb{E}(\varepsilon_2^\times)\mathbb{E}(\varepsilon_2^+)
\end{aligned}$$

mit den Termen

$$\begin{aligned}
T_1 &= \mathbb{E}\left(\frac{X_1X_2}{X_1^2+X_2^2}\right)\left(\mathbb{E}\left(\frac{G_1^2}{G_1^2+G_2^2}\right) - \mathbb{E}\left(\frac{G_2^2}{G_1^2+G_2^2}\right)\right)\\
&\quad + \mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\right)\left(\mathbb{E}\left(\frac{X_2^2}{X_1^2+X_2^2}\right) - \mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right)\right)
\end{aligned}$$

sowie

$$\begin{aligned}
T_2 &= \mathbb{E}\left(\frac{G_1^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1X_2}{X_1^2+X_2^2}\right), & T_3 &= \mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right),\\
T_4 &= \mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_2^2}{X_1^2+X_2^2}\right), & T_5 &= \mathbb{E}\left(\frac{G_2^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1X_2}{X_1^2+X_2^2}\right).
\end{aligned}$$

Da die Zufallsgrößen X_1 und X_2 einerseits identisch und andererseits symmetrisch zum Nullpunkt verteilt sind, haben wir

$$\mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right) = \mathbb{E}\left(\frac{X_2^2}{X_1^2+X_2^2}\right) \tag{2.31}$$

und

$$\mathbb{E}\left(\frac{X_1X_2}{X_1^2+X_2^2}\right) = 0\,. \tag{2.32}$$

Somit verschwinden in T_1 der erste Summand wegen (2.32) und der zweite Summand wegen (2.31). Ebenso folgen $T_2 = T_5 = 0$ mit (2.32). Wegen (2.28) lassen sich nun die übrigen Summanden zusammenfassen, so dass wir zusammen mit der Tatsache, dass sich $T_3 = T_4$ aus (2.31) ergibt, schließlich wie behauptet

$$\mathrm{Cov}\,(\mathfrak{E}_1,\mathfrak{E}_2) = (T_4-T_3)\mu_\times^2u^2 + 2(T_4-T_3)\mu_\times\mu_+u^2 = 0$$

erhalten.

(iii) Wiederum aufgrund der Unabhängigkeit sowie mit (2.28) ergibt sich für $\mathbb{V}\,(\mathfrak{E}_1)$ zunächst

$$\begin{aligned}
&\mathbb{E}\left(\frac{\left((G_1X_1+G_2X_2)\varepsilon_1^+ + G_1X_1\varepsilon_1^\times + G_2X_2\varepsilon_2^\times\right)^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right) = \mathbb{E}\left(\frac{\left(G_1X_1(\varepsilon_1^+ + \varepsilon_1^\times) + G_2X_2(\varepsilon_1^+ + \varepsilon_2^\times)\right)^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)\\
&= \left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right)\left(\mathbb{E}\left(\frac{G_1^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right) + \mathbb{E}\left(\frac{G_2^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_2^2}{X_1^2+X_2^2}\right)\right)u^2\\
&\quad + 2\left(\sigma_+^2 + \mu_+^2 + 2\mu_\times\mu_+ + \mu_\times^2\right)\mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1X_2}{X_1^2+X_2^2}\right)u^2\,.
\end{aligned}$$

Mit Hilfe von (2.32) verschwindet der letzte Summand. Wiederum wegen (2.31) sowie aufgrund der Linearität des Erwartungswertes folgt dann weiter

$$\mathbb{E}\left(\frac{G_1^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right) + \mathbb{E}\left(\frac{G_2^2}{G_1^2+G_2^2}\right)\mathbb{E}\left(\frac{X_2^2}{X_1^2+X_2^2}\right) = \mathbb{E}\left(\frac{X_1^2}{X_1^2+X_2^2}\right) = \frac{1}{2}$$

und schließlich

$$\mathbb{V}\left(\mathfrak{E}_1\right) \;=\; \frac{1}{2}\left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right) u^2 \;.$$

In analoger Weise ergibt sich

$$\mathbb{V}\left(\mathfrak{E}_2\right) \;=\; \frac{1}{2}\left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right) u^2$$

und damit auch

$$\mathbb{E}(\|\mathfrak{E}\|_2^2) \;=\; \mathbb{E}\left(\mathfrak{E}_1^2\right) + \mathbb{E}\left(\mathfrak{E}_2^2\right) \;=\; \mathbb{V}\left(\mathfrak{E}_1\right) + \mathbb{V}\left(\mathfrak{E}_2\right) \;=\; \left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right) u^2 \;. \qquad \blacksquare$$

Nun sind wir in der Lage, auch Aussagen über das Verhalten des absoluten Fehlers zu treffen.

Folgerung 2.21 (vgl. [75], Conclusion 2.8)**.** (i) *Mit den Voraussetzungen und Bezeichnungen von Lemma* 2.20 *ergibt sich*

$$\mathbb{E}\left(\boldsymbol{\Delta}\right) \;=\; \mathbf{0} \;, \qquad \mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right) \mathbb{E}\left(\|\mathfrak{E}\|_2^2\right)$$

für den Vektor der absoluten Rundungsfehler

$$\boldsymbol{\Delta} \;:=\; \mathrm{fl}(A\mathbf{X}) - A\mathbf{X} \;.$$

(ii) *Sei* $n \in \mathbb{N},\ n \geq 4$ *und gerade. Weiter sei* $\mathbf{X} := (X_l)_{l=0}^{n-1}$ *mit identisch und symmetrisch zum Nullpunkt verteilten Zufallsvariablen* X_l *aus* $\mathbb{G}$ *gegeben und*

$$B \;:=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k)}$$

mit $A^{(k)},\ k = 0, \ldots, \frac{n}{2} - 1$, *der Gestalt* (2.27)*. Für jeden Block* $A^{(k)}$ *und zugehörigen Teilvektor* $\mathbf{X}^{(k)} := (X_{2k}, X_{2k+1})^{\mathrm{T}}$ *seien die Voraussetzungen aus Lemma* 2.20 *erfüllt. Dann gilt*

$$\mathbb{E}\left(\boldsymbol{\Delta}\right) \;=\; \mathbf{0} \;, \qquad \mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\|B\mathbf{X}\|_2^2\right) \left(\sigma_\times^2 + \sigma_+^2 + (\mu_\times + \mu_+)^2\right) u^2 \;.$$

für den Vektor der absoluten Rundungsfehler

$$\boldsymbol{\Delta} \;:=\; \mathrm{fl}(B\mathbf{X}) - B\mathbf{X} \;.$$

Beweis: (i) Nach (2.29) besitzt Δ die Komponenten $\|A\mathbf{X}\|_2\ \mathfrak{E}_k$ $(k = 1, 2)$. Nach Lemma 2.20 (i) und (ii) wissen wir, dass einerseits $\mathfrak{E}_k$ $(k = 1, 2)$ Erwartungswert Null besitzt und andererseits die Kovarianzen $\mathrm{Cov}\left(\mathfrak{E}_k, \|A\mathbf{X}\|_2\right)$ $(k = 1, 2)$ verschwinden. Demnach folgt auch

$$\mathbb{E}\left(\|A\mathbf{X}\|_2\ \mathfrak{E}_k\right) \;=\; \mathrm{Cov}\left(\mathfrak{E}_k, \|A\mathbf{X}\|_2\right) \;+\; \mathbb{E}\left(\|A\mathbf{X}\|_2\right) \mathbb{E}\left(\mathfrak{E}_k\right) \;=\; 0 \qquad (k = 1, 2).$$

Mit der Symmetrie der Verteilungen von X_1 und X_2 haben wir außerdem $\mathbb{E}(X_1 X_2) = 0$ analog zu (2.32). Demzufolge und aufgrund der Unabhängigkeit sowie mit (2.28) ergeben sich dann

$$\begin{aligned}
\mathbb{E}\left(\|A\mathbf{X}\|_2^2\ \mathfrak{E}_1^2\right) \;&=\; \mathbb{E}\left(\left(G_1 X_1(\varepsilon_1^+ + \varepsilon_1^\times) + G_2 X_2(\varepsilon_1^+ + \varepsilon_2^\times)\right)^2\right) \\
&=\; \left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right) \left(\mathbb{E}(G_1^2 X_1^2) + \mathbb{E}(G_2^2 X_2^2)\right) u^2
\end{aligned}$$

und

$$\begin{aligned}
\mathbb{E}\left(\|A\mathbf{X}\|_2^2\ \mathfrak{E}_2^2\right) \;&=\; \mathbb{E}\left(\left(G_1 X_2(\varepsilon_2^+ + \varepsilon_3^\times) - G_2 X_1(\varepsilon_2^+ + \varepsilon_4^\times)\right)^2\right) \\
&=\; \left(\sigma_+^2 + \sigma_\times^2 + (\mu_+ + \mu_\times)^2\right) \left(\mathbb{E}(G_1^2 X_2^2) + \mathbb{E}(G_2^2 X_1^2)\right) u^2 \;.
\end{aligned}$$

Mit den Rechenregeln für den Erwartungswert und (2.30) folgt dann

$$\mathbb{E}(G_1^2 X_1^2) + \mathbb{E}(G_2^2 X_2^2) + \mathbb{E}(G_1^2 X_2^2) + \mathbb{E}(G_2^2 X_1^2) \;=\; \mathbb{E}\left(\left(G_1^2 + G_2^2\right)\left(X_1^2 + X_2^2\right)\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)$$

und nach Lemma 2.20 (iii) schließlich

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\left(\|A\mathbf{X}\|_2\;\mathfrak{E}_1\right)^2\right) + \mathbb{E}\left(\left(\|A\mathbf{X}\|_2\;\mathfrak{E}_2\right)^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)\mathbb{E}\left(\|\mathfrak{E}\|_2^2\right)\;.$$

(ii) Aufgrund der Blockdiagonalgestalt von B sowie mit der Definition und der Linearität des Erwartungswertes erhalten wir unter Verwendung von Aussage (i) einerseits

$$\mathbb{E}\left(\mathrm{fl}(B\mathbf{X}) - B\mathbf{X}\right) \;=\; \begin{pmatrix} \mathbb{E}\left(\mathrm{fl}\left(A^{(0)}\mathbf{X}^{(0)}\right) - A^{(0)}\mathbf{X}^{(0)}\right) \\ \vdots \\ \mathbb{E}\left(\mathrm{fl}\left(A^{(\frac{n}{2}-1)}\mathbf{X}^{(\frac{n}{2}-1)}\right) - A^{(\frac{n}{2}-1)}\mathbf{X}^{(\frac{n}{2}-1)}\right) \end{pmatrix} \;=\; \mathbf{0}$$

und zusammen mit Lemma 2.20 (iii) andererseits

$$\begin{aligned}
\mathbb{E}\left(\|\,\mathrm{fl}(B\mathbf{X}) - B\mathbf{X}\|_2^2\right) \;&=\; \sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(\left\|\mathrm{fl}\left(A^{(k)}\mathbf{X}^{(k)}\right) - A^{(k)}\mathbf{X}^{(k)}\right\|_2^2\right) \\
&=\; \sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(\left\|A^{(k)}\mathbf{X}^{(k)}\right\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2 + (\mu_\times + \mu_+)^2\right)u^2 \\
&=\; \mathbb{E}\left(\|B\mathbf{X}\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2 + (\mu_\times + \mu_+)^2\right)u^2\;. \qquad \blacksquare
\end{aligned}$$

Da wir es bei unseren Algorithmen mit einer Vielzahl von Hintereinanderausführungen von blockweisen Drehungen zu tun haben, sind die Voraussetzungen in Lemma 2.20 an die Verteilungen von X_1 und X_2 nicht in jedem Fall realistisch. Die obigen Aussagen bleiben im Wesentlichen jedoch richtig, wenn wir stattdessen die Erwartungswerte aus (2.28) sämtlich als Null annehmen.

Satz 2.22. *Sei $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ mit Zufallsvariablen X_1, X_2 aus $\mathbb{G}$ gegeben. Weiterhin sei die Matrix A wie in* (2.27) *definiert. Angenommen, dass*

$$\mathrm{fl}(A\mathbf{X}) \;=\; \begin{pmatrix} (G_1X_1 + G_2X_2)(1+\varepsilon_1^+) + (G_1X_1)\varepsilon_1^\times + (G_2X_2)\varepsilon_2^\times \\ (G_1X_2 - G_2X_1)(1+\varepsilon_2^+) + (G_1X_2)\varepsilon_3^\times - (G_2X_1)\varepsilon_4^\times \end{pmatrix} \tag{2.33}$$

mit paarweise unabhängigen Zufallsgrößen $\varepsilon_1^\times, \varepsilon_2^\times, \varepsilon_3^\times, \varepsilon_4^\times, \varepsilon_1^+, \varepsilon_2^+$ gelte, wobei

$$\left.\begin{array}{rclrcll} \mathbb{E}(\varepsilon_j^\times) &=& 0\;, & \mathbb{V}(\varepsilon_j^\times) &=& \sigma_\times^2 u^2 & (j=1,2,3,4), \\ \mathbb{E}(\varepsilon_j^+) &=& 0\;, & \mathbb{V}(\varepsilon_j^+) &=& \sigma_+^2 u^2 & (j=1,2) \end{array}\right\} \tag{2.34}$$

erfüllt werde. Darüber hinaus seien X_1, X_2, G_1, G_2 von den relativen Fehlern unabhängig und der Zufallsvektor $\mathfrak{E} = (\mathfrak{E}_1, \mathfrak{E}_2)^{\mathrm{T}}$ wie in (2.29) *definiert. Dann gelten die folgenden Aussagen:*

(i) *Die Komponenten $\mathfrak{E}_1, \mathfrak{E}_2$ von $\mathfrak{E}$ sind untereinander und mit $\|A\mathbf{X}\|_2$ unkorreliert. Weiterhin sind*

$$\mathbb{E}\left(\mathfrak{E}_1\right) = \mathbb{E}\left(\mathfrak{E}_2\right) = 0\;.$$

(ii) *Die Komponenten $\mathfrak{E}_1, \mathfrak{E}_2$ von $\mathfrak{E}$ besitzen im Allgemeinen nicht dieselbe Varianz. Dennoch gilt*

$$\mathbb{E}(\|\mathfrak{E}\|_2^2) \;=\; \left(\sigma_\times^2 + \sigma_+^2\right)u^2\;.$$

(iii) *Der Vektor der absoluten Rundungsfehler*

$$\boldsymbol{\Delta} \;:=\; \mathrm{fl}(A\mathbf{X}) - A\mathbf{X}$$

besitzt Erwartungswert Null. Weiterhin gilt

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2\right)u^2\;.$$

(iv) *Die Aussagen in* (iii) *bleiben gültig, wenn A anstelle von* (2.27) *Blockdiagonalgestalt mit Blöcken* (2.27) *besitzt und für jeden einzelnen Block und entsprechende Teilvektoren die Voraussetzungen wie zuvor erfüllt sind.*

Beweis: (i) Analog zum Beweis von Lemma 2.20 (i) ist für die aus (2.30) resultierenden Darstellungen

$$\begin{aligned}\mathfrak{E}_1 &= \frac{G_1X_1}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_1^+ + \varepsilon_1^\times\right) + \frac{G_2X_2}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_1^+ + \varepsilon_2^\times\right)\ ,\\ \mathfrak{E}_2 &= \frac{G_1X_2}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_2^+ + \varepsilon_3^\times\right) - \frac{G_2X_1}{\sqrt{G_1^2+G_2^2}\sqrt{X_1^2+X_2^2}}\left(\varepsilon_2^+ + \varepsilon_4^\times\right)\end{aligned}$$

der Komponenten des Fehlervektors $\mathfrak{E}$ ersichtlich, dass mit (2.34) aufgrund der Unabhängigkeit

$$\mathbb{E}\left(\mathfrak{E}_1\right) = \mathbb{E}\left(\mathfrak{E}_2\right) = 0$$

folgt. Mit dem gleichen Argument ergeben sich dann auch

$$\begin{aligned}\mathrm{Cov}\left(\mathfrak{E}_1, \|A\mathbf{X}\|_2\right) &= \mathbb{E}\left((G_1X_1+G_2X_2)\varepsilon_1^+ + G_1X_1\varepsilon_1^\times + G_2X_2\varepsilon_2^\times\right) - \mathbb{E}(\mathfrak{E}_1)\mathbb{E}\left(\|A\mathbf{X}\|_2\right)\\ &= \mathbb{E}\left(G_1X_1+G_2X_2\right)\mu_+ \ +\ \mathbb{E}\left(G_1X_1\right)\mu_\times \ +\ \mathbb{E}\left(G_2X_2\right)\mu_\times \ -\ 0 \ =\ 0\end{aligned}$$

und

$$\mathrm{Cov}\left(\mathfrak{E}_2, \|A\mathbf{X}\|_2\right) = \mathbb{E}\left((G_1X_2-G_2X_1)\varepsilon_2^+ + G_1X_2\varepsilon_3^\times - G_2X_1\varepsilon_4^\times\right) - \mathbb{E}(\mathfrak{E}_2)\mathbb{E}\left(\|A\mathbf{X}\|_2\right) \ =\ 0$$

sowie mit den Bezeichnungen aus dem Beweis von Lemma 2.20 (ii) schließlich auch

$$\begin{aligned}&\mathrm{Cov}\left(\mathfrak{E}_1, \mathfrak{E}_2\right)\\ &= \mathbb{E}\left(\frac{\left((G_1X_1+G_2X_2)\varepsilon_1^+ + G_1X_1\varepsilon_1^\times + G_2X_2\varepsilon_2^\times\right)\left((G_1X_2-G_2X_1)\varepsilon_2^+ + G_1X_2\varepsilon_3^\times - G_2X_1\varepsilon_4^\times\right)}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)\\ &= T_1\mu_+^2 \ +(T_2-T_3+T_4-T_5)\,\mu_\times^2 \ +((T_2+T_4)-(T_3+T_5)+(T_2-T_3)+(T_4-T_5))\,\mu_\times\mu_+ \ =\ 0\ .\end{aligned}$$

(ii) Wiederum aufgrund der Unabhängigkeit, der Linearität des Erwartungswertes sowie mit (2.34) ergeben sich zunächst die Varianzen

$$\begin{aligned}\mathbb{V}\left(\mathfrak{E}_1\right) &= \mathbb{E}\left(\frac{\left(G_1X_1(\varepsilon_1^+ + \varepsilon_1^\times) + G_2X_2(\varepsilon_1^+ + \varepsilon_2^\times)\right)^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right) \ -\ \left(\mathbb{E}\left(\mathfrak{E}_1\right)\right)^2\\ &= \mathbb{E}\left(\frac{G_1^2X_1^2+G_2^2X_2^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)\left(\sigma_+^2+\sigma_\times^2\right)u^2 \ +\ 2\,\mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\cdot\frac{X_1X_2}{X_1^2+X_2^2}\right)\sigma_+^2u^2\end{aligned}$$

und

$$\begin{aligned}\mathbb{V}\left(\mathfrak{E}_2\right) &= \mathbb{E}\left(\frac{\left(G_1X_2(\varepsilon_2^+ + \varepsilon_3^\times) - G_2X_1(\varepsilon_2^+ + \varepsilon_4^\times)\right)^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right) \ -\ \left(\mathbb{E}\left(\mathfrak{E}_2\right)\right)^2\\ &= \mathbb{E}\left(\frac{G_1^2X_2^2+G_2^2X_1^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)\left(\sigma_+^2+\sigma_\times^2\right)u^2 \ -\ 2\,\mathbb{E}\left(\frac{G_1G_2}{G_1^2+G_2^2}\cdot\frac{X_1X_2}{X_1^2+X_2^2}\right)\sigma_+^2u^2\ ,\end{aligned}$$

welche im Allgemeinen verschieden sind. Wegen

$$\mathbb{E}\left(\frac{G_1^2X_1^2+G_2^2X_2^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right)+\mathbb{E}\left(\frac{G_1^2X_2^2+G_2^2X_1^2}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right) \ =\ \mathbb{E}\left(\frac{(G_1^2+G_2^2)(X_1^2+X_2^2)}{(G_1^2+G_2^2)(X_1^2+X_2^2)}\right) \ =\ 1$$

erhalten wir nun wie behauptet

$$\mathbb{E}(\|\mathfrak{E}\|_2^2) \;=\; \mathbb{V}\left(\mathfrak{E}_1\right) + \mathbb{V}\left(\mathfrak{E}_2\right) \;=\; \left(\sigma_\times^2 + \sigma_+^2\right) u^2 \;.$$

(iii) Nach (i) wissen wir, dass einerseits die Kovarianzen $\mathrm{Cov}\left(\mathfrak{E}_k, \|A\mathbf{X}\|_2\right)$ $(k = 1, 2)$ aufgrund der Unkorreliertheit verschwinden und andererseits die Erwartungswerte $\mathfrak{E}_k$ $(k = 1, 2)$ Null sind. Somit ergibt sich auch

$$\mathbb{E}\left(\|A\mathbf{X}\|_2\ \mathfrak{E}_k\right) \;=\; \mathrm{Cov}\left(\mathfrak{E}_k, \|A\mathbf{X}\|_2\right) \;+\; \mathbb{E}\left(\|A\mathbf{X}\|_2\right)\mathbb{E}\left(\mathfrak{E}_k\right) \;=\; 0 \qquad (k = 1, 2) \;.$$

Da $\|A\mathbf{X}\|_2\ \mathfrak{E}_k$ $(k = 1, 2)$ nach (2.29) genau die Komponenten von Δ sind, folgt nun ebenso

$$\mathbb{E}(\boldsymbol{\Delta}) \;=\; \mathbf{0} \;.$$

Aufgrund der Unabhängigkeit sowie mit (2.34) ergeben sich weiter

$$\begin{aligned}\mathbb{E}\left(\|A\mathbf{X}\|_2^2\ \mathfrak{E}_1^2\right) &= \mathbb{E}\left(\left(G_1X_1(\varepsilon_1^+ + \varepsilon_1^\times) + G_2X_2(\varepsilon_1^+ + \varepsilon_2^\times)\right)^2\right) \\ &= \mathbb{E}\left(G_1^2X_1^2 + G_2^2X_2^2\right)\left(\sigma_+^2 + \sigma_\times^2\right)u^2 \;+\; 2\,\mathbb{E}(G_1G_2X_1X_2)\;\sigma_+^2\;u^2\end{aligned}$$

und

$$\begin{aligned}\mathbb{E}\left(\|A\mathbf{X}\|_2^2\ \mathfrak{E}_2^2\right) &= \mathbb{E}\left(\left(G_1X_2(\varepsilon_2^+ + \varepsilon_3^\times) - G_2X_1(\varepsilon_2^+ + \varepsilon_4^\times)\right)^2\right) \\ &= \mathbb{E}\left(G_1^2X_2^2 + G_2^2X_1^2\right)\left(\sigma_+^2 + \sigma_\times^2\right)u^2 \;-\; 2\,\mathbb{E}(G_1G_2X_1X_2)\;\sigma_+^2\;u^2 \;.\end{aligned}$$

Wiederum mit (2.30) und somit wegen

$$\mathbb{E}\left(G_1^2X_1^2 + G_2^2X_2^2\right) + \mathbb{E}\left(G_1^2X_2^2 + G_2^2X_1^2\right) \;=\; \mathbb{E}\left(\left(G_1^2 + G_2^2\right)\left(X_1^2 + X_2^2\right)\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)$$

und (ii) erhalten wir schließlich

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\left(\|A\mathbf{X}\|_2\ \mathfrak{E}_1\right)^2\right) + \mathbb{E}\left(\left(\|A\mathbf{X}\|_2\ \mathfrak{E}_2\right)^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)\mathbb{E}\left(\|\mathfrak{E}\|_2^2\right) \;.$$

(iv) Der Beweis verwendet wiederum die Linearität des Erwartungswertes und verläuft analog zum Beweis von Folgerung 2.21 (ii). ■

Mit Hilfe der Jensen-Ungleichung aus Satz A.25 angewandt auf die konvexe Funktion $\varphi(\xi) = \xi^2$ liefert Satz 2.22 nun die

Folgerung 2.23. *Sei $n \in \mathbb{N},\ n \geq 4$ und gerade. Weiter sei $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{G}$ gegeben und*

$$B \;:=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k)} \tag{2.35}$$

mit $A^{(k)} \in \mathbb{G}^{2\times 2},\ k = 0, \ldots, \frac{n}{2} - 1$, der Gestalt (2.27). Für jeden Block $A^{(k)}$ und zugehörigen Teilvektor $\mathbf{X}^{(k)} := (X_{2k}, X_{2k+1})^{\mathrm{T}}$ seien die Voraussetzungen aus Satz 2.22 erfüllt. Dann gilt

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\leq\; \sqrt{\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right)} \;=\; \sqrt{\mathbb{E}\left(\|B\mathbf{X}\|_2^2\right)}\sqrt{\sigma_\times^2 + \sigma_+^2}u \;.$$

für den Vektor der absoluten Rundungsfehler

$$\boldsymbol{\Delta} \;:=\; \mathrm{fl}(B\mathbf{X}) - B\mathbf{X} \;.$$

Beweis: Für eine konvexe Funktion φ und eine reelle Zufallsvariable Y mit endlichen Momenten lautet die Jensen-Ungleichung

$$\varphi\left(\mathbb{E}(Y)\right) \;\leq\; \mathbb{E}\left(\varphi(Y)\right) \;. \tag{2.36}$$

Mit $\varphi(\xi) = \xi^2$ und $Y = \|\boldsymbol{\Delta}\|_2$ erhalten wir zusammen mit Satz 2.22 (iv) demnach

$$\left(\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right)\right)^2 \;\leq\; \mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;=\; \mathbb{E}\left(\|B\mathbf{X}\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2\right)u^2 \;.$$

Aufgrund der Monotonie des Erwartungswertes folgt aus $Y \geq 0$ dann auch $\mathbb{E}(Y) \geq 0$ und somit die Behauptung nach Wurzelziehen. ■

Offenbar benötigen wir bei allen vorangegangenen Aussagen den Erwartungswert

$$\mathbb{E}\left(\left\|\begin{pmatrix} G_1 & G_2 \\ -G_2 & G_1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right\|_2^2\right) \;=\; \mathbb{E}\left((G_1^2+G_2^2)(X_1^2+X_2^2)\right)$$

für Zufallsgrößen G_1, G_2, X_1, X_2 aus $\mathbb{G}$ oder eine geeignete Abschätzung. Da es sich in den Anwendungen um Drehungen handelt, sind die Zufallsgrößen G_1 und G_2 durch die Beziehung

$$G_1^2+G_2^2 \;=\; 1+\varepsilon_{\|\cdot\|_2^2} \qquad \left(\left|\varepsilon_{\|\cdot\|_2^2}\right| \leq 2u+u^2\right) \tag{2.37}$$

über eine weitere Zufallsvariable $\varepsilon_{\|\cdot\|_2^2}$ miteinander verbunden, wobei $\varepsilon_{\|\cdot\|_2^2}$ im Idealfall von X_1, X_2 stochastisch unabhängig ist sowie eine der beiden Annahmen

$$\mathbb{E}\left(\varepsilon_{\|\cdot\|_2^2}\right) \;=\; 0\,, \tag{2.37a}$$

$$\varepsilon_{\|\cdot\|_2^2} \;\leq\; 0 \tag{2.37b}$$

gilt. Insbesondere erhalten wir dann aufgrund der stochastischen Unabhängigkeit von konstanten Zufallsgrößen bzw. mit der Linearität und Monotonie des Erwartungswertes

$$\mathbb{E}\left((G_1^2+G_2^2)(X_1^2+X_2^2)\right) \;=\; \mathbb{E}\left(X_1^2+X_2^2\right)\left(1+\mathbb{E}\left(\varepsilon_{\|\cdot\|_2^2}\right)\right) \;=\; \mathbb{E}\left(X_1^2+X_2^2\right) \tag{2.38a}$$

oder

$$\mathbb{E}\left((G_1^2+G_2^2)(X_1^2+X_2^2)\right) \;=\; \mathbb{E}\left(\left(1+\varepsilon_{\|\cdot\|_2^2}\right)(X_1^2+X_2^2)\right) \;\leq\; \mathbb{E}\left(X_1^2+X_2^2\right)\,. \tag{2.38b}$$

Mit den Forderungen (2.37) und (2.37a) bzw. (2.37b) an jeden zufälligen Block $A^{(k)}$ aus (2.35) können wir nun auch Hintereinanderausführungen von blockweisen Drehungen betrachten.

Lemma 2.24. *Sei $n \in \mathbb{N},\ n \geq 4$ und gerade. Weiter sei $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{G}$ gegeben und es seien*

$$B^{(s)} \;:=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(s,k)} \qquad (s=0,\ldots,r-1) \tag{2.39}$$

zufällige Blockdiagonalmatrizen mit Blöcken

$$A^{(s,k)} \;:=\; \begin{pmatrix} G_1^{(s,k)} & G_2^{(s,k)} \\ -G_2^{(s,k)} & G_1^{(s,k)} \end{pmatrix} \qquad \left(k=0,\ldots,\tfrac{n}{2}-1,\quad s=0,\ldots,r-1\right), \tag{2.40}$$

deren zufällige Einträge $G_1^{(s,k)}, G_2^{(s,k)}$ aus $\mathbb{G}$ jeweils (2.37) mit einer Zufallsvariable $\varepsilon_{\|\cdot\|_2^2}^{(s,k)}$ erfüllen. Rekursiv definieren wir nun die zufälligen Vektoren

$$\begin{aligned} \mathbf{X}^{(0)} \;&:=\; \mathbf{X}\,, \\ \mathbf{X}^{(s+1)} \;&:=\; \mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right)\,, \quad s=0,\ldots,r-1, \end{aligned}$$

über die Blockdiagonalmatrizen $B^{(s)}$, wobei für jeden Block $A^{(s,k)}$ und zugehörigen Teilvektor $\mathbf{X}^{(s,k)} := (X_{2k}^{(s)}, X_{2k+1}^{(s)})^{\mathrm{T}}$ die Voraussetzungen aus Satz 2.22 erfüllt seien. Dann gilt für den Vektor der insgesamt auftretenden absoluten Rundungsfehler

$$\boldsymbol{\Delta} \;:=\; \mathbf{X}^{(r)} - \left(\prod_{s=0}^{r-1} B^{(s)}\right)\mathbf{X} \tag{2.41}$$

die Ungleichung

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;\leq\; 2^{\lceil \log_2(r)\rceil}\sum_{s=0}^{r-1}(1+u)^{2(r-s-1)}\mathbb{E}\left(\left\|\mathbf{X}^{(s)}\right\|_2^2\right)\left(\sigma_\times^2+\sigma_+^2\right)u^2\,, \tag{2.42a}$$

falls für jedes Tupel (s,k) die Zufallsvariable $\varepsilon_{\|\cdot\|_2^2}^{(s,k)}$ von den Zwischenergebnissen $X_{2k}^{(s)}, X_{2k+1}^{(s)}$ stochastisch unabhängig ist und (2.37a) *erfüllt. Bei Gültigkeit von* (2.37b) *haben wir die schärfere Abschätzung*

$$\mathbb{E}\left(\|\mathbf{\Delta}\|_2^2\right) \;\leq\; 2^{\lceil \log_2(r)\rceil} \sum_{s=0}^{r-1} \mathbb{E}\left(\left\|\mathbf{X}^{(s)}\right\|_2^2\right) \left(\sigma_\times^2 + \sigma_+^2\right) u^2 \;. \tag{2.42b}$$

Beweis: Mittels der Teleskopsumme

$$\begin{aligned}
\mathbf{\Delta} &= \mathrm{fl}\left(B^{(r-1)}\mathbf{X}^{(r-1)}\right) - B^{(r-1)}\mathbf{X}^{(r-1)} + B^{(r-1)}\left(\mathrm{fl}\left(B^{(r-2)}\mathbf{X}^{(r-2)}\right) - \left(\prod_{s=0}^{r-2} B^{(s)}\right)\mathbf{X}\right) \;=\; \ldots \\
&= \sum_{s=0}^{r-1}\left(\prod_{m=s+1}^{r-1} B^{(m)}\right)\left(\mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right) - B^{(s)}\mathbf{X}^{(s)}\right) \\
&= \sum_{s=0}^{r-1}\left(\bigoplus_{k=0}^{\frac{n}{2}-1}\left(\prod_{m=s+1}^{r-1} A^{(m,k)}\right)\right)\begin{pmatrix} \mathrm{fl}\left(A^{(s,1)}\mathbf{X}^{(s,1)}\right) - A^{(s,1)}\mathbf{X}^{(s,1)} \\ \vdots \\ \mathrm{fl}\left(A^{\left(s,\frac{n}{2}-1\right)}\mathbf{X}^{\left(s,\frac{n}{2}-1\right)}\right) - A^{\left(s,\frac{n}{2}-1\right)}\mathbf{X}^{\left(s,\frac{n}{2}-1\right)} \end{pmatrix}
\end{aligned}$$

erhalten wir aufgrund der speziellen Struktur und der Linearität des Erwartungswertes zunächst

$$\mathbb{E}\left(\|\mathbf{\Delta}\|_2^2\right) \;=\; \sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(\left\|\sum_{s=0}^{r-1}\left(\prod_{m=s+1}^{r-1} A^{(m,k)}\right)\left(\mathrm{fl}\left(A^{(s,k)}\mathbf{X}^{(s,k)}\right) - A^{(s,k)}\mathbf{X}^{(s,k)}\right)\right\|_2^2\right) \;.$$

Verwenden wir nun, dass nach Voraussetzung

$$\left(G_1^{(s,k)}\right)^2 + \left(G_2^{(s,k)}\right)^2 \;=\; 1 + \varepsilon_{\|\cdot\|_2^2}^{(s,k)} \qquad \left(|\varepsilon_{\|\cdot\|_2^2}^{(s,k)}| \leq 2u + u^2,\; k = 0,\ldots,\tfrac{n}{2}-1,\; s = 1,\ldots,r\right)$$

für jeweils eine im ersten Fall von $X_{2k}^{(s)}, X_{2k+1}^{(s)}$ stochastisch unabhängige Zufallsgröße $\varepsilon_{\|\cdot\|_2^2}^{(s,k)}$ mit Erwartungswert $\mathbb{E}(\varepsilon_{\|\cdot\|_2^2}^{(s,k)}) = 0$ und im zweiten Fall mit $\varepsilon_{\|\cdot\|_2^2}^{(s,k)} \leq 0$ gilt, so ergibt sich in beiden Fällen

$$\mathbb{E}\left(\left\|\begin{pmatrix} G_1^{(s,k)} & G_2^{(s,k)} \\ -G_2^{(s,k)} & G_1^{(s,k)} \end{pmatrix}\mathbf{X}^{(s,k)}\right\|_2^2\right) \;\leq\; \mathbb{E}\left(\left\|\mathbf{X}^{(s,k)}\right\|_2^2\right) \tag{2.43}$$

analog zu (2.38a) bzw. (2.38b). Für die Spektralnorm der Matrizen $A^{(m,k)}$ folgt nach (2.37) außerdem

$$\left\|A^{(m,k)}\right\|_2^2 \;=\; 1 + \varepsilon_{\|\cdot\|_2^2}^{(s,k)} \;\leq\; (1+u)^2$$

und daher

$$\left\|\prod_{m=s+1}^{r-1} A^{(m,k)}\right\|_2^2 \;\leq\; (1+u)^{2(r-s-1)} \;. \tag{2.44a}$$

Im zweiten Fall ergibt sich sogar

$$\left\|\prod_{m=s+1}^{r-1} A^{(m,k)}\right\|_2^2 \;\leq\; 1 \;. \tag{2.44b}$$

Zusammen mit der für jede von einem Skalarprodukt induzierte Norm gültigen Parallelogrammgleichung

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 \;=\; 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right) \tag{2.45}$$

und der Nichtnegativität einer Norm sowie mit der Linearität und Monotonie des Erwartungswertes und Satz 2.22 erhalten wir schließlich

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{\Delta}\|_2^2\right) &\leq 2^{\lceil \log_2(r)\rceil} \sum_{s=0}^{r-1}\sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(\left\|\left(\prod_{m=s+1}^{r-1} A^{(m,k)}\right)\mathrm{fl}\left(A^{(s,k)}\mathbf{X}^{(s,k)}\right) - A^{(s,k)}\mathbf{X}^{(s,k)}\right\|_2^2\right) \\
&\leq 2^{\lceil \log_2(r)\rceil} \sum_{s=0}^{r-1}(1+u)^{2(r-s-1)} \sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(\left\|A^{(s,k)}\mathbf{X}^{(s,k)}\right\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2\right) u^2 \\
&\leq 2^{\lceil \log_2(r)\rceil} \sum_{s=0}^{r-1}(1+u)^{2(r-s-1)} \mathbb{E}\left(\left\|\mathbf{X}^{(s)}\right\|_2^2\right)\left(\sigma_\times^2 + \sigma_+^2\right) u^2
\end{aligned}$$

und somit Behauptung (2.42a). Im zweiten Fall entfallen wegen (2.44b) die Faktoren $(1+u)^{2(r-s-1)}$, so dass sich analog Behauptung (2.42b) ergibt. ■

Im vorangegangenen Lemma haben wir eine Abschätzung für das zweite Moment der Norm des absoluten Rundungsfehlers hergeleitet, innerhalb derer auf der rechten Seite noch Ausdrücke der Gestalt $\mathbb{E}\left(\|\operatorname{fl}(A\mathbf{X})\|_2^2\right)$ mit einer Matrix A wie in Satz 2.22 (iv) vorkommen. Diese wollen wir nun noch durch ein Vielfaches von $\mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)$ ausdrücken.

Lemma 2.25. (i) *Mit den Bezeichnungen und Voraussetzungen aus Satz* 2.22 *gilt für die Norm des Ergebnisvektors*

$$\mathbb{E}\left(\|\operatorname{fl}(A\mathbf{X})\|_2^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)\Big(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\Big)\;. \tag{2.46}$$

(ii) *Die Aussage aus* (i) *bleibt richtig, wenn A anstelle von* (2.27) *Blockdiagonalgestalt mit Blöcken* (2.27) *besitzt und für jeden einzelnen Block und entsprechende Teilvektoren die Voraussetzungen aus Satz* 2.22 *erfüllt sind.*

Beweis: (i) Die Anwendung von (2.29) führt in Zusammenhang mit der Linearität des Erwartungswertes zunächst auf die Gleichung

$$\begin{aligned}\mathbb{E}\left(\|\operatorname{fl}(A\mathbf{X})\|_2^2\right) &= \mathbb{E}\Big((A\mathbf{X}+\|A\mathbf{X}\|_2\mathfrak{E})^{\mathrm{T}}(A\mathbf{X}+\|A\mathbf{X}\|_2\mathfrak{E})\Big)\\ &= \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)+\mathbb{E}\Big((\|A\mathbf{X}\|_2\mathfrak{E})^{\mathrm{T}}\|A\mathbf{X}\|_2\mathfrak{E}\Big)+2\mathbb{E}\left(\|A\mathbf{X}\|_2\mathfrak{E}^{\mathrm{T}}A\mathbf{X}\right)\;.\end{aligned}$$

Zusammen mit Satz 2.22 (iii) ergibt sich für den zweiten Term

$$\mathbb{E}\Big((\|A\mathbf{X}\|_2\mathfrak{E})^{\mathrm{T}}\|A\mathbf{X}\|_2\mathfrak{E}\Big) \;=\; \mathbb{E}\left(\|\Delta\|_2^2\right) \;=\; \mathbb{E}\left(\|A\mathbf{X}\|_2^2\right)\mathbb{E}\left(\|\mathfrak{E}\|_2^2\right)\;.$$

Für den dritten Term erhalten wir analog zum Beweis von Satz 2.22 (i)

$$\begin{aligned}\mathbb{E}\left(\|A\mathbf{X}\|_2\mathfrak{E}^{\mathrm{T}}A\mathbf{X}\right) &= \mathbb{E}\left(\|A\mathbf{X}\|_2\mathfrak{E}_1(G_1X_1+G_2X_2)\;+\;\|A\mathbf{X}\|_2\mathfrak{E}_2(G_1X_2-G_2X_1)\right)\\ &= \mathbb{E}\left(\left[G_1X_1(\varepsilon_1^+ +\varepsilon_1^\times)+G_2X_2(\varepsilon_1^+ +\varepsilon_2^\times)\right](G_1X_1+G_2X_2)\right)\\ &\qquad +\;\mathbb{E}\left(\left[G_1X_2(\varepsilon_2^+ +\varepsilon_3^\times)-G_2X_1(\varepsilon_2^+ +\varepsilon_4^\times)\right](G_1X_2-G_2X_1)\right)\\ &= 0\;,\end{aligned}$$

da die Zufallsgrößen G_1, G_2, X_1, X_2 nach Voraussetzung von den relativen Fehlern $\varepsilon_j^+,\ j=1,2$, und $\varepsilon_k^\times, k=1,,2,3,4$, stochastisch unabhängig sind und letztere nach (2.34) verschwindenden Erwartungswert besitzen. Wegen $\mathbb{E}\left(\|\mathfrak{E}\|_2^2\right)=\left(\sigma_\times^2+\sigma_+^2\right)u^2$ folgt nun die Behauptung.
(ii) Der Beweis verwendet wiederum die Linearität des Erwartungswertes und verläuft analog zum Beweis von Folgerung 2.21 (ii). ■

Abschließend erhalten wir als Hauptergebnis nun den folgenden

Satz 2.26. *Mit den Bezeichnungen und Voraussetzungen aus Lemma* 2.24 *ergibt sich für den Erwartungswert des euklidischen Normquadrates des Vektors* (2.41) *der insgesamt auftretenden Rundungsfehler die Abschätzung*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;\le\; 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)r\left(\sigma_\times^2+\sigma_+^2\right)u^2\Big(1\;+\;\mathcal{O}(u)\Big)^2\;, \tag{2.47a}$$

falls für jedes Tupel (s,k) die Zufallsvariable $\varepsilon_{\|\cdot\|_2^2}^{(s,k)}$ von den Zwischenergebnissen $X_{2k}^{(s)}, X_{2k+1}^{(s)}$ stochastisch unabhängig ist und (2.37a) *erfüllt. Bei Gültigkeit von* (2.37b) *haben wir die schärfere Abschätzung*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \;\le\; 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)r\left(\sigma_\times^2+\sigma_+^2\right)u^2\Big(1\;+\;\mathcal{O}(u^2)\Big)\;. \tag{2.47b}$$

Darüber hinaus gilt

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\le\; \sqrt{\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)}\;r\sqrt{\sigma_\times^2+\sigma_+^2}\cdot u\big(1+\mathcal{O}(u)\big) \tag{2.48}$$

Beweis: Die Ungleichungen (2.47a) und (2.47b) gehen aus den aus Lemma 2.24 bekannten Abschätzungen (2.42a) und (2.42b) hervor, indem wiederholt Lemma 2.25 auf die Erwartungswerte $\mathbb{E}\left(\left\|\mathbf{X}^{(s)}\right\|_2^2\right)$ angewandt wird. Mit den Ungleichungen (2.46) und (2.43) folgt dann für einen solchen Summanden nacheinander

$$\begin{aligned}
\mathbb{E}\left(\left\|\mathbf{X}^{(s)}\right\|_2^2\right) = \mathbb{E}\left(\left\|B^{(s-1)}\mathbf{X}^{(s-1)}\right\|_2^2\right)\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right) &\leq \mathbb{E}\left(\left\|\mathbf{X}^{(s-1)}\right\|_2^2\right)\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right) \\
&\leq \mathbb{E}\left(\left\|\mathbf{X}^{(s-2)}\right\|_2^2\right)\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right)^2 \\
&\leq \ldots \\
&\leq \mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right)^s .
\end{aligned}$$

Zusammen mit Ungleichung (2.42a) liefert dies die Ungleichungskette

$$\begin{aligned}
\mathbb{E}\left(\left\|\boldsymbol{\Delta}\right\|_2^2\right) &\leq 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\left(\sum_{s=0}^{r-1}(1+u)^{2(r-s-1)}\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right)^s\right)\left(\sigma_\times^2+\sigma_+^2\right)u^2 \\
&\leq 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\left(\sum_{s=0}^{r-1}\left(\frac{1+\left(\sigma_\times^2+\sigma_+^2\right)u^2}{1+u}\right)^s\right)\left(\sigma_\times^2+\sigma_+^2\right)(1+u)^{2(r-1)}u^2 \\
&\leq 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right) r\left(\sigma_\times^2+\sigma_+^2\right)u^2\left(1+\sum_{k=1}^{r-1}\binom{r-1}{k}u^k\right)^2 ,
\end{aligned}$$

wobei wir im letzten Schritt den für beliebige $\alpha, \beta \in \mathbb{R}$ und $n \in \mathbb{N}_0$ gültigen binomischen Lehrsatz

$$(\alpha+\beta)^n = \sum_{k=0}^{n}\binom{n}{k}\alpha^k\beta^{n-k}$$

und die Beziehung

$$0 \leq \frac{1+\left(\sigma_\times^2+\sigma_+^2\right)u^2}{1+u} \leq 1 ,$$

verwendet haben. Letztere resultiert aus den Annahmen, dass die Größen $\sigma_\times^2$ und σ_+^2 Werte im Intervall $[0,1]$ annehmen und $0 < u \leq \frac{1}{2}$ gilt. Entsprechend den Rechenregeln für $\mathcal{O}(u^2)$ ergibt sich demnach Behauptung (2.47a) im Fall (2.37a). Mittels geometrischer Summenformel gewinnen wir nun in analoger Weise die Ungleichungskette

$$\begin{aligned}
\mathbb{E}\left(\left\|\boldsymbol{\Delta}\right\|_2^2\right) &\leq 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\left(\sum_{s=0}^{r-1}\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right)^s\right)\left(\sigma_\times^2+\sigma_+^2\right)u^2 \\
&= 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\frac{\left(1+\left(\sigma_\times^2+\sigma_+^2\right)u^2\right)^r-1}{\left(\sigma_\times^2+\sigma_+^2\right)u^2}\left(\sigma_\times^2+\sigma_+^2\right)u^2 \\
&= 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right)\left(r\left(\sigma_\times^2+\sigma_+^2\right)u^2+\sum_{k=2}^{r}\binom{r}{k}\left(\sigma_\times^2+\sigma_+^2\right)^k u^{2k}\right) \\
&= 2^{\lceil\log_2(r)\rceil}\mathbb{E}\left(\left\|\mathbf{X}\right\|_2^2\right) r\left(\sigma_\times^2+\sigma_+^2\right)u^2\left(1+\sum_{k=1}^{r-1}\frac{1}{j+1}\binom{r-1}{k}\left(\sigma_\times^2+\sigma_+^2\right)^k u^{2k}\right)
\end{aligned}$$

aus (2.42b) für den Fall (2.37b), welche Behauptung (2.47b) liefert.
Betrachten wir jetzt analog zum Beweis von Lemma 2.24 die Norm der Teleskopsumme

$$\boldsymbol{\Delta} = \sum_{s=0}^{r-1}\left(\prod_{m=s+1}^{r-1}B^{(m)}\right)\left(\mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right)-B^{(s)}\mathbf{X}^{(s)}\right) ,$$

wenden die Dreiecksungleichung an und ziehen die Linearität sowie die Monotonie des Erwartungswertes hinzu, dann gelangen wir mit Hilfe der Jensen-Ungleichung (2.36) angewandt auf die konvexe Funktion $\varphi(\xi) = \xi^2$ und mit der Monotonie der Wurzel wegen (2.37) zu

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{\Delta}\|_2\right) &\leq \sum_{s=0}^{r-1} \mathbb{E}\left(\left\|\Big(\prod_{m=s+1}^{r-1} B^{(m)}\Big)\left(\mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right) - B^{(s)}\mathbf{X}^{(s)}\right)\right\|_2\right) \\
&\leq \sum_{s=0}^{r-1} (1+u)^{r-s-1} \cdot \mathbb{E}\left(\left\|\left(\mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right) - B^{(s)}\mathbf{X}^{(s)}\right)\right\|_2\right) \\
&\leq \sum_{s=0}^{r-1} (1+u)^{r-s-1} \cdot \sqrt{\mathbb{E}\left(\left\|\left(\mathrm{fl}\left(B^{(s)}\mathbf{X}^{(s)}\right) - B^{(s)}\mathbf{X}^{(s)}\right)\right\|_2^2\right)} \quad .
\end{aligned}$$

Wiederum mit (2.46) und (2.43) liefert dies die Ungleichungskette

$$\begin{aligned}
\mathbb{E}\left(\|\mathbf{\Delta}\|_2\right) &\leq \sqrt{\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)} \left(\sum_{s=0}^{r-1} (1+u)^{r-s-1} \left(1 + \left(\sigma_\times^2 + \sigma_+^2\right) u^2\right)^{\frac{s}{2}}\right) \sqrt{\sigma_\times^2 + \sigma_+^2} \cdot u \\
&\leq \sqrt{\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)} \left(\sum_{s=0}^{r-1} \left(\frac{\sqrt{1+\left(\sigma_\times^2+\sigma_+^2\right)u^2}}{1+u}\right)^s\right) \sqrt{\sigma_\times^2 + \sigma_+^2} \cdot u(1+u)^{r-1} \\
&\leq \sqrt{\mathbb{E}\left(\|\mathbf{X}\|_2^2\right)} r \sqrt{\sigma_\times^2 + \sigma_+^2} \cdot u \left(1 + \sum_{k=1}^{r-1} \binom{r-1}{k} u^k\right) ,
\end{aligned}$$

wobei wir wiederum den binomischen Lehrsatz und die Abschätzung

$$\sqrt{1 + \left(\sigma_\times^2 + \sigma_+^2\right) u^2} \;\leq\; 1 + u$$

verwendet haben, welche für genügend kleine u erfüllt ist. Somit folgt die Behauptung (2.48). ■

Für eine stochastische Rundungsfehleranalyse der in Abschnitt 1.3 vorgestellten Algorithmen in Gleitkomma-Arithmetik haben wir nun alle grundlegenden Aussagen bereitgestellt.

Bemerkung 2.27. (i) Im Unterabschnitt 2.2.2 haben wir in Lemma 2.20 und in Satz 2.22 zwei ähnliche Modelle für den Vektor der absoluten Rundungsfehler untersucht, welche bei der Multiplikation eines Vektors mit einer Drehmatrix auftreten. Gemeinsam ist beiden Modellen, dass die Größen $\varepsilon_j^\times$ $(j = 1,2,3,4)$ und ε_j^+ $(j = 1,2)$ als paarweise unabhängig mit identischen Erwartungswerten $\mu_\times u$ (bzw. $\mu_+ u$) und identischen Varianzen $\sigma_\times^2 u^2$ (bzw. $\sigma_+^2 u^2$) angenommen werden. Beim ersten Modell wird darüber hinaus vorausgesetzt, dass die Komponenten X_1, X_2 des zu drehenden Vektors $\mathbf{X}$ identisch, unabhängig und symmetrisch zum Nullpunkt verteilt sind, wobei die Konstanten $\mu_\times, \mu_+$ ungleich Null sein dürfen. Für die Hintereinanderausführung mehrerer Drehungen hat sich dieses Modell jedoch als ungeeignet herausgestellt. Demzufolge basieren alle weiteren Ergebnisse auf dem zweiten Modell aus Satz 2.22, welches neben $\mu_\times = \mu_+ = 0$ nur die sehr viel schwächere Forderung enthält, dass die Komponenten X_1, X_2 des Vektors und die Einträge G_1, G_2 der Drehmatrix von den Größen $\varepsilon_j^\bullet$ unabhängig sind. Insbesondere dürfen die Komponenten X_1, X_2 des Eingangsvektors korreliert sein, was bei Anwendungen häufig der Fall ist. Bei der zweiten Variante verwenden wir darüber hinaus die für die Einträge G_1, G_2 von Drehmatrizen der Gestalt (2.27) sinnvolle Modellannahme (2.37).

(ii) Da an die Komponenten X_1, X_2 des zu drehenden Vektors $\mathbf{X}$ kaum Forderungen gestellt werden, können wir die Aussagen aus Lemma 2.24 und damit auch aus Satz 2.26 entsprechend übertragen, wenn in der Matrixfaktorisierung neben Blockdiagonalmatrizen (2.39) mit Blöcken der Gestalt (2.40) auch Permutationsmatrizen oder Vorzeichenskalierungsmatrizen auftreten. □

2.3 Festkomma-Arithmetik nach v. Neumann und Goldstine

Das in [67] beschriebene Modell zur Festkomma-Arithmetik, welches bereits auf J. von Neumann und H.H. Goldstine [41] zurückgeht, verwendet die Betragsvorzeichendarstellung einer Festkomma-Zahl.

Für ein $q \in \mathbb{N}$ wird die Menge

$$\mathbb{M}_q := \left\{ m \in [-1,1] : m = \operatorname{sign}(m) \sum_{k=1}^{q} \mu_k 2^{-k}, \quad \mu_k \in \{0,1\} \right\} \cup \{-1,1\} \tag{2.49}$$

betrachtet [15, S. 670]. Offenbar enthält $\mathbb{M}_q$ sämtliche reelle Zahlen aus dem Intervall $[-1,1]$, die eine dyadische Entwicklung besitzen, welche spätestens nach dem q-ten Reihenglied abbricht. Darüber hinaus ist für ein Element $m \in \mathbb{M}_q$ leicht das additive Inverse zu bestimmen. In Abbildung 2.5 ist exemplarisch die Menge $\mathbb{M}_7$ graphisch dargestellt.

Abbildung 2.5: Festkomma-Zahlenmenge $\mathbb{M}_7$.

Offenbar liegen die Elemente von $\mathbb{M}_q$ äquidistant im Intervall $[-1,1]$, wobei zwei benachbarte Elemente den Abstand 2^{-q} besitzen. Fassen wir die Menge $\mathbb{M}_q \setminus \{-1,1\}$ als Spezialfall einer Gleitkomma-Menge $\mathbb{G}(\beta,\tau,\gamma_{\min},\gamma_{\max})$ mit den Parametern $\beta = 2, \tau = q$ sowie $\gamma_{\min} = \gamma_{\max} = 0$ auf, so ist in diesem Fall die im Abschnitt 2.1 eingeführte Rundungseinheit

$$u = 2^{-q} \; . \tag{2.50}$$

Wählen wir für die Approximation einer reellen Zahl $r \in [-1,1]$ in $\mathbb{M}_q$ die Abbildung $\operatorname{fix} : [-1,1] \longrightarrow \mathbb{M}_q$, welche durch

$$\operatorname{fix}(r) \;:=\; \begin{cases} r & \text{falls } r \;\in\; \{-1,1\} \; , \\ \pm \sum\limits_{k=1}^{q} \mu_k 2^{-k} & \text{falls } r \;=\; \pm \sum\limits_{k=1}^{\infty} \mu_k 2^{-k}, \quad \mu_k \in \{0,1\}, \end{cases} \tag{2.51}$$

definiert wird,[1] dann gilt für beliebiges $r \in [-1,1]$ die Beziehung

$$\operatorname{fix}(r) \;=\; r - \delta_r \qquad (|\delta_r| \le u,\; r\delta_r \ge 0,\; |\operatorname{fix}(r)| \le |r|). \tag{2.52}$$

Dies ist einerseits durch die jeweilige Monotonie der Partialsummen innerhalb der dyadischen Reihenentwicklung als auch durch die Abschätzung

$$|\delta_r| \;=\; \sum_{k=q+1}^{\infty} \mu_k \; 2^{-k} \;\le\; \sum_{k=q+1}^{\infty} 2^{-k} \;=\; 2^{-q}$$

des Reihenrestes δ_r begründet. Desweiteren gilt für beliebiges $r \in [-1,1]$ offenbar $\operatorname{fix}(-r) = -\operatorname{fix}(r)$. In Anlehnung an [41] bezeichnen wir im Folgenden Elemente $m \in \mathbb{M}_q$ als *digitale Zahlen*, behalten es uns jedoch vor, digitale Zahlen wiederum als reelle Zahlen aufzufassen. Einer der wesentlichen Unterschiede im Vergleich zum Modell der Gleitkomma-Arithmetik ist, dass die exakt ausgeführte Addition (bzw. Subtraktion) von zwei digitalen Zahlen $m_1, m_2 \in \mathbb{M}_q$ wiederum eine Zahl $m \in \mathbb{M}_q$ ergibt, solange kein Überlauf auftritt, d.h. $m \in [-1,1]$ erfüllt bleibt. Demnach ist die Addition (bzw. Subtraktion) in $\mathbb{M}_q$ unter den genannten Voraussetzungen rundungsfehlerfrei ausführbar. Insbesondere ist die Addition in $\mathbb{M}_q$ assoziativ, d.h. bei der Addition von l verschiedenen digitalen Zahlen $m_1, \ldots, m_l \in \mathbb{M}_q$ entsteht unabhängig von der Summationsreihenfolge stets dasselbe Ergebnis, vorausgesetzt, dass kein Überlauf stattfindet.
Die Multiplikation von zwei digitalen Zahlen $m_1, m_2 \in \mathbb{M}_q$ führt wegen $|m_1|, |m_2| \le 1$ nicht aus dem Intervall $[-1,1]$ heraus. Jedoch beansprucht das exakte Ergebnis $m_1 m_2$ in der Regel $2q$ Nachkommastellen, d.h. es liegt nur noch in der Menge $\mathbb{M}_{2q}$. In Anlehnung an [41] wollen wir nun annehmen, dass

[1] Dabei sei hier von den beiden möglichen Reihendarstellungen von r diejenige ausgewählt, für die nicht gilt: Es gibt ein $K \in \mathbb{N}$, so dass $\mu_k = 1$ für alle $k \ge K$ ist.

intern das exakte Ergebnis berechnet werden kann, welches anschließend – als exakte Zahl aufgefasst – mittels fix in die Menge $\mathbb{M}_q$ abgebildet wird. Für diese in [41] als Pseudo-Multiplikation bezeichnete Operation schreiben wir im Folgenden $m_1 \times m_2$. Somit ergeben sich für beliebige $m, \tilde{m} \in \mathbb{M}_q$ die Modellannahmen

$$\boxed{\begin{array}{lcll} \mathrm{fix}(m+\tilde{m}) & = & m+\tilde{m}\ , & \text{falls} \quad |m+\tilde{m}| \le 1; \\ m \times \tilde{m} & = & m\tilde{m} - \delta_{m\tilde{m}}\ , & (|\delta_{m\tilde{m}}| \le u, \quad m\tilde{m}\delta_{m\tilde{m}} \ge 0, \quad |m \times \tilde{m}| \le |m\tilde{m}|)\ . \end{array}} \tag{2.53}$$

Bemerkung 2.28. Alternativ zur Vorzeichenbetragsdarstellung werden Festkomma-Zahlen häufig auch mittels Zweierkomplement ausgedrückt. Dazu wird für ein $q \in \mathbb{N}$ die Menge

$$\mathbb{F}_q := \left\{ f \in [-1,1[\ :\ f = -\varphi_0 + \sum_{k=1}^{q} \varphi_k 2^{-k}, \quad \varphi_k \in \{0,1\}, \quad k = 0,\ldots,q \right\} \cup \{1\} \tag{2.54}$$

betrachtet, welche dieselben Elemente wie $\mathbb{M}_q$ enthält. Im Vergleich zu $\mathbb{M}_q$ besitzt die Null in $\mathbb{F}_q$ eine eindeutige Darstellung. Das additive Inverse eines Elementes $f \in \mathbb{F}_q \setminus \{-1,1\}$ ergibt sich nach Addition des Bitinversen, d.h. des Elementes $\tilde{f} \in \mathbb{F}_q$ mit den Bits $\tilde{\varphi}_k := 1 - \varphi_k$ und $u = 2^{-q}$, denn mit der Identität

$$\sum_{k=1}^{q} 2^{-k} = 1 - 2^{-q}$$

gilt

$$f + \tilde{f} + u \ = \ -\varphi_0 + \sum_{k=1}^{q} \varphi_k 2^{-k} - (1-\varphi_0) + \sum_{k=1}^{q} (1-\varphi_k) 2^{-k} + 2^{-q} \ = \ 0\ .$$

Wählen wir für die Approximation einer reellen Zahl $r \in [-1,1]$ in $\mathbb{F}_q$ die Abbildung $\mathrm{fit} : [-1,1] \longrightarrow \mathbb{F}_q$, welche durch

$$\mathrm{fit}(r) \ := \ \begin{cases} r & \text{falls } r \ = \ 1, \\ -\varphi_0 + \sum\limits_{k=1}^{q} \varphi_k 2^{-k} & \text{falls } r \ = \ -\varphi_0 + \sum\limits_{k=1}^{\infty} \varphi_k 2^{-k} < 1, \quad \varphi_k \in \{0,1\}, \end{cases}$$

definiert wird, dann gilt mit $u = 2^{-q}$ für beliebiges $r \in [-1,1]$ die Beziehung

$$\mathrm{fit}(r) \ = \ r - \varepsilon_r \qquad (0 \le \varepsilon_r \le u, \ \ \mathrm{fit}(r) \le r)\ .$$

Aufgrund der Tatsache jedoch, dass im Allgemeinen $|\,\mathrm{fit}(r)| \ne |\,\mathrm{fit}(-r)|$ ist, werden wir diese Darstellung von Festkomma-Zahlen im Weiteren nicht genauer untersuchen. □

Da wir vordergründig Matrix-Vektor-Multiplikationen betrachten wollen, erweitern wir den Definitionsbereich der Abbildung fix vom Intervall $[-1,1]$ in kanonischer Weise auf die Menge $[-1,1]^{n\times n}$ der $(n \times n)$-Matrizen mit Einträgen aus $[-1,1]$ bzw. auf die Menge $[-1,1]^n$ der Vektoren der Länge n mit Komponenten aus $[-1,1]$. Für eine Matrix $A = (a_{ij})_{i,j=0}^{n-1} \in [-1,1]^{n\times n}$ bzw. für einen Vektor $\mathbf{x} = (x_i)_{i=0}^{n-1} \in [-1,1]^n$ sei demnach die Abbildung fix element- bzw. komponentenweise durch

$$\mathrm{fix}(A) \ = \ (\mathrm{fix}(a_{ij}))_{i,j=0}^{n-1} \qquad \text{bzw.} \qquad \mathrm{fix}(\mathbf{x}) \ = \ (\mathrm{fix}(x_i))_{i=0}^{n-1}$$

definiert. Zur besseren Unterscheidung kennzeichnen wir im Folgenden Elemente aus $\mathbb{M}_q$ mit $\hat{x}$ und mit $\hat{\mathbf{x}}$ bzw. $\hat{A}$ entsprechend Elemente aus $\mathbb{M}_q^n$ bzw. $\mathbb{M}_q^{n\times n}$. Es ergibt sich nun

Folgerung 2.29. *Unter den Modellannahmen* (2.53) *gilt für das innere Pseudo-Produkt zweier Vektoren* $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{M}_q^n$*, für die*

$$\sum_{i=0}^{n-1} |\hat{x}_i \hat{y}_i| \ \le \ 1 \tag{2.55}$$

erfüllt ist, die Beziehung

$$\hat{\mathbf{x}}^{\mathrm{T}} \times \hat{\mathbf{y}} \ := \ \sum_{i=0}^{n-1} \hat{x}_i \times \hat{y}_i \ = \ \sum_{i=0}^{n-1} \hat{x}_i \hat{y}_i \ - \ \delta_{\hat{\mathbf{x}}^{\mathrm{T}} \times \hat{\mathbf{y}}} \qquad \left(|\delta_{\hat{\mathbf{x}}^{\mathrm{T}} \times \hat{\mathbf{y}}}| \le nu, \quad \sum_{i=0}^{n-1} |\hat{x}_i \times \hat{y}_i| \le \sum_{i=0}^{n-1} |\hat{x}_i \hat{y}_i| \right)$$

bzw. für das Matrix-Vektor-Pseudo-Produkt $\hat{A}\hat{\mathbf{x}}$ *mit* $\hat{A} \in \mathbb{M}_q^{n\times n}$ *und* $\hat{\mathbf{x}} \in \mathbb{M}_q^n$ *– vorausgesetzt, die Bedingung* (2.55) *ist jeweils für* $\hat{\mathbf{x}}$ *und für jede Zeile* $\hat{\mathbf{y}}^{\mathrm{T}}$ *von* $\hat{A}$ *erfüllt – die Beziehung*

$$\hat{A} \times \hat{\mathbf{x}} \;:=\; \left(\sum_{j=0}^{n-1} \hat{a}_{ij} \times \hat{x}_j\right)_{i=0}^{n-1} \;=\; \hat{A}\hat{\mathbf{x}} - \boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}} \qquad \left(\|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_\infty \le nu,\;\; \|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_2 \le \sqrt{n^3}u\right) . \quad (2.56)$$

Hierbei wird durch die Voraussetzung (2.55) sicher gestellt, dass in keinem Rechenschritt Überlauf auftritt, d.h. dass sich alle Zwischenergebnisse – als reelle Zahlen aufgefasst – im Intervall $[-1, 1]$ befinden. Aufgrund des schnellen Anstiegs der Norm des Fehlervektors $\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}$ bei Erhöhung der Transformationslänge n wird in [41] vorgeschlagen, für die Zwischenergebnisse doppelte Genauigkeit zu verwenden, d.h. erst nach Aufsummieren der exakten Produkte, welche sich in $\mathbb{M}_{2q}$ befinden, nach $\mathbb{M}_q$ abzubilden. Durch diese zusätzliche Annahme, welche eine nur wenig stärkere Forderung an die entsprechende Hardware als die zur Erfüllung der Modellannahmen (2.53) darstellt, lässt sich der Rundungsfehler merklich reduzieren. In Anlehnung an [41] bezeichnen wir diese Methode zur Ermittlung einer Approximation des inneren Produktes zweier Vektoren $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{M}_q^n$ als *doppelt genaues inneres Pseudo-Produkt* und schreiben abkürzend $\hat{\mathbf{x}}^{\mathrm{T}} \odot \hat{\mathbf{y}} \;:=\; {\sum}' \hat{x}_i\hat{y}_i$. Offenbar ergibt sich dann

Folgerung 2.30. *Unter der Modellannahme, dass alle Zwischenschritte doppelt genau (d.h. exakt) ausgeführt werden, gilt für das doppelt genaue innere Pseudo-Produkt zweier Vektoren* $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{M}_q^n$*, für die* (2.55) *erfüllt ist, die Beziehung*

$$\hat{\mathbf{x}}^{\mathrm{T}} \odot \hat{\mathbf{y}} \;=\; {\sum_{i=0}^{n-1}}' \hat{x}_i\hat{y}_i \;=\; \sum_{i=0}^{n-1} \hat{x}_i\hat{y}_i \;-\; \delta_{\hat{\mathbf{x}}^{\mathrm{T}}\odot\hat{\mathbf{y}}} \qquad \left(|\delta_{\hat{\mathbf{x}}^{\mathrm{T}}\odot\hat{\mathbf{y}}}| \le u,\;\; \left|{\sum_{i=0}^{n-1}}' \hat{x}_i\hat{y}_i\right| \le \left|\sum_{i=0}^{n-1} \hat{x}_i\hat{y}_i\right|\right) \quad (2.57)$$

bzw. für das doppelt genaue Matrix-Vektor-Pseudo-Produkt $\hat{A} \odot \hat{\mathbf{x}}$ *mit* $\hat{A} \in \mathbb{M}_q^{n\times n}$ *und* $\hat{\mathbf{x}} \in \mathbb{M}_q^n$ *– vorausgesetzt, die Bedingung* (2.55) *ist jeweils für* $\hat{\mathbf{x}}$ *und für jede Zeile* $\hat{\mathbf{y}}^{\mathrm{T}}$ *von* $\hat{A}$ *erfüllt – die Beziehung*

$$\hat{A} \odot \hat{\mathbf{x}} \;:=\; \left({\sum_{i=0}^{n-1}}' \hat{a}_{ij}\hat{x}_j\right)_{i=0}^{n-1} \;=\; \hat{A}\hat{\mathbf{x}} - \boldsymbol{\delta}_{\hat{A}\odot\hat{\mathbf{x}}} \qquad \left(\|\boldsymbol{\delta}_{\hat{A}\odot\hat{\mathbf{x}}}\|_\infty \le u,\;\; \|\boldsymbol{\delta}_{\hat{A}\odot\hat{\mathbf{x}}}\|_2 \le \sqrt{n}u\right) . \quad (2.58)$$

Für spezielle Matrizen lassen sich die Abschätzungen in (2.56) und (2.58) weiter verbessern. Bei der Multiplikation mit den in Abschnitt 1.2 eingeführten Permutationsmatrizen tritt offenbar kein Rundungsfehler auf. Genauso verhält es sich bei der Multiplikation mit der in Abschnitt 1.1 definierten Vorzeichenmatrix Σ_n. Insbesondere haben wir dabei verwendet, dass sich Permutationsmatrizen und Vorzeichenmatrizen bereits in $\mathbb{M}_q^{n\times n}$ befinden. Ein weiterer Spezialfall sind Matrizen der Form

$$\hat{Q}_2 := \begin{pmatrix} \hat{a} & \hat{b} \\ -\hat{b} & \hat{a} \end{pmatrix} \qquad (\hat{a}, \hat{b} \in \mathbb{M}_q,\; \hat{a} \cdot \hat{b} \neq 0) , \quad (2.59)$$

welche skalierte ebene Drehungen beschreiben, wegen $\hat{Q}_2^{\mathrm{T}}\hat{Q}_2 = (\hat{a}^2 + \hat{b}^2)I_2$ fast orthogonal sind, d.h. orthogonal bis auf einen positiven Faktor, und für deren Spektralnorm somit

$$\|\hat{Q}_2\|_2 = \|\Sigma_2\hat{Q}_2\|_2 = \sqrt{\hat{a}^2 + \hat{b}^2} \quad (2.60)$$

gilt. Die Abschätzungen in (2.56) und (2.58) schreiben sich für $\hat{Q}_2 \in \mathbb{M}_q^{2\times 2}$ dann als

$$\|\boldsymbol{\delta}_{\hat{Q}_2\times\hat{\mathbf{x}}}\|_\infty \le 2u , \quad \|\boldsymbol{\delta}_{\hat{Q}_2\times\hat{\mathbf{x}}}\|_2 \le 2\sqrt{2}u , \quad \|\boldsymbol{\delta}_{\hat{Q}_2\odot\hat{\mathbf{x}}}\|_\infty \le u , \quad \|\boldsymbol{\delta}_{\hat{Q}_2\odot\hat{\mathbf{x}}}\|_2 \le \sqrt{2}u . \quad (2.61)$$

Fordern wir für die Einträge $\hat{a}, \hat{b}$ von $\hat{Q}_2$ zusätzlich Nichtnegativität, folgen bei der Multiplikation

$$\hat{Q}_2 \times \hat{\mathbf{x}} \;=\; \begin{pmatrix} \hat{a}\times\hat{x}_0 + \hat{b}\times\hat{x}_1 \\ \hat{a}\times\hat{x}_1 - \hat{b}\times\hat{x}_0 \end{pmatrix} \;=\; \hat{Q}_2\hat{\mathbf{x}} - \begin{pmatrix} \delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1} \\ \delta_{\hat{a}\hat{x}_1} - \delta_{\hat{b}\hat{x}_0} \end{pmatrix} \qquad \left(\left\|\begin{pmatrix} \delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1} \\ \delta_{\hat{a}\hat{x}_1} - \delta_{\hat{b}\hat{x}_0} \end{pmatrix}\right\|_\infty \;\le\; 2u\right)$$

für ein $\hat{\mathbf{x}} = (\hat{x}_0, \hat{x}_1)^{\mathrm{T}}$ nach (2.53) die Beziehungen

$$x_0 \cdot \delta_{\hat{a}\hat{x}_0} \;\ge\; 0 , \quad x_0 \cdot \delta_{\hat{b}\hat{x}_0} \;\ge\; 0 , \quad x_1 \cdot \delta_{\hat{a}\hat{x}_1} \;\ge\; 0 , \quad x_1 \cdot \delta_{\hat{b}\hat{x}_1} \;\ge\; 0 . \quad (2.62)$$

Ist nun $\hat{x}_0\hat{x}_1 \leq 0$, dann folgt auch $\delta_{\hat{a}\hat{x}_0} \cdot \delta_{\hat{b}\hat{x}_1} \leq 0$ und somit $|\delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1}| \leq u$. Andernfalls ist $\hat{x}_0\hat{x}_1 \geq 0$, so dass dann $\delta_{\hat{a}\hat{x}_1} \cdot \delta_{\hat{b}\hat{x}_0} \geq 0$ und folglich $|\delta_{\hat{a}\hat{x}_1} - \delta_{\hat{b}\hat{x}_0}| \leq u$ erfüllt ist. Somit gilt für Matrizen (2.59) mit nichtnegativen $\hat{a}$, $\hat{b}$ sogar die schärfere Abschätzung

$$\|\boldsymbol{\delta}_{\hat{Q}_2 \times \hat{\mathbf{x}}}\|_\infty \leq 2u \ , \quad \|\boldsymbol{\delta}_{\hat{Q}_2 \times \hat{\mathbf{x}}}\|_2 \leq \sqrt{5}u \ . \tag{2.63}$$

Damit ergibt sich die

Satz 2.31. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und $\varphi_k \in \left]0, \frac{\pi}{4}\right]$ $(k = 0, \ldots, n_1 - 1)$. Weiter sei $\hat{\mathbf{x}} \in \mathbb{M}_q^n$, so dass alle Teilvektoren $\hat{\mathbf{x}}^{(k)} := (\hat{x}_{2k}, \hat{x}_{2k+1})^{\mathrm{T}}$ $(k = 0, \ldots, n_1 - 1)$ der Abschätzung $\|\hat{\mathbf{x}}^{(k)}\|_2 \leq 1$ genügen. Dann existieren Fehlervektoren $\boldsymbol{\delta}^{\times}$ und $\boldsymbol{\delta}^{\odot}$ mit*

$$\left(\bigoplus_{k=0}^{n_1-1} \hat{Q}_2(\varphi_k)\right) \times \hat{\mathbf{x}} \ = \ \left(\bigoplus_{k=0}^{n_1-1} \hat{Q}_2(\varphi_k)\right) \hat{\mathbf{x}} \ + \ \boldsymbol{\delta}^{\times} \quad \left(\|\boldsymbol{\delta}^{\times}\|_\infty \ \leq \ 2u \ , \ \ \|\boldsymbol{\delta}^{\times}\|_2 \ \leq \ \sqrt{\tfrac{5n}{2}}u\right), \tag{2.64a}$$

$$\left(\bigoplus_{k=0}^{n_1-1} \hat{Q}_2(\varphi_k)\right) \odot \hat{\mathbf{x}} \ = \ \left(\bigoplus_{k=0}^{n_1-1} \hat{Q}_2(\varphi_k)\right) \hat{\mathbf{x}} \ + \ \boldsymbol{\delta}^{\odot} \quad \left(\|\boldsymbol{\delta}^{\odot}\|_\infty \ \leq \ u \ , \ \ \|\boldsymbol{\delta}^{\odot}\|_2 \ \leq \ \sqrt{n}u\right). \tag{2.64b}$$

Beweis: Zunächst halten wir fest, dass aufgrund der speziellen Struktur n_1 voneinander unabhängige Matrix-Vektor-Multiplikationen mit Matrizen der Form (2.59) vorliegen, deren Einträge den Ungleichungsketten $0 \leq \hat{a}_k \leq \cos\varphi_k \leq 1$ und $0 \leq \hat{b}_k \leq \sin\varphi_k \leq 1$ genügen. Insbesondere gilt für die Spektralnormen

$$\left\|\hat{Q}_2(\varphi_k)\right\|_2 \ = \ \sqrt{\hat{a}_k^2 + \hat{b}_k^2} \ \leq \ 1 \tag{2.65}$$

und weiter

$$\left\|\hat{Q}_2(\varphi_k)\hat{\mathbf{x}}^{(k)}\right\|_\infty \ \leq \ \left\|\hat{Q}_2(\varphi_k)\hat{\mathbf{x}}^{(k)}\right\|_2 \ \leq \ \|\hat{\mathbf{x}}^{(k)}\|_2 \ .$$

Somit liegen unter den gegebenen Voraussetzungen sämtliche Einträge im Intervall $[-1, 1]$. Um nun die Folgerungen 2.29 und 2.30 anwenden zu können, bleibt noch die Bedingung (2.55) zu überprüfen. Aufgrund der Blockdiagonalgestalt beinhaltet die Summe in (2.55) für jede Zeile jeweils nur zwei Summanden, die sich mittels Cauchy-Schwarz-Ungleichung durch

$$\begin{aligned} |\hat{a}_k||\hat{x}_{2k}| + |\hat{b}_k||\hat{x}_{2k+1}| \ &\leq \ \|\hat{\mathbf{x}}^{(k)}\|_2 \sqrt{\hat{a}_k^2 + \hat{b}_k^2} \ , \\ |-\hat{b}_k||\hat{x}_{2k}| + |\hat{a}_k||\hat{x}_{2k+1}| \ &\leq \ \|\hat{\mathbf{x}}^{(k)}\|_2 \sqrt{\hat{b}_k^2 + \hat{a}_k^2} \ , \end{aligned}$$

$k = 0, \ldots, n_1 - 1$, abschätzen lassen. Wegen (2.65) und mit der Voraussetzung $\|\hat{\mathbf{x}}^{(k)}\|_2 \leq 1$, $k = 0, \ldots, n_1 - 1$, ist (2.55) jedoch stets erfüllt, so dass (2.64b) direkt aus (2.58) folgt. Mit

$$\|\boldsymbol{\delta}^{\times}\|_2^2 = \sum_{k=0}^{n_1-1} \|\boldsymbol{\delta}_{\hat{Q}_2(\varphi_k) \times \hat{\mathbf{x}}}\|_2^2$$

ergibt sich aus (2.63) nun ebenso (2.64a). ■

Basierend auf den Ergebnissen aus Satz 2.31 werden in Kapitel 4 dann entsprechende Abschätzungen für den gesamten Rundungsfehler der uns interessierenden Algorithmen aus Abschnitt 1.3 hergeleitet.

2.4 Stochastisches Modell für Festkomma-Arithmetik

Im vorangegangenen Abschnitt haben wir die Menge $\mathbb{M}_q$ der Festkomma-Zahlen sowie die Abbildung $\mathrm{fix} : [-1, 1] \to \mathbb{M}_q$ eingeführt. Weiterhin haben wir das Modell (2.53) auf das Skalarprodukt zweier Vektoren sowie Matrix-Vektor-Multiplikation angewandt und Fehlerschranken für die resultierenden Rundungsfehler bestimmt. In diesem Abschnitt beschäftigen wir uns nun mit dem durchschnittlichen Verhalten des Rundungsfehlers. Dazu fassen wir sämtliche Größen als Zufallsvariablen (vgl. Definition A.21) auf. Ein kurzer Überblick zu den Grundlagen der Maß- und Wahrscheinlichkeitstheorie wird in Anhang A.1 zur Verfügung gestellt. Um Zufallsgrößen von deterministischen Größen abzugrenzen, verwenden wir im Folgenden für Zufallsgrößen große Buchstaben. Wir betrachten nun den bei Anwendung des Einbettungsoperators (2.51) entstehenden Approximationsfehler unter der Annahme, dass die Eingangsgröße auf dem Intervall $[-1, 1]$ gleichverteilt ist. In diesem Fall können die Verteilungen explizit angegeben werden.

Lemma 2.32. *Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathfrak{S}, P)$ und eine auf dem Intervall $[-1,1]$ gleichverteilte Zufallsvariable $X : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Weiterhin sei $q \in \mathbb{N}$ fest gewählt und die in (2.51) definierte Abbildung* fix *auf $\mathbb{R} \setminus [-1,1]$ beliebig fortgesetzt. Dann gelten:*

(i) *Die durch $\hat{X} := \mathrm{fix}(X)$ definierte Zufallsvariable $\hat{X} : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ ist diskret verteilt auf der Menge*

$$A_q \;:=\; \left\{ k \cdot 2^{-q} \;\middle|\; k = -2^q + 1, \ldots, 2^q - 1 \right\} \;=\; \mathbb{M}_q \setminus \{-1, 1\}$$

mit Wahrscheinlichkeitsmaß

$$\hat{X}(P) \;=\; \frac{1}{2^{q+1}} \sum_{k=0}^{2^q-1} \left(\eth_{k \cdot 2^{-q}} + \eth_{-k \cdot 2^{-q}} \right) \;, \tag{2.66}$$

wobei $\eth_a$ das in (A.44) definierte Dirac-Maß mit Masse im Punkt $a \in \mathbb{R}$ bezeichnet.

(ii) *Die Zufallsgröße $\mathfrak{d} : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, definiert durch $\mathfrak{d} := X - \mathrm{fix}(X)$, ist gleichverteilt auf dem Intervall $[-2^{-q}, 2^{-q}]$. Entsprechend ist $|\mathfrak{d}|$ gleichverteilt auf dem Intervall $[0, 2^{-q}]$.*

Der **Beweis** ist in Anhang A.2 zu finden. ■

Bei Kenntnis der Verteilung können die charakteristischen Größen Erwartungswert und Varianz exakt berechnet werden.

Folgerung 2.33. *Unter den Voraussetzungen und mit den Bezeichnungen aus Lemma 2.32 ergeben sich die Erwartungswerte $\mathbb{E}(X) = \mathbb{E}\big(\hat{X}\big) = \mathbb{E}(\mathfrak{d}) = 0$ und $\mathbb{E}(|\mathfrak{d}|) = 2^{-(q+1)}$ sowie die Varianzen $\mathbb{V}(X) = \frac{1}{3}$, $\mathbb{V}(\mathfrak{d}) = \frac{1}{3}\, 2^{-2q}$, $\mathbb{V}(|\mathfrak{d}|) = \frac{1}{12}\, 2^{-2q}$ und*

$$\mathbb{V}\big(\hat{X}\big) \;=\; \frac{1}{3} \left(1 - 2^{-q}\right) \left(1 - 2^{-(q+1)}\right) .$$

Beweis: Aufgrund der nach Lemma 2.32 stetigen Gleichverteilung der Zufallsgrößen X und $\mathfrak{d}$ sowie $|\mathfrak{d}|$ folgen aus (A.58) und (A.59) mit $b = -a = 1$ für X bzw. mit $b = -a = 2^{-q}$ für $\mathfrak{d}$ bzw. mit $b = 2^{-q}$ und $a = 0$ für $|\mathfrak{d}|$ sofort die Behauptungen. Aus Lemma 2.32 geht weiter hervor, dass $\hat{X}$ eine diskrete Zufallsvariable mit Bildmaß (2.66) ist. Analog zu den Berechnungen nach Definition A.39 ergeben sich daher

$$\mathbb{E}(\hat{X}) \;=\; \frac{1}{2^{q+1}} \sum_{k=0}^{2^q-1} \left((k \cdot 2^{-q}) + (-k \cdot 2^{-q}) \right) \;=\; 0$$

und

$$\mathbb{V}(\hat{X}) \;=\; \frac{1}{2^{q+1}} \sum_{k=0}^{2^q-1} \left((k \cdot 2^{-q})^2 + (-k \cdot 2^{-q})^2 \right) \;=\; \frac{1}{2^{3q}} \cdot \frac{(2^q - 1) 2^q (2^{q+1} - 1)}{6} \;=\; \frac{\left(1 - 2^{-q}\right) \left(1 - 2^{-(q+1)}\right)}{3} ,$$

wobei wir die per Induktion leicht zu überprüfende Summenformel

$$\sum_{k=1}^{N-1} k^2 \;=\; \frac{(N-1) N (2N-1)}{6}$$

verwendet haben. ■

Wie im Anhang A.1 ab Lemma A.41 zu erkennen ist, erfordert die tatsächliche Bestimmung der Verteilung der bei der Addition von Zufallsvariablen neu entstehenden Zufallsgröße einen erheblichen Aufwand, wenn wir sie in einer allgemeinen geschlossenen Form darstellen möchten. Daher werden wir uns im Folgenden damit zufrieden geben, lediglich die charakteristischen Größen Erwartungswert und Varianz bestimmen zu wollen oder Ungleichungen für diese herzuleiten.

Wie zuvor im Fall der Gleitkomma-Arithmetik betrachten wir nun Matrix-Vektor-Multiplikationen mit Matrizen der Gestalt (2.59).

Lemma 2.34. *Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in (2.49) und (2.50) definiert. Weiter enthalte der Vektor $\mathbf{X} = (X_1, X_2)^{\mathrm{T}}$ die Zufallsvariablen X_1, X_2 mit Werten in $\mathbb{M}_q$, für welche die Zufallsvariable $\Xi := \|\mathbf{X}\|_2^2$ nur Werte im Intervall $[0,1]$ annehmen kann. Darüber hinaus sei die Matrix*

$$A \;:=\; \begin{pmatrix} M_1 & M_2 \\ -M_2 & M_1 \end{pmatrix} \tag{2.67}$$

mit Zufallsvariablen M_1 und M_2 aus $\mathbb{M}_q$ definiert, wobei $M_1^2 + M_2^2 = 1 - \delta_A$ mit $0 \leq \delta_A \leq 2u$ erfüllt sei. Angenommen, es gelten bei der in Abschnitt 2.3 vorgestellten einfach bzw. doppelt genauen Matrix-Vektor-Multiplikation die Beziehungen

$$A \times \mathbf{X} \;=\; A\mathbf{X} + \mathfrak{D}_{A\times\mathbf{X}} \qquad \textit{mit} \qquad \mathfrak{D}_{A\times\mathbf{X}} \;:=\; \begin{pmatrix} \delta_{11} + \delta_{22} \\ \delta_{12} - \delta_{21} \end{pmatrix} \;, \tag{2.68a}$$

$$A \odot \mathbf{X} \;=\; A\mathbf{X} + \mathfrak{D}_{A\odot\mathbf{X}} \qquad \textit{mit} \qquad \mathfrak{D}_{A\odot\mathbf{X}} \;:=\; \begin{pmatrix} \delta_{1,\odot} \\ \delta_{2,\odot} \end{pmatrix} \;, \tag{2.68b}$$

wobei die Zufallsgrößen $\delta_{11}, \delta_{22}, \delta_{12}, \delta_{21}$ bzw. $\delta_{1,\odot}, \delta_{2,\odot}$ jeweils den Erwartungswert $\mathbb{E}(\delta_{rs}) = 0$ bzw. $\mathbb{E}(\delta_{v,\odot}) = 0$ sowie die Varianz $\mathbb{V}(\delta_{rs}) = \sigma_{rs}^2 u^2$ bzw. $\mathbb{V}(\delta_{v,\odot}) = \sigma_{v,\odot}^2 u^2$ $(r, s, v = 1, 2)$ besitzen. Dann verschwinden die Erwartungswerte $\mathbb{E}\left(\mathfrak{D}_{A\times\mathbf{X}}\right)$, $\mathbb{E}\left(\mathfrak{D}_{A\odot\mathbf{X}}\right)$ und es gelten die Ungleichungen

$$\mathbb{E}\left(\|\mathfrak{D}_{A\times\mathbf{X}}\|_2^2\right) \;\leq\; 2\left(\sigma_{11}^2 + \sigma_{22}^2 + \sigma_{12}^2 + \sigma_{21}^2\right) u^2 \;, \tag{2.69a}$$

$$\mathbb{E}\left(\|\mathfrak{D}_{A\odot\mathbf{X}}\|_2^2\right) \;\leq\; 2\left(\sigma_{1,\odot}^2 + \sigma_{2,\odot}^2\right) u^2 \;. \tag{2.69b}$$

Stimmen die Kovarianzen $\mathrm{Cov}(\delta_{11}, \delta_{22})$ und $\mathrm{Cov}(\delta_{12}, \delta_{21})$ überein bzw. sind zusätzlich die Größen δ_{rs}, $r, s = 1, 2$, und $\delta_{1,\odot}, \delta_{2,\odot}$ jeweils unabhängig identisch verteilt (u.i.v.) mit Varianz $\sigma^2 u^2$ bzw. $\sigma_\odot^2 u^2$, so werden die Gleichungen

$$\mathbb{E}\left(\|\mathfrak{D}_{A\times\mathbf{X}}\|_2^2\right) \;=\; \begin{cases} \left(\sigma_{11}^2 + \sigma_{22}^2 + \sigma_{12}^2 + \sigma_{21}^2\right) u^2 \;, & \textit{falls } \mathrm{Cov}(\delta_{11}, \delta_{22}) \;=\; \mathrm{Cov}(\delta_{12}, \delta_{21}) \;, \\ 4\sigma^2 u^2 \;, & \textit{falls } \delta_{jk},\ j, k = 1, 2, \textit{ u.i.v.}, \end{cases} \tag{2.70a}$$

$$\mathbb{E}\left(\|\mathfrak{D}_{A\odot\mathbf{X}}\|_2^2\right) \;=\; \begin{cases} \left(\sigma_{1,\odot}^2 + \sigma_{2,\odot}^2\right) u^2 \;, & \textit{falls } \mathrm{Cov}\left(\delta_{1,\odot}, \delta_{2,\odot}\right) \;=\; 0 \;, \\ 2\sigma_\odot^2 u^2 \;, & \textit{falls } \delta_{1,\odot}, \delta_{2,\odot} \textit{ u.i.v.}, \end{cases} \tag{2.70b}$$

erfüllt.

Beweis: Wegen $\Xi \in [0, 1]$ sowie $\delta_A \in [0, 2u]$ und somit

$$\|A\mathbf{X}\|_2^2 \;=\; (M_1^2 + M_2^2)(X_1^2 + X_2^2) \;\leq\; (1 - \delta_A) \cdot \Xi \;\leq\; \Xi$$

wird sicher gestellt, dass die beiden Komponenten von $A\mathbf{X}$ jeweils wieder im Intervall $[-1, 1]$ liegen. Mittels elementarer Rechenregeln für den Erwartungswert ergibt sich zunächst

$$\begin{aligned} \mathbb{E}\left(\|\mathfrak{D}_{A\times\mathbf{X}}\|_2^2\right) \;&=\; \mathbb{E}\left((\delta_{11} + \delta_{22})^2 + (\delta_{12} - \delta_{21})^2\right) \\ &=\; \left(\sigma_{11}^2 + \sigma_{22}^2 + \sigma_{12}^2 + \sigma_{21}^2\right) u^2 + 2\mathbb{E}(\delta_{11}\delta_{22}) - 2\mathbb{E}(\delta_{12}\delta_{21}) \end{aligned}$$

und analog

$$\mathbb{E}\left(\|\mathfrak{D}_{A\odot\mathbf{X}}\|_2^2\right) \;=\; \left(\sigma_{1,\odot}^2 + \sigma_{2,\odot}^2\right) u^2 + 2\mathbb{E}\left(\delta_{1,\odot}\delta_{2,\odot}\right) \;.$$

Unter Berücksichtigung der Monotonie des Erwartungswertes sowie mittels $2|ab| \leq a^2 + b^2$ erhalten wir dann die behaupteten Ungleichungen (2.69a) und (2.69b). Stimmen die in der ersten Gleichungskette auftretenden Kovarianzen überein, so entfallen dort die gemischten Terme. Darüber hinaus können wir im Falle unabhängig identisch verteilter Zufallsgrößen alle weiteren Terme zusammenfassen, womit sich (2.70a) und (2.70b) ergeben. ■

Aufgrund ihrer speziellen Struktur, können wir die Ergebnisse aus Lemma 2.34 sofort auf eine Blockdiagonalmatrix mit Blöcken der Gestalt (2.67) übertragen.

Lemma 2.35. *Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in (2.49) und (2.50) definiert. Sei $n \in \mathbb{N}$, $n \geq 4$ und gerade. Weiter sei der Vektor $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{M}_q$ gegeben, so dass die Werte der Zufallsvariablen $\Xi_k = X_{2k}^2 + X_{2k+1}^2$, $k = 0, \ldots, \frac{n}{2} - 1$, im Intervall $[0, 1]$ enthalten sind. Darüber hinaus sei die Blockdiagonalmatrix*

$$B \;:=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k)} \qquad \left(A^{(k)} \;:=\; \begin{pmatrix} M_1^{(k)} & M_2^{(k)} \\ -M_2^{(k)} & M_1^{(k)} \end{pmatrix}\right) \tag{2.71}$$

mit Zufallsvariablen $M_1^{(k)}$ *und* $M_2^{(k)}$ *aus* $\mathbb{M}_q$ *definiert, für welche*

$$(M_1^{(k)})^2 + (M_2^{(k)})^2 = 1 - \delta_{A^{(k)}}$$

mit $0 \leq \delta_{A^{(k)}} \leq 2u,\ k = 0, \ldots, \frac{n}{2} - 1$, *erfüllt sei. Angenommen, für jedes* $k = 0, \ldots, \frac{n}{2} - 1$ *gelten*

$$A^{(k)} \times \mathbf{X}^{(k)} = A^{(k)}\mathbf{X}^{(k)} + \mathfrak{D}_{A^{(k)}\times\mathbf{X}^{(k)}} \qquad mit \qquad \mathfrak{D}_{A^{(k)}\times\mathbf{X}^{(k)}} := \begin{pmatrix} \delta_{11}^{(k)} + \delta_{22}^{(k)} \\ \delta_{12}^{(k)} - \delta_{21}^{(k)} \end{pmatrix}, \tag{2.72a}$$

$$A^{(k)} \odot \mathbf{X}^{(k)} = A^{(k)}\mathbf{X}^{(k)} + \mathfrak{D}_{A^{(k)}\odot\mathbf{X}^{(k)}} \qquad mit \qquad \mathfrak{D}_{A^{(k)}\odot\mathbf{X}^{(k)}} := \begin{pmatrix} \delta_{1,\odot}^{(k)} \\ \delta_{2,\odot}^{(k)} \end{pmatrix} \tag{2.72b}$$

mit den Teilvektoren $\mathbf{X}^{(k)} = (X_{2k}, X_{2k+1})$ *und den Zufallsgrößen* $\delta_{11}^{(k)}, \delta_{22}^{(k)}, \delta_{12}^{(k)}, \delta_{21}^{(k)}$ *bzw.* $\delta_{1,\odot}^{(k)}, \delta_{2,\odot}^{(k)}$, *welche jeweils den Erwartungswert* $\mathbb{E}(\delta_{rs}^{(k)}) = 0$ *bzw.* $\mathbb{E}(\delta_{v,\odot}^{(k)}) = 0$ *sowie die Varianz* $\mathbb{V}(\delta_{rs}^{(k)}) = (\sigma_{rs}^{(k)})^2 u^2$ *bzw.* $\mathbb{V}(\delta_{v,\odot}^{(k)}) = (\sigma_{v,\odot}^{(k)})^2 u^2$ $(r, s, v = 1, 2)$ *besitzen. Dann erfüllen die Fehlervektoren*

$$\mathfrak{D}_{B\times\mathbf{X}} := B \times \mathbf{X} - B\mathbf{X}\ , \tag{2.73a}$$

$$\mathfrak{D}_{B\odot\mathbf{X}} := B \odot \mathbf{X} - B\mathbf{X}\ , \tag{2.73b}$$

einerseits

$$\mathbb{E}\left(\mathfrak{D}_{B\times\mathbf{X}}\right) = \mathbb{E}\left(\mathfrak{D}_{B\odot\mathbf{X}}\right) = \mathbf{0} \tag{2.74}$$

und andererseits die Ungleichungen

$$\mathbb{E}\left(\|\mathfrak{D}_{B\times\mathbf{X}}\|_2^2\right) \leq 2\sum_{k=0}^{\frac{n}{2}-1}\left((\sigma_{11}^{(k)})^2 + (\sigma_{22}^{(k)})^2 + (\sigma_{12}^{(k)})^2 + (\sigma_{21}^{(k)})^2\right)u^2\ , \tag{2.75a}$$

$$\mathbb{E}\left(\|\mathfrak{D}_{B\odot\mathbf{X}}\|_2^2\right) \leq 2\sum_{k=0}^{\frac{n}{2}-1}\left((\sigma_{1,\odot}^{(k)})^2 + (\sigma_{2,\odot}^{(k)})^2\right)u^2. \tag{2.75b}$$

Im Fall $\mathrm{Cov}\left(\delta_{11}^{(k)}, \delta_{22}^{(k)}\right) = \mathrm{Cov}\left(\delta_{12}^{(k)}, \delta_{21}^{(k)}\right)$ *bzw.* $\mathbb{E}\left(\delta_{1,\odot}^{(k)}\delta_{2,\odot}^{(k)}\right) = 0$ *für alle* $k = 0, \ldots, \frac{n}{2} - 1$ *gilt*

$$\mathbb{E}\left(\|\mathfrak{D}_{B\times\mathbf{X}}\|_2^2\right) = \sum_{k=0}^{\frac{n}{2}-1}\left((\sigma_{11}^{(k)})^2 + (\sigma_{22}^{(k)})^2 + (\sigma_{12}^{(k)})^2 + (\sigma_{21}^{(k)})^2\right)u^2\ , \tag{2.76a}$$

$$\mathbb{E}\left(\|\mathfrak{D}_{B\odot\mathbf{X}}\|_2^2\right) = \sum_{k=0}^{\frac{n}{2}-1}\left((\sigma_{1,\odot}^{(k)})^2 + (\sigma_{2,\odot}^{(k)})^2\right)u^2. \tag{2.76b}$$

Sind zusätzlich $\delta_{rs}^{(k)},\ r, s = 1, 2$ *bzw.* $\delta_{1,\odot}^{(k)}, \delta_{2,\odot}^{(k)}$ *für alle* $k = 0, \ldots, \frac{n}{2} - 1$ *unabhängig identisch verteilt mit Varianz* $\sigma^2 u^2$ *bzw.* $\sigma_\odot^2 u^2$, *so werden die Gleichungen*

$$\mathbb{E}\left(\|\mathfrak{D}_{B\times\mathbf{X}}\|_2^2\right) = 2n\sigma^2 u^2\ , \tag{2.77a}$$

$$\mathbb{E}\left(\|\mathfrak{D}_{B\odot\mathbf{X}}\|_2^2\right) = n\sigma_\odot^2 u^2. \tag{2.77b}$$

erfüllt.

Beweis: Alle Aussagen ergeben sich aus Lemma 2.34 aufgrund der Blockdiagonalgestalt (2.71) und unter Verwendung von

$$\|\mathfrak{D}_{B\times\mathbf{X}}\|_2^2 = \sum_{k=0}^{\frac{n}{2}-1}\|\mathfrak{D}_{A^{(k)}\times\mathbf{X}^{(k)}}\|_2^2\ , \tag{2.78a}$$

$$\|\mathfrak{D}_{B\odot\mathbf{X}}\|_2^2 = \sum_{k=0}^{\frac{n}{2}-1}\|\mathfrak{D}_{A^{(k)}\odot\mathbf{X}^{(k)}}\|_2^2 \tag{2.78b}$$

sowie der Linearität des Erwartungswertes. ■

Die Aussagen aus Lemma 2.35 bleiben im Wesentlichen erhalten, wenn anstelle der Blockdiagonalmatrix B in (2.71) eine Matrix betrachtet wird, welche durch Permutations- oder Vorzeichenskalierungsmatrizen auf die Gestalt (2.71) gebracht werden kann.

Folgerung 2.36. *Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in* (2.49) *und* (2.50) *definiert. Sei $n \in \mathbb{N},\ n \geq 4$ und gerade. Weiter sei der Vektor $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{M}_q$ gegeben. Sind nun U und V orthogonale Matrizen aus $\mathbb{M}_q^{n\times n}$, so dass eine Multiplikation mit ihnen lediglich Permutationen sowie zeilen- bzw. spaltenweise Vorzeichenskalierungen bewirkt und sind die Voraussetzungen aus Lemma* 2.35 *für eine Matrix B wie in* (2.71) *und den Vektor $\mathbf{Y} := V\mathbf{X}$ erfüllt, so gelten die Aussagen aus Lemma* 2.35 *ebenso für die Fehlervektoren*

$$\begin{aligned} \mathfrak{D}_{UBV\times\mathbf{X}} &:= UBV \times \mathbf{X} - UBV\mathbf{X}\ , \\ \mathfrak{D}_{UBV\odot\mathbf{X}} &:= UBV \odot \mathbf{X} - UBV\mathbf{X}\ . \end{aligned}$$

Beweis: Zunächst halten wir fest, dass Permutationen und spaltenweise Vorzeichenskalierungen eines Vektors keine Auswirkungen auf den Rundungsfehler haben. Aufgrund der Linearität des Erwartungswertes ergibt sich wegen

$$\mathfrak{D}_{UBV\times\mathbf{X}} \ = \ U\left(BV \times \mathbf{X} - BV\mathbf{X}\right) \ = \ U\left(B \times V\mathbf{X} - BV\mathbf{X}\right) \ = \ U\mathfrak{D}_{B\times V\mathbf{X}}$$

dann $\mathbb{E}\left(\mathfrak{D}_{UBV\times\mathbf{X}}\right) = U\mathbb{E}\left(\mathfrak{D}_{B\times V\mathbf{X}}\right) = \mathbf{0}$ und analog $\mathbb{E}\left(\mathfrak{D}_{UBV\odot\mathbf{X}}\right) = \mathbf{0}$. Alle weiteren Behauptungen sind eine Konsequenz aus Lemma 2.35 angewandt auf die Blockdiagonalmatrix B und den Vektor $V\mathbf{X}$ unter Verwendung von $\|U\mathfrak{D}_{B\times V\mathbf{X}}\|_2 = \|\mathfrak{D}_{B\times V\mathbf{X}}\|_2$ bzw. $\|U\mathfrak{D}_{B\odot V\mathbf{X}}\|_2 = \|\mathfrak{D}_{B\odot V\mathbf{X}}\|_2$. Hierbei wird implizit verwendet, dass die Werte der Zufallsgrößen $\Upsilon_k := Y_{2k}^2 + Y_{2k+1}^2,\ k = 0,\ldots,\frac{n}{2}-1$, nach Voraussetzung im Intervall $[0,1]$ liegen. ■

Nun können wir auch Hintereinanderausführungen von blockweisen Drehungen betrachten.

Satz 2.37. *Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in* (2.49) *und* (2.50) *definiert. Sei $n \in \mathbb{N},\ n \geq 4$ und gerade. Weiter sei $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{M}_q$ gegeben, so dass die Zufallsvariablen $\Xi_k = X_{2k}^2 + X_{2k+1}^2,\ k = 0,\ldots,\frac{n}{2}-1$, nur Werte im Intervall $[0,1]$ annehmen. Darüber hinaus seien für $\imath = 0,\ldots,\jmath-1$ mit einem $\jmath \in \mathbb{N}$ die Blockdiagonalmatrizen*

$$B^{(\imath)} := \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k,\imath)} \qquad \left(A^{(k,\imath)} := \begin{pmatrix} M_1^{(k,\imath)} & M_2^{(k,\imath)} \\ -M_2^{(k,\imath)} & M_1^{(k,\imath)} \end{pmatrix}\right) \tag{2.79}$$

mit Zufallsvariablen $M_1^{(k,\imath)}$ und $M_2^{(k,\imath)}$ aus $\mathbb{M}_q$ definiert, für welche jeweils

$$(M_1^{(k,\imath)})^2 + (M_2^{(k,\imath)})^2 = 1 - \delta_{A^{(k,\imath)}} \tag{2.80}$$

mit $0 \leq \delta_{A^{(k,\imath)}} \leq 2u,\ k = 0,\ldots,\frac{n}{2}-1,\ \imath = 0,\ldots,\jmath-1$, gelte.
Rekursiv definieren wir nun die zufälligen Vektoren

$$\begin{aligned} \mathbf{X}^{(0)} &:= \mathbf{X}\ , \\ \mathbf{X}^{(\imath+1)} &:= B^{(\imath)} \times \mathbf{X}^{(\imath)}\ , \quad \imath = 0,\ldots,\jmath-1, \end{aligned} \qquad \begin{aligned} \mathbf{X}^{(0,\odot)} &:= \mathbf{X}\ , \\ \mathbf{X}^{(\imath+1,\odot)} &:= B^{(\imath)} \odot \mathbf{X}^{(\imath,\odot)}\ , \quad \imath = 0,\ldots,\jmath-1, \end{aligned}$$

über die Blockdiagonalmatrizen (2.79)*. Weiter gäbe es für jede Blockdiagonalmatrix $B^{(\imath)}$ und den zugehörigen Vektor $\mathbf{X}^{(\imath)}$ bzw. $\mathbf{X}^{(\imath,\odot)}$ Zufallsgrößen $\delta_{11}^{(k,\imath)}$, $\delta_{22}^{(k,\imath)}$, $\delta_{12}^{(k,\imath)}$, $\delta_{21}^{(k,\imath)}$ bzw. $\delta_{1,\odot}^{(k,\imath)}, \delta_{2,\odot}^{(k,\imath)}$, welche jeweils den Erwartungswert $\mathbb{E}(\delta_{rs}^{(k,\imath)}) = 0$ bzw. $\mathbb{E}(\delta_{v,\odot}^{(k,\imath)}) = 0$ sowie die Varianz $\mathbb{V}(\delta_{rs}^{(k,\imath)}) = (\sigma_{rs}^{(k,\imath)})^2u^2$ bzw. $\mathbb{V}(\delta_{v,\odot}^{(k,\imath)}) = (\sigma_{v,\odot}^{(k,\imath)})^2u^2,\ r,s,v = 1,2$, besitzen und* (2.72a) *bzw.* (2.72b) *genügen. Dann gelten für den Vektor der insgesamt auftretenden absoluten Rundungsfehler*

$$\boldsymbol{\Delta} := \mathbf{X}^{(\jmath)} - \left(\prod_{\imath=0}^{\jmath-1} B^{(\imath)}\right)\mathbf{X} \qquad \textit{bzw.} \qquad \boldsymbol{\Delta}^{\odot} := \mathbf{X}^{(\jmath,\odot)} - \left(\prod_{\imath=0}^{\jmath-1} B^{(\imath)}\right)\mathbf{X} \tag{2.81}$$

die Ungleichungen

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \leq 2^{\lceil\log_2(\jmath)\rceil+1} \sum_{\imath=0}^{\jmath-1}\sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{11}^{(k,\imath)})^2 + (\sigma_{22}^{(k,\imath)})^2 + (\sigma_{12}^{(k,\imath)})^2 + (\sigma_{21}^{(k,\imath)})^2\right) u^2\ , \tag{2.82a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2^2\right) \leq 2^{\lceil\log_2(\jmath)\rceil+1} \sum_{\imath=0}^{\jmath-1}\sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{1,\odot}^{(k,\imath)})^2 + (\sigma_{2,\odot}^{(k,\imath)})^2\right) u^2\ . \tag{2.82b}$$

Im Fall $\mathrm{Cov}\left(\delta_{11}^{(k,\imath)}, \delta_{22}^{(k,\imath)}\right) = \mathrm{Cov}\left(\delta_{12}^{(k,\imath)}, \delta_{21}^{(k,\imath)}\right)$ *bzw.* $\mathbb{E}\left(\delta_{1,\odot}^{(k,\imath)}\delta_{2,\odot}^{(k,\imath)}\right) = 0$ *für alle* $k = 0, \ldots, \frac{n}{2} - 1$ *und alle* $\imath = 0, \ldots, \jmath - 1$ *verschärfen sich die Ungleichungen zu*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \leq 2^{\lceil \log_2(\jmath)\rceil} \sum_{\imath=0}^{\jmath-1} \sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{11}^{(k,\imath)})^2 + (\sigma_{22}^{(k,\imath)})^2 + (\sigma_{12}^{(k,\imath)})^2 + (\sigma_{21}^{(k,\imath)})^2\right) u^2 \ , \tag{2.83a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2^2\right) \leq 2^{\lceil \log_2(\jmath)\rceil} \sum_{\imath=0}^{\jmath-1} \sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{1,\odot}^{(k,\imath)})^2 + (\sigma_{2,\odot}^{(k,\imath)})^2\right) u^2 \ . \tag{2.83b}$$

Sind zusätzlich $\delta_{rs}^{(k,\imath)}, r, s = 1, 2,$ *bzw.* $\delta_{1,\odot}^{(k,\imath)}, \delta_{2,\odot}^{(k,\imath)}$ *für alle* $k = 0, \ldots, \frac{n}{2} - 1$ *und alle* $\imath = 0, \ldots, \jmath - 1$ *unabhängig identisch verteilt mit Varianz* $\sigma^2 u^2$ *bzw.* $\sigma_{\odot}^2 u^2$*, so werden die Ungleichungen*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) \leq \jmath \cdot 2^{\lceil \log_2(\jmath)\rceil + 1} \cdot n\sigma^2 u^2 \ , \tag{2.84a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2^2\right) \leq \jmath \cdot 2^{\lceil \log_2(\jmath)\rceil} \cdot n\sigma_{\odot}^2 u^2 \tag{2.84b}$$

erfüllt.

Beweis: Zunächst halten wir fest, dass sich der Fehlervektor $\boldsymbol{\Delta}$ bzw. $\boldsymbol{\Delta}^{\odot}$ aufgrund der Blockdiagonalgestalt (2.79) aller beteiligten Matrixfaktoren in $\frac{n}{2}$ voneinander getrennte Teilvektoren

$$\begin{pmatrix} X_{2k}^{(\jmath)} \\ X_{2k+1}^{(\jmath)} \end{pmatrix} - \left(\prod_{\imath=0}^{\jmath-1} A^{(k,\imath)}\right) \begin{pmatrix} X_{2k} \\ X_{2k+1} \end{pmatrix} \qquad \text{bzw.} \qquad \begin{pmatrix} X_{2k}^{(\jmath,\odot)} \\ X_{2k+1}^{(\jmath,\odot)} \end{pmatrix} - \left(\prod_{\imath=0}^{\jmath-1} A^{(k,\imath)}\right) \begin{pmatrix} X_{2k} \\ X_{2k+1} \end{pmatrix} \ ,$$

$k = 0, \ldots, \frac{n}{2} - 1$, aufsplitten lässt und dass wegen (2.80) die Größen

$$\Xi_k^{(\imath)} := \left(X_{2k}^{(\imath)}\right)^2 + \left(X_{2k+1}^{(\imath)}\right)^2 \qquad \text{bzw.} \qquad \Xi_k^{(\imath,\odot)} := \left(X_{2k}^{(\imath,\odot)}\right)^2 + \left(X_{2k+1}^{(\imath,\odot)}\right)^2 \tag{2.85}$$

für alle $k = 0, \ldots, \frac{n}{2} - 1$ und für alle $\imath = 1, \ldots, \jmath$ keine Werte außerhalb des Intervalls $[0, 1]$ annehmen. Demnach sind alle auftretenden Matrix-Vektor-Multiplikationen in Festkomma-Arithmetik ohne Überlauf ausführbar und die Vektoren $\mathbf{X}^{(\imath)}$ bzw. $\mathbf{X}^{(\imath,\odot)}$ für alle $\imath = 0, \ldots, \jmath$ wohldefiniert. Desweiteren ist für jedes $\imath = 0, \ldots, \jmath - 1$ Lemma 2.35 jeweils auf die Matrix $B^{(\imath)}$ und den Vektor $\mathbf{X}^{(\imath)}$ bzw. $\mathbf{X}^{(\imath,\odot)}$ anwendbar. Mit der Teleskopsumme

$$\boldsymbol{\Delta} = \sum_{\imath=0}^{\jmath-1} \left(\prod_{m=\imath+1}^{\jmath-1} B^{(m)}\right) \left(B^{(\imath)} \times \mathbf{X}^{(\imath)} - B^{(\imath)}\mathbf{X}^{(\imath)}\right) = \sum_{\imath=0}^{\jmath-1} \left(\prod_{m=\imath+1}^{\jmath-1} B^{(m)}\right) \mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}} \ ,$$

erhalten wir nach mehrmaliger Anwendung der Parallelogrammgleichung (2.45), Gleichung (2.80) sowie der Linearität und Monotonie des Erwartungswertes zunächst

$$\begin{aligned} \mathbb{E}\left(\|\boldsymbol{\Delta}\|_2^2\right) &= \mathbb{E}\left(\left\|\sum_{\imath=0}^{\jmath-1} \left(\prod_{m=\imath+1}^{\jmath-1} B^{(m)}\right) \mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}}\right\|_2^2\right) \leq 2^{\lceil \log_2(\jmath)\rceil} \sum_{\imath=0}^{\jmath-1} \mathbb{E}\left(\left\|\left(\prod_{m=\imath+1}^{\jmath-1} B^{(m)}\right) \mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}}\right\|_2^2\right) \\ &\leq 2^{\lceil \log_2(\jmath)\rceil} \sum_{\imath=0}^{\jmath-1} \mathbb{E}\left(\left\|\mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}}\right\|_2^2\right) \ , \end{aligned}$$

wobei im letzten Schritt verwendet wird, dass sich

$$\left\|\prod_{m=\imath+1}^{\jmath-1} B^{(m)}\right\|_2^2 = \left\|\prod_{m=\imath+1}^{\jmath-1} \left(\bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k,m)}\right)\right\|_2^2 \leq \prod_{m=\imath+1}^{\jmath-1} \left(\max_{k=0,\ldots,\frac{n}{2}-1} (1 - \delta_{A^{(k,\imath)}})\right) \leq 1 \tag{2.86}$$

aus Gleichung (2.80) ergibt. Analog folgt auch

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2^2\right) \leq 2^{\lceil \log_2(\jmath)\rceil} \sum_{\imath=0}^{\jmath-1} \mathbb{E}\left(\left\|\mathfrak{D}_{B^{(\imath)} \odot \mathbf{X}^{(\imath,\odot)}}\right\|_2^2\right) \ .$$

Die Behauptungen können nun aus Lemma 2.35 abgeleitet werden. ■

Lassen wir bei den Faktoren $B^{(\imath)}$ aus (2.81) auch Matrizen zu, welche durch Permutations- oder Vorzeichenskalierungsmatrizen auf die Gestalt (2.79) gebracht werden können, so benötigen wir zusätzliche Voraussetzungen. Für die Wohldefiniertheit der Zwischenergebnisse ist erforderlich, dass die Zufallsgrößen (2.85) nur Werte im Intervall $[0,1]$ annehmen.

Satz 2.38. *Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in* (2.49) *und* (2.50) *definiert. Sei $n \in \mathbb{N}$, $n \geq 4$ und gerade. Weiter sei $\mathbf{X} := (X_l)_{l=0}^{n-1}$ mit Zufallsvariablen X_l aus $\mathbb{M}_q$ gegeben, so dass $\Xi := \|\mathbf{X}\|_2^2$ nur Werte im Intervall $[0,1]$ annimmt. Für $\imath = 0, \ldots, \jmath - 1$ mit einem $\jmath \in \mathbb{N}$ seien $U^{(\imath)}, V^{(\imath)} \in \mathbb{M}_q^{n \times n}$ orthogonale Matrizen, so dass eine Multiplikation mit ihnen lediglich Permutationen sowie zeilen- bzw. spaltenweise Vorzeichenskalierungen bewirkt. Dann bleiben die Aussagen aus Satz* 2.37 *richtig, wenn anstelle der Blockdiagonalmatrizen $B^{(\imath)}$ aus* (2.79) *Faktoren der Gestalt $U^{(\imath)}B^{(\imath)}V^{(\imath)}$ auftreten.*

Beweis: Aufgrund der Orthogonalität der Matrizen $U^{(\imath)}, V^{(\imath)}$ und (2.80) ergibt sich zunächst

$$\left\| U^{(\imath)} B^{(\imath)} V^{(\imath)} \mathbf{X}^{(\imath)} \right\|_2^2 \;\leq\; \max_{k=0,\ldots,\frac{n}{2}-1} (1 - \delta_{A^{(k,\imath)}}) \|\mathbf{X}^{(\imath)}\|_2^2 \;\leq\; \|\mathbf{X}^{(\imath)}\|_2^2 \,. \tag{2.87}$$

Nach den Eigenschaften der Festkomma-Arithmetik folgen dann sowohl $\|\mathbf{X}^{(\imath+1)}\|_2^2 \leq \|\mathbf{X}^{(\imath)}\|_2^2$ als auch $\|\mathbf{X}^{(\imath+1,\odot)}\|_2^2 \leq \|\mathbf{X}^{(\imath,\odot)}\|_2^2$ für alle $\imath = 0, \ldots, \jmath - 1$. Insbesondere haben die Zufallsgrößen (2.85) nur Werte im Intervall $[0,1]$, da dies nach Voraussetzung für $\Xi = \|\mathbf{X}^{(0)}\|_2^2 = \|\mathbf{X}^{(0,\odot)}\|_2^2$ erfüllt ist. Demnach sind alle Zwischenergebnisse wohldefiniert. Auf die Behauptungen schließen wir nun analog zum Beweis von Satz 2.37 und mittels Folgerung 2.36. ■

Satz 2.39. *Unter den Voraussetzungen von Satz* 2.37 *oder* 2.38 *erfüllt der Vektor* (2.81) *der insgesamt auftretenden absoluten Rundungsfehler die Ungleichungen*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{2 \sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{11}^{(k,\imath)})^2 + (\sigma_{22}^{(k,\imath)})^2 + (\sigma_{12}^{(k,\imath)})^2 + (\sigma_{21}^{(k,\imath)})^2 \right)} \cdot u \,, \tag{2.88a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2\right) \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{2 \sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{1,\odot}^{(k,\imath)})^2 + (\sigma_{2,\odot}^{(k,\imath)})^2 \right)} \cdot u \,. \tag{2.88b}$$

Im Fall $\mathrm{Cov}\left(\delta_{11}^{(k,\imath)}, \delta_{22}^{(k,\imath)}\right) = \mathrm{Cov}\left(\delta_{12}^{(k,\imath)}, \delta_{21}^{(k,\imath)}\right)$ *bzw.* $\mathbb{E}\left(\delta_{1,\odot}^{(k,\imath)} \delta_{2,\odot}^{(k,\imath)}\right) = 0$ *für* $k = 0, \ldots, \frac{n}{2} - 1$ *und* $\imath = 0, \ldots, \jmath - 1$ *verschärfen sich die Ungleichungen zu*

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{\sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{11}^{(k,\imath)})^2 + (\sigma_{22}^{(k,\imath)})^2 + (\sigma_{12}^{(k,\imath)})^2 + (\sigma_{21}^{(k,\imath)})^2 \right)} \cdot u \,, \tag{2.89a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2\right) \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{\sum_{k=0}^{\frac{n}{2}-1} \left((\sigma_{1,\odot}^{(k,\imath)})^2 + (\sigma_{2,\odot}^{(k,\imath)})^2 \right)} \cdot u \,. \tag{2.89b}$$

Sind zusätzlich $\delta_{rs}^{(k)}$, $r,s = 1,2$, bzw. $\delta_{1,\odot}^{(k)}, \delta_{2,\odot}^{(k)}$ für alle $k = 0, \ldots, \frac{n}{2} - 1$ unabhängig identisch verteilt mit Varianz $\sigma^2 u^2$ bzw. $\sigma_{\odot}^2 u^2$, so gelten die Ungleichungen

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\leq\; \jmath\sqrt{2n} \cdot \sigma u \,, \tag{2.90a}$$

$$\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2\right) \;\leq\; \jmath\sqrt{n} \cdot \sigma_{\odot} u \,. \tag{2.90b}$$

Beweis: Wenden wir die Dreiecksungleichung der euklidischen Norm auf die Teleskopsumme aus dem Beweis von Satz 2.37 an und ziehen die Monotonie und Linearität des Erwartungswertes heran, erhalten wir zusammen mit der Jensen-Ungleichung (2.36) wie bereits in Folgerung 2.23 argumentiert zunächst

$$\mathbb{E}\left(\|\boldsymbol{\Delta}\|_2\right) \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{\mathbb{E}\left(\left\| \left(\prod_{m=\imath+1}^{\jmath-1} B^{(m)} \right) \mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}} \right\|_2^2 \right)} \;\leq\; \sum_{\imath=0}^{\jmath-1} \sqrt{\mathbb{E}\left(\left\| \mathfrak{D}_{B^{(\imath)} \times \mathbf{X}^{(\imath)}} \right\|_2^2 \right)} \,,$$

wobei im letzten Schritt wiederum (2.86) Berücksichtigung findet. Eine analoge Abschätzung ergibt sich mittels (2.87) für den Fall, dass Faktoren der Gestalt $U^{(\imath)}B^{(\imath)}V^{(\imath)}$ wie in Satz 2.38 vorhanden sind, bzw. jeweils auch für $\mathbb{E}\left(\|\boldsymbol{\Delta}^{\odot}\|_2\right)$. Die Behauptungen ergeben sich nun aus Lemma 2.35 bzw. analog zu Satz 2.38. ■

Bemerkung 2.40. In den Sätzen 2.37 bis 2.39 sind obere Schranken für den Erwartungswert der nichtnegativen Zufallsgrößen $\|\mathbf{\Delta}\|_2$ und $\|\mathbf{\Delta}^{\odot}\|_2$ sowie deren Quadrate angegeben. Für die Wahrscheinlichkeit, dass diese Zufallsgrößen vorgegebene positive Werte nicht überschreiten, liefert Folgerung A.27 einige Aussagen. Bezeichnen $\mu_{\mathbf{\Delta}}$ bzw. $\sigma^2_{\mathbf{\Delta}}$ den Erwartungswert bzw. die Varianz von $\|\mathbf{\Delta}\|_2$ und sind K_μ bzw. K_σ obere Schranken der ersten beiden Momente von $\|\mathbf{\Delta}\|_2$, so ergeben sich für ein $a > 0$ die Ungleichungen

$$P\Big(\|\mathbf{\Delta}\|_2^{-1}([0,a])\Big) \;\geq\; 1 - \frac{1}{a}\mathbb{E}\big(\|\mathbf{\Delta}\|_2\big) \;\geq\; 1 - \frac{K_\mu}{a}$$

aus der Markov-Ungleichung (A.38) und

$$P\Big(\|\mathbf{\Delta}\|_2^{-1}([0,a])\Big) \;\geq\; 1 - \frac{1}{a^2}\mathbb{E}\big(\|\mathbf{\Delta}\|_2^2\big) \;\geq\; 1 - \frac{K_\sigma}{a^2}$$

aus der Chebyshev-Markovschen Ungleichung (A.40) für $m = 2$. □

Bemerkung 2.41 (Zusammenfassung von Kapitel 2)**.** In diesem Kapitel haben wir die Eigenschaften der bei algebraischen Operationen auftretenden Rundungsfehler sowohl in Gleitkomma-Arithmetik als auch für Festkomma-Arithmetik untersucht.
Ausgehend vom Wilkinson-Modell (2.6) betrachtet Abschnitt 2.1 zunächst in den Folgerungen 2.6 und 2.7 die Summe von n Gleitkomma-Zahlen bzw. das Skalarprodukt von zwei Vektoren der Länge n für beliebiges $n \in \mathbb{N}$ unter Verwendung verschiedener Summationsweisen. Unter Berücksichtigung der Besetzungsstruktur einer Matrix liefern Lemma 2.8, Folgerung 2.9 und schließlich Lemma 2.10 Ergebnisse für Matrix-Vektor-Multiplikationen. Dabei spielt die komponentenweise Abschätzung (2.14) eine Schlüsselrolle. Unter Verwendung der in [43, Lemma 5.1] angegebenen Ungleichung (2.19) erhalten wir in Satz 2.12 das Hauptergebnis von Abschnitt 2.1.
Da die Abschätzungen aus Abschnitt 2.1 insbesondere den ungünstigsten Fall abdecken und nicht berücksichtigen, dass die Größen $\varepsilon^\bullet$ aus dem Wilkinson-Modell (2.6) in der Regel kleiner als die Rundungseinheit u aus (2.5) ausfallen, wird in Abschnitt 2.2 eine stochastische Rundungsfehleranalyse für die Gleitkomma-Arithmetik durchgeführt. Dazu betrachten wir sowohl Gleitkomma-Zahlen als auch auftretende Rundungsfehler als Zufallsgrößen und verwenden die Modellannahmen (2.20) und (2.21). Die in Unterabschnitt 2.2.1 gewonnenen Ergebnisse aus Lemma 2.14 und Lemma 2.16 können – obwohl sie bereits die stochastische Unabhängigkeit aller beteiligten Zufallsvariablen voraussetzen – nur durch weitere Forderungen in den Folgerungen 2.15 und 2.17 auf eine überschaubare Form gebracht werden. Somit sind die Voraussetzungen aus Satz 2.18 für eine Matrix-Vektor-Multiplikation mit einer vollbesetzten Matrix nur in Spezialfällen erfüllt. Als Alternative wird in Lemma 2.20 die ursprünglich für die Multiplikation komplexer Zahlen entwickelte Theorie aus [75, Proposition 2.3] unter Berücksichtigung der speziellen Struktur (2.27) aufgegriffen, welche unter geringfügiger Abwandlung der Voraussetzungen in Gestalt von Satz 2.22 die Basis für die weiteren Betrachtungen in Unterabschnitt 2.2.2 darstellt. Neu ist hierbei, dass auf die Unkorreliertheit der Eingangsdaten verzichtet werden kann. Mit den sinnvollen Annahmen (2.37) und (2.37a) bzw. (2.37b) werden in Lemma 2.24 und Lemma 2.25 die Hauptergebnisse aus Abschnitt 2.2 vorbereitet, welche in Satz 2.26 zusammengefasst sind.
Ausgehend von dem bereits auf J. von Neumann und H.H. Goldstine [41] zurückgehenden Modell (2.53) liefert Abschnitt 2.3 in den Folgerungen 2.29 und 2.30 Abschätzungen für den beim Skalarprodukt in einfach und doppelt genauer Festkomma-Arithmetik auftretenden absoluten Fehler. Die daraus resultierenden Ergebnisse für Matrix-Vektor-Multiplikationen mit Matrizen der Gestalt (2.59) lassen sich unter Berücksichtigung von (2.62) weiter zu (2.63) verschärfen. In Satz 2.31 steht schließlich das Hauptergebnis von Abschnitt 2.3.
Analog zur Gleitkomma-Arithmetik fassen wir in Abschnitt 2.4 alle auftretenden Größen als Zufallsvariablen auf. Zunächst werden in Lemma 2.32 und Folgerung 2.33 die Verteilung, der Erwartungswert und die Varianz des Eingangsfehlers bestimmt, falls die Eingangsgröße auf dem Intervall $[-1, 1]$ gleichverteilt ist. Anschließend liefern Lemma 2.34 und Lemma 2.35 Ungleichungen für den Erwartungswert des Normquadrates des Fehlervektors, welcher sowohl in einfach als auch in doppelt genauer Festkomma-Arithmetik bei einer Matrix-Vektor-Multiplikation mit einer Matrix der Gestalt (2.67) bzw. mit einer Blockdiagonalmatrix mit Blöcken der Gestalt (2.67) auftritt. Folgerung 2.36 berücksichtigt zusätzlich Matrizen, welche mittels Permutations- und Vorzeichenskalierungsmatrizen auf die vorher genannte Blockdiagonalgestalt überführt werden können. Die Hauptergebnisse aus Abschnitt 2.4 sind in den Sätzen 2.37, 2.38 und 2.39 zu finden. Dabei ist die Forderung, dass $\Xi := \|\mathbf{X}\|_2^2$ nur Werte im Intervall $[0, 1]$ annimmt, bei der Anwendung auf die Algorithmen 1.15 – 1.22 wesentlich. □

3 Numerische Stabilität in Gleitkomma-Arithmetik

Numerische Verfahren liefern meist anstelle der gesuchten, exakten Ergebnisse lediglich Näherungswerte. Im Wesentlichen lässt sich das damit begründen, dass sich nur endlich viele reelle Zahlen im Computer exakt darstellen lassen. Neben den beim Einlesen der zu verarbeitenden Daten auftretenden Eingangsfehlern entstehen noch zusätzliche Rundungsfehler, die aus der verwendeten Arithmetik resultieren.
Im Folgenden wollen wir die in Kapitel 1 hergeleiteten Algorithmen 1.15 – 1.22 bezüglich ihres Rundungsfehlerverhaltens bei Verwendung von Gleitkomma-Arithmetik untersuchen. Dafür benötigen wir zunächst allgemeine Kriterien, nach denen ein Algorithmus als numerisch stabil eingestuft wird. In dem Überblicksartikel [54] werden die Begriffe der normweisen Rückwärts- und Vorwärtsstabilität für den deterministischen Fall sowie der Begriff der durchschnittlichen Rückwärtsstabilität für den stochastischen Fall eingeführt. Für die Algorithmen 1.15 und 1.16 sind entsprechende Konstanten bereits bekannt (vgl. [54, Examples 8.10 und 8.11]), die jedoch noch von zusätzlichen Parametern abhängen, in welche die Genauigkeit der vorberechneten Matrix-Einträge eingehen. In Abschnitt 3.1 werden jetzt für alle Algorithmen aus Abschnitt 1.3 allgemeine Stabilitätskonstanten nur unter Hinzuziehen der Modellannahme (2.6) hergeleitet. Insbesondere werden bei den Einzelschrittabschätzungen jeweils die in Lemma 3.7 bewiesenen Ungleichungen (3.21) mit bestmöglichen Konstanten verwendet. Im Spezialfall einer nur skalierte Butterfly-Matrizen enthaltenden Blockdiagonalmatrix gewinnt ebenso die in Lemma 3.14 gefundene Ungleichung (3.40b) bei der Einzelschrittabschätzung an Bedeutung.
Da bei der Herleitung der Stabilitätskonstanten die Vorzeichen der einzelnen Rundungsfehler größtenteils keine Berücksichtigung finden und – obwohl in jedem Schritt die Einzelabschätzungen vielleicht scharf sind – vermutlich kein Eingangsvektor existiert, für den in jedem Schritt Gleichheit eintritt, sind die ermittelten Stabilitätskonstanten nicht selten um Einiges größer als notwendig. Um das Fehlerverhalten besser verstehen und damit vorhersagen zu können, bietet es sich an, zusätzlich einen stochastischen Zugang zum Rundungsfehler zu wählen. Erste auf einem einfachen stochastischen Modell basierende Analysen finden sich in [41, S. 1027 und S. 1036], welche die in [24, S. 48] genannte und bereits in [68, S. 31 ff] begründete „Faustregel“ bestätigt, die da lautet:

> Ist $f(n)u$ mit einer dimensionsabhängigen Konstante $f(n)$ eine obere Schranke, welche wir für den Rundungsfehler mit deterministischen Abschätzungen (worst case error) gewinnen konnten, so wird sich der durchschnittliche Fehler (average case error) in der Größenordnung $\sqrt{f(n)}u$ bewegen. Begründet werden kann dies unter der Annahme, dass alle Rundungsfehler unabhängige Zufallsvariablen sind, und unter Heranziehen des zentralen Grenzwertsatzes.

In Anlehnung daran führen wir in Abschnitt 3.2 eine Analyse des durchschnittlichen Rundungsfehlers durch, welche auf dem in Abschnitt 2.2 vorgestellten stochastischen Modell der Gleitkomma-Arithmetik und insbesondere auf den in Unterabschnitt 2.2.2 hergeleiteten Ergebnissen beruht. Insbesondere ist hervorzuheben, dass wir an die Verteilung des Eingangsvektors keine speziellen Forderungen stellen. Dafür erhalten wir jedoch im Gegensatz zu [54, Theorem 8.4], in dem alle Komponenten des Eingangsvektors als unkorreliert und mit gleicher Varianz sowie mit Erwartungswert Null angenommen werden, keine explizite Schätzung für den Erwartungswert des Normquadrates des absoluten Fehlers in Form einer Gleichung, sondern nur eine obere Abschätzung.

3.1 Deterministische Rundungsfehleranalyse

Die hier betrachteten Algorithmen zu speziellen diskreten trigonometrischen Transformationen liefern stets eine Approximation $\tilde{\mathbf{y}}$ zu einer Matrix-Vektor-Multiplikation $\mathbf{y} := A\mathbf{x}$ einer speziellen Matrix $A \in R_{\mathbb{G}}^{n\times n} \subset \mathbb{R}^{n\times n}$ mit einem Vektor $\mathbf{x} \in R_{\mathbb{G}}^{n} \subset \mathbb{R}^{n}$. Das erhaltene Ergebnis $\tilde{\mathbf{y}}$ kann nun als exaktes Ergebnis der Transformation eines gestörten Vektors $\tilde{\mathbf{x}} := \mathbf{x} + \Delta\mathbf{x}$ aufgefasst werden (vgl. Abb. 3.1).

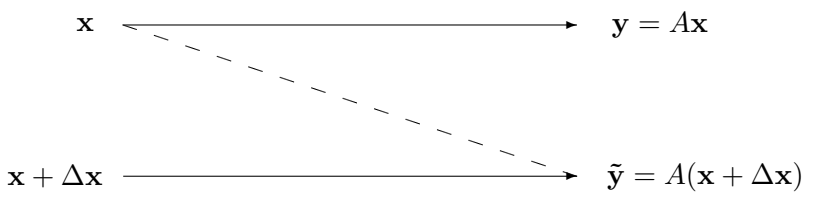


Abbildung 3.1: Modellierung des Vorwärts- und Rückwärtsfehlers.

In Anlehnung daran bezeichnen wir die Differenz $\tilde{\mathbf{y}} - \mathbf{y}$ als *Vorwärtsfehler* und $\Delta\mathbf{x}$ als *Rückwärtsfehler* der Matrix-Vektor-Multiplikation $\mathbf{y} = A\mathbf{x}$. Besitzen diese Fehler dieselbe Größenordnung wie der Eingangsfehler, so sprechen wir von einem *stabilen* Algorithmus, genauer:

Definition 3.1 (vgl. [54], Abschnitt 8.1)**.** Ein in Gleitkomma-Arithmetik implementierter Algorithmus für eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ ($\mathbf{x} \in R_{\mathbb{G}}^n$) heißt *normweise rückwärtsstabil*, falls eine Konstante $k_n > 0$ mit $k_n u \ll 1$ existiert, so dass für alle $\mathbf{x} \in R_{\mathbb{G}}^n$ mit $A\mathbf{x} \in R_{\mathbb{G}}^n$ die Abschätzung

$$\| \Delta\mathbf{x} \|_2 \;\leq\; \left(k_n u + \mathcal{O}\left(u^2\right)\right) \| \mathbf{x} \|_2 \tag{3.1}$$

erfüllt wird. Bezeichnet $\mathbf{y}$ das exakte Ergebnis von $A\mathbf{x}$ und $\tilde{\mathbf{y}}$ den tatsächlich berechneten Vektor, so heißt ein in Gleitkomma-Arithmetik implementierter Algorithmus zur Berechnung von $A\mathbf{x}$ ($\mathbf{x} \in R_{\mathbb{G}}^n$) *normweise vorwärtsstabil*, falls eine Konstante $k_n > 0$ mit $k_n u \ll 1$ existiert, so dass für alle $\mathbf{x} \in R_{\mathbb{G}}^n$ mit $A\mathbf{x} \in R_{\mathbb{G}}^n$ die Abschätzung

$$\| \tilde{\mathbf{y}} - \mathbf{y} \|_2 \;\leq\; \left(k_n u + \mathcal{O}\left(u^2\right)\right) \| \mathbf{x} \|_2 \tag{3.2}$$

erfüllt wird.

Die in Definition 3.1 vorkommende Konstante k_n für die Rückwärtsstabilität (3.1) muss im Allgemeinen nicht mit der Konstanten k_n für die Vorwärtsstabilität (3.2) übereinstimmen. Im Fall einer orthogonalen Matrix A gilt wegen $(A\mathbf{x})^{\mathrm{T}}(A\mathbf{x}) = \mathbf{x}^{\mathrm{T}}(A^{\mathrm{T}}A)\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{x}$ jedoch

$$\|\tilde{\mathbf{y}} - \mathbf{y}\|_2 \;=\; \|A(\Delta\mathbf{x})\|_2 \;=\; \|\Delta\mathbf{x}\|_2 \;. \tag{3.3}$$

Durch die Bedingung $A\mathbf{x} \in R_{\mathbb{G}}^n$ sollen Über- und Unterlauf explizit ausgeschlossen werden. Zu beachten ist, dass bei der Definition 3.1 die Komponenten des Eingangsvektors keine Gleitkomma-Zahlen sind. Würde bereits $\mathbf{x} \in \mathbb{G}^n$ vorausgesetzt, dann ließen sich gegebenenfalls kleinere Stabilitätskonstanten k_n bestimmen.
Mit den in (2.12) eingeführten Bezeichnungen und mit Hilfe der im Abschnitt 2.1 hergeleiteten Aussagen erhalten wir sofort Abschätzungen für das Skalarprodukt von zwei beliebigen Vektoren der Länge n und ebenso für eine Matrix-Vektor-Multiplikation mit einer beliebigen $(n \times n)$-Matrix.

Lemma 3.2. *Zu gegebener Matrix $A \in R_{\mathbb{G}}^{n\times n}$ und gegebenem Vektor $\mathbf{x} \in R_{\mathbb{G}}^n$ seien $\hat{A} = \mathrm{fl}(A)$ und $\hat{\mathbf{x}} = \mathrm{fl}(\mathbf{x})$ die entsprechenden Approximationen innerhalb der in Abschnitt* 2.1 *definierten Gleitkomma-Arithmetik. Weiter erfülle jede Zeile $\hat{\mathbf{y}}^{\mathrm{T}}$ von $\hat{A}$ die Bedingung*

$$|\hat{\mathbf{y}}|^{\mathrm{T}}|\hat{\mathbf{x}}| \in R_{\mathbb{G}} \;. \tag{3.4}$$

Dann gelten folgende Aussagen:

(i) *Ist $\|\hat{\mathbf{x}}\|_1 \in R_{\mathbb{G}}$ erfüllt, dann gilt*

$$\left|\mathrm{fl}(\hat{\mathbf{x}}^{\mathrm{T}}\mathbf{1}) - \mathbf{x}^{\mathrm{T}}\mathbf{1}\right| \;\leq\; \begin{cases} n\|\mathbf{x}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei sequentieller Summation,} \\ (\lceil \log_2 n \rceil + 1)\, \|\mathbf{x}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei Kaskaden-Summation.} \end{cases} \tag{3.5}$$

(ii) *Für das Skalarprodukt einer beliebigen Zeile $\mathbf{y}^{\mathrm{T}}$ von A mit $\mathbf{x}$ gilt*

$$\left|\mathrm{fl}(\hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{y}}) - \mathbf{x}^{\mathrm{T}}\mathbf{y}\right| \;\leq\; \begin{cases} (n+2)\|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei sequentieller Summ.,} \\ (\lceil \log_2 n \rceil + 3)\, \|\mathbf{x} \circ \mathbf{y}\|_1 \left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei Kaskaden-Summation.} \end{cases} \tag{3.6}$$

Im Fall $\mathbf{x} = \hat{\mathbf{x}}$ *ergeben sich die Abschätzungen*

$$\left|\mathrm{fl}(\hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{y}}) - \mathbf{x}^{\mathrm{T}}\mathbf{y}\right| \;\leq\; \begin{cases} (n+1)\|\mathbf{x}\circ\mathbf{y}\|_1\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei sequentieller Summ.,} \\ (\lceil\log_2 n\rceil + 2)\,\|\mathbf{x}\circ\mathbf{y}\|_1\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei Kaskaden-Summation.} \end{cases} \tag{3.7}$$

(iii) *Bezeichnet* $\rho < n$ *die Anzahl der Nichtnullelemente einer Zeile* $\mathbf{y}^{\mathrm{T}}$ *von* A*, so gelten die Abschätzungen aus* (3.6) *und* (3.7) *in diesem Fall mit* ρ *anstelle von* n*.*

(iv) *Enthält jede Zeile* $\mathbf{y}^{\mathrm{T}}$ *von* A *höchstens* ρ *Nichtnullelemente, so ergeben sich bei der Matrix-Vektor-Multiplikation* $A\mathbf{x}$ *die komponentenweisen Abschätzungen*

$$\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right| \;\leq\; \begin{cases} (\rho+2)\,|A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei sequentieller Summation,} \\ (\lceil\log_2 \rho\rceil + 3)\,|A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei Kaskaden-Summation.} \end{cases} \tag{3.8}$$

Im Fall $\mathbf{x} = \hat{\mathbf{x}}$ *ergeben sich die Abschätzungen*

$$\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right| \;\leq\; \begin{cases} (\rho+1)\,|A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei sequentieller Summation,} \\ (\lceil\log_2 \rho\rceil + 2)\,|A||\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right) & \textit{bei Kaskaden-Summation.} \end{cases} \tag{3.9}$$

Beweis: (i) Die Behauptung ergibt sich direkt aus Folgerung 2.6 angewandt auf $\hat{\mathbf{x}}$, mit Dreiecksungleichung und der Tatsache, dass nach Modellannahme (2.6) die komponentenweise Ungleichung $|\hat{\mathbf{x}} - \mathbf{x}| \leq |\mathbf{x}|u$ und damit $\|\hat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1\,(1 + \mathcal{O}(u))$ gilt.

(ii) Nach Modellannahme (2.6) existieren Vektoren $\boldsymbol{\varepsilon}_{\mathbf{x}} \in \mathbb{R}^n$ bzw. $\boldsymbol{\varepsilon}_{\mathbf{y}} \in \mathbb{R}^n$ mit $\hat{\mathbf{x}} = (\mathbf{1} + \boldsymbol{\varepsilon}_{\mathbf{x}}) \circ \mathbf{x}$ und $\|\boldsymbol{\varepsilon}_{\mathbf{x}}\|_\infty \leq u$ bzw. $\hat{\mathbf{y}} = (\mathbf{1} + \boldsymbol{\varepsilon}_{\mathbf{y}}) \circ \mathbf{y}$ und $\|\boldsymbol{\varepsilon}_{\mathbf{y}}\|_\infty \leq u$. Demnach lässt sich die Differenz

$$\begin{aligned} \hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{y}} - \mathbf{x}^{\mathrm{T}}\mathbf{y} &= \left((\mathbf{1} + \boldsymbol{\varepsilon}_{\mathbf{x}}) \circ \mathbf{x}\right)^{\mathrm{T}}\left((\mathbf{1} + \boldsymbol{\varepsilon}_{\mathbf{y}}) \circ \mathbf{y}\right) - \mathbf{x}^{\mathrm{T}}\mathbf{y} \\ &= \left(\boldsymbol{\varepsilon}_{\mathbf{x}} \circ \mathbf{x}\right)^{\mathrm{T}}\mathbf{y} \;+\; \mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\varepsilon}_{\mathbf{y}} \circ \mathbf{y}\right) \;+\; \left(\boldsymbol{\varepsilon}_{\mathbf{x}} \circ \mathbf{x}\right)^{\mathrm{T}}\left(\boldsymbol{\varepsilon}_{\mathbf{y}} \circ \mathbf{y}\right) \end{aligned}$$

betragsmäßig durch $(2u + u^2)\|\mathbf{x} \circ \mathbf{y}\|_1$ abschätzen. Mit der komponentenweisen Ungleichungskette

$$|\hat{\mathbf{x}} \circ \hat{\mathbf{y}}| \;\leq\; |\mathbf{x} \circ \mathbf{y}| + |\hat{\mathbf{x}} \circ \hat{\mathbf{y}} - \mathbf{x} \circ \mathbf{y}| \;\leq\; (1 + 2u + u^2)|\mathbf{x} \circ \mathbf{y}|$$

folgt ebenso $\|\hat{\mathbf{x}} \circ \hat{\mathbf{y}}\|_1 \leq \|\mathbf{x} \circ \mathbf{y}\|_1\,(1 + \mathcal{O}(u))$. Die Ungleichungen (3.6) ergeben sich nun mittels Dreiecksungleichung und Folgerung 2.7 angewandt auf $\hat{\mathbf{x}}$ und $\hat{\mathbf{y}}$. Gilt zusätzlich $\mathbf{x} = \hat{\mathbf{x}}$, so lässt sich die Differenz $\hat{\mathbf{x}}^{\mathrm{T}}\hat{\mathbf{y}} - \mathbf{x}^{\mathrm{T}}\mathbf{y}$ aufgrund des Wegfalls der $\boldsymbol{\varepsilon}_{\mathbf{x}}$-Terme betragsmäßig durch $u\|\mathbf{x} \circ \mathbf{y}\|_1$ abschätzen. Die Gültigkeit von (3.7) folgt dann analog.

(iii) Die Behauptung ergibt sich gemäß der Bemerkung im Anschluss an Folgerung 2.7.

(iv) Bezeichnet $\mathbf{a}_j^{\mathrm{T}}$ für $j = 0, \ldots, n-1$ die j-te Zeile von A und entsprechend $\hat{\mathbf{a}}_j^{\mathrm{T}}$ die j-te Zeile von $\hat{A}$, so ist $\mathrm{fl}(\hat{\mathbf{a}}_j^{\mathrm{T}}\hat{\mathbf{x}}) - \mathbf{a}_j^{\mathrm{T}}\mathbf{x}$ die j-te Komponente von $\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}$. Beachten wir noch, dass $\|\mathbf{a}_j \circ \mathbf{x}\|_1 = |\mathbf{a}_j|^{\mathrm{T}}|\mathbf{x}|$ der j-ten Komponente von $|A||\mathbf{x}|$ entspricht, ergeben sich die Behauptungen aus (iii). ■

In [54, Lemma 8.4] sind ebenfalls Abschätzungen für $\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right|$ in komplexwertiger Arithmetik angegeben, wobei es sich genau genommen um den Fall $\mathbf{x} = \hat{\mathbf{x}}$ handelt. Zu beachten ist, dass dort nach Voraussetzung der absolute Rundungsfehler der Elemente von $\hat{A}$ durch ηu (vgl. [54, (8.40)]) für ein festes $\eta > 0$ gleichmäßig beschränkt ist. Ersetzen wir in den Abschätzungen aus [54, Lemma 8.4] den Term $\eta\tilde{A}|\mathbf{x}|\left(u + \mathcal{O}(u^2)\right)$ durch $|A||\mathbf{x}|\left(u + \mathcal{O}(u^2)\right)$ und setzen $\mu_{\mathbb{C}} = 1$, dann stimmen diese Abschätzungen mit (3.9) überein.

Bemerkung 3.3. Wird auf die komponentenweise Ungleichung

$$|\mathbf{y}| \;\leq\; \alpha|A||\mathbf{x}| \qquad\qquad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^n,\; A \in \mathbb{R}^{n\times n},\; \alpha > 0)$$

zeilenweise die Cauchy-Schwarz-Ungleichung angewandt und quadrieren sowie summieren wir anschließend alle so entstandenen Ungleichungen, ergibt sich offenbar

$$\|\mathbf{y}\|_2^2 \;\leq\; \alpha^2\|A\|_F^2\|\mathbf{x}\|_2^2$$

mit der Frobeniusnorm

$$\|A\|_F := \sqrt{\sum_{j,k=0}^{n-1} a_{jk}^2} \, . \tag{3.10}$$

Für eine vollbesetzte Matrix wie etwa eine der Kosinus- und Sinusmatrizen aus Abschnitt 1.1 folgen aus Lemma 3.2 (iv) somit die Abschätzungen

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; \begin{cases} (n-1+\beta)\,\|A\|_F \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{bei sequentieller Summation,} \\ (\lceil \log_2 n \rceil + \beta)\,\|A\|_F \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{bei Kaskaden-Summation,} \end{cases} \tag{3.11}$$

wobei $\beta = 2$ im Fall $\mathbf{x} = \hat{\mathbf{x}}$ und $\beta = 3$ sonst. Stimmt nun A mit einer der orthogonalen Matrizen aus Definition 1.1 überein, so ergibt sich für die Frobeniusnorm $\|A\|_F = \sqrt{n}$. Aus (3.11) erhalten wir dann gemäß Definition 3.1 die Stabilitätskonstanten $k_n = \sqrt{n}(n-1+\beta)$ bei sequentieller Summation und $k_n = \sqrt{n}\,(\lceil \log_2 n \rceil + \beta)$ bei Kaskaden-Summation. Tabelle 3.1 listet für ausgewählte n die expliziten Werte der Stabilitätskonstanten auf, welche sich im Fall $\mathbf{x} = \hat{\mathbf{x}}$ nach (3.11) bei direkter Multiplikation mit einer der vollbesetzten Matrizen aus Definition 1.1 ergeben. □

n	8	16	32	64	128	256	512
k_n bei seq. Sum.	25.4558	68.0000	186.6762	520.0000	1459.4684	4112.0000	11607.8649
k_n bei Kas.-Sum.	14.1421	24.0000	39.5980	64.0000	101.8234	160.0000	248.9016

Tabelle 3.1: Stabilitätskonstanten im Fall $\mathbf{x} = \hat{\mathbf{x}}$ für „naiv implementierte“ DCT- und DST-Algorithmen gemäß Definition 3.1. Selbst im günstigeren Fall der Kaskaden-Summation verzeichnen wir einen enormen Zuwachs bei den k_n. Ein Vergleich mit den am Ende dieses Abschnittes angegebenen Werten in Tabelle 3.2 offenbart, dass die schnellen Algorithmen 1.15 – 1.22 nicht nur einen geringeren Rechenaufwand verursachen, sondern spätestens ab $n = 64$ als wesentlich stabiler anzusehen sind.

Ein schneller Algorithmus für eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ ($\mathbf{x} \in \mathbb{R}^n$) mit vollbesetzter Matrix $A \in \mathbb{R}^{n\times n}$ basiert üblicherweise auf einer Faktorisierung von A der Gestalt

$$A \;=\; \prod_{m=1}^{\nu} A^{(m)} \;:=\; A^{(\nu)} \ldots A^{(2)} A^{(1)} \tag{3.12}$$

mit Matrizen $A^{(m)} \in \mathbb{R}^{n\times n}$, $m = 1, \ldots, \nu$, welche im Idealfall dünnbesetzt sind. Daher sind wir insbesondere an Abschätzungen für den Rundungsfehler interessiert, der bei einer Matrix-Vektor-Multiplikation $A\mathbf{x}$ ($\mathbf{x} \in \mathbb{R}^n$) mit einer dünnbesetzten Matrix $A \in \mathbb{R}^{n\times n}$ auftritt.

Lemma 3.4. *Zu gegebener Matrix $A \in R_{\mathbb{G}}^{n\times n}$ und gegebenem Vektor $\mathbf{x} \in R_{\mathbb{G}}^n$ seien $\hat{A} = \mathrm{fl}(A)$ und $\hat{\mathbf{x}} = \mathrm{fl}(\mathbf{x})$ die entsprechenden Approximationen innerhalb der in Abschnitt 2.1 definierten Gleitkomma-Arithmetik. Weiter erfülle jede Zeile $\hat{\mathbf{y}}^{\mathrm{T}}$ von $\hat{A}$ die Bedingung (3.4). Enthält jede Zeile $\mathbf{y}^{\mathrm{T}}$ von A höchstens zwei Nichtnullelemente und ist $B = (b_{j,k})_{j,k=0}^{n-1}$ durch*

$$b_{j,k} \;:=\; \begin{cases} 1 \;, & \text{falls } a_{j,k} \notin \{0, \pm 1\} \;, \\ 0 \;, & \text{falls } a_{j,k} \in \{0, \pm 1\} \end{cases}$$

definiert, so ergibt sich bei der Matrix-Vektor-Multiplikation die komponentenweise Abschätzung

$$\left| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right| \;\leq\; |A\mathbf{x}|u + \Big(2|A| + B \circ |A|\Big)|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right). \tag{3.13}$$

Im Fall $\mathbf{x} = \hat{\mathbf{x}}$ gilt

$$\left| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right| \;\leq\; |A\mathbf{x}|u \;+\; \left(|A| + B \circ |A|\right)|\mathbf{x}| \left(u \;+\; \mathcal{O}(u^2)\right) \;. \tag{3.14}$$

Beweis: Zunächst halten wir fest, dass wegen $|\hat{A} - A| \leq B \circ |A| u$ insbesondere

$$|\hat{A}| \leq |A|(1+u) \tag{3.15}$$

gilt. Analog folgt $|\hat{\mathbf{x}}| \leq |\mathbf{x}|(1+u)$ aus $|\hat{\mathbf{x}} - \mathbf{x}| \leq u|\mathbf{x}|$. Wenden wir Folgerung 2.9 auf $\hat{A}$ und $\hat{\mathbf{x}}$ an, erhalten wir mit Dreiecksungleichung

$$\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\hat{\mathbf{x}}\right| \;\leq\; |\hat{A}\hat{\mathbf{x}}|u \;+\; |\hat{A}||\hat{\mathbf{x}}|\left(u \;+\; \mathcal{O}(u^2)\right) \;+\; |(\hat{A} - A)\hat{\mathbf{x}}| \,.$$

Wegen $|\hat{A}\hat{\mathbf{x}}| \;\leq\; |A\hat{\mathbf{x}}| + (B \circ |A|)\,|\hat{\mathbf{x}}|u$ und $|(\hat{A} - A)\hat{\mathbf{x}}| \leq (B \circ |A|)\,|\hat{\mathbf{x}}|u$ folgt dann weiter

$$\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\hat{\mathbf{x}}\right| \;\leq\; |A\hat{\mathbf{x}}|u + |\hat{A}||\hat{\mathbf{x}}|\left(u \;+\; \mathcal{O}(u^2)\right) \;+\; (B \circ |A|)\,|\hat{\mathbf{x}}|\left(u \;+\; \mathcal{O}(u^2)\right).$$

Für $\mathbf{x} = \hat{\mathbf{x}}$ ergibt sich unter Beachtung von (3.15) die Abschätzung (3.14). Andernfalls folgt mit Hilfe der Vorbemerkung, mit (2.8) und wegen $|A\hat{\mathbf{x}}| \leq |A\mathbf{x}| \;+\; |A||\mathbf{x}|u$ über die Dreiecksungleichung

$$\left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right| \;\leq\; \left|\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\hat{\mathbf{x}}\right| \;+\; |A||\mathbf{x}|u \;\leq\; |A\mathbf{x}|u + \Big(2|A| + B \circ |A|\Big)|\mathbf{x}|\left(u \;+\; \mathcal{O}(u^2)\right). \qquad \blacksquare$$

Ersetzen wir in [54, Lemma 8.4] den Ausdruck $\eta\tilde{A}|\mathbf{x}|$ durch $(B \circ |A|)\,|\mathbf{x}|u$ und verwenden wir aufgrund der reellen Arithmetik $\mu_{\mathbb{R}} = 1$ anstelle von $\mu_{\mathbb{C}} = \frac{4}{3}\sqrt{3}$, so stimmt die dort angegebene Abschätzung für Matrizen mit maximal zwei Nichtnulleinträgen im Fall $\mathbf{x} = \hat{\mathbf{x}}$ mit (3.14) überein. Im Unterschied zu unserer Abschätzung sind jedoch zusätzliche Informationen in Bezug auf A bzw. $\hat{A}$ notwendig, da für η eine gleichmäßige obere Schranke für den Betrag aller Einträge von $A - \hat{A}$ bekannt sein muss. Dies sind aber genau die absoluten Fehler und nicht die dem Wilkinson-Modell (2.6) zugrunde liegenden relativen Fehler.
Wir betrachten nun den Spezialfall von Blockdiagonalmatrizen mit Blöcken der Gestalt (1.24).

3.1.1 Stabilitätskonstanten unter Einbeziehung spezieller Blockdiagonalgestalt

Mit Hilfe der in (2.19) angegebenen optimalen Ungleichung und Satz 2.12 aus Abschnitt 2.1 können wir nun nachstehende Fehlerabschätzungen gewinnen.

Folgerung 3.5. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n\times n}$ eine direkte Summe von Drehmatrizen der Gestalt* (1.24) *und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin seien $\hat{\mathbf{x}} := \mathrm{fl}(\mathbf{x})$ und $\hat{A} := \mathrm{fl}(A)$ die entsprechenden Approximationen innerhalb der in Abschnitt* 2.1 *definierten Gleitkomma-Arithmetik. Falls $|\hat{A}||\hat{\mathbf{x}}| \in R_{\mathbb{G}}^{n}$ erfüllt ist, gilt*

$$\left\|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right\|_2 \;\leq\; \left(\tfrac{4}{3}\sqrt{3} + 2\sqrt{2}\right)\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right). \tag{3.16}$$

Ist zusätzlich noch $\mathbf{x} = \hat{\mathbf{x}}$ erfüllt, gilt sogar

$$\left\|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right\|_2 \;\leq\; \left(\tfrac{4}{3}\sqrt{3} + \sqrt{2}\right)\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right). \tag{3.17}$$

Beweis: Analog zum Beweis von Lemma 2.10 folgt nun aus der komponentenweisen Abschätzung (3.13) wiederum mit Dreiecksungleichung, mit Satz 2.12 und unter Berücksichtigung von $B \circ |A| \leq |A|$ die Ungleichungskette

$$\begin{aligned}\left\|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right\|_2 \;&\leq\; \Big(\big\||A\mathbf{x}|\;+\;|A||\mathbf{x}|\big\|_2 \;+\; \big\|2|A||\mathbf{x}|\big\|_2\Big)\left(u \;+\; \mathcal{O}(u^2)\right)\\ &\leq\; \Big(\tfrac{4}{3}\sqrt{3}\|A\|_2 \;+\; 2\,\big\||A|\big\|_2\Big)\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right)\;,\end{aligned}$$

wobei die Gültigkeit von $\big\||\mathbf{x}|\big\|_2 = \|\mathbf{x}\|_2$ sowie $\|A\mathbf{x}\|_2 \leq \|A\|_2\|\mathbf{x}\|_2$ verwendet worden ist. Da nun eine ähnliche Argumentation wie nach Folgerung 2.9 unter Beachtung der Orthogonalität von Drehmatrizen die Ungleichung

$$\big\||A|\big\|_2 \;\leq\; \sqrt{2}\|A\|_2 \;=\; \sqrt{2} \tag{3.18}$$

liefert, folgt die Behauptung in (3.16). Die Gültigkeit von (3.17) ergibt sich analog mit (3.14). $\blacksquare$

Bemerkung 3.6. (i) Mit der Abschätzung aus Lemma 2.10 hätten wir für den Fall $\mathbf{x} = \hat{\mathbf{x}}$ die etwas größere Konstante $1 + 2\sqrt{2}$ erhalten.
(ii) In [54, Example 8.1] wird die Konstante $1 + \sqrt{2} + 2\eta$ angegeben. Dies liegt daran, dass in [54, Theorem 8.3] die komponentenweise Ungleichung $|A - \hat{A}| \leq B\eta u$ anstelle von $|A - \hat{A}| \leq B \circ |A|u$ vorausgesetzt ist und in diesem Fall

$$\|B\|_2 \;=\; \left\| \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right\|_2 \;=\; 2 \;,$$

also

$$\left\| |A - \hat{A}| \right\|_2 \;\leq\; 2\eta u \tag{3.19}$$

gilt, wohingegen in unserem Fall $\left\| |A| \right\|_2 \leq \sqrt{2}$ bei der Herleitung von (3.17) verwendet wird. Ziehen wir – wiederum nur für den Fall $\mathbf{x} = \hat{\mathbf{x}}$ – ebenfalls die zusätzliche Information heran, dass ηu eine gleichmäßige obere Schranke des jeweiligen Betrages der Einträge von $A - \hat{A}$ ist und somit die Spektralnorm der fast orthogonalen Matrix $A - \hat{A}$ durch $\sqrt{2}\eta u$ beschränkt bleibt, dann erhalten wir wiederum mit Satz 2.12 die Konstante

$$\tfrac{4}{3}\sqrt{3} + \sqrt{2}\eta \;. \tag{3.20}$$

(iii) Die Konstante (3.20) ergibt sich als Spezialfall von [43, Lemma 5.2 (iii)] mit $b = 1$ und $c_2 = \eta$.
(iv) Im bestmöglichen Fall, also bei direktem Aufruf aller Einträge von A bzw. $\hat{A}$, ist $\eta = \frac{1}{2}$ und somit $\frac{1}{2}\sqrt{2} + \frac{4}{3}\sqrt{3} < 2 + \sqrt{2}$ als Konstante zu gewinnen. □

Die Abschätzungen aus Folgerung 3.5 lassen sich jedoch noch weiter verbessern, indem zu (2.19) ähnlich optimale Ungleichungen gefunden werden.

Lemma 3.7. *Seien $a, b, c, d \in \mathbb{R}$ beliebig. Dann gelten die Ungleichungen*

$$\left.\begin{aligned} (3|ac| + 3|bd| + |ac - bd|)^2 + (3|ad| + 3|bc| + |ad + bc|)^2 &\leq \tfrac{128}{5}(a^2 + b^2)(c^2 + d^2) \;, \\ (2|ac| + 2|bd| + |ac - bd|)^2 + (2|ad| + 2|bc| + |ad + bc|)^2 &\leq \tfrac{27}{2}(a^2 + b^2)(c^2 + d^2) \;. \end{aligned}\right\} \tag{3.21}$$

Dabei sind die Konstanten jeweils optimal.
Beweis: O.B.d.A. können wir annehmen, dass $a, b, c, d \geq 0$ gilt. Weiterhin können wir annehmen, dass $ac \geq bd$ erfüllt ist, ansonsten benennen wir das Quadrupel (a, b, c, d) zu (b, a, d, c) um. Damit vereinfachen sich die Behauptungen zu

$$\begin{aligned} \left(ac + \tfrac{bd}{2}\right)^2 + (ad + bc)^2 &\leq \tfrac{8}{5}(a^2 + b^2)(c^2 + d^2) \;, \\ \left(ac + \tfrac{bd}{3}\right)^2 + (ad + bc)^2 &\leq \tfrac{3}{2}(a^2 + b^2)(c^2 + d^2) \;, \end{aligned}$$

welche zu den beiden Ungleichungen

$$\begin{aligned} 0 &\leq \left(ac - \tfrac{3bd}{2}\right)^2 + (ad - bc)^2 \;, \\ 0 &\leq \left(ac - \tfrac{5bd}{3}\right)^2 + (ad - bc)^2 \end{aligned}$$

äquivalent sind. Diese sind jedoch offensichtlich erfüllt. Da die erste Ungleichung beispielsweise für

$$a = c = 1, \quad b = d = \sqrt{\tfrac{2}{3}}$$

zu einer Gleichung wird und ebenso die zweite Ungleichung für

$$a = c = 1, \quad b = d = \sqrt{\tfrac{3}{5}} \;,$$

sind die Konstanten $\frac{128}{5}$ und $\frac{27}{2}$ bestmöglich. ■

Indem wir nun analog zu Satz 2.12 die Ungleichungen (3.21) anwenden, erhalten wir verbesserte Abschätzungen für die euklidische Norm des Fehlervektors.

Folgerung 3.8. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n\times n}$ eine direkte Summe von Drehmatrizen der Form* (1.24) *und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin seien $\hat{\mathbf{x}} := \mathrm{fl}(\mathbf{x})$ und $\hat{A} := \mathrm{fl}(A)$ die entsprechenden Approximationen innerhalb der in Abschnitt* 2.1 *definierten Gleitkomma-Arithmetik. Falls $|\hat{A}||\hat{\mathbf{x}}| \in R_{\mathbb{G}}^{n}$ erfüllt ist, gilt*

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; 8\sqrt{\tfrac{2}{5}}\, \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right). \tag{3.22}$$

Ist zusätzlich noch $\mathbf{x} = \hat{\mathbf{x}}$ erfüllt, gilt sogar

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; \tfrac{3}{2}\sqrt{6}\, \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right). \tag{3.23}$$

Beweis: Aus der komponentenweisen Abschätzung (3.13) und unter Berücksichtigung von $B \circ |A| \leq |A|$ erhalten wir zunächst

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; \left(\left\| |A\mathbf{x}| \;+\; 3|A||\mathbf{x}| \right\|_2 \right) \left(u \;+\; \mathcal{O}(u^2)\right) .$$

Aufgrund der Blockstruktur von A und der besonderen Form (1.24) der einzelnen Blöcke können wir jeweils die erste Ungleichung aus (3.21) anwenden und erhalten

$$\begin{aligned}
\left\| |A\mathbf{x}| \;+\; 3|A||\mathbf{x}| \right\|_2^2 \;&=\; \sum_{k=0}^{\frac{n}{2}-1} \left(3|\cos(\varphi_k)x_{2k+1}| + 3|\sin(\varphi_k)x_{2k}| + |\cos(\varphi_k)x_{2k+1} + \sin(\varphi_k)x_{2k}|\right)^2 \\
&\qquad + \sum_{k=0}^{\frac{n}{2}-1} \left(3|\cos(\varphi_k)x_{2k}| + 3|\sin(\varphi_k)x_{2k+1}| + |\cos(\varphi_k)x_{2k} - \sin(\varphi_k)x_{2k+1}|\right)^2 \\
&\leq\; \frac{128}{5} \sum_{k=0}^{\frac{n}{2}-1} \left((\cos(\varphi_k))^2 + (\sin(\varphi_k))^2 \right) \left(x_{2k}^2 + x_{2k+1}^2\right) \;=\; \frac{128}{5} \sum_{k=0}^{n-1} x_k^2 \,.
\end{aligned}$$

Offensichtlich folgt dann (3.22), wenn wir auf beiden Seiten die Wurzel ziehen. Genauso ergibt sich (3.23) aus der komponentenweisen Abschätzung (3.14), wenn wir die zweite Ungleichung aus (3.21) blockweise anwenden. ■

Mit Hilfe von Folgerung 3.8 können wir nun die eigentliche Rundungsfehleranalyse durchführen.

Satz 3.9. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n\times n}$ und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin existiere eine Faktorisierung* (3.12) *von A, wobei jede Matrix $A^{(m)},\ m = 1, \ldots, \nu$ eine direkte Summe von Drehmatrizen der Gestalt* (1.24) *ist. Unter Ausschluss von Über- und Unterlauf genügt der Vorwärtsfehler*

$$\Delta\mathbf{y} \;:=\; \mathrm{fl}\left(\hat{A}^{(\nu)}\, \mathrm{fl}\left(\ldots \hat{A}^{(2)}\, \mathrm{fl}\left(\hat{A}^{(1)}\hat{\mathbf{x}}\right)\right)\right) - A\mathbf{x}$$

der Ungleichung

$$\|\Delta\mathbf{y}\|_2 \;\leq\; \left(k_n u + \mathcal{O}(u^2)\right) \|\mathbf{x}\|_2$$

mit $k_n := \frac{3}{2}\sqrt{6}\,(\nu - 1) + 8\sqrt{\frac{2}{5}}$. Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt die Ungleichung bereits mit $k_n = \frac{3}{2}\sqrt{6}\,\nu$.

Beweis: Zunächst definieren wir mit $\mathbf{y}^{(0)} := \mathbf{x}$ und $\tilde{\mathbf{y}}^{(0)} := \mathrm{fl}(\mathbf{x})$ die Vektoren

$$\mathbf{y}^{(m)} := A^{(m)}\mathbf{y}^{(m-1)},\ m = 1, \ldots, \nu,$$

der exakten Zwischenergebnisse, für die nach Voraussetzung $\mathbf{y}^{(m)} \in R_{\mathbb{G}}^{n}$ gilt, und entsprechend die Vektoren

$$\tilde{\mathbf{y}}^{(m)} := \mathrm{fl}\left(\hat{A}^{(m)}\tilde{\mathbf{y}}^{(m-1)}\right),\ m = 1, \ldots, \nu, \tag{3.24}$$

der tatsächlich auftretenden Zwischenergebnisse. Dann lässt sich der Vorwärtsfehler $\Delta\mathbf{y} = \tilde{\mathbf{y}}^{(\nu)} - \mathbf{y}^{(\nu)}$ durch die Teleskopsumme

$$\begin{aligned}
\Delta\mathbf{y} \;&=\; \tilde{\mathbf{y}}^{(\nu)} - A^{(\nu)}\tilde{\mathbf{y}}^{(\nu-1)} + A^{(\nu)}\left(\tilde{\mathbf{y}}^{(\nu-1)} - A^{(\nu-1)}\mathbf{y}^{(\nu-2)}\right) \;=\; \ldots \\
&=\; \sum_{m=2}^{\nu} \left(\prod_{k=m+1}^{\nu} A^{(k)} \right) \left(\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)}\right) \;+\; \left(\prod_{k=2}^{\nu} A^{(k)} \right) \left(\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}\right)
\end{aligned}$$

darstellen. Da wir Über- und Unterlauf nach Voraussetzung ausschließen, kann wegen (3.24) jeweils Folgerung 3.8 auf die Differenzen $\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)},\ m = 2,\ldots,\nu$, und $\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}$ angewandt werden. Aufgrund der Orthogonalität aller beteiligten Matrizen ergibt sich nun

$$\begin{aligned}\|\Delta\mathbf{y}\|_2 &\leq \sum_{m=2}^{\nu}\left(\prod_{k=m+1}^{\nu}\left\|A^{(k)}\right\|_2\right)\left\|\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)}\right\|_2 \;+\; \left(\prod_{k=2}^{\nu}\left\|A^{(k)}\right\|_2\right)\left\|\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}\right\|_2 \\ &\leq \sum_{m=2}^{\nu}\tfrac{3}{2}\sqrt{6}\,\|\tilde{\mathbf{y}}^{(m-1)}\|_2\left(u \;+\; \mathcal{O}(u^2)\right) + \begin{cases}\frac{3}{2}\sqrt{6}\,\|\mathbf{y}^{(0)}\|_2\left(u \;+\; \mathcal{O}(u^2)\right)\ , & \mathbf{x} = \hat{\mathbf{x}} \\ 8\sqrt{\frac{2}{5}}\,\|\mathbf{y}^{(0)}\|_2\left(u \;+\; \mathcal{O}(u^2)\right)\ , & \text{sonst.}\end{cases}\end{aligned}$$

Wegen $\|\mathbf{y}^{(m)}\|_2 = \|\mathbf{x}\|_2,\ m = 0,\ldots,\nu$, und somit auch $\|\tilde{\mathbf{y}}^{(m)}\|_2 \leq \|\mathbf{x}\|_2(1 + \mathcal{O}(u))$ (Induktion über m und wiederholte Anwendung von Dreiecksungleichung sowie von Folgerung 3.8) erhalten wir dann die Behauptung. ■

Bemerkung 3.10. (i) Aufgrund der allgemeinen Formulierung von Satz 3.9 ist zunächst nicht ersichtlich, in welcher Weise die angegebenen Stabilitätskonstanten k_n von n abhängen bzw. wie der genaue Zusammenhang zur Anzahl ν der Matrix-Faktoren ist. Dies wird erst klar, wenn die von n abhängigen Faktorisierungen aus Kapitel 1 explizit angewandt werden (vgl. Satz 3.12).

(ii) Die in [54, Theorem 8.3] für den oben betrachteten Fall mit $\mathbf{x} = \hat{\mathbf{x}}$ angegebene Stabilitätskonstante ist

$$\nu(1 + \sqrt{2}) + 2\sum_{m=1}^{\nu}\eta_m\ , \tag{3.25}$$

wobei jeweils die Elemente von $|\hat{A}^{(m)} - A^{(m)}|$ durch $\eta_m u$ beschränkt sind. Somit kann k_n durch zusätzliche Annahmen an die Genauigkeit der vorberechneten Elemente innerhalb der beteiligten Matrizen $A^{(m)},\ m = 1,\ldots,\nu$, variiert werden. Bei direktem Aufruf, d.h. $\eta_m = \frac{1}{2}$ für alle m, ergibt sich aus (3.25) die Stabilitätskonstante

$$k_n = \nu(2 + \sqrt{2})\ , \tag{3.26}$$

wobei ν die Anzahl der beteiligten Matrizen ist.

(iii) Verwenden wir im Beweis von Satz 3.9 für den Fall $\mathbf{x} = \hat{\mathbf{x}}$ anstelle von (3.23) die in Bemerkung 3.6 (ii) gewonnene Konstante (3.20), so erhalten wir die verbesserte Stabilitätskonstante

$$k_n \;=\; \frac{4}{3}\sqrt{3}\,\nu + \sqrt{2}\sum_{m=1}^{\nu}\eta_m \tag{3.27}$$

bzw.

$$k_n \;=\; \nu\left(\frac{1}{2}\sqrt{2} + \frac{4}{3}\sqrt{3}\right) \tag{3.28}$$

bei direktem Aufruf, d.h., falls $\eta_m = \frac{1}{2}$ für alle m garantiert werden kann. □

Bevor wir mit Hilfe von Satz 3.9 die entsprechenden Stabilitätskonstanten für die Algorithmen aus Abschnitt 1.3 herleiten, geben wir einige Verallgemeinerungen von Satz 3.9 an.

Folgerung 3.11. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n\times n}$ und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin existiere eine Faktorisierung* (3.12) *von A.*

(i) *Ist jede der Matrizen $A^{(m)},\ m = 1,\ldots,\nu$, eine direkte Summe von Drehmatrizen der Form* (1.24)*, dann gelten dieselben Fehlerabschätzungen wie in Satz* 3.9 *auch für A^{T} und die entsprechend transponierte Faktorisierung.*

(ii) *Ist jede der Matrizen $A^{(m)},\ m = 1,\ldots,\nu$, durch Permutations- oder Vorzeichenskalierungsmatrizen in eine direkte Summe von Drehmatrizen der Gestalt* (1.24) (*mit gegebenenfalls auch Winkel $\varphi = 0$*) *überführbar, dann gelten dieselben Fehlerabschätzungen wie in Satz* 3.9*.*

Beweis: Offensichtlich folgen die Behauptungen aus $Q_2(\varphi)^{\mathrm{T}} = \Sigma_2 Q_2(\varphi)\Sigma_2$, $\|\Sigma_2\|_2 = 1$ sowie aus den Anmerkungen nach Lemma 2.8. ■

Entsprechend Folgerung 3.11 ergeben sich nun die einzelnen Stabilitätskonstanten.

Satz 3.12. *Seien $t \geq 2$, $n = 2^t$ sowie $\mathbf{x} \in R_{\mathbb{G}}^n$ gegeben.*

(i) *Für die Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ mittels der Algorithmen* 1.15 *und* 1.18 *ergibt sich die Stabilitätskonstante*

$$k_n = 3\sqrt{6}\,(\log_2(n) - 1) + 8\sqrt{\tfrac{2}{5}}\ . \tag{3.29a}$$

Für $\mathbf{x} = \hat{\mathbf{x}}$ ist

$$k_n = \tfrac{3}{2}\sqrt{6}\,(2\log_2(n) - 1)\ . \tag{3.29b}$$

(ii) *Für die Berechnung der Matrix-Vektor-Multiplikationen $C_n^{\mathrm{II}}\mathbf{x}$ und $C_n^{\mathrm{III}}\mathbf{x}$ mittels der Algorithmen* 1.16 *und* 1.17 *ergibt sich die Stabilitätskonstante*

$$k_n = \tfrac{3}{2}\sqrt{6}\,(3\log_2(n) - 4) + 8\sqrt{\tfrac{2}{5}} \tag{3.29c}$$

im Fall $t \geq 3$ und $k_4 = \frac{3}{2}\sqrt{6} + 8\sqrt{\frac{2}{5}}$. Für $\mathbf{x} = \hat{\mathbf{x}}$ ist $k_4 = 3\sqrt{6}$, und im Fall $t \geq 3$ gilt

$$k_n = \tfrac{9}{2}\sqrt{6}\,(\log_2(n) - 1)\ . \tag{3.29d}$$

(iii) *Für die Berechnung der Matrix-Vektor-Multiplikationen $C_n^{\mathrm{II}}\mathbf{x}$ und $C_n^{\mathrm{III}}\mathbf{x}$ mittels der Algorithmen* 1.19 *und* 1.20 *mit $k = 0$ ergibt sich die Stabilitätskonstante*

$$k_n = \tfrac{3}{2}\sqrt{6}\,(2\log_2(n) - 3) + 8\sqrt{\tfrac{2}{5}}\ . \tag{3.29e}$$

Für $\mathbf{x} = \hat{\mathbf{x}}$ ist

$$k_n = 3\sqrt{6}\,(\log_2(n) - 1)\ . \tag{3.29f}$$

(iv) *Für die Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ mittels der Algorithmen* 1.19 *und* 1.20 *mit $k = 1$ ergeben sich dieselben Stabilitätskonstanten wie bei* (i).

(v) *Für die Berechnung der Matrix-Vektor-Multiplikationen $S_n^{\mathrm{II}}\mathbf{x}$ und $S_n^{\mathrm{III}}\mathbf{x}$ mittels der Algorithmen* 1.21 *und* 1.22 *mit $k = 0$ ergeben sich dieselben Stabilitätskonstanten wie bei* (iii).

(vi) *Für die Berechnung der Matrix-Vektor-Multiplikation $S_n^{\mathrm{IV}}\mathbf{x}$ mittels der Algorithmen* 1.21 *und* 1.22 *mit $k = 1$ ergeben sich dieselben Stabilitätskonstanten wie bei* (i).

Beweis: Es genügt jeweils, die Voraussetzungen von Folgerung 3.11 zu überprüfen.
(i) Die beiden Algorithmen zur Berechnung der DCT-IV(n) basieren auf den Faktorisierungen (1.32) und (1.33) bzw. auf den dazu transponierten Gleichungen (1.37) und (1.39). Nach Folgerung 3.11 genügt es also, die ersten beiden Faktorisierungen zu betrachten. Da die Matrizen $P_n^{(s)}$, $s = 0, \ldots, t-1$ sowie U_n als Permutationsmatrizen keinen Beitrag zum Rundungsfehler leisten, können sie vernachlässigt werden. Somit verursachen in (1.33) lediglich die drei Blockdiagonalmatrizen $I_2 \oplus T_2(0)$, $T_4(0)$ und $D_4^{(0)}$ Rundungsfehler. Da sie jeweils eine direkte Summe von Drehmatrizen der Form (1.24) enthalten, ergibt sich die Stabilitätskonstante k_4 jeweils entsprechend Satz 3.9 mit $\nu = 3 = 2\log_2(4) - 1$. Analog sind $2t-1$ Faktoren in (1.32) enthalten, die im Sinn von Satz 3.9 bzw. Folgerung 3.11 Rundungsfehler verursachen. Somit ergibt sich dann k_n jeweils mit $\nu = 2t - 1$ zu

$$k_n = \begin{cases} \frac{3}{2}\sqrt{6}\,(2t-1)\ , & \mathbf{x} = \hat{\mathbf{x}}\ , \\ \frac{3}{2}\sqrt{6}\,(2t-2) + 8\sqrt{\frac{2}{5}}\ , & \text{sonst.} \end{cases}$$

(ii) In den Faktorisierungen (1.35) und (1.38) für die Matrizen C_n^{III} und C_n^{II} haben wir für $n = 2^t$ mit $t \geq 3$ offenbar $t + 2(t-2) + 1 = 3(t-1)$ Matrizen, welche im Sinn von Satz 3.9 bzw. Folgerung 3.11 Rundungsfehler verursachen, wie aus ihrer Definition in Lemma 1.12 hervorgeht. Somit ergibt sich die Stabilitätskonstante aus Satz 3.9 jeweils mit $\nu = 3(t-1)$. Für den Fall $t = 2$ tragen nur die Matrizen $T_4(0)$ und $\tilde{D}_4^{(0)}$ zum Rundungsfehler bei, so dass sich nach Satz 3.9 bzw. Folgerung 3.11 die entsprechende Stabilitätskonstante aus Satz 3.9 jeweils mit $\nu = 2$ ergibt.

(iii) Für $k = 0$ basieren die Algorithmen 1.19 und 1.20 auf den Faktorisierungen (1.45), die jeweils $2t-1$ Matrizen beinhalten. Aus ihrer Definition in (1.43) und den Darstellungen (1.25), (1.27) und (1.28) mittels spezieller Blockdiagonalmatrizen ist zu erkennen, dass auch hier die Voraussetzungen für Folgerung 3.11 erfüllt sind. Insbesondere handelt es sich bei $A_n(\boldsymbol{\beta}_0)$ und $A_n(\boldsymbol{\beta}_0)^{\mathrm{T}}$ lediglich um Permutationsmatrizen. Demzufolge ergibt sich die Stabilitätskonstante aus Satz 3.9 jeweils mit $\nu = 2(t-1)$.

(iv) Ähnlich wie bei (iii) basieren die Algorithmen 1.19 und 1.20 für $k = 1$ auf den Faktorisierungen (1.48), die ebenfalls je $2t-1$ analog zu (1.43) definierte Matrizen beinhalten. Somit sind auch hier wegen der Darstellungen (1.25), (1.27) und (1.28) die Voraussetzungen für Folgerung 3.11 erfüllt, so dass sich die jeweilige Stabilitätskonstante aus Satz 3.9 mit $\nu = 2t-1$ ergibt.

(v) Analog zu (iii) sind für Folgerung 3.11 die Voraussetzungen an die für $k = 0$ den beiden Algorithmen zugrunde liegenden Faktorisierungen (1.55) erfüllt, so dass sich die jeweilige Stabilitätskonstante aus Satz 3.9 mit $\nu = 2(t-1)$ ergibt. Dabei wird wiederum verwendet, dass die einzelnen Blöcke der in (1.53) definierten Matrizen die Darstellungen (1.27) und (1.28) besitzen und es sich bei $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)$ bzw. $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)^{\mathrm{T}}$ lediglich um eine Permutationsmatrix handelt.

(vi) Die Algorithmen 1.21 und 1.22 basieren für $k = 1$ auf den Faktorisierungen (1.60), deren Faktoren analog zu (1.53) definiert sind. Entsprechend folgt auch hier wiederum die Anwendbarkeit von Folgerung 3.11, so dass sich die jeweilige Stabilitätskonstante aus Satz 3.9 mit $\nu = 2t-1$ ergibt. ■

Offenbar besitzen die in Abschnitt 1.3 vorgestellten Algorithmen ein ähnliches Rundungsfehlerverhalten, was nicht zuletzt darauf beruht, dass die zugrunde liegenden Faktorisierungen auf orthogonalen dünnbesetzten Matrizen beruhen und ähnlich viele Faktoren enthalten.

Bemerkung 3.13. (i) Wenden wir [54], Theorem 8.3, auf den hier vorgestellten Algorithmus 1.15 an, so ergibt sich im Fall $\mathbf{x} = \hat{\mathbf{x}}$ als Stabilitätskonstante

$$k_n \;=\; (2\log_2(n)-1)(1+\sqrt{2}) + 2 \sum_{s=1}^{2\log_2(n)-1} \eta_s \;.$$

Hierbei wird jedoch anstelle von (2.6) angenommen, dass sich die absoluten Fehler sämtlicher Einträge innerhalb einer Matrix $A^{(s)}$ betragsmäßig durch $\eta_s u$ abschätzen lassen. Gehen wir im optimalen Fall wiederum von einem direkten Aufruf aus, ergibt sich jeweils $\eta_s = \frac{1}{2}$ und somit die Stabilitätskonstante

$$k_n \;=\; (2\log_2(n)-1)(2+\sqrt{2}) \;\approx\; 3.414214 \cdot (2\log_2(n)-1) \;, \tag{3.30}$$

welche wegen $\frac{3}{2}\sqrt{6} \;\approx\; 3.674235$ im Vergleich zu der aus dem Modell (2.6) gewonnenen Konstante (3.29b) etwas kleiner ist. Unter adäquaten Voraussetzungen liefert (3.28) für diesen Fall jedoch die kleinere Schranke

$$k_n \;=\; (2\log_2(n)-1)\left(\frac{1}{2}\sqrt{2}+\frac{4}{3}\sqrt{3}\right) \;\approx\; 3,016508 \cdot (2\log_2(n)-1). \tag{3.31}$$

(ii) In [54, Example 8.10] ist für eine leicht abgewandelte Variante von Algorithmus 1.15 im Fall $\mathbf{x} = \hat{\mathbf{x}}$ als Stabilitätskonstante

$$k_n = (2+\sqrt{2})\log_2(n) - 1 + 2 \sum_{s=3}^{\log_2(n)+1} c_{8n,s} + 2c_{8n,\log_2(n)+3}$$

angegeben, wobei $c_{8n,s}u$ jeweils als obere Schranken des absoluten Vorberechnungsfehlers für die jeweiligen Sinus- und Kosinuswerte vorausgesetzt werden. Unter der Annahme, dass alle Sinus- und

Kosinuswerte tabelliert sind und direkt abgerufen werden können, sind diese Schranken alle jeweils $\frac{1}{2}u$, so dass sich damit die wegen $\frac{3+\sqrt{2}}{2} \approx 2.207107$ im Vergleich zu (3.31) bessere Konstante

$$k_n \;=\; (2+\sqrt{2})\log_2(n) - 1 + 2(\log_2(n)-1)\frac{1}{2} + 1 \;=\; \left(\frac{3+\sqrt{2}}{2}\right)(2\log_2(n)-1) + \frac{1+\sqrt{2}}{2} \qquad (3.32)$$

ergibt. Diese Verbesserung beruht im Wesentlichen auf der Modifizierung von Algorithmus 1.15, bei der anstelle der Matrizen $T_n(0)$ zunächst die Butterfly-Matrizen $\sqrt{2}T_n(0)$ verwendet werden, welche keine Rundungsfehler durch Multiplikationen verursachen, und am Ende durch eine entsprechende Potenz von $\sqrt{2}$ dividiert wird. In der speziellen Struktur von skalierten Butterfly-Matrizen befindet sich somit noch Potenzial zur Verbesserung der in Satz 3.12 hergeleiteten Stabilitätskonstanten.

(iii) Bei der Herleitung der Stabilitätskonstanten in [54, Example 8.11] ist ein Rechenfehler aufgetreten. Wenden wir [54, Theorem 8.3] auf die dort angegebene von (1.35) etwas abweichende Faktorisierung für die Matrix C_n^{III} an, so ergibt sich für $\hat{\mathbf{x}} = \mathbf{x}$ die Stabilitätskonstante

$$\begin{aligned} k_n &= t(1+\sqrt{2}+2c_{4n,3}) + \sum_{s=4}^{t+2}(1+\sqrt{2}+2c_{4n,s}) + (t-2)\left(1+\sqrt{2}+\frac{4}{\sqrt{2}}c_{4n,3}\right) \\ &= t(1+\sqrt{2}) + 2tc_{4n,3} + (t-1)(1+\sqrt{2}) + 2\sum_{s=4}^{t+2} c_{4n,s} + (t-2)(1+\sqrt{2}) + (t-2)2\sqrt{2}c_{4n,3} \\ &= \log_2(n)(1+\sqrt{2})(3+2c_{4n,3}) + 2\sum_{s=4}^{\log_2(n)+2} c_{4n,s} - 3(1+\sqrt{2}) - 4\sqrt{2}c_{4n,3}\ . \end{aligned}$$

Unter der Annahme, dass alle Sinus- und Kosinuswerte tabelliert sind und direkt abgerufen werden können, sind die entsprechenden Konstanten wiederum jeweils $\frac{1}{2}$ und somit

$$k_n \;=\; \log_2(n)(5+4\sqrt{2}) - 5\sqrt{2} - 4 \;\approx\; 10.656854\cdot\log_2(n) - 11.071067\ .$$

Hierbei ist zu beachten, dass es sich in [54], Example 8.11, um eine Faktorisierung in fast orthogonale Matrizen handelt. Die Anwendung von [54, Theorem 8.3] auf die nur orthogonale Matrizen enthaltende Faktorisierung (1.35) liefert für den Fall $\hat{\mathbf{x}} = \mathbf{x}$ die Stabilitätskonstante

$$k_n \;=\; 3(\log_2(n)-1)(1+\sqrt{2}) + 2\sum_{s=1}^{3(\log_2(n)-1)} \eta_s\ .$$

Unter der bei direktem Aufruf aller Sinus- und Kosinuswerte sinnvollen Annahme $\eta_s = \frac{1}{2}$ für alle s ergibt sich dann aufgrund von $\frac{9}{2}\sqrt{6} \approx 11.022704$ die im Vergleich zu (3.29d) nur leicht bessere Konstante

$$k_n \;=\; 3(\log_2(n)-1)(2+\sqrt{2}) \;\approx\; 10.242641\cdot(\log_2(n)-1)\ . \qquad (3.33)$$

Letztere ist jedoch nicht nur aus dem Modell (2.6), sondern mit Hilfe zusätzlicher Voraussetzungen gewonnen worden. Unter adäquaten Annahmen liefert (3.28) die kleinere Konstante

$$k_n \;=\; \left(\frac{3}{2}\sqrt{2}+4\sqrt{3}\right)(\log_2(n)-1) \;\approx\; 9.049524\cdot(\log_2(n)-1)\ . \qquad (3.34)$$

(iv) Für die Algorithmen 1.19 – 1.22 mit $k = 0$ ergibt sich für $\mathbf{x} = \hat{\mathbf{x}}$ nach [54], Theorem 8.3, jeweils die Stabilitätskonstante

$$k_n \;=\; 2(\log_2(n)-1)(1+\sqrt{2}) + 2\sum_{s=1}^{2(\log_2(n)-1)} \eta_s\ .$$

Unter der optimalen Annahme $\eta_s = \frac{1}{2}$ für alle s, die beispielsweise durch direkten Aufruf erreicht werden kann, ergibt sich dann die im Vergleich zu (3.29f) etwas kleinere Konstante

$$k_n \;=\; 2(\log_2(n)-1)(2+\sqrt{2}) \;\approx\; 6.828427\cdot(\log_2(n)-1)\ . \qquad (3.35)$$

Diese Verbesserung ist jedoch wiederum nicht nur aus dem Modell (2.6), sondern nur mit zusätzlichen Annahmen zu erreichen. Unter vergleichbaren Voraussetzungen liefert (3.28) in diesem Fall die Stabilitätskonstante

$$k_n \;=\; 2(\log_2(n)-1)\left(\frac{1}{2}\sqrt{2}+\frac{4}{3}\sqrt{3}\right) \;\approx\; 6.033016\cdot(\log_2(n)-1)\;, \tag{3.36}$$

welche noch kleiner als (3.35) ist. □

Die mit Hilfe der neuen optimalen Ungleichungen aus Lemma 3.7 gewonnenen Fehlerabschätzungen in Folgerung 3.8 und schließlich in Satz 3.9 sind die Hauptergebnisse dieses Unterabschnittes. Obwohl sie im Wesentlichen nur auf dem Wilkinson-Modell (2.6) beruhen, sind die gewonnenen Stabilitätskonstanten mit denen aus [54], Theorem 8.3, vergleichbar. Dabei werden in [54] zusätzliche Parameter für den absoluten Vorberechnungsfehler zur Verfügung gestellt. Unter adäquater Abwandlung der Voraussetzungen von Folgerung 3.5 sind wir wegen $\frac{4}{3}\sqrt{3} < 1+\sqrt{2}$ in der Lage, im Vergleich zu [54], Theorem 8.3, verbesserte Konstanten anzugeben.

3.1.2 Stabilitätskonstanten unter Berücksichtigung skalierter Butterfly-Matrizen

Bei der Herleitung der Stabilitätskonstanten in [54] ist an manchen Stellen verwendet worden, dass die Matrizen $T_n(0)$ mittels Permutationen auf Blockdiagonalgestalt überführt werden können, wobei die Blöcke jeweils skalierte Butterfly-Matrizen der Gestalt

$$s\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} \qquad\qquad (s\neq 0) \tag{3.37}$$

sind. Eine reine Butterfly-Matrix ($s=1$) liefert für einen Vektor $\mathbf{x}=(x_0,x_1)^{\mathrm{T}}\in\mathbb{G}^2$ wegen

$$\begin{aligned}\mathrm{fl}(x_0+x_1) &= (x_0+x_1)(1+\varepsilon^+)\;,\\ \mathrm{fl}(x_0-x_1) &= (x_0-x_1)(1+\varepsilon^-)\end{aligned}$$

mit $|\varepsilon^+|,|\varepsilon^-|\leq u$ nach Umstellen, Übergang zu Beträgen, Quadrieren, Aufsummieren der resultierenden Ungleichungen und anschließendem Wurzelziehen die zu [43, Lemma 5.2 (i)] äquivalente Abschätzung

$$\left\|\mathrm{fl}\left(\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\mathbf{x}\right)-\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\mathbf{x}\right\|_2 \;\leq\; \sqrt{2}\|\mathbf{x}\|_2 u\;.$$

Da innerhalb der Gleitkomma-Arithmetik keine Distributivität garantiert wird, gilt für $a,b,c\in\mathbb{G}$ im Allgemeinen

$$\mathrm{fl}(\mathrm{fl}(ca)+\mathrm{fl}(cb)) \;\neq\; \mathrm{fl}(c\,\mathrm{fl}(a+b))\;. \tag{3.38}$$

Unter Anwendung von (2.6) ergeben sich bei der Matrix-Vektor-Multiplikation für eine mit $s\in\mathbb{G}$ skalierte Butterfly-Matrix (3.37) und für einen Vektor $\mathbf{x}=(x_0,x_1)^{\mathrm{T}}\in\mathbb{G}^2$ die Komponenten

$$\begin{aligned}\mathrm{fl}(s\,\mathrm{fl}(x_0+x_1)) &= s\,\mathrm{fl}(x_0+x_1)(1+\varepsilon_1^\times) \;=\; s(x_0+x_1)(1+\varepsilon_1^\times+\varepsilon_1^+ +\varepsilon_1^\times\varepsilon_1^+)\;,\\ \mathrm{fl}(s\,\mathrm{fl}(x_0-x_1)) &= s\,\mathrm{fl}(x_0-x_1)(1+\varepsilon_2^\times) \;=\; s(x_0-x_1)(1+\varepsilon_2^\times+\varepsilon_2^+ +\varepsilon_2^\times\varepsilon_2^+)\end{aligned}$$

mit $|\varepsilon_1^\times|,|\varepsilon_1^+|,|\varepsilon_2^\times|,|\varepsilon_2^+|\leq u$. Somit erfüllt der Fehlervektor die komponentenweise Abschätzung

$$\left|\mathrm{fl}\left(s\,\mathrm{fl}\left(\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\mathbf{x}\right)\right) - \begin{pmatrix}s & s\\ s & -s\end{pmatrix}\mathbf{x}\right| \;\leq\; \left|\begin{pmatrix}s & s\\ s & -s\end{pmatrix}\mathbf{x}\right|\left(2u+\mathcal{O}(u^2)\right)\;. \tag{3.39}$$

Bevor wir nun (3.39) auf Matrix-Vektor-Multiplikationen von Blockdiagonalmatrizen mit Blöcken der Form (3.37) anwenden, leiten wir noch zwei zu (2.19) ähnliche Ungleichungen her.

Lemma 3.14. *Seien $a,b,s\in\mathbb{R}$ beliebig. Dann gelten*

$$(|as|+|bs|+|as-bs|)^2+(|as|+|bs|+|as+bs|)^2 \;\leq\; 2\Big(3+\sqrt{5}\Big)\Big(a^2+b^2\Big)s^2\;, \tag{3.40a}$$

$$(|as|+|bs|+2|as-bs|)^2+(|as|+|bs|+2|as+bs|)^2 \;\leq\; 2\Big(7+\sqrt{13}\Big)\Big(a^2+b^2\Big)s^2\;. \tag{3.40b}$$

Dabei sind die Konstanten $2(3+\sqrt{5})$ und $2(7+\sqrt{13})$ optimal.

Beweis: O.B.d.A. können wir annehmen, dass $a \geq b \geq 0$ sowie $s > 0$ gelten. Da s^2 auf beiden Seiten als Faktor ausgeklammert werden kann, vereinfachen sich die Behauptungen wegen $s > 0$ zu

$$\begin{aligned} a^2 \ + \ (a+b)^2 \ &\leq \ \tfrac{3+\sqrt{5}}{2}(a^2+b^2) \ , \\ \left(a - \tfrac{b}{3}\right)^2 \ + \ (a+b)^2 \ &\leq \ \tfrac{2(7+\sqrt{13})}{9}(a^2+b^2) \ . \end{aligned}$$

Die jeweilige Äquivalenz zu den offensichtlich gültigen Ungleichungen

$$\begin{aligned} 0 \ &\leq \ \left(a - \tfrac{1+\sqrt{5}}{2}b\right)^2 \ , \\ 0 \ &\leq \ \left(a - \tfrac{2+\sqrt{13}}{3}b\right)^2 \end{aligned}$$

lässt sich leicht überprüfen. Da die Ungleichung (3.40a) beispielsweise für $a = 1$,

$$b = \tfrac{2}{1+\sqrt{5}} = \tfrac{\sqrt{5}-1}{2}$$

und beliebiges s auf beiden Seiten den Wert $\left(10 + 2\sqrt{5}\right) s^2$ annimmt und da ebenso die Ungleichung (3.40b) beispielsweise für $a = 1$,

$$b = \tfrac{3}{2+\sqrt{13}} = \tfrac{\sqrt{13}-2}{3}$$

und beliebiges s auf beiden Seiten den Wert $\frac{4}{9}\left(65 - \sqrt{13}\right) s^2$ annimmt, ist sowohl die Konstante $2(3 + \sqrt{5})$ als auch die Konstante $2(7 + \sqrt{13})$ bestmöglich. ■

Im Vergleich zu Satz 2.12 und Folgerung 3.8 erhalten wir für diesen Spezialfall nun die folgenden verbesserten Ergebnisse.

Lemma 3.15. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n \times n}$ eine Blockdiagonalmatrix mit identischen fast orthogonalen Blöcken wie in* (3.37) *definiert und $\mathbf{x} \in R_{\mathbb{G}}^n$. Weiter seien $\hat{A} := \mathrm{fl}(A)$ und $\hat{\mathbf{x}} = \mathrm{fl}(\mathbf{x})$ die entsprechenden Gleitkomma-Approximationen und es gelte $|\hat{A}||\hat{\mathbf{x}}| \in R_{\mathbb{G}}$. Darüber hinaus werde* (3.38) *bzw.* (3.39) *berücksichtigt. Dann sind folgende Aussagen gültig.*

(i) *Der Fehlervektor erfüllt die komponentenweise Abschätzung*

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \ \leq \ \|A\|_2 \ \|\mathbf{x}\|_2 \left(2\sqrt{3+\sqrt{5}}u \ + \ \mathcal{O}(u^2)\right). \tag{3.41}$$

(ii) *Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt*

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \ \leq \ \|A\|_2 \ \|\mathbf{x}\|_2 \left(\sqrt{7+\sqrt{13}}u \ + \ \mathcal{O}(u^2)\right). \tag{3.42}$$

(iii) *Im Fall $A = \hat{A}$ sowie $\mathbf{x} = \hat{\mathbf{x}}$ genügt der Fehlervektor der komponentenweisen Abschätzung*

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \ \leq \ \|A\|_2 \ \|\mathbf{x}\|_2 \left(2u \ + \ \mathcal{O}(u^2)\right). \tag{3.43}$$

Beweis: Aus (3.39) und damit insbesondere aus $|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - \hat{A}\hat{\mathbf{x}}| \leq |\hat{A}\hat{\mathbf{x}}|\left(2u \ + \ \mathcal{O}(u^2)\right)$ folgt wie im Beweis zu Lemma 2.10 sofort

$$\left\| \mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - \hat{A}\hat{\mathbf{x}} \right\|_2 \ \leq \ \|\hat{A}\|_2 \ \|\hat{\mathbf{x}}\|_2 \left(2u \ + \ \mathcal{O}(u^2)\right),$$

womit sich für $A = \hat{A}$ sowie $\mathbf{x} = \hat{\mathbf{x}}$ bereits die Aussage (iii) ergibt. Verwenden wir $|\hat{A}\hat{\mathbf{x}}| \leq |A\hat{\mathbf{x}}| + |A||\hat{\mathbf{x}}|u$, so ergibt sich mit Dreiecksungleichung

$$\left|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\hat{\mathbf{x}}\right| \ \leq \ |\hat{A}\hat{\mathbf{x}}|\left(2u \ + \ \mathcal{O}(u^2)\right) + |A||\hat{\mathbf{x}}|u \ \leq \ \left(2|A\hat{\mathbf{x}}| + |A||\hat{\mathbf{x}}|\right)\left(u \ + \ \mathcal{O}(u^2)\right).$$

Für $\hat{\mathbf{x}} = \mathbf{x}$ folgt dann wie im Beweis von Satz 2.12 mit Ungleichung (3.40b) die Aussage (ii), wobei hier verwendet wird, dass die Spektralnorm einer Matrix der Form (3.37) genau $\sqrt{2s^2}$ ist. Andernfalls erhalten wir mit Dreiecksungleichung sowie mit $|A\hat{\mathbf{x}}| \leq |A\mathbf{x}| + |A||\mathbf{x}|u$ und $|\hat{\mathbf{x}}| \leq |\mathbf{x}|(1+u)$ die komponentenweise Ungleichungskette

$$\left|\,\mathrm{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x}\right| \ \leq \ \left(2|A\hat{\mathbf{x}}| + |A||\hat{\mathbf{x}}|\right)\left(u \ + \ \mathcal{O}(u^2)\right) + |A||\mathbf{x}|u \ \leq \ \left(|A\mathbf{x}| + |A||\mathbf{x}|\right)\left(2u \ + \ \mathcal{O}(u^2)\right) \ .$$

Die Aussage (i) folgt nun wie im Beweis von Satz 2.12 aus Ungleichung (3.40a) analog zuvor. ■

Verwenden wir in Lemma 3.15 für den Fall $\mathbf{x} = \hat{\mathbf{x}}$ wiederum wie in [54] vorgeschlagen eine gleichmäßige obere Schranke ηu für die Einträge von $|A - \hat{A}|$ und beachten, dass dies aufgrund der identischen Blöcke (3.37) zu der Bedingung $|s - \hat{s}| \leq \eta u$ äquivalent ist, ergibt sich aus $|\hat{A}\hat{\mathbf{x}}| \leq |A\mathbf{x}| + B\eta|\mathbf{x}|u$ mit B wie in Lemma 3.4 definiert und aus der daraus resultierenden komponentenweisen Ungleichung

$$\left| \operatorname{fl}(\hat{A}\hat{\mathbf{x}}) - A\hat{\mathbf{x}} \right| \;\leq\; |A\mathbf{x}| \left(2u \;+\; \mathcal{O}(u^2)\right) + |(A - \hat{A})\mathbf{x}|$$

analog zum Beweis von Lemma 2.10 nach Übergang zu Normen über Dreiecksungleichung somit

$$\left\| \operatorname{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; \|A\|_2 \|\mathbf{x}\|_2 \left(2u \;+\; \mathcal{O}(u^2)\right) + \eta \left\| \bigoplus_{k=1}^{\frac{n}{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \right\|_2 \|\mathbf{x}\|_2 u \;.$$

Wegen $\|A\|_2 = \sqrt{2}|s|$ erhalten wir nun die zu [43], Lemma 5.2 (ii), äquivalente Abschätzung

$$\left\| \operatorname{fl}(\hat{A}\hat{\mathbf{x}}) - A\mathbf{x} \right\|_2 \;\leq\; \left(2\sqrt{2}|s| + \sqrt{2}\eta \;+\; \mathcal{O}(u)\right) u \left\|\mathbf{x}\right\|_2 \;. \tag{3.44}$$

Für orthogonale Matrizen A haben wir in diesem Fall die Konstante $2 + \sqrt{2}\eta$ gewonnen. Mit Hilfe von Lemma 3.15 lassen sich nun die Stabilitätskonstanten von Satz 3.9 für Spezialfälle noch verbessern.

Satz 3.16. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n \times n}$ und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin existiere eine Faktorisierung* (3.12) *von A, wobei jede Matrix $A^{(m)},\ m = 1, \ldots, \nu$ eine direkte Summe von Drehmatrizen der Form* (1.24) *ist. Besitzen $\mu \leq \nu$ der Blockdiagonalmatrizen $A^{(m)},\ m = 1, \ldots, \nu$, die Gestalt*

$$A^{(m)} \;=\; \bigoplus_{k=1}^{\frac{n}{2}} Q_2\left(\tfrac{\pi}{4}\right) \tag{3.45}$$

und findet (3.38) *bei der Implementierung Berücksichtigung, so genügt der Vorwärtsfehler*

$$\Delta\mathbf{y} \;:=\; \operatorname{fl}\left(\hat{A}^{(\nu)} \operatorname{fl}\left(\ldots \hat{A}^{(2)} \operatorname{fl}\left(\hat{A}^{(1)}\hat{\mathbf{x}}\right)\right)\right) - A\mathbf{x}$$

unter Ausschluss von Über- und Unterlauf der Ungleichung $\|\Delta\mathbf{y}\|_2 \;\leq\; \left(k_n u + \mathcal{O}(u^2)\right) \|\mathbf{x}\|_2$ *mit*

$$k_n \;=\; \begin{cases} \sqrt{7 + \sqrt{13}}\,(\mu - 1) \;+\; 2\sqrt{3 + \sqrt{5}} & \text{für } \mu = \nu, \\ \frac{3}{2}\sqrt{6}\,(\nu - \mu) \;+\; \sqrt{7 + \sqrt{13}}(\mu - 1) \;+\; 2\sqrt{3 + \sqrt{5}} & \text{für } \mu < \nu \text{ und (3.45) für } m = 1, \\ \frac{3}{2}\sqrt{6}\,(\nu - \mu - 1) \;+\; \sqrt{7 + \sqrt{13}}\mu \;+\; 8\sqrt{\frac{2}{5}} & \text{sonst.} \end{cases}$$

Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt die Ungleichung bereits mit

$$k_n \;=\; \tfrac{3}{2}\sqrt{6}\,(\nu - \mu) \;+\; \sqrt{7 + \sqrt{13}}\mu \;.$$

Beweis: Analog zum Beweis von Satz 3.9 lässt sich $\|\Delta\mathbf{y}\|_2$ über die Ungleichung

$$\|\Delta\mathbf{y}\|_2 \;\leq\; \sum_{m=2}^{\nu} \left\| \tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)} \right\|_2 \;+\; \left\| \tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)} \right\|_2$$

abschätzen. Da sich die Zwischenergebnisse $\tilde{\mathbf{y}}^{(m-1)}$ bereits in $\mathbb{G}^n$ befinden, ergeben sich nach Folgerung 3.8 und Lemma 3.15 für $m \geq 2$ weiterhin

$$\left\| \tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)} \right\|_2 \;\leq\; \begin{cases} \sqrt{7 + \sqrt{13}}\ \|\tilde{\mathbf{y}}^{(m-1)}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{im Fall (3.45)}, \\ \frac{3}{2}\sqrt{6}\ \|\tilde{\mathbf{y}}^{(m-1)}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{sonst.} \end{cases}$$

Ebenso ergibt sich

$$\left\| \tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)} \right\|_2 \;\leq\; \begin{cases} 2\sqrt{3 + \sqrt{5}}\ \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{im Fall (3.45)}, \\ 8\sqrt{\frac{2}{5}}\ \|\mathbf{x}\|_2 \left(u \;+\; \mathcal{O}(u^2)\right) & \text{sonst} \end{cases}$$

bzw. für $\mathbf{x} = \hat{\mathbf{x}}$ dann

$$\left\|\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}\right\|_2 \;\leq\; \begin{cases} \sqrt{7+\sqrt{13}}\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right) & \text{im Fall (3.45),} \\ \frac{3}{2}\sqrt{6}\,\|\mathbf{x}\|_2\left(u \;+\; \mathcal{O}(u^2)\right) & \text{sonst.} \end{cases}$$

Mit $\|\tilde{\mathbf{y}}^{(m)}\|_2 \leq \|\mathbf{x}\|_2(1+\mathcal{O}(u))$ folgen nun alle Behauptungen. ■

Sind zusätzlich gleichmäßige obere Schranken $\eta_m u$ für die Einträge von $|A^{(m)} - \hat{A}^{(m)}|$, $m = 1, \ldots, \nu$, bekannt, so erhalten wir mittels der nach Folgerung 3.5 gewonnenen Konstante $\frac{4}{3}\sqrt{3} + 2\eta_m$ sowie mit der für orthogonale Matrizen aus (3.44) resultierenden Konstante $2 + \sqrt{2}\eta_m$ im Fall $\mathbf{x} = \hat{\mathbf{x}}$ unter denselben Voraussetzungen wie in Satz 3.16 als Stabilitätskonstante

$$k_n = \frac{4}{3}\sqrt{3}(\nu - \mu) + 2\mu + \sqrt{2}\sum_{m=1}^{\nu}\eta_m \tag{3.46}$$

bzw.

$$k_n = \frac{4}{3}\sqrt{3}(\nu - \mu) + 2\mu + \frac{1}{2}\sqrt{2}\;\nu \tag{3.47}$$

bei direktem Aufruf, d.h., falls $\eta_m = \frac{1}{2}$ für alle m garantiert werden kann. Wie Satz 3.9 lässt sich nun ebenso Satz 3.16 weiter verallgemeinern, wenn berücksichtigt wird, dass Permutations- und Vorzeichenskalierungsmatrizen keinen Einfluss auf Rundungsfehler ausüben.

Folgerung 3.17. *Sei $n \in \mathbb{N}$ gerade, $A \in R_{\mathbb{G}}^{n\times n}$ und $\mathbf{x} \in R_{\mathbb{G}}^{n}$. Weiterhin existiere eine Faktorisierung* (3.12) *von A.*

(i) *Ist jede der Matrizen $A^{(m)}$, $m = 1, \ldots, \nu$, eine direkte Summe von Drehmatrizen der Form* (1.24) *und besitzen $\mu \leq \nu$ dieser Matrizen die Darstellung* (3.45)*, dann gelten dieselben Fehlerabschätzungen wie in Satz* 3.16 *auch für A^{T} und die entsprechend transponierte Faktorisierung, wobei hier die Rollen von $A^{(1)}$ und $A^{(\nu)}$ vertauscht sind.*

(ii) *Ist jede der Matrizen $A^{(m)}$, $m = 1, \ldots, \nu$, durch Permutations- oder Vorzeichenskalierungsmatrizen in eine direkte Summe von Drehmatrizen der Form* (1.24) *überführbar (wobei gegebenenfalls auch Winkel $\varphi = 0$ zugelassen ist) und enthalten $\mu \leq \nu$ dieser Matrizen nur Blöcke* (1.24) *mit $\varphi = \frac{\pi}{4}$ oder $\varphi = 0$, dann gelten dieselben Fehlerabschätzungen wie in Satz* 3.16*.*

Beweis: Die Behauptungen ergeben sich wie für Folgerung 3.11. ■

Folgerung 3.17 erlaubt es uns nun, für die Algorithmen aus Abschnitt 1.3 Stabilitätskonstanten herzuleiten, indem wir noch mehr auf die Struktur der beteiligten Matrixfaktoren eingehen.

Satz 3.18. *Seien $t \geq 2$, $n = 2^t$ sowie $\mathbf{x} \in R_{\mathbb{G}}^{n}$ gegeben.*

(i) *Für die Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ mittels der Algorithmen* 1.15 *und* 1.18 *ergibt sich für $t \geq 3$ die Stabilitätskonstante*

$$k_n \;=\; \left(\tfrac{3}{2}\sqrt{6} + \sqrt{7+\sqrt{13}}\right)(\log_2(n) - 1) \;+\; \begin{cases} \frac{3}{2}\sqrt{6}\,, & \textit{falls } \mathbf{x} = \hat{\mathbf{x}}\,, \\ 8\sqrt{\frac{2}{5}} & \textit{sonst.} \end{cases}$$

Im Fall $t = 2$ erhalten wir

$$k_4 \;=\; \tfrac{3}{2}\sqrt{6} \;+\; 2\sqrt{7+\sqrt{13}} \;+\; \begin{cases} 0\,, & \textit{falls } \mathbf{x} = \hat{\mathbf{x}}\,, \\ 8\sqrt{\frac{2}{5}} - \frac{3}{2}\sqrt{6} & \textit{für Algorithmus } 1.15\,, \\ 2\sqrt{3+\sqrt{5}} - \sqrt{7+\sqrt{13}} & \textit{für Algorithmus } 1.18\,. \end{cases}$$

(ii) *Für die Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *mittels der Algorithmen* 1.16 *und* 1.17 *ergibt sich für* $t \geq 3$ *die Stabilitätskonstante*

$$k_n \;=\; \left(\tfrac{3}{2}\sqrt{6}+2\sqrt{7+\sqrt{13}}\right)(\log_2(n)-1) \;+\; \begin{cases} 0\;, & \textit{falls } \mathbf{x}=\hat{\mathbf{x}}\;, \\ 8\sqrt{\tfrac{2}{5}}-\tfrac{3}{2}\sqrt{6} & \textit{für Algorithmus } 1.16\;, \\ 2\sqrt{3+\sqrt{5}}-\sqrt{7+\sqrt{13}} & \textit{für Algorithmus } 1.17\;. \end{cases}$$

Im Fall $t=2$ *erhalten wir*

$$k_4 \;=\; \begin{cases} \tfrac{3}{2}\sqrt{6}\;+\;\sqrt{7+\sqrt{13}}\;, & \textit{falls } \mathbf{x}=\hat{\mathbf{x}}\;, \\ \sqrt{7+\sqrt{13}}\;+\;8\sqrt{\tfrac{2}{5}} & \textit{für Algorithmus } 1.16\;, \\ \tfrac{3}{2}\sqrt{6}\;+\;2\sqrt{3+\sqrt{5}} & \textit{für Algorithmus } 1.17\;. \end{cases}$$

(iii) *Für die Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *mittels der Algorithmen* 1.19 *und* 1.20 *mit* $k=0$ *ergibt sich die Stabilitätskonstante*

$$k_n \;=\; \left(\tfrac{3}{2}\sqrt{6}+\sqrt{7+\sqrt{13}}\right)(\log_2(n)-1) \;+\; \begin{cases} 0\;, & \textit{falls } \mathbf{x}=\hat{\mathbf{x}}\;, \\ 2\sqrt{3+\sqrt{5}}-\sqrt{7+\sqrt{13}} & \textit{sonst}\;. \end{cases}$$

(iv) *Für die Berechnung der Matrix-Vektor-Multiplikation* $C_n^{\mathrm{IV}}\mathbf{x}$ *mittels der Algorithmen* 1.19 *und* 1.20 *mit* $k=1$ *ergibt sich die Stabilitätskonstanten*

$$k_n \;=\; \tfrac{3}{2}\sqrt{6}\log_2(n)\;+\;\sqrt{7+\sqrt{13}}(\log_2(n)-1)\;+\begin{cases} 0\;, & \textit{falls } \mathbf{x}=\hat{\mathbf{x}}\;, \\ 8\sqrt{\tfrac{2}{5}}-\tfrac{3}{2}\sqrt{6} & \textit{für Alg. } 1.19\;, \\ 2\sqrt{3+\sqrt{5}}-\sqrt{7+\sqrt{13}} & \textit{für Alg. } 1.20\;. \end{cases}$$

(v) *Für die Berechnung der Matrix-Vektor-Multiplikationen* $S_n^{\mathrm{II}}\mathbf{x}$ *und* $S_n^{\mathrm{III}}\mathbf{x}$ *mittels der Algorithmen* 1.21 *und* 1.22 *mit* $k=0$ *ergeben sich dieselben Stabilitätskonstanten wie bei* (iii).

(vi) *Für die Berechnung der Matrix-Vektor-Multiplikation* $S_n^{\mathrm{IV}}\mathbf{x}$ *mittels der Algorithmen* 1.21 *und* 1.22 *mit* $k=1$ *ergeben sich die Stabilitätskonstanten*

$$k_n \;=\; \tfrac{3}{2}\sqrt{6}\log_2(n)\;+\;\sqrt{7+\sqrt{13}}(\log_2(n)-1)\;+\begin{cases} 0\;, & \textit{falls } \mathbf{x}=\hat{\mathbf{x}}\;, \\ 2\sqrt{3+\sqrt{5}}-\sqrt{7+\sqrt{13}} & \textit{für Alg. } 1.21\;, \\ 8\sqrt{\tfrac{2}{5}}-\tfrac{3}{2}\sqrt{6} & \textit{für Alg. } 1.22\;. \end{cases}$$

Beweis: Es genügt, jeweils die Anzahl der skalierten Butterfly-Matrizen zu ermitteln und zu überprüfen, ob die erste Matrix eine solche ist.

(i) Die beiden Algorithmen zur Berechnung der DCT-IV(n) basieren auf den Faktorisierungen (1.32) und (1.33) bzw. auf den dazu transponierten Gleichungen (1.37) und (1.39). Von den betreffenden Matrizen sind nur die $B_n^{(s)}$ skalierte Butterfly-Matrizen. Somit ist für $t \geq 3$ Satz 3.16 bzw. Folgerung 3.17 mit $\nu = 2t-1$ und $\mu = t-1$ anzuwenden, wobei (3.45) für $m=1$ nicht erfüllt ist. Für $t=2$ ist $\nu = 3$ und $\mu = 2$, wobei in der ersten Gleichung von (1.39) der erste wesentliche Matrixfaktor höchstens skalierte Butterfly-Blöcke enthält.

(ii) Aus den entsprechenden Faktorisierungen (1.35) und (1.38) für $t \geq 3$ geht hervor, dass Satz 3.16 bzw. Folgerung 3.17 mit $\nu = 3(t-1)$ und $\mu = 2(t-1)$ angewandt werden kann, wobei der erste wesentliche Matrixfaktor in (1.38) höchstens skalierte Butterfly-Blöcke besitzt. Für $t=2$ ist hier $\nu = 2$

sowie $\mu = 1$ zu wählen, wobei in der zweiten Gleichung von (1.39) der erste wesentliche Matrixfaktor höchstens skalierte Butterfly-Blöcke enthält.

(iii) Für $k = 0$ basieren die Algorithmen 1.19 und 1.20 auf den Faktorisierungen (1.45). Von den jeweils $2(t-1)$ wesentlichen Matrixfaktoren ist $T_n(\boldsymbol{\beta}_0)$ mittels Permutationen und Vorzeichenskalierungsmatrizen auf die Gestalt (3.45) überführbar. Genauso bewirken die Matrizen $A_n(\boldsymbol{\beta}_s),\ s = 1, \ldots, t-2$, höchstens skalierte Butterfly-Operationen. Somit lässt sich in diesem Fall Satz 3.16 bzw. Folgerung 3.17 mit $\nu = 2(t-1)$ und $\mu = t-1$ anwenden.

(iv) Für $k = 1$ basieren die Algorithmen 1.19 und 1.20 auf den Faktorisierungen (1.48). Da $A_n(\boldsymbol{\gamma}_0)$ im Gegensatz zu $A_n(\boldsymbol{\beta}_0)$ keine Permutationsmatrix ist, sondern höchstens skalierte Butterfly-Operationen bewirkt, ist hier ein wesentlicher Matrixfaktor mehr als bei (iii) vorhanden. Jedoch enthält nun $T_n(\boldsymbol{\gamma}_0)$ auch Drehmatrizen für $\varphi \in]0, \frac{\pi}{4}[$. Somit findet Satz 3.16 bzw. Folgerung 3.17 hier mit $\nu = 2t-1$ und $\mu = t-1$ Anwendung.

(v) Da die Algorithmen 1.21 und 1.22 für $k = 0$ auf den Faktorisierungen (1.55) basieren, $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)$ lediglich eine Permutationsmatrix ist und die Matrizen $\breve{T}_n(\breve{\boldsymbol{\beta}}_0)$ und $\breve{A}_n(\breve{\boldsymbol{\beta}}_s),\ s = 1, \ldots, t-2$ höchstens skalierte Butterfly-Operationen, kann hier analog zu (iii) Satz 3.16 bzw. Folgerung 3.17 mit $\nu = 2(t-1)$ und $\mu = t-1$ angewandt werden.

(vi) Da die Algorithmen 1.21 und 1.22 für $k = 1$ auf den Faktorisierungen (1.60) basieren, folgt die Behauptung analog zu (iv). ■

Bemerkung 3.19. (i) Die Stabilitätskonstante in Satz 3.18 (i) für die Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ mittels der Algorithmen 1.15 und 1.18 im Fall $\mathbf{x} = \hat{\mathbf{x}}$ ist wegen

$$\tfrac{3}{2}\sqrt{6} + \sqrt{7+\sqrt{13}} \;\approx\; 6.930851 \tag{3.48}$$

und $3\sqrt{6} \approx 7.348469$ im Vergleich mit der in Satz 3.12 (i) gewonnenen Konstante (3.29b) besser. Darüber hinaus wird – selbst für den bestmöglichen Fall $\eta = \frac{1}{2}$ – die aus der Anwendung von [54, Lemma 8.3] resultierende Abschätzung (3.30) unterboten (vgl. Bemerkung 3.13 (i)). Unter adäquaten Voraussetzungen liefert (3.46) mit $\nu = 2t-1$ und $\mu = t-1$ sogar

$$k_n \;=\; \left(\tfrac{4}{3}\sqrt{3} + 2 + \sqrt{2}\right)\log_2(n) - \left(2 + \tfrac{1}{2}\sqrt{2}\right) \;\approx\; 5.723615 \cdot \log_2(n) - 2.707107\ . \tag{3.49}$$

(ii) Betrachten wir nun wiederum die leicht abgewandelte Variante von Algorithmus 1.15 im Fall $\mathbf{x} = \hat{\mathbf{x}}$ aus [54, Example 8.10] haben wir uns aufgrund von (3.48) und $3 + \sqrt{2} \approx 4.414214$ mit der Konstanten aus Satz 3.18 (i) vergleichsweise wenig an die Schranke (3.32) heran bewegen können. Das liegt zum einen daran, dass wir die Multiplikation mit $s = \frac{1}{\sqrt{2}} = \cos\left(\frac{\pi}{4}\right) = \sin\left(\frac{\pi}{4}\right)$ bei den skalierten Butterfly-Matrizen mit Blöcken der Gestalt (3.37) zwar unter Berücksichtigung von (3.38) als in günstigerer Weise durchgeführt angenommen haben, jedoch diese Multiplikationen in [54, Example 8.10] erst am Ende als eine solche zusammengefasst auftreten. Aufgrund dessen erscheinen bei der Herleitung unserer Abschätzungen mehr Terme für diese Multiplikation, so dass die Konstante insgesamt größer ist. Da es sich jedoch nicht um denselben Algorithmus handelt, ist dies nicht verwunderlich.

(iii) Für die Algorithmen 1.16 und 1.17 liefert Folgerung 3.18 (ii) im Fall $\mathbf{x} = \hat{\mathbf{x}}$ wegen

$$\tfrac{3}{2}\sqrt{6} + 2\sqrt{7+\sqrt{13}} \;\approx\; 10.187468 \tag{3.50}$$

und $\frac{9}{2}\sqrt{6} \approx 11.022704$ im Vergleich zu (3.29d) aus Satz 3.12 (ii) eine Verbesserung. Ebenso gewinnt im Vergleich zu (3.34) die aus (3.46) mit $\nu = 3(t-1)$ und $\mu = 2(t-1)$ (vgl. Beweis von Satz 3.18 (ii)) sowie jeweils $\eta_m = \frac{1}{2}$ resultierende Konstante

$$k_n \;=\; \left(\tfrac{4}{3}\sqrt{3} + 4 + \tfrac{3}{2}\sqrt{2}\right)\left(\log_2(n) - 1\right) \;\approx\; 8.430721 \cdot \left(\log_2(n) - 1\right)\ . \tag{3.51}$$

(iv) Schließlich bedeuten die Aussagen aus Satz 3.18 (iii)-(vi) für die Algorithmen 1.19 – 1.22 analog zu (i) jeweils eine Verbesserung von Satz 3.12 (iii)-(vi). Darüber hinaus liefert (3.46) im Fall $\mathbf{x} = \hat{\mathbf{x}}$ mit

$\mu = t-1$ und $\eta_m = \frac{1}{2}$ sowie $\nu = 2(t-1)$ für $k=0$ bzw. $\nu = 2t-1$ für $k=1$ die Konstanten

$$k_n \;=\; \begin{cases} \left(\frac{4}{3}\sqrt{3}+2+\sqrt{2}\right)\left(\log_2(n)-1\right) & \text{im Fall } k=0\ , \\ \left(\frac{4}{3}\sqrt{3}+2+\sqrt{2}\right)\log_2(n)-2-\frac{1}{2}\sqrt{2} & \text{im Fall } k=1\ . \end{cases} \tag{3.52}$$

Wegen

$$\tfrac{4}{3}\sqrt{3}+2+\sqrt{2} \approx 5.723615 \tag{3.53}$$

sind die Konstanten (3.52) ebenfalls kleiner als (3.36).

(v) Eine Faktorisierung der Matrix $\sqrt{n}C_n^{\mathrm{II}}$ bzw. $\sqrt{n}C_n^{\mathrm{IV}}$ für eine Zweierpotenz $n=2^t$ in fast orthogonale dünnbesetzte Matrizen liegt den Algorithmen 3.1 und 3.2 in [43] zugrunde. Sie lässt sich jeweils aus (1.45) bzw. (1.48) gewinnen, indem alle Faktoren $T_n(\boldsymbol{\beta}_s)$ bzw. $T_n(\boldsymbol{\gamma}_s)$, $s=0,\ldots,t-1$, mit $\sqrt{2}$ multipliziert werden. Da sowohl $\sqrt{2}C_2^{\mathrm{II}}$ als auch $\sqrt{2}T_n(0)$ (vgl. (1.16)) reine Butterfly-Operationen enthalten, kann die Anzahl der Multiplikationen deutlich reduziert werden. Im Gegenzug enthalten die Matrizen $\sqrt{2}C_2^{\mathrm{IV}}$ und $\sqrt{2}T_n(1)$ (vgl. (1.18)) anstelle von Drehungen nun skalierte Drehungen. Mit den Ungleichungen aus [43, Lemma 5.2 (i) und (iii)] wird unter Berücksichtigung der abschließenden Skalierung mit $1/\sqrt{n}$ in [43, Theorem 5.3 und 5.4] die Stabilitätskonstante

$$k_n \;=\; \left(\tfrac{4}{3}\sqrt{3}+\tfrac{1}{2}\sqrt{2}+3\right)\left(\log_2(n)-1\right) \;\approx\; 6.016508\cdot\left(\log_2(n)-1\right) \tag{3.54}$$

für die Algorithmen 3.1 und 3.2 aus [43] hergeleitet. Dabei wird vorausgesetzt, dass sowohl $\sqrt{2}$ als auch sämtliche Matrixeinträge $\sqrt{2}\cos(\alpha)$ und $\sqrt{2}\sin(\alpha)$ für

$$\alpha \in \left\{ \frac{(2k+1)\pi}{2^{s+3}} \;\middle|\; k=0,\ldots,2^s-1;\ s=0,\ldots,\log_2(n)-2 \right\} \tag{3.55}$$

mit jeweils einem betragsmäßig durch u beschränkten absoluten Fehler vorberechnet werden. Aus (3.53) erkennen wir jedoch, dass die Stabilitätskonstante (3.52) nicht unterboten wird.

(vi) Eine modifizierte diskrete Kosinustransformation vom Typ II (MDCT-II) wird in [43, Algorithmus 3.4] angegeben, welche auf einer lediglich dünnbesetzte Matrizen enthaltenden Faktorisierung beruht, die durch iterierte Anwendung von

$$\sqrt{n}C_n^{\mathrm{II}} \;=\; P_n^T\left(I_{n_1}\oplus\sqrt{2}P_{n_1}^T A_{n_1}(1)\right)\left(\sqrt{n_1}C_{n_1}^{\mathrm{II}}\oplus\left(\sqrt{n_2}C_{n_2}^{\mathrm{II}}\oplus\sqrt{n_2}C_{n_2}^{\mathrm{II}}\right)T_{n_1}(1)\right)\sqrt{2}T_n(0) \tag{3.56}$$

entsteht. Einerseits vermeiden wir hier die skalierten Drehungen wie bei (v) und wegen $\sqrt{2}P_{n_1}^T A_{n_1}(1)$ (vgl. (1.17)) treten weitere reine Butterfly-Operationen auf. Jedoch sind die beteiligten Matrizen mit wenigen Ausnahmen nicht einmal mehr fast orthogonal. Demzufolge ist die in [43, Theorem 5.6] ermittelte Stabilitätskonstante nur noch in der Größenordnung $\sqrt{n}\log_2(n)$. □

In diesem Unterabschnitt haben wir die Abschätzungen aus dem vorangegangenen Unterabschnitt noch weiter verschärfen können, in dem wir für Blockdiagonalmatrizen mit identischen Blöcken der Gestalt (3.37) unter Berücksichtigung von (3.38) bessere komponentenweise zu lesende Ungleichungen verwenden. Analog Lemma 3.7 in Unterabschnitt 3.1.1 sind die optimalen Ungleichungen in Lemma 3.14 die wesentlichen Werkzeuge, um die Hauptergebnisse in Lemma 3.15 und schließlich in Satz 3.16 zu erhalten. Die gesuchten Stabilitätskonstanten für die in Abschnitt 1.3 vorgestellten Algorithmen finden sich in Satz 3.18. Darüber hinaus liefert (3.46) mit den Bezeichnungen aus Satz 3.16 kleinere Schranken sowohl im Vergleich zu [54, Theorem 8.3], obwohl die Voraussetzungen für den betrachteten Fall angepasst sind, als auch im Vergleich zu [43, Theorem 5.3 und 5.4]. Abschließend sind die Konstanten aus den Sätzen 3.12 und 3.18 sowie aus (3.26), (3.28) und (3.47) in Tabelle 3.2 für ausgewählte t bzw. n zusammengestellt. Nach einem Vergleich mit Tabelle 3.1 gelangen wir zu dem Schluss, dass die Anwendung der in Abschnitt 1.3 vorgestellten schnellen Algorithmen im Hinblick auf die Genauigkeit selbst schon für kleine n einer „naiven“ Implementierung vorzuziehen ist.

Stabilitätskonstanten im Fall $\mathbf{x} = \hat{\mathbf{x}}$		Satz 3.12	(3.26)	(3.28)	Satz 3.18	(3.47)
DCT-II(n) mit Alg. 1.19	k_8	14.6969	13.6569	12.0660	13.8617	11.4472
	k_{16}	22.0454	20.4853	18.0990	20.7926	17.1708
sowie	k_{32}	29.3939	27.3137	24.1321	27.7234	22.8945
	k_{64}	36.7423	34.1421	30.1651	34.6543	28.6181
DCT-III(n) mit Alg. 1.20	k_{128}	44.0908	40.9706	36.1981	41.5851	34.3417
DST-III(n) mit Alg. 1.21	k_{256}	51.4393	47.7990	42.2311	48.5160	40.0653
DST-II(n) mit Alg. 1.22	k_{512}	58.7878	54.6274	48.2641	55.4468	45.7889
	k_{1024}	66.1362	61.4558	54.2971	62.3777	51.5125
mit $\nu = 2(t-1)$	k_{2048}	73.4847	68.2843	60.3302	69.3085	57.2361
bzw. $\mu = t-1$	k_{4096}	80.8332	75.1127	66.3632	76.2394	62.9598
DCT-IV(n) mit Alg. 1.15	k_8	18.3712	17.0711	15.0825	17.5359	14.4637
Alg. 1.18	k_{16}	25.7196	23.8995	21.1156	24.4668	20.1874
Alg. 1.19	k_{32}	33.0681	30.7279	27.1486	31.3976	25.9110
Alg. 1.20	k_{64}	40.4166	37.5563	33.1816	38.3285	31.6346
sowie	k_{128}	47.7650	44.3848	39.2146	45.2593	37.3582
	k_{256}	55.1135	51.2132	45.2476	52.1902	43.0818
DST-IV(n) mit Alg. 1.21	k_{512}	62.4620	58.0416	51.2806	59.1210	48.8054
Alg. 1.22	k_{1024}	69.8105	64.8701	57.3136	66.0519	54.5290
mit $\nu = 2t-1$	k_{2048}	77.1589	71.6985	63.3467	72.9827	60.2527
bzw. $\mu = t-1$	k_{4096}	84.5074	78.5269	69.3797	79.9136	65.9763
DCT-III(n) mit Alg. 1.16	k_8	22.0454	20.4853	18.0990	20.3749	16.8614
	k_{16}	33.0681	30.7279	27.1486	30.5624	25.2922
sowie	k_{32}	44.0908	40.9706	36.1981	40.7499	33.7229
	k_{64}	55.1135	51.2132	45.2476	50.9373	42.1536
DCT-II(n) mit Alg. 1.17	k_{128}	66.1362	61.4558	54.2971	61.1248	50.5843
	k_{256}	77.1589	71.6985	63.3467	71.3123	59.0150
mit $\nu = 3(t-1)$	k_{512}	88.1816	81.9411	72.3962	81.4997	67.4458
bzw. $\mu = 2(t-1)$	k_{1024}	99.2043	92.1838	81.4457	91.6872	75.8765

Tabelle 3.2: Stabilitätskonstanten im Fall $\mathbf{x} = \hat{\mathbf{x}}$ für die DCT- und DST-Algorithmen 1.15 – 1.22.

3.2 Stochastische Rundungsfehleranalyse

Die in Abschnitt 3.1 erhaltenen Stabilitätskonstanten basieren auf der Verwendung oberer Schranken für die bei jedem Rechenschritt auftretenden relativen Rundungsfehler. Im Allgemeinen können wir jedoch davon ausgehen, dass selbst im schlechtesten Fall die oberen Schranken nicht bei jedem Rechenschritt angenommen werden. Somit werden die Abschätzungen aus Abschnitt 3.1 erwartungsgemäß pessimistischer ausfallen als notwendig. Daher interessiert uns neben oberen Schranken auch eine Vorhersage für die mittlere Fehlernorm sowie ein zugehöriges Konfidenzintervall. Beides erhalten wir gegebenenfalls im Zuge einer stochastischen Rundungsfehleranalyse. Dazu werden sowohl alle auftretenden relativen Rundungsfehler als auch alle verwendeten Gleitkommazahlen als reelle Zufallsvariablen mit endlichem Erwartungswert und endlicher Varianz aufgefasst. Die genaue Definition sowie die Eigenschaften von reellen Zufallsvariablen sind im Anhang A.1 zu finden.

Bemerkung 3.20. Über die spezielle Verteilung der auftretenden relativen Fehler ist wenig bekannt. Daher wird beispielsweise in [8] anstelle des relativen Fehlers (dort mit r_3 bezeichnet) der standardisierte Fehler (dort mit r_1 bezeichnet) und der fraktionale Fehler (dort mit r_2 bezeichnet) untersucht, der bei der Multiplikation (einschließlich erneuter Normalisierung) zweier normalisierter Gleitkomma-Zahlen auftritt. Die entsprechenden Angaben zu Erwartungswerten $\mu+\mathcal{O}(u^3)$ und Varianzen $\mathbb{V}+\mathcal{O}(u^3)$ in Abhängigkeit von der gewählten Rundungsart sowie vom Zeitpunkt der Normalisierung finden sich in [8, Theorem 2.1] und der zugehörigen Tabelle. Sind die Exponenten der Faktoren fest gewählt, erhalten wir sofort auch Schätzwerte für Erwartungswert und Varianz des zugehörigen absoluten Fehlers.

Der relative Fehler erfüllt zwar aufgrund der Normalisierung der Mantissen der einzelnen Faktoren die Beziehung

$$r_1 \;\leq\; r_3 \;\leq\; \beta^2 r_1 \;,$$

jedoch sind die entsprechenden zugrunde liegenden Wahrscheinlichskeitsdichten völlig verschieden. Ebenso wird in [75, Remark 2.2] eingeräumt, dass wir über den fraktionalen Fehler η, definiert durch

$$\mathrm{fl}(g_1 \bullet g_2) \;=\; (g_1 \bullet g_2) + \zeta(g_1 \bullet g_2)\eta \qquad (|\eta| \;\leq\; u), \qquad \zeta(g) \;:=\; \begin{cases} \operatorname{sign}(g) \max\limits_{z\in\mathbb{Z}\,:\, z\leq \log_2(|g|)} 2^z \;, & g \neq 0, \\ 0 \;, & g = 0, \end{cases}$$

ein realistischeres Modell als (2.6) erhalten, sich der entsprechende Erwartungswert sowie die Varianz jedoch für nicht wenige Klassen von Verteilungen für den relativen Fehler jeweils nur um einen Faktor nahe Eins unterscheiden. Dabei fällt auf, dass sich die Definitionen des fraktionalen Fehlers in [8] und [75] um eine Zweierpotenz unterscheiden. □

Zunächst definieren wir analog zu Abschnitt 3.1 bzw. an Abbildung 3.1 orientierend die durchschnittliche numerische (Rückwärts-)Stabilität mit Hilfe des Erwartungswertes in Anlehnung an [54].

Definition 3.21 (vgl. [54], Abschnitt 8.1)**.** Ein Algorithmus für eine Matrix-Vektor-Multiplikation $C\mathbf{X} \in R_{\mathbb{G}}^n$ mit regulärer Matrix $C \in R_{\mathbb{G}}^{n\times n}$ und mit einem Zufallsvektor $\mathbf{X} \in R_{\mathbb{G}}^n$ heißt *durchschnittlich normweise rückwärtsstabil*, falls eine Konstante $\breve{k}_n > 0$ mit $\breve{k}_n u \ll 1$ und

$$\mathbb{E}\left(\| \Delta\mathbf{X} \|_2^2\right) \;\leq\; \left(\breve{k}_n^2 u^2 + \mathcal{O}\left(u^3\right)\right) \mathbb{E}\left(\| \mathbf{X} \|_2^2\right) \tag{3.57}$$

existiert. Dabei bezeichnet $\Delta\mathbf{X}$ den Rückwärtsfehler gemäß Abbildung 3.1.

Ohne uns auf eine genaue Verteilung der einzelnen Zufallsgrößen festzulegen, haben wir in Abschnitt 2.2 nach Möglichkeiten gesucht, ein auf (2.6) abgestimmtes stochastisches Modell zu entwickeln, welches wir auf die Algorithmen aus Abschnitt 1.3 anwenden können. Die zunächst in Unterabschnitt 2.2.1 geforderte Annahme der stochastischen Unabhängigkeit an die Eingangsdaten erweist sich als ungeeignet, da sie im Allgemeinen bereits nach einer Matrix-Vektor-Multiplikation nicht mehr gewährleistet werden kann. Überdies ist der Fall stochastisch unabhängiger Daten für die Anwendung – beispielsweise im Hinblick auf Grauwerte eines natürlichen Bildes – eher uninteressant. Unter Hinzunahme der speziellen Blockdiagonalgestalt der auftretenden Matrizen hat sich der bereits von Zeuner [75] verwendete Ansatz zur Modellierung des bei der Multiplikation zweier komplexer Zahlen auftretenden relativen Fehlers als günstig erwiesen. Auf einer Variation von [75, Proposition 2.3] basierend haben wir in Unterabschnitt 2.2.2 durch Satz 2.26 und durch die abschließende Bemerkung 2.27 alle wesentlichen Werkzeuge zu einer stochastischen Rundungsfehleranalyse in Gleitkomma-Arithmetik zur Verfügung. Die entsprechenden Voraussetzungen wollen wir in folgendem Modell zusammenfassen.

Modell 3.22. Sei $t \in \mathbb{N}$ und $n = 2^t$. Eine Matrix-Vektor-Multiplikation $A\mathbf{X} \in R_{\mathbb{G}}^n$ für einen Zufallsvektor $\mathbf{X} = (X_k)_{k=0}^{n-1}$ mit Werten in $\mathbb{G}^n$, die auf einer Faktorisierung der Gestalt

$$A \;=\; \prod_{m=1}^{\nu} A^{(m)} \;:=\; A^{(\nu)} \ldots A^{(2)} A^{(1)} \tag{3.58}$$

mit Faktoren $A^{(m)} \in \mathbb{G}^{n\times n}$ basiert, genüge den folgenden Annahmen.

(M1) Die Matrizen $A^{(m)} \in \mathbb{G}^{n\times n}$ besitzen bis auf Permutationen und Vorzeichenskalierungen Blockdiagonalgestalt, d.h., es existieren Matrizen $U^{(m)},\ V^{(m)} \in \mathbb{G}^{n\times n}$, welche lediglich Permutationen und zeilen- bzw. spaltenweise Vorzeichenskalierungen bewirken, und Matrizen

$$B^{(m)} \;=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(m,k)} \qquad (m = 1, \ldots, \nu) \tag{3.59}$$

mit Blöcken

$$A^{(m,k)} \;:=\; \begin{pmatrix} G_1^{(m,k)} & G_2^{(m,k)} \\ -G_2^{(m,k)} & G_1^{(m,k)} \end{pmatrix} \qquad \left(k = 0, \ldots, \tfrac{n}{2} - 1, \quad m = 1, \ldots, \nu\right) \;, \tag{3.60}$$

so dass $A^{(m)} = U^{(m)} B^{(m)} V^{(m)}$ gilt.

(M2) Für $k = 0, \dots, \frac{n}{2} - 1, \quad m = 1, \dots, \nu$ existieren reelle Zufallsvariablen $\varepsilon^{(m,k)}_{\|\cdot\|_2^2}$ mit

$$\left(G_1^{(m,k)}\right)^2 + \left(G_1^{(m,k)}\right)^2 \;=\; 1 + \varepsilon^{(m,k)}_{\|\cdot\|_2^2} \qquad \left(\left|\varepsilon^{(m,k)}_{\|\cdot\|_2^2}\right| \le 2u + u^2\right) \;, \tag{3.61}$$

und es gelte

$$\mathbb{E}\left(\varepsilon^{(m,k)}_{\|\cdot\|_2^2}\right) = \;\; 0 \tag{3.61a}$$

oder

$$\varepsilon^{(m,k)}_{\|\cdot\|_2^2} \le \;\; 0. \tag{3.61b}$$

(M3) Zu den Blockdiagonalmatrizen $B^{(m)}$ und den rekursiv durch

$$\begin{array}{lcllcll} \mathbf{Y}^{(0)} & := & V^{(1)}\mathbf{X} \;, & & \mathbf{Z}^{(1)} & := & \mathrm{fl}\left(B^{(1)}\mathbf{Y}^{(0)}\right) \;, \\ \mathbf{Y}^{(m)} & := & V^{(m+1)}U^{(m)}\mathbf{Z}^{(m)} \;, & & \mathbf{Z}^{(m+1)} & := & \mathrm{fl}\left(B^{(m+1)}\mathbf{Y}^{(m)}\right) \;, \qquad (m = 1, \dots, \nu - 1) \\ \mathbf{Y}^{(\nu)} & := & U^{(\nu)}\mathbf{Z}^{(\nu)} & & & & \end{array} \tag{3.62}$$

definierten Zwischenergebnissen $\mathbf{Y}^{(m)}$ existieren paarweise unabhängige Zufallsgrößen

$$\varepsilon^{(m,k)}_{1,\times}, \varepsilon^{(m,k)}_{2,\times}, \varepsilon^{(m,k)}_{3,\times}, \varepsilon^{(m,k)}_{4,\times}, \varepsilon^{(m,k)}_{1,+}, \varepsilon^{(m,k)}_{2,+} \tag{3.63}$$

mit

$$\left.\begin{array}{lcllcll} \mathbb{E}(\varepsilon^{(m,k)}_{j,\times}) & = & 0 \;, & \mathbb{V}(\varepsilon^{(m,k)}_{j,\times}) & = & \sigma_\times^2 u^2 & (j = 1,2,3,4), \\ \mathbb{E}(\varepsilon^{(m,k)}_{j,+}) & = & 0 \;, & \mathbb{V}(\varepsilon^{(m,k)}_{j,+}) & = & \sigma_+^2 u^2 & (j = 1,2) \end{array}\right\} \tag{3.64}$$

und

$$\begin{aligned} &\mathrm{fl}\left(A^{(m,k)}\begin{pmatrix} Y^{(m-1)}_{2k} \\ Y^{(m-1)}_{2k+1}\end{pmatrix}\right) \\ &= \begin{pmatrix} (G_1^{(m,k)}Y^{(m-1)}_{2k} + G_2^{(m,k)}Y^{(m-1)}_{2k+1})(1+\varepsilon^{(m,k)}_{1,+}) + (G_1^{(m,k)}Y^{(m-1)}_{2k})\varepsilon^{(m,k)}_{1,\times} + (G_2^{(m,k)}Y^{(m-1)}_{2k+1})\varepsilon^{(m,k)}_{2,\times} \\ (G_1^{(m,k)}Y^{(m-1)}_{2k+1} - G_2^{(m,k)}Y^{(m-1)}_{2k})(1+\varepsilon^{(m,k)}_{2,+}) + (G_1^{(m,k)}Y^{(m-1)}_{2k+1})\varepsilon^{(m,k)}_{3,\times} - (G_2^{(m,k)}Y^{(m-1)}_{2k})\varepsilon^{(m,k)}_{4,\times} \end{pmatrix} \end{aligned}$$

für $k = 0, \dots, \frac{n}{2} - 1, \quad m = 1, \dots, \nu$, wobei jede der Größen $G_1^{(m,k)}, G_2^{(m,k)}, Y^{(m-1)}_{2k}, Y^{(m-1)}_{2k+1}$ von den Größen (3.63) unabhängig sei. □

Da Modell 3.22 genau die wesentlichen Voraussetzungen von Satz 2.22, Lemma 2.24 und Satz 2.26 in sich vereint, können wir nun Konstanten $\breve{k}_n$ gemäß Definition 3.21 für jeden der Algorithmen aus Abschnitt 1.3 angeben.

Satz 3.23. *Sei $t \ge 2, n = 2^t$. Genügt die Implementierung in Gleitkomma-Arithmetik allen Annahmen aus Modell* 3.22*, dann sind die Algorithmen* 1.15 – 1.22 *gemäß* (3.57) *durchschnittlich rückwärtsstabil mit Stabilitätskonstanten*

$$\breve{k}_n \;=\; \begin{cases} \sqrt{2^{\lceil \log_2(2(\log_2(n)-1))\rceil} \cdot 2(\log_2(n)-1) \cdot \left(\sigma_\times^2 + \sigma_+^2\right)} & \textit{für} \begin{cases} \text{DCT-II}(n) & \textit{mit Alg.}\ 1.19\ , \\ \text{DCT-III}(n) & \textit{mit Alg.}\ 1.20\ , \\ \text{DST-III}(n) & \textit{mit Alg.}\ 1.21\ , \\ \text{DST-II}(n) & \textit{mit Alg.}\ 1.22\ , \end{cases} \\ \sqrt{2^{\lceil \log_2(2\log_2(n)-1)\rceil} \cdot (2\log_2(n)-1) \cdot \left(\sigma_\times^2 + \sigma_+^2\right)} & \textit{für} \begin{cases} \text{DCT-IV}(n) & \textit{mit Alg.}\ 1.15\ , \\ \text{DCT-IV}(n) & \textit{mit Alg.}\ 1.18\ , \\ \text{DCT-IV}(n) & \textit{mit Alg.}\ 1.19\ , \\ \text{DCT-IV}(n) & \textit{mit Alg.}\ 1.20\ , \\ \text{DST-IV}(n) & \textit{mit Alg.}\ 1.21\ , \\ \text{DST-IV}(n) & \textit{mit Alg.}\ 1.22\ , \end{cases} \\ \sqrt{2^{\lceil \log_2(3(\log_2(n)-1))\rceil} \cdot 3(\log_2(n)-1) \cdot \left(\sigma_\times^2 + \sigma_+^2\right)} & \textit{für} \begin{cases} \text{DCT-III}(n) & \textit{mit Alg.}\ 1.16\ , \\ \text{DCT-II}(n) & \textit{mit Alg.}\ 1.17\ . \end{cases} \end{cases}$$

Beweis: Im ersten Fall basieren die Algorithmen 1.19 – 1.22 für $k = 0$ auf den Faktorisierungen (1.45) bzw. (1.55), welche jeweils aus $2\log_2(n) - 1$ Faktoren bestehen. Insbesondere bezeichnen $A_n(\boldsymbol{\beta}_0)$ und $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)$ Permutationsmatrizen. Wegen (1.43) bzw. (1.53) und aufgrund von (1.25) – (1.29) besitzen die restlichen $2\log_2(n) - 2$ Faktoren jeweils einen Repräsentanten gemäß Modellannahme (M1). Daher ergibt sich die erste Stabilitätskonstante als direkte Anwendung von Satz 2.26 mit $r = 2(\log_2(n) - 1)$.

Im zweiten Fall basieren die Algorithmen 1.15 und 1.18 auf den in Lemma 1.11 und Folgerung 1.13 angegebenen Faktorisierungen (1.32) und (1.37), deren Repräsentation in der Gleitkomma-Arithmetik jeweils genau $2\log_2(n) - 1$ Faktoren gemäß Modellannahme (M1) besitzen. Ebenso enthalten die Produkte (1.48) bzw. (1.60), welche den Algorithmen 1.19 – 1.22 für $k = 1$ zugrunde liegen, jeweils genau $2\log_2(n) - 1$ analog zu (1.43) bzw. (1.53) definierte Matrizen als Faktoren, von denen jeder einzelne der Modellannahme (M1) entspricht. Daher ergibt sich die zweite Stabilitätskonstante, wenn wir Satz 2.26 mit $r = 2\log_2(n) - 1$ anwenden.

Schließlich werden die Algorithmen 1.16 und 1.17 durch die Faktorisierungen (1.35) bzw. (1.38) beschrieben, so dass die Behauptung wiederum aus Satz 2.26 mit $r = 3(\log_2(n) - 1)$ folgt. ■

In Tabelle 3.3 sind nun die Stabilitätskonstanten k_n bzw. $\breve{k}_n$ aus Satz 3.12, Satz 3.18 und Satz 3.23 für ausgewählte n mit den in Anhang A.3 geschätzten Grössen $\sigma_\times = \sigma_+ = 0.425$ angegeben.

n	8	16	32	64	128	256	512	1024
G1: k_n aus Satz 3.12	14.6969	22.0454	29.3939	36.7423	44.0908	51.4393	58.7878	66.1362
k_n aus Satz 3.18	13.8617	20.7926	27.7234	34.6543	41.5851	48.5160	55.4468	62.3777
$\breve{k}_n$ aus Satz 3.23	2.4042	4.1641	4.8083	7.6026	8.3283	8.9956	9.6167	14.4250
G2: k_n aus Satz 3.12	18.3712	25.7196	33.0681	40.4166	47.7650	55.1135	62.4620	69.8105
k_n aus Satz 3.18	17.5359	24.4668	31.3976	38.3285	45.2593	52.1902	59.1210	66.0519
$\breve{k}_n$ aus Satz 3.23	3.8013	4.4978	7.2125	7.9737	8.6683	9.3113	14.0186	14.8203
G3: k_n aus Satz 3.12	22.0454	33.0681	44.0908	55.1135	66.1362	77.1589	88.1816	99.2043
k_n aus Satz 3.18	20.3749	30.5624	40.7499	50.9373	61.1248	71.3123	81.4997	91.6872
$\breve{k}_n$ aus Satz 3.23	4.1641	7.2125	8.3283	9.3113	14.4250	15.5808	16.6565	17.6669

Tabelle 3.3: Stabilitätskonstanten im Fall $\mathbf{x} = \hat{\mathbf{x}}$ für die DCT- und DST-Algorithmen 1.15 – 1.22 gemäß der Definitionen 3.1 und 3.21 sowie nach der Gruppeneinteilung aus Satz 3.23. Dabei werden in der Gruppe G1 die Algorithmen 1.19 – 1.22 für $k = 0$, in der Gruppe G2 die Algorithmen 1.16 und 1.17 und in der Gruppe G3 alle verbleibenden Algorithmen zusammengefasst.

Bemerkung 3.24 (Zusammenfassung von Kapitel 3)**.** Für die Algorithmen 1.15 – 1.22, mit denen die diskreten trigonometrischen Transformationen aus Definition 1.14 schnell ausgeführt werden können, haben wir in den Abschnitten 3.1 und 3.2 Stabilitätskonstanten k_n bzw. $\breve{k}_n$ gemäß Definitionen 3.1 und 3.21 ermittelt. Die Grundlage der Untersuchungen bilden dabei die den Algorithmen 1.15 – 1.22 zugrunde liegenden Faktorisierungen der Gestalt (3.12). Die Orthogonalität der beteiligten Matrizen wird dabei in großem Maße ausgenutzt. Für die deterministische Rundungsfehleranalyse sind die auf den neuen Ungleichungen (3.21) bzw. (3.40a) und (3.40b) basierenden Hauptergebnisse in den Sätzen 3.12 und 3.18 zu finden. Letzterer berücksichtigt, dass einige der Matrixfaktoren höchstens skalierte Butterfly-Matrizen enthalten, und verwendet die im Falle einer günstigen Implementierung gültige Abschätzung (3.39). Hierbei wird noch einmal deutlich, dass die in einem Zahlenkörper üblicherweise gültige Distributivität in Rechner-Arithmetiken keineswegs gültig sein muss.

Fordern wir von einer Implementierung der Algorithmen 1.15 – 1.22 darüber hinaus, dass alle Matrix-Einträge tabelliert sind und somit für das ε aus Satz 2.1 jeweils $|\varepsilon| \leq \frac{u}{2}$ gilt, dann liefern (3.27) bzw. (3.46) kleinere Konstanten im Vergleich zu der in [54, Theorem 8.3] ermittelten Größe (3.25).

Für die auf den Annahmen aus Modell 3.22 basierende stochastische Rundungsfehleranalyse sind die Hauptergebnisse in Satz 3.23 zusammengefasst. Neu ist hierbei, dass Modell 3.22 im Vergleich zu [54, (M3)] darauf verzichtet, an die Eingangsdaten stark einschränkende Forderungen wie Unkorreliertheit oder gar stochastische Unabhängigkeit zu erheben. □

4 Numerische Stabilität in Festkomma-Arithmetik

Ebenso wie in Gleitkomma-Arithmetik besitzen in Festkomma-Arithmetik nur endlich viele reelle Zahlen eine exakte Darstellung, so dass auch hier neben den Eingangsfehlern die bei arithmetischen Operationen auftretenden Rundungsfehler zu berücksichtigen sind. Aus diesem Grund untersuchen wir in diesem Kapitel das Rundungsfehlerverhalten der Algorithmen 1.15 – 1.22 in Festkomma-Arithmetik. Dazu betrachten wir erneut den Vorwärts- bzw. Rückwärtsfehler aus Abbildung 3.1 und führen – um wiederum ein Vergleichskriterium zur Verfügung zu haben – in Definition 4.6 den Begriff der numerischen Stabilität in Festkomma-Arithmetik ein.

4.1 Deterministische Rundungsfehleranalyse

In diesem Abschnitt werden wir das in Abschnitt 2.3 eingeführte Modell zur Festkomma-Arithmetik zunächst auf Matrix-Vektor-Multiplikationen anwenden, bei welchen die Matrizen direkte Summen von Drehmatrizen $Q_2(\varphi)$ mit einem Drehwinkel $\varphi \in \left]0, \frac{\pi}{4}\right]$ sind. Aus den Faktorisierungen (1.25) – (1.29) ist ersichtlich, dass diese Betrachtungen genügen, um daraufhin Abschätzungen für die in Abschnitt 1.3 angegebenen Algorithmen herzuleiten.

4.1.1 Fehlerabschätzungen für Matrix-Vektor-Multiplikationen

Im Abschnitt 2.3 haben wir bereits im Fall allgemeiner Matrizen $\hat{A} \in \mathbb{M}_q^{n\times n}$ und Vektoren $\hat{\mathbf{x}} \in \mathbb{M}_q^n$ Abschätzungen für die entstehenden Rundungsfehler $\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}} := \hat{A}\hat{\mathbf{x}} - \hat{A}\times\hat{\mathbf{x}}$ beim einfach genauen Matrix-Vektor-Pseudo-Produkt und $\boldsymbol{\delta}_{\hat{A}\odot\hat{\mathbf{x}}} := \hat{A}\hat{\mathbf{x}} - \hat{A}\odot\hat{\mathbf{x}}$ beim doppelt genauen Matrix-Vektor-Pseudo-Produkt angegeben, solange kein Überlauf auftritt. Von Interesse sind nun die Gesamtfehler

$$\boldsymbol{\delta}_{A\mathbf{x}} \;:=\; A\mathbf{x} \;-\; \hat{A}\times\hat{\mathbf{x}}\,, \tag{4.1}$$

$$\boldsymbol{\delta}^{\odot}_{A\mathbf{x}} \;:=\; A\mathbf{x} \;-\; \hat{A}\odot\hat{\mathbf{x}}\,. \tag{4.2}$$

Wir betrachten zunächst den Fall des einfach genauen Matrix-Vektor-Pseudo-Produktes. Offenbar lässt sich der Fehlervektor $\boldsymbol{\delta}_{A\mathbf{x}}$ in der Form

$$\boldsymbol{\delta}_{A\mathbf{x}} \;=\; A\,(\mathbf{x}-\hat{\mathbf{x}}) \;+\; \left(A-\hat{A}\right)\hat{\mathbf{x}} \;+\; \boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}} \tag{4.3}$$

ausdrücken und für seine Norm die folgenden Abschätzungen angeben.

Lemma 4.1. *Seien $n \in \mathbb{N}$, $A \in [-1,1]^{n\times n}$ und $\mathbf{x} \in [-1,1]^n$ derart, dass jede Zeile $\mathbf{y}^{\mathrm{T}}$ von A der Bedingung*

$$\sum_{i=0}^{n-1} |x_i y_i| \le 1 \tag{4.4}$$

genügt. Dann existieren für den in (4.1) *definierten Gesamtfehler $\boldsymbol{\delta}_{A\mathbf{x}}$ für das Matrix-Vektor-Produkt Ax die normweisen Abschätzungen*

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty \;\le\; (\|A\|_\infty \;+\; n\|\mathbf{x}\|_\infty + n)\,u\,, \tag{4.5a}$$

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 \;\le\; \left(\|A\|_2 \;+\; \sqrt{n}\|\mathbf{x}\|_2 + n\right)\sqrt{n}u\,, \tag{4.5b}$$

wobei $\|A\|_\infty$ die Zeilensummennorm und $\|A\|_2$ die Spektralnorm der Matrix A bezeichnet.

Beweis: Die Voraussetzung (4.4) für $\mathbf{x}$ und jede Zeile $\mathbf{y}^{\mathrm{T}}$ von A impliziert wegen (2.52) sofort die Gültigkeit von (2.55) für $\hat{\mathbf{x}} := \mathrm{fix}(\mathbf{x})$ und jede Zeile $\hat{\mathbf{y}}^{\mathrm{T}} := \mathrm{fix}\left(\mathbf{y}^{\mathrm{T}}\right)$ von $\hat{A} := \mathrm{fix}(A)$. Demzufolge kann Überlauf ausgeschlossen und Folgerung 2.29 angewandt werden. Mit (2.56) und wegen $|a_{ij} - \hat{a}_{ij}| \le$

$u,\ i,j = 0,\dots,n-1$, sowie $|x_k - \hat{x}_k| \le u,\ |\hat{x}_k| \le |x_k|,\ k = 0,\dots,n-1$, ergeben sich dann mit der Dreiecksungleichung für die Maximumnorm

$$\begin{aligned}\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty &\le \|A\,(\mathbf{x}-\hat{\mathbf{x}})\,\|_\infty + \|(A-\hat{A})\hat{\mathbf{x}}\|_\infty + \|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_\infty \\ &\le \|A\|_\infty\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty \ + \ \|A-\hat{A}\|_\infty\|\hat{\mathbf{x}}\|_\infty \ + \ nu \\ &\le (\|A\|_\infty \ + \ n\|\mathbf{x}\|_\infty + n)\,u\end{aligned}$$

und ebenso für die euklidische Norm

$$\begin{aligned}\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 &\le \|A\,(\mathbf{x}-\hat{\mathbf{x}})\,\|_2 + \|(A-\hat{A})\hat{\mathbf{x}}\|_2 + \|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_2 \\ &\le \|A\|_2\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \ + \ \|A-\hat{A}\|_2\|\hat{\mathbf{x}}\|_2 \ + \ n^{\frac{3}{2}}u \\ &\le \left(\|A\|_2 \ + \ \sqrt{n}\|\mathbf{x}\|_2 + n\right)\sqrt{n}u\ .\end{aligned}$$

Im letzten Schritt haben wir dabei verwendet, dass

$$\|\mathbf{c}\|_\infty \ = \ |c|\ , \qquad \|\mathbf{c}\|_2 \ = \ \sqrt{n}|c|$$

für einen Vektor $\mathbf{c} = c\mathbf{1} \in \mathbb{R}^n$ und entsprechend

$$\|C\|_\infty \ = \ \|C\|_2 \ = \ n|c|$$

für eine Matrix $C \in \mathbb{R}^{n\times n}$ mit $C = (c)_{j,k=0}^{n-1}$, also mit konstanten Einträgen c gilt. ■

Offenbar lassen sich die Ungleichungen (4.5a) und (4.5b) im Falle von dünnbesetzten Matrizen in den einzelnen Summanden noch erheblich verbessern. Uns interessieren insbesondere Matrix-Vektor-Multiplikationen mit einer direkten Summe von Drehmatrizen $Q_2(\varphi)$. Mit (2.64a) ergibt sich demnach

Satz 4.2. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und $\varphi_k \in \left]0, \frac{\pi}{4}\right]$ $(k = 0,\dots,n_1-1)$. Weiter sei $\mathbf{x} \in [-1,1]^n$, so dass alle Teilvektoren $\mathbf{x}^{(k)} := (x_{2k}, x_{2k+1})^{\mathrm{T}}$ $(k = 0,\dots,n_1-1)$ der Abschätzung $\|\mathbf{x}^{(k)}\|_2 \le 1$ genügen. Dann erfüllt der in* (4.1) *definierte Gesamtfehler $\boldsymbol{\delta}_{A\mathbf{x}}$ für das Matrix-Vektor-Produkt $A\mathbf{x}$ mit der Matrix*

$$A \ := \ \bigoplus_{k=0}^{n_1-1} Q_2(\varphi_k) \tag{4.6}$$

die normweisen Abschätzungen

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty \ \le \ (\|A\|_\infty \ + \ 2\|\mathbf{x}\|_\infty + 2)\,u\ , \tag{4.7a}$$

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 \ \le \ \left(\sqrt{n} \ + \ \sqrt{2}\|\mathbf{x}\|_2 + \sqrt{5n_1}\right)u \tag{4.7b}$$

mit $\|A\|_\infty := \sin(\alpha) + \cos(\alpha) \le \sqrt{2}$, wobei $\alpha := \max\limits_{k=0,\dots,n_1-1} \varphi_k$, sowie mit $\|\mathbf{x}\|_\infty \le 1$ und $\|\mathbf{x}\|_2 \le \sqrt{n_1}$.

Beweis: Aufgrund der Orthogonalität von Drehmatrizen gilt zunächst

$$\|Q_2(\varphi_k)\|_2 \ = \ \sqrt{(\cos(\varphi_k))^2 + (\sin(\varphi_k))^2} \ = \ 1 \qquad (k = 0,\dots,n_1-1). \tag{4.8}$$

Mittels Cauchy-Schwarz-Ungleichung resultieren dann für $k = 0,\dots,n_1-1$ die Abschätzungen

$$\begin{aligned}|\cos(\varphi_k)||x_{2k}| + |\sin(\varphi_k)||x_{2k+1}| &\le \|\mathbf{x}^{(k)}\|_2\ , \\ |-\sin(\varphi_k)||x_{2k}| + |\cos(\varphi_k)||x_{2k+1}| &\le \|\mathbf{x}^{(k)}\|_2\ .\end{aligned}$$

Es folgt nun, dass die Bedingung $\|\mathbf{x}^{(k)}\|_2 \le 1$ $(k = 0,\dots,n_1-1)$ hinreichend ist, um (4.4) zu gewährleisten, also insbesondere um Überlauf ausschließen zu können. Weiterhin ergeben sich aufgrund der speziellen Gestalt (4.6) von A für die entsprechenden Normen

$$\|A\|_2 \ = \ \max_{k=0,\dots,n_1-1} \|Q_2(\varphi_k)\|_2 \ = \ 1$$

und

$$\|A\|_\infty \;=\; \max_{k=0,\ldots,n_1-1} \|Q_2(\varphi_k)\|_\infty \;=\; \max_{k=0,\ldots,n_1-1} (|\sin(\varphi_k)| + |\cos(\varphi_k)|) \;\leq\; \sqrt{2}\;,$$

wobei im letzten Schritt die Cauchy-Schwarz-Ungleichung auf die Vektoren $(1,1)^{\mathrm{T}}$ und $(\cos(\varphi),\sin(\varphi))^{\mathrm{T}}$ mit beliebigem $\varphi \in \left]0,\frac{\pi}{4}\right]$ angewandt worden ist. Da die Funktion $f(x) = \sin(x) + \cos(x)$ auf $\left]0,\frac{\pi}{4}\right[$ wegen $f'(x) = \cos(x) - \sin(x) > 0$ auf $\left]0,\frac{\pi}{4}\right[$ streng monoton wachsend ist, haben wir einerseits Gleichheit genau dann, wenn $\varphi_k = \frac{\pi}{4}$ für mindestens ein $k \in \{0,\ldots,n_1-1\}$ erfüllt ist, und andererseits

$$\max_{k=0,\ldots,n_1-1} (|\sin(\varphi_k)| + |\cos(\varphi_k)|) \;=\; \sin(\alpha) + \cos(\alpha)\;.$$

Weiterhin ist die Differenzmatrix $A - \hat{A}$ wegen (2.52) und der speziellen Gestalt (4.6) von A eine direkte Summe von fast orthogonalen Matrizen mit Einträgen, welche betragsmäßig kleiner oder gleich u sind. Zusammen mit (2.60) ergeben sich für die Normen dann analog

$$\|A - \hat{A}\|_2 \;\leq\; \sqrt{2}u\;, \qquad \|A - \hat{A}\|_\infty \;\leq\; 2u\;.$$

Schließlich erhalten wir aus (4.3) wegen $|x_k - \hat{x}_k| \leq u$, $|\hat{x}_k| \leq |x_k|$, $k = 0,\ldots,n-1$, und wiederum mittels Dreiecksungleichung die Abschätzungen

$$\begin{aligned}
\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty &\leq \|A\,(\mathbf{x} - \hat{\mathbf{x}})\,\|_\infty + \|(A - \hat{A})\hat{\mathbf{x}}\|_\infty + \|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_\infty \\
&\leq (\|A\|_\infty \;+\; 2\|\mathbf{x}\|_\infty + 2)\,u \\
&\leq (\sin(\alpha) + \cos(\alpha) + 4)\,u
\end{aligned}$$

sowie

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 \;\leq\; \|A\,(\mathbf{x} - \hat{\mathbf{x}})\,\|_2 + \|(A - \hat{A})\hat{\mathbf{x}}\|_2 + \|\boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}\|_2 \;\leq\; \left(\sqrt{n} \;+\; \sqrt{2}\|\mathbf{x}\|_2 + \sqrt{5n_1}\right) u\;,$$

wobei die Ergebnisse aus (2.64a) mit $\boldsymbol{\delta}^\times = \boldsymbol{\delta}_{\hat{A}\times\hat{\mathbf{x}}}$ eingegangen sind. Mit der Voraussetzung $\|\mathbf{x}^{(k)}\|_2 \leq 1$ für $k = 0,\ldots,n_1-1$ ergibt sich wegen

$$\|\mathbf{x}\|_2^2 \;=\; \sum_{k=0}^{n_1-1} \|\mathbf{x}^{(k)}\|_2^2 \tag{4.9}$$

zusätzlich $\|\mathbf{x}\|_2 \;\leq\; \sqrt{n_1}$. ■

Der letzte Schritt im Beweis von Lemma 4.1 impliziert, dass sich die euklidische Norm des in (4.1) definierten Gesamtfehlers durch

$$\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 \;\leq\; (2\sqrt{2} + \sqrt{5})\sqrt{n_1}u \tag{4.10}$$

abschätzen lässt.
Nun wollen wir uns dem Fall des doppelt genauen Matrix-Vektor-Pseudo-Produktes zuwenden. Analog zu (4.3) lässt sich der in (4.2) definierte Gesamtfehler $\boldsymbol{\delta}^{\odot}_{A\mathbf{x}}$ in der Form

$$\boldsymbol{\delta}^{\odot}_{A\mathbf{x}} \;=\; A\,(\mathbf{x} - \hat{\mathbf{x}}) \;+\; \left(A - \hat{A}\right)\hat{\mathbf{x}} \;+\; \boldsymbol{\delta}_{\hat{A}\odot\hat{\mathbf{x}}} \tag{4.11}$$

ausdrücken und für seine Norm analog wie in Lemma 4.1 die folgenden Abschätzungen angeben.

Lemma 4.3. *Seien $n \in \mathbb{N}$, $A \in [-1,1]^{n\times n}$ und $\mathbf{x} \in [-1,1]^n$ derart, dass jede Zeile $\mathbf{y}^{\mathrm{T}}$ von A der Bedingung* (4.4) *genügt. Dann existieren für den in* (4.2) *definierten Gesamtfehler $\boldsymbol{\delta}^{\odot}_{A\mathbf{x}}$ für das Matrix-Vektor-Produkt $A\mathbf{x}$ die normweisen Abschätzungen*

$$\|\boldsymbol{\delta}^{\odot}_{A\mathbf{x}}\|_\infty \;\leq\; (\|A\|_\infty \;+\; n\|\mathbf{x}\|_\infty + 1)\,u\;, \tag{4.12a}$$

$$\|\boldsymbol{\delta}^{\odot}_{A\mathbf{x}}\|_2 \;\leq\; \left(\|A\|_2 \;+\; \sqrt{n}\|\mathbf{x}\|_2 + 1\right)\sqrt{n}u\;, \tag{4.12b}$$

wobei $\|A\|_\infty$ die Zeilensummennorm und $\|A\|_2$ die Spektralnorm der Matrix A bezeichnet.

Der **Beweis** verläuft analog zu Lemma 4.1, wobei (2.64b) anstelle von (2.64a) verwendet wird. ■

Wie die Ungleichungen (4.5a) und (4.5b) lassen sich die Abschätzungen (4.12a) und (4.12b) im Falle von dünnbesetzten Matrizen weiter verbessern. Für Matrix-Vektor-Multiplikationen mit einer direkten Summe von Drehmatrizen $Q_2(\varphi)$ ergibt sich demnach

Satz 4.4. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und $\varphi_k \in \left]0, \frac{\pi}{4}\right]$ $(k = 0, \ldots, n_1 - 1)$. Weiter sei $\mathbf{x} \in [-1, 1]^n$, so dass alle Teilvektoren $\mathbf{x}^{(k)} := (x_{2k}, x_{2k+1})^{\mathrm{T}}$ $(k = 0, \ldots, n_1 - 1)$ der Abschätzung $\|\mathbf{x}^{(k)}\|_2 \leq 1$ genügen. Dann erfüllt der in* (4.2) *definierte Gesamtfehler $\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}$ für das Matrix-Vektor-Produkt $A\mathbf{x}$ mit der Matrix A, definiert wie in* (4.6)*, die normweisen Abschätzungen*

$$\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_\infty \;\leq\; (\|A\|_\infty \;+\; 2\|\mathbf{x}\|_\infty + 1)\, u \;, \tag{4.13a}$$

$$\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_2 \;\leq\; \left(2\sqrt{n} \;+\; \sqrt{2}\|\mathbf{x}\|_2\right) u \tag{4.13b}$$

mit $\|A\|_\infty := \sin(\alpha) + \cos(\alpha) \leq \sqrt{2}$, wobei $\alpha := \max\limits_{k=0,\ldots,n_1-1} \varphi_k$, sowie mit $\|\mathbf{x}\|_\infty \leq 1$ und $\|\mathbf{x}\|_2 \leq \sqrt{n_1}$.

Der **Beweis** ergibt sich analog zu Satz 4.2, wobei (2.58) anstelle von (2.56) verwendet wird. ■

Verwenden wir wiederum, dass sich aus den Voraussetzungen die Abschätzung $\|\mathbf{x}\|_2 \leq \sqrt{n_1}$ ergibt, ist

$$\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_2 \;\leq\; 3\sqrt{n}\, u \tag{4.14}$$

die zu (4.10) analoge Aussage für die euklidische Norm des in (4.2) definierten Gesamtfehlers $\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}$.

In den folgenden Unterabschnitten wollen wir die Ergebnisse der Sätze 4.2 und 4.4 auf die Faktorisierungen, welche den Algorithmen 1.15 – 1.22 zugrunde liegen, wiederholt anwenden. Dabei tritt die Schwierigkeit auf, dass die Bedingung $\|\mathbf{x}^{(k)}\|_2 \leq 1$ für $\mathbf{x}^{(k)} := (x_{2k}, x_{2k+1})^{\mathrm{T}}$ $(k = 0, \ldots, n_1 - 1)$ nach einer Permutation der Komponenten von $\mathbf{x}$ verletzt sein kann. Da für eine beliebige orthogonale Matrix $Q \in \mathbb{R}^{n\times n}$ stets $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ gilt, kann jedoch durch die Bedingung

$$\|\mathbf{x}\|_2 \;\leq\; 1 \tag{4.15}$$

sicher gestellt werden, dass bei einem mehrfachen Matrix-Vektor-Produkt

$$\mathbf{y} \;=\; Q^{(\nu)} \left(\ldots \left(Q^{(1)} \mathbf{x} \right) \ldots \right)$$

mit orthogonalen Matrizen $Q^{(1)}, \ldots, Q^{(\nu)} \in \mathbb{R}^{n\times n}$ wegen (4.9) die Voraussetzungen der Sätze 4.2 und 4.4 für jedes Zwischenergebnis $\mathbf{y}^{(j)} := Q^{(j)}\mathbf{y}^{(j-1)}$, $j = 1, \ldots, \nu - 1$ mit $\mathbf{y}^{(0)} := \mathbf{x}$ erfüllt bleiben.
Abschließend formulieren wir demnach die Hauptergebnisse für direkte Summen von Drehmatrizen mit der Voraussetzung (4.15).

Satz 4.5. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und $\varphi_k \in \left]0, \frac{\pi}{4}\right]$ $(k = 0, \ldots, n_1 - 1)$. Weiter sei*

$$\alpha := \max_{k=0,\ldots,n_1-1} \varphi_k \tag{4.16}$$

und gelte (4.15) *für $\mathbf{x} \in [-1, 1]^n$. Dann erfüllen die in* (4.1) *und* (4.2) *definierten Gesamtfehler $\boldsymbol{\delta}_{A\mathbf{x}}$ und $\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}$ für das Matrix-Vektor-Produkt $A\mathbf{x}$ mit der Matrix A, definiert wie in* (4.6)*, die normweisen Abschätzungen*

$$\left.\begin{array}{lcll}
\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty & \leq & (\cos(\alpha) + \sin(\alpha) \;+\; 2\|\mathbf{x}\|_\infty + 2)\, u & \leq \quad (\sqrt{2} + 4)u \;, \\
\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_\infty & \leq & (\cos(\alpha) + \sin(\alpha) \;+\; 2\|\mathbf{x}\|_\infty + 1)\, u & \leq \quad (\sqrt{2} + 3)u \;, \\
\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 & \leq & \left(\sqrt{n} \;+\; \sqrt{2}\|\mathbf{x}\|_2 + \sqrt{\frac{5n}{2}}\right) u & \leq \quad \left(\sqrt{2} + \left(1 + \sqrt{\frac{5}{2}}\right)\sqrt{n}\right) u \;, \\
\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_2 & \leq & \left(2\sqrt{n} \;+\; \sqrt{2}\|\mathbf{x}\|_2\right) u & \leq \quad \left(\sqrt{2} + 2\sqrt{n}\right) u \;.
\end{array}\right\} \tag{4.17}$$

Ist zusätzlich $\mathbf{x} = \hat{\mathbf{x}}$, gelten sogar

$$\left.\begin{array}{lcll}
\|\boldsymbol{\delta}_{A\mathbf{x}}\|_\infty & \leq & (2\|\mathbf{x}\|_\infty + 2)\, u & \leq \quad 4u \;, \\
\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_\infty & \leq & (2\|\mathbf{x}\|_\infty + 1)\, u & \leq \quad 3u \;,
\end{array}\right.
\qquad
\left.\begin{array}{lcl}
\|\boldsymbol{\delta}_{A\mathbf{x}}\|_2 & \leq & \left(\sqrt{2}\|\mathbf{x}\|_2 + \sqrt{\frac{5n}{2}}\right) u \;, \\
\|\boldsymbol{\delta}_{A\mathbf{x}}^{\odot}\|_2 & \leq & \left(\sqrt{n} \;+\; \sqrt{2}\|\mathbf{x}\|_2\right) u \;.
\end{array}\right\} \tag{4.18}$$

Beweis: Die Behauptungen in (4.17) folgen sofort aus den Sätzen 4.2 und 4.4 unter Verwendung der nach Voraussetzung gültigen Beziehung $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq 1$. Berücksichtigen wir in den jeweiligen Beweisen noch, dass der erste Term aus (4.3) bzw. (4.11) im Fall $\mathbf{x} = \hat{\mathbf{x}}$ wegfällt, so ergeben sich ebenso die Ungleichungen (4.18). ■

Für einen beliebigen Vektor $\mathbf{x} \in [-1,1]^n$ ist die Bedingung (4.15) üblicherweise nicht erfüllt, so dass eine komponentenweise Skalierung benötigt wird. Um Überlauf gänzlich ausschließen zu können, müssen die einzelnen Komponenten vor der Ausführung der Algorithmen (mindestens) durch $\sqrt{n}$ geteilt werden. Demzufolge sind die einzelnen Komponenten des Ergebnisvektors nach der Ausführung der Algorithmen 1.15 – 1.22 mit $\sqrt{n}$ zu multiplizieren. Entsprechend müssen die oberen Schranken für die Normen der Fehlervektoren aus Satz 4.5 angepasst werden.

4.1.2 Numerische Stabilität von schnellen Algorithmen

In Abschnitt 3.1 wird definiert, wann ein in Gleitkomma-Arithmetik implementierter Algorithmus für eine Matrix-Vektor-Multiplikation numerisch stabil ist. Analog zur Definition 3.1 können wir ebenso für die Festkomma-Arithmetik Stabilitätsbegriffe einführen.

Definition 4.6. Ein Algorithmus für eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ ($\mathbf{x} \in [-1,1]^n$) mit einer regulären Matrix $A \in [-1,1]^{n\times n}$ heißt *in einfach genauer Festkomma-Arithmetik normweise rückwärtsstabil*, falls Konstanten $\kappa_n^{\times} > 0$ und $\lambda_n^{\times} > 0$ existieren, so dass für alle $\mathbf{x} \in [-1,1]^n$ die Abschätzung

$$\| \Delta\mathbf{x} \|_2 \;\leq\; \left(\kappa_n^{\times} + \lambda_n^{\times}\|\mathbf{x}\|_2\right)\left(u + \mathcal{O}\left(u^2\right)\right) \tag{4.19}$$

erfüllt wird, wobei $\Delta\mathbf{x}$ der Rückwärtsfehler aus Abbildung 3.1 ist. Bezeichnet $\mathbf{y}$ das exakte Ergebnis von $A\mathbf{x}$ und $\tilde{\mathbf{y}}$ den tatsächlich berechneten Vektor, so heißt ein Algorithmus zur Berechnung von $A\mathbf{x}$ mit $A \in [-1,1]^{n\times n}$, $\mathbf{x} \in [-1,1]^n$ *in einfach genauer Festkomma-Arithmetik normweise vorwärtsstabil*, falls Konstanten $\kappa_n^{\times} > 0$ und $\lambda_n^{\times} > 0$ existieren, so dass für alle $\mathbf{x} \in [-1,1]^n$ die Abschätzung

$$\| \tilde{\mathbf{y}} - \mathbf{y} \|_2 \;\leq\; \left(\kappa_n^{\times} + \lambda_n^{\times}\|\mathbf{x}\|_2\right)\left(u + \mathcal{O}\left(u^2\right)\right) \tag{4.20}$$

gilt. Verwendet der Algorithmus bei der Berechnung von $A\mathbf{x}$ das doppelt genaue Matrix-Vektor-Pseudo-Produkt, kennzeichnen wir die entsprechenden Stabilitätskonstanten mit $\kappa_n^{\odot}$ und $\lambda_n^{\odot}$.

Im Fall orthogonaler Matrizen sind die Begriffe der Vorwärts- und Rückwärtsstabilität wegen (3.3) offenbar äquvalent, im Allgemeinen müssen die Konstanten $\kappa_n^{\times}$ bzw. $\lambda_n^{\times}$ aus (4.19) und (4.20) jedoch nicht übereinstimmen. Mit dem oben vereinbarten Stabilitätsbegriff können wir nun analog zu Abschnitt 3.1 Algorithmen für eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ ($\mathbf{x} \in [-1,1]^n$) mit einer allgemeinen vollbesetzten orthogonalen Matrix $A \in [-1,1]^{n\times n}$ untersuchen, welche auf einer Faktorisierung von A der Gestalt

$$A \;=\; \prod_{m=1}^{\nu} A^{(m)} \;:=\; A^{(\nu)} \ldots A^{(2)} A^{(1)} \tag{4.21}$$

mit dünnbesetzten orthogonalen Matrizen $A^{(m)} \in [-1,1]^{n\times n}$, $m = 1, \ldots, \nu$, beruht. Dabei ist zu beachten, dass der Ergebnisvektor $\mathbf{y} := A\mathbf{x}$ im Allgemeinen nicht mehr in $[-1,1]^n$ liegt. Demzufolge sollte der Vektor um einen Faktor skaliert werden, dass die Bedingung (4.15) vor Ausführen der iterierten Matrix-Vektor-Multiplikation erfüllt ist.

Bemerkung 4.7. (i) Es ist zu beachten, dass im Gegensatz zum exakten Matrix-Vektor-Produkt beim einfach genauen Matrix-Vektor-Pseudo-Produkt nicht klar ist, ob tatsächlich für alle $\mathbf{x} \in [-1,1]^n$ mit (4.15) die Bedingung $\|\hat{Q} \times \hat{\mathbf{x}}\|_2 \leq \|\mathbf{x}\|_2$ für Matrizen $Q \in [-1,1]^{n\times n}$ der Gestalt (4.6) erfüllt wird. Nach dem Modell (2.53) folgt für eine fast orthogonale Matrix $\hat{Q}_2$ der Gestalt (2.59) mit $\hat{a}^2 + \hat{b}^2 \leq 1$ zunächst mit Hilfe der Cauchy-Schwarz-Ungleichung

$$\begin{aligned}
\|\hat{Q}_2 \times \hat{\mathbf{x}}\|_2^2 &= \left(\hat{a} \times \hat{x}_0 + \hat{b} \times \hat{x}_1\right)^2 + \left(-\hat{b} \times \hat{x}_0 + \hat{a} \times \hat{x}_1\right)^2 \\
&= \left(\hat{a}\hat{x}_0 + \hat{b}\hat{x}_1 - \left(\delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1}\right)\right)^2 + \left(-\hat{b}\hat{x}_0 + \hat{a}\hat{x}_1 - \left(-\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right)\right)^2 \\
&\leq \|\hat{Q}_2\hat{\mathbf{x}}\|_2^2 + 2\|\hat{Q}_2\hat{\mathbf{x}}\|_2\sqrt{\left(\delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1}\right)^2 + \left(-\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right)^2} + \left(\delta_{\hat{a}\hat{x}_0} + \delta_{\hat{b}\hat{x}_1}\right)^2 + \left(-\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right)^2 .
\end{aligned}$$

Fordern wir zusätzlich $\hat{a}, \hat{b} \geq 0$, dann folgt mit (2.63) weiter

$$\|\hat{Q}_2 \times \hat{\mathbf{x}}\|_2 \;\leq\; \|\hat{Q}_2\hat{\mathbf{x}}\|_2 + \sqrt{5}u \;\leq\; \sqrt{\hat{a}^2 + \hat{b}^2} \cdot \|\hat{\mathbf{x}}\|_2 + \sqrt{5}u \; .$$

Beachten wir jedoch von Anfang an die Beziehungen (2.62), so ergibt sich im Fall $\hat{x}_0\hat{x}_1 \geq 0$ die schärfere Abschätzung

$$\begin{aligned}\|\hat{Q}_2 \times \hat{\mathbf{x}}\|_2^2 &\leq \left(\hat{a}\hat{x}_0 + \hat{b}\hat{x}_1\right)^2 + \left(-\hat{b}\hat{x}_0 + \hat{a}\hat{x}_1 - \left(-\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right)\right)^2 \\ &\leq \|\hat{Q}_2\hat{\mathbf{x}}\|_2^2 + 2\|\hat{Q}_2\hat{\mathbf{x}}\|_2 \cdot \left| -\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right| + \left(-\delta_{\hat{b}\hat{x}_0} + \delta_{\hat{a}\hat{x}_1}\right)^2 \leq \left(\|\hat{Q}_2\hat{\mathbf{x}}\|_2 + u\right)^2.\end{aligned}$$

Für $\hat{x}_0\hat{x}_1 \leq 0$ können in analoger Weise die Fehlerterme im zweiten Summanden weggelassen werden. Insgesamt haben wir somit für eine fast orthogonale Matrix $\hat{Q}_2$ der Gestalt (2.59) mit $\hat{a}, \hat{b} > 0$ und $\hat{a}^2 + \hat{b}^2 \leq 1$ sowie für beliebiges $\mathbf{x} \in [-1,1]^2$ mit (4.15) nur die Abschätzung

$$\|\hat{Q}_2 \times \hat{\mathbf{x}}\|_2 \leq \sqrt{\hat{a}^2 + \hat{b}^2} \cdot \|\hat{\mathbf{x}}\|_2 + u \leq 1 + u \tag{4.22}$$

zur Verfügung. Daher können wir nicht sicher sein, ob bei einer Drehung von $\hat{Q}_2 \times \hat{\mathbf{x}}$ der Bereich $[-1,1]^2$ nicht verlassen würde. Für eine Matrix $Q \in [-1,1]^{n\times n}$ der Gestalt (4.6) sowie Vektoren $\mathbf{x} \in [-1,1]^n$ mit (4.15) ergibt sich dann in ähnlicher Weise und wiederum mit der Cauchy-Schwarz-Ungleichung die Abschätzung

$$\|\hat{Q} \times \hat{\mathbf{x}}\|_2^2 \leq \|\hat{Q}\hat{\mathbf{x}}\|_2^2 + 2\|\hat{Q}\hat{\mathbf{x}}\|_2 \cdot \sqrt{\frac{n}{2}}\, u + \frac{n}{2}\, u^2 \leq \left(\|\hat{Q}\hat{\mathbf{x}}\|_2 + \sqrt{\frac{n}{2}}\, u\right)^2. \tag{4.23}$$

(ii) Beim doppelt genauen Matrix-Vektor-Pseudo-Produkt treten die bei (i) erwähnten Schwierigkeiten nicht auf, denn zusammen mit (2.57) haben wir in diesem Fall

$$\|\hat{Q} \odot \hat{\mathbf{x}}\|_2 \leq \|\hat{Q}\hat{\mathbf{x}}\|_2 \leq \|\hat{Q}\|_2 \cdot \|\hat{\mathbf{x}}\|_2 \leq \|\mathbf{x}\|_2$$

für jeden Vektor $\mathbf{x} \in [-1,1]^n$ mit (4.15) und Matrizen $Q \in [-1,1]^{n\times n}$ der Gestalt (4.6). □

Demzufolge betrachten wir zunächst den Fall, für den die Bedingung (4.15) nirgendwo verletzt wird.

Satz 4.8. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und besitze $A \in [-1,1]^{n\times n}$ eine Faktorisierung* (4.21) *mit Matrizen $A^{(m)} \in [-1,1]^{n\times n}$, $m = 1, \ldots, \nu$, der Gestalt* (4.6). *Weiterhin erfülle $\mathbf{x} \in [-1,1]^n$ die Bedingung* (4.15) *und sei $\hat{\mathbf{x}}$ die Approximation in $\mathbb{M}_q^n$ für ein $q \in \mathbb{N}$. Dann gelten mit $u = 2^{-q}$ die folgenden Aussagen:*

(i) *Unter der Annahme, dass alle berechneten Zwischenergebnisse die Bedingung* (4.15) *erfüllen, genügt der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\times} := \hat{A}^{(\nu)} \times \left(\hat{A}^{(\nu-1)} \times \ldots \left(\hat{A}^{(2)} \times \left(\hat{A}^{(1)} \times \hat{\mathbf{x}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.24}$$

des iterierten einfach genauen Matrix-Vektor-Pseudo-Produktes der Ungleichung

$$\|\Delta\mathbf{y}^{\times}\|_2 \leq \kappa_{n,\nu}^{\times} u + \lambda_{n,\nu}^{\times} \|\mathbf{x}\|_2 u + \mu_{n,\nu}^{\times} u^2$$

mit $\kappa_{n,\nu}^{\times} := \nu\sqrt{\frac{5n}{2}} + \sqrt{n}$, $\lambda_{n,\nu}^{\times} := \nu\sqrt{2}$ sowie $\mu_{n,\nu}^{\times} := \frac{\nu(\nu-1)}{2}\sqrt{n}$. Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt die Ungleichung bereits mit $\kappa_{n,\nu}^{\times} = \nu\sqrt{\frac{5n}{2}}$.

(ii) *Der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\odot} := \hat{A}^{(\nu)} \odot \left(\hat{A}^{(\nu-1)} \odot \ldots \left(\hat{A}^{(2)} \odot \left(\hat{A}^{(1)} \odot \hat{\mathbf{x}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.25}$$

des iterierten doppelt genauen Matrix-Vektor-Pseudo-Produktes genügt der Ungleichung

$$\|\Delta\mathbf{y}^{\odot}\|_2 \leq \kappa_{n,\nu}^{\odot} u + \lambda_{n,\nu}^{\odot} \|\mathbf{x}\|_2 u$$

mit $\kappa_{n,\nu}^{\odot} := (\nu+1)\sqrt{n}$ und $\lambda_{n,\nu}^{\odot} := \nu\sqrt{2}$. Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt die Ungleichung bereits mit $\kappa_{n,\nu}^{\odot} = \nu\sqrt{n}$.

Beweis: (i) Zunächst definieren wir mit $\mathbf{y}^{(0)} := \mathbf{x}$ und $\tilde{\mathbf{y}}^{(0)} := \hat{\mathbf{x}}$ die Vektoren

$$\mathbf{y}^{(m)} := A^{(m)}\mathbf{y}^{(m-1)} \ , \qquad m = 1, \ldots, \nu,$$

der exakten Zwischenergebnisse und entsprechend die Vektoren

$$\tilde{\mathbf{y}}^{(m)} := \hat{A}^{(m)} \times \tilde{\mathbf{y}}^{(m-1)} \ , \qquad m = 1, \ldots, \nu, \tag{4.26}$$

der tatsächlich auftretenden Zwischenergebnisse, was nach Voraussetzung möglich ist. (Es ist zu beachten, dass für alle Vektoren aus (4.26) die Bedingung (4.15) vorausgesetzt worden ist, um Überlauf bei der Berechnung aller Zwischenergebnisse entsprechend Bemerkung 4.7 auszuschließen.) Der Vorwärtsfehler $\Delta\mathbf{y} = \tilde{\mathbf{y}}^{(\nu)} - \mathbf{y}^{(\nu)}$ lässt sich nun durch die Teleskopsumme

$$\begin{aligned}
\Delta\mathbf{y}^{\times} &= \tilde{\mathbf{y}}^{(\nu)} - A^{(\nu)}\tilde{\mathbf{y}}^{(\nu-1)} + A^{(\nu)}\left(\tilde{\mathbf{y}}^{(\nu-1)} - A^{(\nu-1)}\mathbf{y}^{(\nu-2)}\right) \ = \ \ldots \\
&= \sum_{m=2}^{\nu}\left(\prod_{k=m+1}^{\nu} A^{(k)}\right)\left(\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)}\right) \ + \ \left(\prod_{k=2}^{\nu} A^{(k)}\right)\left(\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}\right)
\end{aligned}$$

mit den Vektoren aus (4.26) darstellen. Da nach Voraussetzung alle Vektoren aus (4.26) die Bedingung (4.15) erfüllen, kann auf die Differenzen $\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)},\ m = 2, \ldots, \nu$, wegen $\tilde{\mathbf{y}}^{(m)} \in \mathbb{M}_q^n$ jeweils (4.18) und auf $\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}$ die Abschätzungen (4.17) angewandt werden. Aufgrund der Orthogonalität aller beteiligten Matrizen ergibt sich nun

$$\begin{aligned}
\|\Delta\mathbf{y}^{\times}\|_2 &\leq \sum_{m=2}^{\nu}\left(\prod_{k=m+1}^{\nu}\left\|A^{(k)}\right\|_2\right)\left\|\tilde{\mathbf{y}}^{(m)} - A^{(m)}\tilde{\mathbf{y}}^{(m-1)}\right\|_2 \ + \ \left(\prod_{k=2}^{\nu}\left\|A^{(k)}\right\|_2\right)\left\|\tilde{\mathbf{y}}^{(1)} - A^{(1)}\mathbf{y}^{(0)}\right\|_2 \\
&\leq \sum_{m=2}^{\nu}\left(\sqrt{2}\ \|\tilde{\mathbf{y}}^{(m-1)}\|_2 + \sqrt{\frac{5n}{2}}\right)u + \left(\sqrt{2}\ \|\mathbf{y}^{(0)}\|_2 + \sqrt{\frac{5n}{2}}\right)u \ + \begin{cases} 0, & \mathbf{x} = \hat{\mathbf{x}}, \\ \sqrt{n}\ u\ , & \text{sonst}, \end{cases}
\end{aligned}$$

sowie $\|\mathbf{y}^{(m)}\|_2 = \|\mathbf{x}\|_2,\ m = 0, \ldots, \nu$. Wie jedoch in Bemerkung 4.7 (i) dargelegt, haben wir mit (4.23) per Induktion nur

$$\|\tilde{\mathbf{y}}^{(m)}\|_2 \ \leq \ \|\tilde{\mathbf{y}}^{(0)}\|_2 + \sqrt{\frac{n}{2}}mu \ \leq \ \|\mathbf{x}\|_2 + \sqrt{\frac{n}{2}}mu \ , \qquad m = 0, \ldots, \nu - 1,$$

und erhalten dann die Behauptung (i).

(ii) Der Beweis verläuft analog, wobei sich $\Delta\mathbf{y}^{\odot}$ von $\Delta\mathbf{y}^{\times}$ darin unterscheidet, dass anstelle von (4.26) die Vektoren

$$\tilde{\mathbf{y}}^{(m)} := \hat{A}^{(m)} \odot \tilde{\mathbf{y}}^{(m-1)} \ , \qquad m = 1, \ldots, \nu, \tag{4.27}$$

verwendet werden. Wie in Bemerkung 4.7 (ii) begründet, genügt es hier, die Bedingung (4.15) nur für $\mathbf{x}$ zu fordern. Aus der mit den entsprechenden Ungleichungen aus (4.18) und (4.17) analog zu (i) herzuleitenden Abschätzung

$$\|\Delta\mathbf{y}^{\odot}\|_2 \ \leq \ \sum_{m=2}^{\nu}\left(\sqrt{2}\ \|\tilde{\mathbf{y}}^{(m-1)}\|_2 + \sqrt{n}\right)u + \left(\sqrt{2}\ \|\mathbf{y}^{(0)}\|_2 + \sqrt{n}\right)u \ + \begin{cases} 0, & \mathbf{x} = \hat{\mathbf{x}}, \\ \sqrt{n}\ u\ , & \text{sonst}, \end{cases}$$

und mit Hilfe der in diesem Fall gültigen Beziehungen

$$\|\tilde{\mathbf{y}}^{(m)}\|_2 \ \leq \ \|\tilde{\mathbf{y}}^{(0)}\|_2 \ \leq \ \|\mathbf{x}\|_2 \ , \qquad m = 0, \ldots, \nu - 1,$$

folgt nun auch die Behauptung (ii). ■

Nun können wir ebenso für den Fall $\mathbf{x} \in [-1, 1]^n$ mit $\|\mathbf{x}\|_2 > 1$ entsprechende Aussagen formulieren, indem wir eine vorherige Skalierung vornehmen und die entsprechende Fehlerfortpflanzung berücksichtigen.

Satz 4.9. *Sei $n \in \mathbb{N}$ gerade, $n_1 = \frac{n}{2}$ und besitze $A \in [-1,1]^{n\times n}$ eine Faktorisierung (4.21) mit Matrizen $A^{(m)} \in [-1,1]^{n\times n}$, $m = 1,\ldots,\nu$, der Gestalt (4.6). Weiterhin seien $\mathbf{x} \in [-1,1]^n$ und $\hat{\mathbf{x}}$ die entsprechende Approximation in $\mathbb{M}_q^n$ für ein $q \in \mathbb{N}$. Dann gelten mit $u = 2^{-q}$ die folgenden Aussagen:*

(i) *Für $\|\hat{\mathbf{x}}\|_2 > 1$ und $\mathbf{w} := \frac{1}{\|\hat{\mathbf{x}}\|_2}\hat{\mathbf{x}}$ sowie unter Ausschluss von Überlauf genügt der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\times} \;:=\; \|\hat{\mathbf{x}}\|_2 \cdot \hat{A}^{(\nu)} \times \left(\hat{A}^{(\nu-1)} \times \ldots \left(\hat{A}^{(2)} \times \left(\hat{A}^{(1)} \times \hat{\mathbf{w}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.28}$$

des iterierten einfach genauen Matrix-Vektor-Pseudo-Produktes mit vorheriger Normskalierung der Ungleichung

$$\|\Delta\mathbf{y}^{\times}\|_2 \;\leq\; \kappa_{n,\nu}^{\times} u + \left(\lambda_{n,\nu}^{\times} u + \mu_{n,\nu}^{\times} u^2\right)\|\mathbf{x}\|_2$$

mit $\kappa_{n,\nu}^{\times} := \sqrt{n}$, $\lambda_{n,\nu}^{\times} := \nu\sqrt{2} + \nu\sqrt{\frac{5n}{2}} + \sqrt{n}$ sowie $\mu_{n,\nu}^{\times} := \frac{\nu(\nu-1)}{2}\sqrt{n}$. Für $\hat{\mathbf{x}} = \mathbf{x}$ gilt die Abschätzung bereits mit $\kappa_{n,\nu}^{\times} = 0$.

(ii) *Für $\|\hat{\mathbf{x}}\|_2 > 1$ und $\mathbf{w} := \frac{1}{\|\hat{\mathbf{x}}\|_2}\hat{\mathbf{x}}$ genügt der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\odot} \;:=\; \|\hat{\mathbf{x}}\|_2 \cdot \hat{A}^{(\nu)} \odot \left(\hat{A}^{(\nu-1)} \odot \ldots \left(\hat{A}^{(2)} \odot \left(\hat{A}^{(1)} \odot \hat{\mathbf{w}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.29}$$

des iterierten doppelt genauen Matrix-Vektor-Pseudo-Produktes mit vorheriger Normskalierung der Ungleichung

$$\|\Delta\mathbf{y}^{\odot}\|_2 \;\leq\; \kappa_{n,\nu}^{\odot} u + \lambda_{n,\nu}^{\odot}\|\mathbf{x}\|_2 u$$

mit $\kappa_{n,\nu}^{\odot} := \sqrt{n}$ und $\lambda_{n,\nu}^{\odot} := (\nu+1)\sqrt{n}+\nu\sqrt{2}$. Für $\mathbf{x} = \hat{\mathbf{x}}$ gilt die Ungleichung bereits mit $\kappa_{n,\nu}^{\odot} = 0$.

Beweis: (i) Wegen $\|\mathbf{w}\|_2 = 1$ sowie mit Hilfe von Satz 4.8 (i) angewandt auf den Vektor $\mathbf{w}$ ergibt sich nacheinander

$$\begin{aligned}
\|\Delta\mathbf{y}^{\times}\|_2 \;&\leq\; \left\|\|\hat{\mathbf{x}}\|_2 \left(\hat{A}^{(\nu)} \times \left(\hat{A}^{(\nu-1)} \times \ldots \left(\hat{A}^{(2)} \times \left(\hat{A}^{(1)} \times \hat{\mathbf{w}}\right)\right)\right) - A\mathbf{w}\right)\right\|_2 + \|A(\hat{\mathbf{x}} - \mathbf{x})\|_2 \\
&\leq\; \|\hat{\mathbf{x}}\|_2 \left(\nu\sqrt{2} + \nu\sqrt{\frac{5n}{2}} + \sqrt{n} + \frac{\nu(\nu-1)}{2}\sqrt{n}u\right) u + \sqrt{n}u
\end{aligned}$$

und mit $\|\hat{\mathbf{x}}\|_2 \leq \|\mathbf{x}\|_2$ demnach die Behauptung. Darüber hinaus entfällt der Term $\sqrt{n}u$, falls $\hat{\mathbf{x}} = \mathbf{x}$.
(ii) Die Behauptung folgt mit Satz 4.8 (ii) analog zum Beweis von (i). ■

Ist vom Eingangsvektor $\mathbf{x}$ lediglich bekannt, dass er sich in der Menge $[-1,1]^n$ befindet, so ist eine von $\mathbf{x}$ unabhängige Skalierung mit $n^{-\frac{1}{2}}$ notwendig.

Satz 4.10. *Mit den Voraussetzungen von Satz 4.9 und $\mathbf{w} := \frac{1}{\sqrt{n}}\hat{\mathbf{x}}$ gelten die folgenden Aussagen:*

(i) *Unter Ausschluss von Überlauf genügt der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\times} \;:=\; \sqrt{n} \cdot \hat{A}^{(\nu)} \times \left(\hat{A}^{(\nu-1)} \times \ldots \left(\hat{A}^{(2)} \times \left(\hat{A}^{(1)} \times \hat{\mathbf{w}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.30}$$

des iterierten einfach genauen Matrix-Vektor-Pseudo-Produktes mit vorheriger gleichmäßiger Skalierung der Ungleichung

$$\|\Delta\mathbf{y}^{\times}\|_2 \;\leq\; \left(\kappa_{n,\nu}^{\times} + \lambda_{n,\nu}^{\times}\|\mathbf{x}\|_2\right) u + \mu_{n,\nu}^{\times} u^2$$

mit $\kappa_{n,\nu}^{\times} := n\nu\sqrt{\frac{5}{2}} + \sqrt{n}$, $\lambda_{n,\nu}^{\times} := \nu\sqrt{2}$ sowie $\mu_{n,\nu}^{\times} := \frac{\nu(\nu-1)}{2}n$. Für $\hat{\mathbf{x}} = \mathbf{x}$ gilt die Abschätzung bereits mit $\kappa_{n,\nu}^{\times} = n\nu\sqrt{\frac{5}{2}}$.

(ii) *Der Vorwärtsfehler*

$$\Delta\mathbf{y}^{\odot} \;:=\; \sqrt{n} \cdot \hat{A}^{(\nu)} \odot \left(\hat{A}^{(\nu-1)} \odot \ldots \left(\hat{A}^{(2)} \odot \left(\hat{A}^{(1)} \odot \hat{\mathbf{w}}\right)\right)\ldots\right) - A\mathbf{x} \tag{4.31}$$

des iterierten doppelt genauen Matrix-Vektor-Pseudo-Produktes mit vorheriger gleichmäßiger Skalierung genügt der Ungleichung

$$\|\Delta \mathbf{y}^{\odot}\|_2 \;\leq\; \kappa^{\odot}_{n,\nu} u + \lambda^{\odot}_{n,\nu} \|\mathbf{x}\|_2 u$$

mit $\kappa^{\odot}_{n,\nu} := (\nu+1)n + \sqrt{n}$ *und* $\lambda^{\odot}_{n,\nu} := \nu\sqrt{2}$. *Für* $\mathbf{x} = \hat{\mathbf{x}}$ *gilt die Ungleichung bereits mit* $\kappa^{\odot}_{n,\nu} = (\nu+1)n$.

Der **Beweis** ergibt sich wie für Satz 4.9 mit Hilfe von Satz 4.8. ■

Die Abschätzungen aus den Sätzen 4.8, 4.9 und 4.10 ermöglichen es nun, Stabilitätsuntersuchungen für die in Abschnitt 1.3 hergeleiteten Algorithmen durchzuführen. Dabei benötigen wir noch, dass Vorzeichenskalierungen sowie Permutationen keinen Einfluss auf den Rundungsfehler haben.

Folgerung 4.11. *Die Aussagen aus den Sätzen* 4.8, 4.9 *und* 4.10 *bleiben richtig, wenn die Matrizen* $A^{(m)} \in [-1,1]^{n\times n},\ m = 1,\ldots,\nu$, *die Gestalt*

$$\bigoplus_{k=0}^{\frac{n}{2}-1} Q_2(\varphi_k) \tag{4.32}$$

mit Winkeln $\varphi_k \in \left[0, \frac{\pi}{4}\right]\ \left(k = 0,\ldots,\frac{n}{2}-1\right)$ *besitzen oder*

(i) *durch Multiplikation mit einer oder mehreren Permutationsmatrizen und/oder*

(ii) *durch Multiplikation mit einer oder mehreren Vorzeichenskalierungsmatrizen und/oder*

(iii) *durch Transponieren*

in die Gestalt (4.32) *überführt werden können.*
Der **Beweis** ist direkt aus dem Modell (2.49) ersichtlich. ■

Es ist zu beachten, dass die Matrizen (4.32) in Folgerung 4.11 auch Blöcke mit der Einheitsmatrix I_2 enthalten dürfen.

4.1.3 Rundungsfehleranalyse für die Algorithmen 1.15 – 1.22

Im Unterabschnitt 4.1.2 haben wir Abschätzungen für die euklidische Norm des Fehlervektors ermittelt, welcher in Festkomma-Arithmetik bei wiederholter Matrix-Vektor-Multiplikation eines Vektors mit Matrizen $Q \in [-1,1]^{n\times n}$ der Gestalt (4.6) auftritt. Dabei unterscheiden wir zwischen einfach genauer und doppelt genauer Ausführung entsprechend (2.56) bzw. (2.58), welches wir jeweils durch den Operanden $\times$ und $\odot$ kennzeichnen.
Wie in Kapitel 3 untersuchen wir nun die Faktorisierungen, auf denen die Algorithmen 1.15 – 1.22 aus Abschnitt 1.3 beruhen. Mittels Folgerung 4.11 genügt es wiederum, lediglich die Anzahl der wesentlichen Faktoren zu bestimmen. Entsprechend der einzelnen Gegebenheiten an den Eingangsvektor erhalten wir nun die folgenden Abschätzungen für die euklidische Norm des jeweiligen Rundungsfehlers. Zunächst gehen wir davon aus, dass sich die Eingangsvektoren bereits in der euklidischen Einheitskugel befinden und somit keine Skalierung notwendig ist.

Satz 4.12 (Fehlerabschätzung für die Algorithmen 1.15 – 1.22 ohne Skalierung)**.** *Seien* $t \geq 2$, $n = 2^t$ *sowie* $\mathbf{x} \in [-1,1]^n$ *gegeben. Weiterhin erfülle* $\mathbf{x}$ *die Bedingung* (4.15). *Dann gilt:*

(i) *Bei der Berechnung der Matrix-Vektor-Multiplikation* $C_n^{\mathrm{IV}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.15 *und* 1.18 *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *die Abschätzungen*

$$\|\Delta \mathbf{y}^{\times}\|_2 \;\leq\; (2\log_2(n)-1)\left(\sqrt{\tfrac{5n}{2}} + \sqrt{2}\|\mathbf{x}\|_2 + (\log_2(n)-1)\sqrt{n}u\right)u \;+\; \begin{cases} 0 & \text{für } \mathbf{x} = \hat{\mathbf{x}}, \\ \sqrt{n}u & \text{sonst,} \end{cases}$$

$$\|\Delta \mathbf{y}^{\odot}\|_2 \;\leq\; (2\log_2(n)-1)\left(\sqrt{n} + \sqrt{2}\|\mathbf{x}\|_2\right)u \;+\; \begin{cases} 0 & \text{für } \mathbf{x} = \hat{\mathbf{x}}, \\ \sqrt{n}u & \text{sonst.} \end{cases}$$

(ii) *Bei der Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.16 *und* 1.17 *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *im Fall* $t \geq 3$ *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 3(\log_2(n)-1)\left(\sqrt{\tfrac{5n}{2}}+\sqrt{2}\|\mathbf{x}\|_2+\tfrac{3\log_2(n)-4}{2}\sqrt{n}u\right)u \;+\; \begin{cases}0 & \text{für } \mathbf{x}=\hat{\mathbf{x}},\\ \sqrt{n}u & \text{sonst,}\end{cases}\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq 3(\log_2(n)-1)\left(\sqrt{n}+\sqrt{2}\|\mathbf{x}\|_2\right)u \;+\; \begin{cases}0 & \text{für } \mathbf{x}=\hat{\mathbf{x}},\\ \sqrt{n}u & \text{sonst.}\end{cases}\end{aligned}$$

Ist $t = 2$*, so ergeben sich*

$$\|\Delta\mathbf{y}^{\times}\|_2 \leq 2\left(\sqrt{10}+\sqrt{2}\|\mathbf{x}\|_2+u\right)u\,, \qquad \|\Delta\mathbf{y}^{\odot}\|_2 \leq \left(4+2\sqrt{2}\|\mathbf{x}\|_2\right)u$$

für $\mathbf{x} = \hat{\mathbf{x}}$*, andernfalls gilt*

$$\|\Delta\mathbf{y}^{\times}\|_2 \leq 2\left(\sqrt{10}+1+\sqrt{2}\|\mathbf{x}\|_2+u\right)u\,, \qquad \|\Delta\mathbf{y}^{\odot}\|_2 \leq \left(6+2\sqrt{2}\|\mathbf{x}\|_2\right)u\,.$$

(iii) *Bei der Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.19 *und* 1.20 *mit Parameter* $k = 0$ *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 2(\log_2(n)-1)\left(\sqrt{\tfrac{5n}{2}}+\sqrt{2}\|\mathbf{x}\|_2+\tfrac{2\log_2(n)-3}{2}\sqrt{n}u\right)u \;+\; \begin{cases}0 & \text{für } \mathbf{x}=\hat{\mathbf{x}},\\ \sqrt{n}u & \text{sonst,}\end{cases}\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq 2(\log_2(n)-1)\left(\sqrt{n}+\sqrt{2}\|\mathbf{x}\|_2\right)u \;+\; \begin{cases}0 & \text{für } \mathbf{x}=\hat{\mathbf{x}},\\ \sqrt{n}u & \text{sonst.}\end{cases}\end{aligned}$$

(iv) *Bei der Berechnung der Matrix-Vektor-Multiplikation* $C_n^{\mathrm{IV}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.19 *und* 1.20 *mit Parameter* $k = 1$ *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *dieselben Abschätzungen wie bei* (i)*.*

(v) *Bei der Berechnung der Matrix-Vektor-Multiplikationen* $S_n^{\mathrm{II}}\mathbf{x}$ *und* $S_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.21 *und* 1.22 *mit Parameter* $k = 0$ *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *dieselben Abschätzungen wie bei* (iii)*.*

(vi) *Bei der Berechnung der Matrix-Vektor-Multiplikation* $S_n^{\mathrm{IV}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.21 *und* 1.22 *mit Parameter* $k = 1$ *ergeben sich für die entsprechend Satz* 4.8 *definierten Rundungsfehler* (4.24) *und* (4.25) *dieselben Abschätzungen wie bei* (i)*.*

Beweis: Es genügt jeweils, die Voraussetzungen von Satz 4.8 zu überprüfen.
(i) Die beiden Algorithmen zur Berechnung der DCT-IV(n) basieren auf den Faktorisierungen (1.32) und (1.33) bzw. auf den dazu transponierten Gleichungen (1.37) und (1.39). Nach Folgerung 4.11 genügt es, die ersten beiden Faktorisierungen zu betrachten. Da die Matrizen $P_n^{(s)}$, $s = 0, \ldots, t-1$ sowie U_n als Permutationsmatrizen keinen Beitrag zum Rundungsfehler leisten, können sie vernachlässigt werden. Somit verursachen in (1.33) lediglich die drei Blockdiagonalmatrizen $I_2 \oplus T_2(0)$, $T_4(0)$ und $D_4^{(0)}$ Rundungsfehler. Aus der Definition bzw. aus (1.25) geht hervor, dass sie durch Permutationen und Vorzeichenskalierungen jeweils auf eine Matrix der Form (4.32) zurückführbar sind. Damit ergeben sich hier die entsprechenden Stabilitätskonstanten jeweils entsprechend Satz 4.8 mit $n = 4$ und $\nu = 3 = 2\log_2(4) - 1$. Analog sind $2t-1$ Faktoren in (1.32) enthalten, die im Sinn von Satz 4.8 bzw. Folgerung 4.11 Rundungsfehler verursachen. Somit findet Satz 4.8 mit $\nu = 2\log_2(n) - 1$ Anwendung.
(ii) In den Faktorisierungen (1.35) und (1.38) für die Matrizen C_n^{III} und C_n^{II} haben wir für $n = 2^t$ mit $t \geq 3$ offenbar $t + 2(t-2) + 1 = 3(t-1)$ Matrizen, welche im Sinn von Satz 4.8 bzw. Folgerung 4.11 Rundungsfehler verursachen, wie aus ihrer Definition in Lemma 1.12 hervorgeht. Somit ergeben sich die Abschätzungen aus Satz 4.8 jeweils mit $\nu = 3(t-1)$. Für den Fall $t = 2$ tragen nur die Matrizen $T_4(0)$ und $\tilde{D}_4^{(0)}$ zum Rundungsfehler bei, so dass hier Satz 4.8 mit $n = 4$ und $\nu = 2$ anzuwenden ist.

(iii) Für $k = 0$ basieren die Algorithmen 1.19 und 1.20 auf den Faktorisierungen (1.45), die jeweils $2t - 1$ Matrizen beinhalten. Aufgrund ihrer Definition in (1.43) und den Darstellungen (1.25), (1.27) und (1.28) mittels spezieller Blockdiagonalmatrizen ist die Anwendung von Folgerung 4.11 möglich. Insbesondere handelt es sich bei $A_n(\boldsymbol{\beta}_0)$ lediglich um eine Permutationsmatrix, so dass wir $\nu = 2(t-1)$ in Satz 4.8 wählen können.

(iv) Ähnlich wie bei (iii) basieren die Algorithmen 1.19 und 1.20 für $k = 1$ auf den Faktorisierungen (1.48), die ebenfalls je $2t - 1$ analog zu (1.43) definierte Matrizen beinhalten. Somit sind auch hier wegen der Darstellungen (1.25), (1.27) und (1.28) die Voraussetzungen für Folgerung 4.11 erfüllt, so dass sich die jeweiligen Abschätzungen aus Satz 4.8 mit $\nu = 2t - 1$ ergeben.

(v) Analog zu (iii) sind für Folgerung 4.11 die Voraussetzungen an die für $k = 0$ den beiden Algorithmen zugrunde liegenden Faktorisierungen (1.55) erfüllt, so dass sich die jeweiligen Abschätzungen aus Satz 4.8 mit $\nu = 2(t-1)$ ergeben. Dabei wird wiederum verwendet, dass die einzelnen Blöcke der in (1.53) definierten Matrizen die Darstellungen (1.27) und (1.28) besitzen und es sich bei $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)$ lediglich um eine Permutationsmatrix handelt.

(vi) Die Algorithmen 1.21 und 1.22 basieren für $k = 1$ auf den Faktorisierungen (1.60), deren Faktoren analog zu (1.53) definiert sind. Entsprechend sind wiederum die Voraussetzungen von Folgerung 4.11 erfüllt, so dass sich die jeweiligen Abschätzungen aus Satz 4.8 mit $\nu = 2t - 1$ ergeben. ■

Die Bedingung (4.15) an den Eingangsvektor $\mathbf{x} \in [-1, 1]^n$ bedeutet für größere n eine äußerst starke Einschränkung. Uns interessiert nun, welche Rundungsfehlerabschätzungen sich für Vektoren

$$\mathbf{x} \in [-1, 1]^n \setminus \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1\} \tag{4.33}$$

ergeben, wenn ihre Komponenten zuvor mit dem Reziproken der euklidischen Norm ihrer Festkomma-Approximation skaliert werden.

Satz 4.13 (Fehlerabschätzung für die Algorithmen 1.15 – 1.22 mit Norm-Skalierung)**.** *Seien $t \geq 2$, $n = 2^t$ sowie $\mathbf{x}$ gemäß (4.33) gegeben. Mit*

$$\varpi_n(\mathbf{x}) \;:=\; \begin{cases} 0 & \text{für } \mathbf{x} = \hat{\mathbf{x}}, \\ \sqrt{n}u & \text{sonst} \end{cases} \tag{4.34}$$

gilt dann:

(i) *Bei der Berechnung der Matrix-Vektor-Multiplikationen $C_n^{\mathrm{II}}\mathbf{x}$ und $C_n^{\mathrm{III}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen* 1.19 *und* 1.20 *mit Parameter $k = 0$ ergeben sich für die entsprechend Satz* 4.9 *definierten Rundungsfehler* (4.28) *und* (4.29) *die Abschätzungen*

$$\begin{aligned} \|\Delta\mathbf{y}^{\times}\|_2 &\leq \left(2(\log_2(n)-1)\Big(\sqrt{2} + \sqrt{\frac{5n}{2}} + \frac{2\log_2(n)-3}{2}\sqrt{n}u\Big) + \sqrt{n}\right)\|\mathbf{x}\|_2 u + \varpi_n(\mathbf{x}) \\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \Big((2\log_2(n)-1)\sqrt{n} + 2(\log_2(n)-1)\sqrt{2}\Big)\|\mathbf{x}\|_2 u + \varpi_n(\mathbf{x}) \end{aligned}$$

Dieselben Aussagen gelten bei der Berechnung der Matrix-Vektor-Multiplikationen $S_n^{\mathrm{II}}\mathbf{x}$ und $S_n^{\mathrm{III}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen 1.21 *und* 1.22 *mit Parameter $k = 0$.*

(ii) *Bei der Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen* 1.15 *und* 1.18 *ergeben sich für die entsprechend Satz* 4.9 *definierten Rundungsfehler* (4.28) *und* (4.29) *die Abschätzungen*

$$\begin{aligned} \|\Delta\mathbf{y}^{\times}\|_2 &\leq \left((2\log_2(n)-1)\Big(\sqrt{2} + \sqrt{\frac{5n}{2}} + (\log_2(n)-1)\sqrt{n}u\Big) + \sqrt{n}\right)\|\mathbf{x}\|_2 u + \varpi_n(\mathbf{x}) \\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \Big(2\log_2(n)\sqrt{n} + (2\log_2(n)-1)\sqrt{2}\Big)\|\mathbf{x}\|_2 u + \varpi_n(\mathbf{x}) \end{aligned}$$

Dieselben Aussagen gelten bei der Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ bzw. $S_n^{\mathrm{IV}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen 1.19 *und* 1.20 *bzw.* 1.21 *und* 1.22 *jeweils mit Parameter $k = 1$.*

(iii) *Bei der Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.16 *und* 1.17 *ergeben sich für die entsprechend Satz* 4.9 *definierten Rundungsfehler* (4.28) *und* (4.29) *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq \left(3(\log_2(n)-1)\left(\sqrt{2}+\sqrt{\frac{5n}{2}}+\frac{3\log_2(n)-4}{2}\sqrt{n}u\right)+\sqrt{n}\right)\|\mathbf{x}\|_2 u+\varpi_n(\mathbf{x})\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \left((3\log_2(n)-2)\sqrt{n}+3(\log_2(n)-1)\sqrt{2}\right)\|\mathbf{x}\|_2 u+\varpi_n(\mathbf{x})\end{aligned}$$

im Fall $t \geq 3$. *Für* $t=2$ *gelten die Ungleichungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 2\left(\sqrt{2}+\sqrt{10}+1+u\right)\|\mathbf{x}\|_2 u+\varpi_4(\mathbf{x})\ ,\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq 2\left(3+\sqrt{2}\right)\|\mathbf{x}\|_2 u+\varpi_4(\mathbf{x})\ .\end{aligned}$$

Der **Beweis** verwendet Satz 4.9 und verläuft analog zum Beweis von Satz 4.12. ■

Die im Satz 4.13 angegebenen Abschätzungen für den auftretenden Rundungsfehler sind zwar denkbar günstig, jedoch ist es in den meisten Anwendungen aufgrund des zusätzlichen Zeitaufwandes wenig sinnvoll, jeweils die euklidische Norm der Approximation des Eingangsvektors auszuwerten. Alternativ bietet sich eine gleichmäßige Skalierung mit dem Reziproken der maximal möglichen euklidischen Norm an, d.h. mit

$$\frac{1}{\max\limits_{\mathbf{x}\in[-1,1]^n}\|\mathbf{x}\|_2} = \frac{1}{\sqrt{n}}\ . \tag{4.35}$$

Dabei haben wir verwendet, dass die Menge $[-1,1]^n$, welche genau der Einheitskugel im $\mathbb{R}^n$ bezüglich der Maximumnorm entspricht, kompakt ist und die euklidische Norm als stetige Funktion auf dieser kompakten Menge ihr Maximum annimmt. Die komponentenweise Multiplikation mit $\frac{1}{\sqrt{n}}$ lässt sich für den Fall, dass $t=\log_2(n)$ eine gerade Zahl ist, jeweils als einfache Bitverschiebung realisieren. Um zu vermeiden, dass sich durch diese Bitverschiebung nur jeweils der Nullvektor ergibt, muss der Parameter q der Festkomma-Zahlenmenge genügend groß gewählt werden. Andernfalls wird der Nullvektor korrekt auf den Nullvektor transformiert und die euklidische Norm des Rundungsfehlers stimmt aufgrund der Orthogonalität der Kosinus- und Sinusmatrizen mit der euklidischen Norm des Eingangsvektors überein.

Satz 4.14 (Fehlerabschätzung für die Algorithmen 1.15 – 1.22 mit gleichmäßiger Skalierung)**.** *Seien* $t \geq 2$, $n=2^t$ *sowie* $\mathbf{x}\in[-1,1]^n$ *gegeben. Mit* (4.34) *gilt dann:*

(i) *Bei der Berechnung der Matrix-Vektor-Multiplikationen* $C_n^{\mathrm{II}}\mathbf{x}$ *und* $C_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.19 *und* 1.20 *mit Parameter* $k=0$ *ergeben sich für die entsprechend Satz* 4.10 *definierten Rundungsfehler* (4.30) *und* (4.31) *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 2(\log_2(n)-1)\left(n\sqrt{\frac{5}{2}}+\frac{2\log_2(n)-3}{2}nu+\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\ ,\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \left((2\log_2(n)-1)n+2(\log_2(n)-1)\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\ .\end{aligned}$$

Dieselben Ungleichungen gelten bei der Berechnung der Matrix-Vektor-Multiplikationen $S_n^{\mathrm{II}}\mathbf{x}$ *und* $S_n^{\mathrm{III}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.21 *und* 1.22 *für den Parameter* $k=0$.

(ii) *Bei der Berechnung der Matrix-Vektor-Multiplikation* $C_n^{\mathrm{IV}}\mathbf{x}$ *in Festkomma-Arithmetik mittels der Algorithmen* 1.15 *und* 1.18 *ergeben sich für die entsprechend Satz* 4.10 *definierten Rundungsfehler* (4.30) *und* (4.31) *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq (2\log_2(n)-1)\left(n\sqrt{\frac{5}{2}}+(\log_2(n)-1)nu+\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\ ,\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \left(2n\log_2(n)+(2\log_2(n)-1)\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\ .\end{aligned}$$

Dieselben Aussagen gelten bei der Berechnung der Matrix-Vektor-Multiplikation $C_n^{\mathrm{IV}}\mathbf{x}$ bzw. $S_n^{\mathrm{IV}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen 1.19 *und* 1.20 *bzw.* 1.21 *und* 1.22 *jeweils für den Parameter $k = 1$.*

(iii) *Bei der Berechnung der Matrix-Vektor-Multiplikationen $C_n^{\mathrm{II}}\mathbf{x}$ und $C_n^{\mathrm{III}}\mathbf{x}$ in Festkomma-Arithmetik mittels der Algorithmen* 1.16 *und* 1.17 *ergeben sich für die entsprechend Satz* 4.10 *definierten Rundungsfehler* (4.30) *und* (4.31) *die Abschätzungen*

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 3(\log_2(n)-1)\left(n\sqrt{\frac{5}{2}}+\frac{3\log_2(n)-4}{2}nu+\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq \left((3\log_2(n)-2)n+3(\log_2(n)-1)\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_n(\mathbf{x})\end{aligned}$$

im Fall $t \geq 3$. Für $t = 2$ gelten die Ungleichungen

$$\begin{aligned}\|\Delta\mathbf{y}^{\times}\|_2 &\leq 2\left(2\sqrt{10}+2u+\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_4(\mathbf{x})\ ,\\ \|\Delta\mathbf{y}^{\odot}\|_2 &\leq 2\left(6+\sqrt{2}\|\mathbf{x}\|_2\right)u+\varpi_4(\mathbf{x})\ .\end{aligned}$$

Der **Beweis** verwendet Satz 4.10 und verläuft analog zum Beweis von Satz 4.12. ■

Zusammenfassend sind nun für die Algorithmen 1.15 – 1.22 im Fall $\hat{\mathbf{x}} = \mathbf{x}$ die Stabilitätskonstanten gemäß Definition 4.6 in Tabelle 4.1 angegeben.

Stabilitätskonstanten aus			Satz 4.12	Satz 4.13	Satz 4.14
DCT-II(n)	mit Alg. 1.19,	$\kappa_n^{\times}$	$2(t-1)\sqrt{\frac{5n}{2}}$	0	$2(t-1)n\sqrt{\frac{5}{2}}$
DCT-III(n)	mit Alg. 1.20,	$\lambda_n^{\times}$	$2(t-1)\sqrt{2}$	$2(t-1)\left(\sqrt{2}+\sqrt{\frac{5n}{2}}\right)+\sqrt{n}$	$2(t-1)\sqrt{2}$
DST-III(n)	mit Alg. 1.21,	$\kappa_n^{\odot}$	$2(t-1)\sqrt{n}$	0	$(2t-1)n$
DST-II(n)	mit Alg. 1.22.	$\lambda_n^{\odot}$	$2(t-1)\sqrt{2}$	$(2t-1)\sqrt{n}+2(t-1)\sqrt{2}$	$2(t-1)\sqrt{2}$
DCT-IV(n)	mit Alg. 1.15,	$\kappa_n^{\times}$	$(2t-1)\sqrt{\frac{5n}{2}}$	0	$(2t-1)n\sqrt{\frac{5}{2}}$
	1.18 – 1.20.	$\lambda_n^{\times}$	$(2t-1)\sqrt{2}$	$(2t-1)\left(\sqrt{2}+\sqrt{\frac{5n}{2}}\right)+\sqrt{n}$	$(2t-1)\sqrt{2}$
DST-IV(n)	mit Alg. 1.21,	$\kappa_n^{\odot}$	$(2t-1)\sqrt{n}$	0	$2tn$
	1.22.	$\lambda_n^{\odot}$	$(2t-1)\sqrt{2}$	$2t\sqrt{n}+(2t-1)\sqrt{2}$	$(2t-1)\sqrt{2}$
DCT-III(n)	mit Alg. 1.16,	$\kappa_n^{\times}$	$3(t-1)\sqrt{\frac{5n}{2}}$	0	$3(t-1)n\sqrt{\frac{5}{2}}$
DCT-II(n)	mit Alg. 1.17.	$\lambda_n^{\times}$	$3(t-1)\sqrt{2}$	$3(t-1)\left(\sqrt{2}+\sqrt{\frac{5n}{2}}\right)+\sqrt{n}$	$3(t-1)\sqrt{2}$
		$\kappa_n^{\odot}$	$3(t-1)\sqrt{n}$	0	$(3t-2)n$
		$\lambda_n^{\odot}$	$3(t-1)\sqrt{2}$	$(3t-2)\sqrt{n}+3(t-1)\sqrt{2}$	$3(t-1)\sqrt{2}$

Tabelle 4.1: Stabilitätskonstanten nach Definition 4.6 für die DCT- und DST-Algorithmen 1.15 – 1.22 im Fall $\mathbf{x} = \hat{\mathbf{x}}$ für einfach und doppelt genaue Festkomma-Arithmetik. Dabei sei $n = 2^t$ mit einer natürlichen Zahl $t \geq 3$. Es fällt auf, dass einerseits $\lambda_n^{\times} = \lambda_n^{\odot}$ in den Sätzen 4.12 und 4.14 gilt und sich andererseits $\lambda_n^{\times} = \lambda_n^{\odot} = 0$ in Satz 4.13 ergibt. Letzteres ist eine direkte Konsequenz der Rückskalierung mit der Norm des Eingangsvektors.

4.2 Stochastische Rundungsfehleranalyse

Die im Abschnitt 2.3 eingeführten Rundungsfehler δ, welche bei den einzelnen Pseudo-Multiplikationen durch die notwendigen Abbildungen nach $\mathbb{M}_q$ entstehen, genügen für feste Zahlen deterministischen Funktionen, welche vom gewählten Modell abhängen. Innerhalb eines Algorithmus treten eine Vielzahl

aufeinander folgender Multiplikationen auf, die darüber hinaus mit Additionen kombiniert werden. Demnach ist es überaus aufwendig, die Entwicklung der auftretenden Rundungsfehler exakt über die jeweiligen deterministischen Funktionen auszudrücken.
Um dennoch Aussagen über die Güte eines Algorithmus treffen zu können, sind in Abschnitt 4.1 obere Schranken hergeleitet worden, die für jeden einzelnen Schritt den schlechtesten Fall berücksichtigen. Diese oberen Schranken überschätzen jedoch – wie nicht anders zu erwarten – die tatsächlich auftretenden Rundungsfehler im Allgemeinen um ein Vielfaches. Daher ist es sinnvoll, den in jedem einzelnen Schritt auftretenden Rundungsfehler als eine Zufallsvariable zu modellieren, um über deren Erwartungswert sowie deren Streuung das Verhalten der nach der Ausführung eines der Algorithmen 1.15 – 1.22 entstandenen Rundungsfehler besser vorhersagen bzw. schätzen zu können. In Anlehnung an die Definitionen 3.21 und 4.6 sowie Abbildung 3.1 vereinbaren wir für die Festkomma-Arithmetik folgende Sprechweisen.

Definition 4.15. Ein Algorithmus für eine Matrix-Vektor-Multiplikation $A\mathbf{X}$ mit einer regulären Matrix $A \in [-1,1]^{n\times n}$ und einem Zufallsvektor $\mathbf{X} \in [-1,1]^n$ mit $\|\mathbf{X}\|_2 \leq 1$ heißt *in einfach genauer Festkomma-Arithmetik durchschnittlich normweise rückwärtsstabil*, falls eine Konstante $\breve{\kappa}_{n,\times} > 0$ mit $\breve{\kappa}_{n,\times} u \ll 1$ und

$$\mathbb{E}\left(\| \Delta\mathbf{X} \|_2^2\right) \;\leq\; \breve{\kappa}_{n,\times}^2 u^2 + \mathcal{O}\left(u^3\right) \tag{4.36}$$

existiert. Verwendet der Algorithmus bei der Berechnung von $A\mathbf{X}$ das doppelt genaue Matrix-Vektor-Pseudo-Produkt, kennzeichnen wir die entsprechende Stabilitätskonstante mit $\breve{\kappa}_{n,\odot}$ und nennen ihn *in doppelt genauer Festkomma-Arithmetik durchschnittlich normweise rückwärtsstabil*.

Mit Hilfe der in Abschnitt 2.4 gewonnenen Sätze 2.37, 2.38 und 2.39 können wir nun Stabilitätskonstanten gemäß Definition 4.15 für die Algorithmen 1.15 – 1.22 in Festkomma-Arithmetik angeben. Insbesondere ist dies auch ohne Kenntnis der genauen Verteilung aller auftretenden Zufallsgrößen und ohne die Forderung nach Unkorreliertheit oder stochastischer Unabhängigkeit der Eingangsdaten $\mathbf{X}$ möglich. Die notwendigen Voraussetzungen fassen wir vorab in folgendem Modell zusammen.

Modell 4.16. Sei $t \in \mathbb{N}$ und $n = 2^t$. Für ein $q \in \mathbb{N}$ seien $\mathbb{M}_q$ und u wie in (2.49) und (2.50) definiert. Eine Matrix-Vektor-Multiplikation $A\mathbf{X}$ für einen Zufallsvektor $\mathbf{X} = (X_k)_{k=0}^{n-1}$ mit Werten in $\mathbb{M}_q^n$, die auf einer Faktorisierung der Gestalt

$$A \;=\; \prod_{\imath=0}^{\jmath-1} A^{(\imath)} \;:=\; A^{(\jmath-1)} \ldots A^{(2)} A^{(0)} \tag{4.37}$$

mit Faktoren $A^{(\imath)} \in \mathbb{M}_q^{n\times n}$ basiert, genüge den folgenden Annahmen.

(A0) Die Zufallsvariable

$$\Xi \;:=\; \sum_{k=0}^{n-1} X_k^2 \tag{4.38}$$

nimmt nur Werte im Intervall $[0,1]$ an.

(A1) Die Matrizen $A^{(\imath)} \in \mathbb{M}_q^{n\times n}$ besitzen bis auf Permutationen und Vorzeichenskalierungen Blockdiagonalgestalt, d.h., es existieren Matrizen $U^{(\imath)},\ V^{(\imath)} \in \mathbb{M}_q^{n\times n}$, welche höchstens Permutationen und/oder zeilen- bzw. spaltenweise Vorzeichenskalierungen bewirken, und Matrizen

$$B^{(\imath)} \;=\; \bigoplus_{k=0}^{\frac{n}{2}-1} A^{(k,\imath)} \qquad (\imath = 1,\ldots,\jmath-1) \tag{4.39}$$

mit Blöcken

$$A^{(k,\imath)} \;:=\; \begin{pmatrix} M_1^{(k,\imath)} & M_2^{(k,\imath)} \\ -M_2^{(k,\imath)} & M_1^{(k,\imath)} \end{pmatrix} \qquad \left(k = 0,\ldots,\tfrac{n}{2}-1,\ \ \imath = 0,\ldots,\jmath-1\right), \tag{4.40}$$

so dass $A^{(\imath)} = U^{(\imath)} B^{(\imath)} V^{(\imath)}$ gilt.

(A2) Für $k = 0, \ldots, \frac{n}{2} - 1, \quad \imath = 1, \ldots, \jmath - 1$ existieren reelle Zufallsvariablen $\delta_{A^{(k,\imath)}}$ mit

$$0 \leq \delta_{A^{(k,\imath)}} \leq 2u \tag{4.41a}$$

und

$$(M_1^{(k,\imath)})^2 + (M_2^{(k,\imath)})^2 = 1 - \delta_{A^{(k,\imath)}} \ . \tag{4.41b}$$

(A3) Zu den Blockdiagonalmatrizen $B^{(\imath)}$ und den rekursiv durch

$$\begin{array}{rclcrcl} \mathbf{Y}^{(0)} & := & V^{(0)}\mathbf{X} \ , & & \mathbf{Z}^{(0)} & := & B^{(0)} \times \mathbf{Y}^{(0)} \ , \\ \mathbf{Y}^{(\imath)} & := & V^{(\imath)}U^{(\imath-1)}\mathbf{Z}^{(\imath-1)} \ , & & \mathbf{Z}^{(\imath)} & := & B^{(\imath)} \times \mathbf{Y}^{(\imath)} \ , \qquad (\imath = 1, \ldots, \jmath - 1) \\ \mathbf{Y}^{(\jmath)} & := & U^{(\jmath-1)}\mathbf{Z}^{(\jmath-1)} & & & & \end{array} \tag{4.42a}$$

bzw.

$$\begin{array}{rclcrcl} \mathbf{Y}^{(0,\odot)} & := & V^{(0)}\mathbf{X} \ , & & \mathbf{Z}^{(0,\odot)} & := & B^{(0)} \odot \mathbf{Y}^{(0,\odot)} \ , \\ \mathbf{Y}^{(\imath,\odot)} & := & V^{(\imath)}U^{(\imath-1)}\mathbf{Z}^{(\imath-1,\odot)} \ , & & \mathbf{Z}^{(\imath,\odot)} & := & B^{(\imath)} \odot \mathbf{Y}^{(\imath,\odot)} \ , \qquad (\imath = 1, \ldots, \jmath - 1) \\ \mathbf{Y}^{(\jmath,\odot)} & := & U^{(\jmath-1)}\mathbf{Z}^{(\jmath-1,\odot)} & & & & \end{array} \tag{4.42b}$$

definierten Zwischenergebnissen $\mathbf{Y}^{(\imath)}$ bzw. $\mathbf{Y}^{(\imath,\odot)}$ existieren Zufallsgrößen $\delta_{11}^{(k,\imath)}$, $\delta_{22}^{(k,\imath)}$, $\delta_{12}^{(k,\imath)}$, $\delta_{21}^{(k,\imath)}$ bzw. $\delta_{1,\odot}^{(k,\imath)}, \delta_{2,\odot}^{(k,\imath)}$ und Zahlen $\sigma, \sigma_\odot \in [0,1]$ mit

$$\left.\begin{array}{rclcrcll} \mathbb{E}(\delta_{rs}^{(k,\imath)}) & = & 0 \ , & & \mathbb{V}(\delta_{rs}^{(k,\imath)}) & \leq & \sigma^2 u^2 & (r,s = 1,2), \\ \mathbb{E}(\delta_{v,\odot}^{(k,\imath)}) & = & 0 \ , & & \mathbb{V}(\delta_{v,\odot}^{(k,\imath)}) & \leq & \sigma_\odot^2 u^2 & (v = 1,2) \end{array}\right\} \tag{4.43}$$

und

$$\begin{pmatrix} Z_{2k}^{(\imath)} \\ Z_{2k+1}^{(\imath)} \end{pmatrix} = A^{(k,\imath)} \begin{pmatrix} Y_{2k}^{(\imath)} \\ Y_{2k+1}^{(\imath)} \end{pmatrix} + \begin{pmatrix} \delta_{11}^{(k,\imath)} + \delta_{22}^{(k,\imath)} \\ \delta_{12}^{(k,\imath)} - \delta_{21}^{(k,\imath)} \end{pmatrix} \ , \tag{4.44a}$$

$$\begin{pmatrix} Z_{2k}^{(\imath,\odot)} \\ Z_{2k+1}^{(\imath,\odot)} \end{pmatrix} = A^{(k,\imath)} \begin{pmatrix} Y_{2k}^{(\imath,\odot)} \\ Y_{2k+1}^{(\imath,\odot)} \end{pmatrix} + \begin{pmatrix} \delta_{1,\odot}^{(k,\imath)} \\ \delta_{2,\odot}^{(k,\imath)} \end{pmatrix} \tag{4.44b}$$

für $k = 0, \ldots, \frac{n}{2} - 1, \quad \imath = 0, \ldots, \jmath - 1$. □

Als Hauptergebnis der stochastischen Rundungsfehleranalyse erhalten wir nun den folgenden

Satz 4.17. *Sei $t \geq 2, n = 2^t$. Genügt die Implementierung in einfach bzw. doppelt genauer Festkomma-Arithmetik allen Annahmen aus Modell* 4.16*, dann sind die Algorithmen* 1.15 – 1.22 *gemäß* (4.36) *in einfach bzw. doppelt genauer Festkomma-Arithmetik durchschnittlich rückwärtsstabil mit Stabilitätskonstanten*

$$\breve{\kappa}_{n,\times} = \begin{cases} \sqrt{2^{\lceil \log_2(2(\log_2(n)-1))\rceil+2} \cdot 2(\log_2(n)-1)n \cdot \sigma^2} & \textit{für} \begin{cases} \text{DCT-II}(n) & \textit{mit Alg.} \ 1.19 \ , \\ \text{DCT-III}(n) & \textit{mit Alg.} \ 1.20 \ , \\ \text{DST-III}(n) & \textit{mit Alg.} \ 1.21 \ , \\ \text{DST-II}(n) & \textit{mit Alg.} \ 1.22 \ , \end{cases} \\ \sqrt{2^{\lceil \log_2(2\log_2(n)-1)\rceil+2} \cdot (2\log_2(n)-1)n \cdot \sigma^2} & \textit{für} \begin{cases} \text{DCT-IV}(n) & \textit{mit Alg.} \ 1.15 \ , \\ \text{DCT-IV}(n) & \textit{mit Alg.} \ 1.18 \ , \\ \text{DCT-IV}(n) & \textit{mit Alg.} \ 1.19 \ , \\ \text{DCT-IV}(n) & \textit{mit Alg.} \ 1.20 \ , \\ \text{DST-IV}(n) & \textit{mit Alg.} \ 1.21 \ , \\ \text{DST-IV}(n) & \textit{mit Alg.} \ 1.22 \ , \end{cases} \\ \sqrt{2^{\lceil \log_2(3(\log_2(n)-1))\rceil+2} \cdot 3(\log_2(n)-1)n \cdot \sigma^2} & \textit{für} \begin{cases} \text{DCT-III}(n) & \textit{mit Alg.} \ 1.16 \ , \\ \text{DCT-II}(n) & \textit{mit Alg.} \ 1.17 \ , \end{cases} \end{cases}$$

bzw.

$$\breve{\kappa}_{n,\odot} = \begin{cases} \sqrt{2^{\lceil \log_2(2(\log_2(n)-1))\rceil+1} \cdot 2(\log_2(n)-1)n \cdot \sigma_\odot^2} & \text{für } \begin{cases} \text{DCT-II}(n) & \text{mit Alg. 1.19 ,} \\ \text{DCT-III}(n) & \text{mit Alg. 1.20 ,} \\ \text{DST-III}(n) & \text{mit Alg. 1.21 ,} \\ \text{DST-II}(n) & \text{mit Alg. 1.22 ,} \end{cases} \\ \sqrt{2^{\lceil \log_2(2\log_2(n)-1)\rceil+1} \cdot (2\log_2(n)-1)n \cdot \sigma_\odot^2} & \text{für } \begin{cases} \text{DCT-IV}(n) & \text{mit Alg. 1.15 ,} \\ \text{DCT-IV}(n) & \text{mit Alg. 1.18 ,} \\ \text{DCT-IV}(n) & \text{mit Alg. 1.19 ,} \\ \text{DCT-IV}(n) & \text{mit Alg. 1.20 ,} \\ \text{DST-IV}(n) & \text{mit Alg. 1.21 ,} \\ \text{DST-IV}(n) & \text{mit Alg. 1.22 ,} \end{cases} \\ \sqrt{2^{\lceil \log_2(3(\log_2(n)-1))\rceil+1} \cdot 3(\log_2(n)-1)n \cdot \sigma_\odot^2} & \text{für } \begin{cases} \text{DCT-III}(n) & \text{mit Alg. 1.16 ,} \\ \text{DCT-II}(n) & \text{mit Alg. 1.17 .} \end{cases} \end{cases}$$

Im Fall $\operatorname{Cov}\left(\delta_{11}^{(k,\imath)}, \delta_{22}^{(k,\imath)}\right) = \operatorname{Cov}\left(\delta_{12}^{(k,\imath)}, \delta_{21}^{(k,\imath)}\right)$ *bzw.* $\mathbb{E}\left(\delta_{1,\odot}^{(k,\imath)} \delta_{2,\odot}^{(k,\imath)}\right) = 0$ *für alle* $k = 0, \ldots, \frac{n}{2} - 1$ *und für alle auftretenden* $\imath$ *verkleinern sich die Stabilitätskonstanten* $\breve{\kappa}_{n,\times}$ *bzw.* $\breve{\kappa}_{n,\odot}$ *um den Faktor* $\frac{1}{\sqrt{2}}$.

Beweis: Im ersten Fall basieren die Algorithmen 1.19 – 1.22 für $k = 0$ auf den Faktorisierungen (1.45) bzw. (1.55), welche jeweils aus $2\log_2(n) - 1$ Faktoren bestehen. Da $A_n(\boldsymbol{\beta}_0)$ bzw. $\breve{A}_n(\breve{\boldsymbol{\beta}}_0)$ Permutationsmatrizen bezeichnen und die restlichen $2\log_2(n) - 2$ Faktoren wegen (1.43) bzw. (1.53) und aufgrund von (1.25) – (1.29) jeweils einen Repräsentanten gemäß Modellannahme (A1) besitzen, ergibt sich die erste Stabilitätskonstante als direkte Anwendung der Ungleichungen (2.82a) bzw. (2.82b) aus Satz 2.37 mit $\jmath = 2(\log_2(n) - 1)$.

Im zweiten Fall basieren die Algorithmen 1.15 und 1.18 auf den in Lemma 1.11 und Folgerung 1.13 angegebenen Faktorisierungen (1.32) und (1.37), deren Repräsentation in der Festkomma-Arithmetik jeweils genau $2\log_2(n) - 1$ Faktoren gemäß Modellannahme (A1) besitzen. Ebenso enthalten die Produkte (1.48) bzw. (1.60), welche den Algorithmen 1.19 – 1.22 für $k = 1$ zugrunde liegen, jeweils genau $2\log_2(n) - 1$ analog zu (1.43) bzw. (1.53) definierte Matrizen als Faktoren, von denen jeder einzelne der Modellannahme (A1) entspricht. Daher ergibt sich die zweite Stabilitätskonstante, wenn wir (2.82a) bzw. (2.82b) aus Satz 2.37 mit $\jmath = 2\log_2(n) - 1$ anwenden.

Da die Algorithmen 1.16 und 1.17 durch die Faktorisierungen (1.35) bzw. (1.38) beschrieben werden, folgt die dritte Stabilitätskonstante aus (2.82a) bzw. (2.82b) jeweils mit $\jmath = 3(\log_2(n) - 1)$.

Gilt zusätzlich $\operatorname{Cov}\left(\delta_{11}^{(k,\imath)}, \delta_{22}^{(k,\imath)}\right) = \operatorname{Cov}\left(\delta_{12}^{(k,\imath)}, \delta_{21}^{(k,\imath)}\right)$ bzw. $\mathbb{E}\left(\delta_{1,\odot}^{(k,\imath)} \delta_{2,\odot}^{(k,\imath)}\right) = 0$ für alle $k = 0, \ldots, \frac{n}{2} - 1$ und für alle auftretenden $\imath$, so ergeben sich die kleineren Stabilitätskonstanten in allen drei Fällen analog aus den Ungleichungen (2.83a) bzw. (2.83b). ■

Bemerkung 4.18 (Zusammenfassung von Kapitel 4)**.** In Unterabschnitt 4.1.1 werden die in Festkomma-Arithmetik bei einer Matrix-Vektor-Multiplikation auftretenden Rundungsfehler für Eingangsvektoren aus $[-1, 1]^n$ unter der Annahme (4.4) untersucht. Dabei betrachten wir wie in Abschnitt 2.3 sowohl einfache als auch doppelte Genauigkeit. Speziell für eine Blockdiagonalmatrix der Gestalt (4.6) und einen Eingangsvektor $\mathbf{x} \in [-1, 1]^n$ mit $\|\mathbf{x}\|_2 \leq 1$ sind die wichtigsten Abschätzungen in Satz 4.5 zu finden. Unterabschnitt 4.1.2 beginnt in Definition 4.6 mit der Einführung des Begriffes der numerischen Rückwärtsstabilität in einfach bzw. doppelt genauer Festkomma-Arithmetik und liefert in den Sätzen 4.8 – 4.10 die Grundlage für die eigentliche Stabilitätsanalyse in Unterabschnitt 4.1.3. Als Hauptergebnis von Abschnitt 4.1 werden in den Sätzen 4.12 – 4.14 entsprechende Stabilitätskonstanten für die Algorithmen 1.15 – 1.22 unter verschiedenen Skalierungsarten angegeben.

Schließlich wird in Abschnitt 4.2 eine stochastische Rundungsfehleranalyse basierend auf Modell 4.16 durchgeführt. Nachdem Definition 4.15 den Begriff der durchschnittlichen numerischen Rückwärtsstabilität in einfach bzw. doppelt genauer Festkomma-Arithmetik einführt, werden entsprechende Stabilitätskonstanten als Hauptergebnis in Satz 4.17 zusammengefasst. □

5 Numerische Ergebnisse

In Kapitel 3 und 4 werden sowohl in Gleit- als auch in Festkomma-Arithmetik numerische Stabilitätsanalysen für die Algorithmen aus Abschnitt 1.3 durchgeführt. Dabei sind sowohl obere Schranken für den ungünstigsten Fall als auch Schätzungen des durchschnittlich auftretenden Fehlers von Interesse. Um die theoretischen Ergebnisse zu überprüfen, sind die Algorithmen aus Abschnitt 1.3 in MATLAB implementiert worden.

5.1 Testrechnungen in Gleitkomma-Arithmetik

Aufgrund der Rundungsfehler steht uns nur in den seltensten Fällen das exakte Ergebnis einer Folge von arithmetischen Operationen in Gleitkomma-Arithmetik zur Verfügung. Um dennoch die Qualität ermittelter Abschätzungen für den bei einem Algorithmus auftretenden relativen Fehler sinnvoll testen zu können, schlägt Higham in [24, S. 454] am Beispiel der FFT folgende Methode vor. Anstelle des relativen Fehlers wird ersatzweise die Größe

$$\varsigma_n := \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} \tag{5.1}$$

betrachtet, wobei $\hat{\mathbf{x}} = \mathrm{fl}\left(F_n^{-1}\,\mathrm{fl}\left(F_n\mathbf{x}\right)\right)$ das in Gleitkomma-Arithmetik berechnete Resultat eines schnellen Algorithmus zur Fourier-Transformation gefolgt von der inversen Fourier-Transformation bezeichnet. Um die Resultate aus Kapitel 3 für die Algorithmen 1.15 – 1.22 zu überprüfen, betrachten wir anstelle der Fourier-Matrix (1.1) die entsprechenden Sinus- und Kosinus-Matrizen (1.2) – (1.7). In den mittels MATLAB erzeugten Abbildungen 5.1 – 5.3 ist nun analog zu [24, Figure 24.1] jeweils die Größe (5.1) für $n = 2^t,\ t = 3, \ldots, 12$ und $M = 100$ zufällige Vektoren mit standard-normalverteilten Komponenten abgetragen.

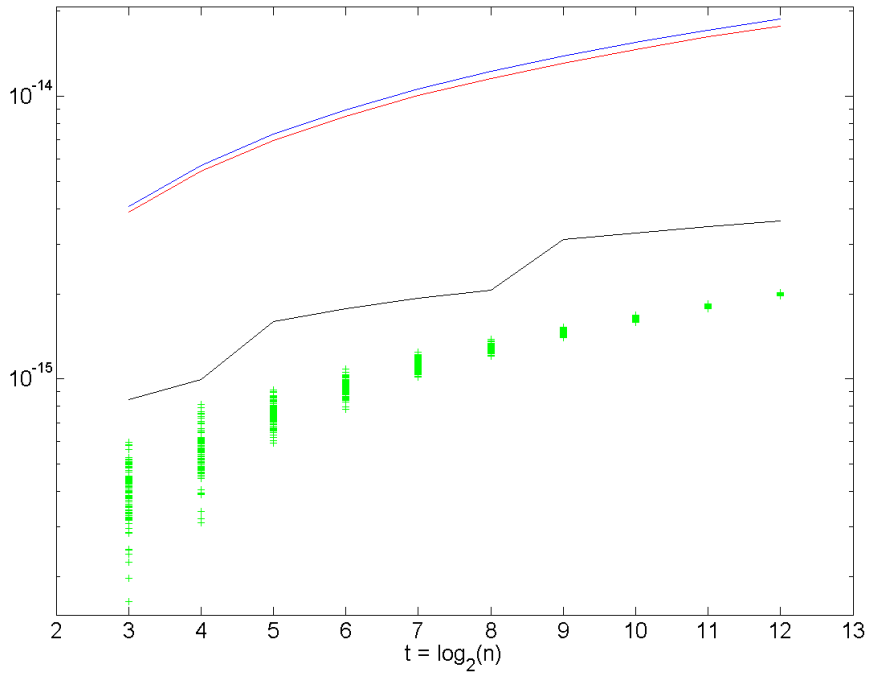


Abbildung 5.1: Relative Fehler aus (5.1) mit $\hat{\mathbf{x}} = \mathrm{fl}(C_n^{\mathrm{IV}}\,\mathrm{fl}(C_n^{\mathrm{IV}}\mathbf{x}))$ bei zweifach ausgeführter DCT-IV(n) mittels Algorithmus 1.15 in Gleitkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 3.12 (blau), 3.18 (rot) und 3.23 (schwarz) abgetragen.

Darüber hinaus finden sich dort die jeweiligen Schranken aus den Sätzen 3.12, 3.18 und 3.23. Zur Generierung der Zufallsvektoren ist der MATLAB-Befehl `randn([n 1])` verwendet worden.

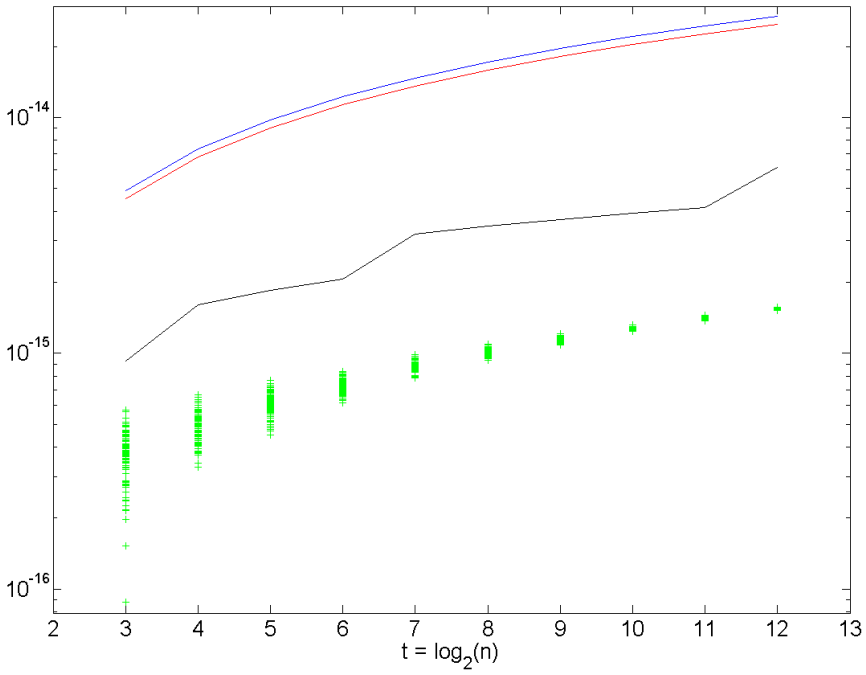


Abbildung 5.2: Relative Fehler aus (5.1) mit $\hat{\mathbf{x}} = \mathrm{fl}(C_n^{\mathrm{II}}\,\mathrm{fl}(C_n^{\mathrm{III}}\mathbf{x}))$ bei der DCT-III(n) mittels Algorithmus 1.16 gefolgt von der DCT-II(n) mittels Algorithmus 1.17 jeweils in Gleitkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 3.12 (blau), 3.18 (rot) und 3.23 (schwarz) abgetragen.

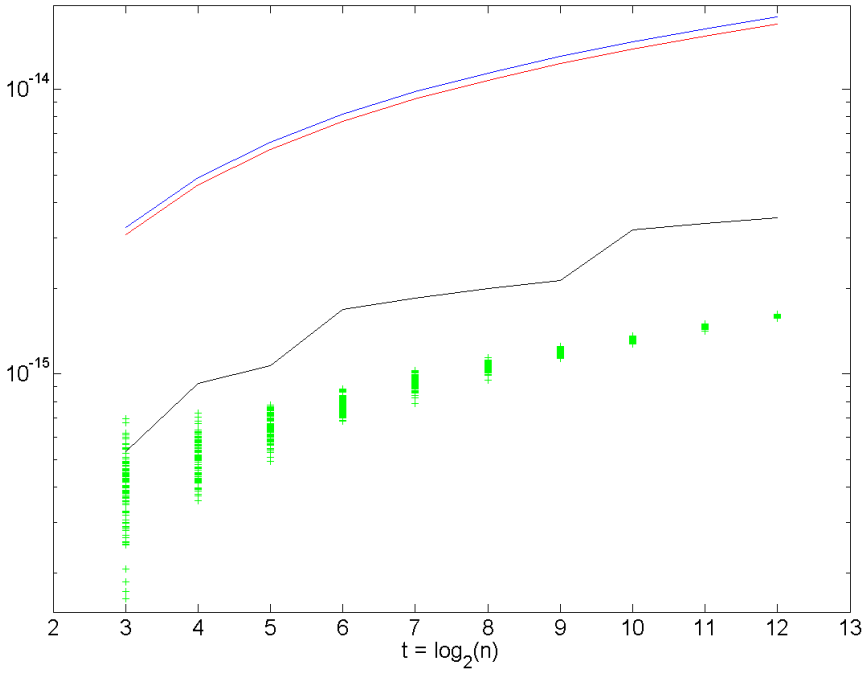


Abbildung 5.3: Relative Fehler aus (5.1) mit $\hat{\mathbf{x}} = \mathrm{fl}(C_n^{\mathrm{III}}\,\mathrm{fl}(C_n^{\mathrm{II}}\mathbf{x}))$ bei der DCT-II(n) mittels Algorithmus 1.19 gefolgt von der DCT-III(n) mittels Algorithmus 1.20 jeweils in Gleitkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 3.12 (blau), 3.18 (rot) und 3.23 (schwarz) abgetragen.

5.2 Simulationen für die Festkomma-Arithmetik

Zur Unterstützung der Aussagen aus Kapitel 4 sind nun alle Algorithmen 1.15 – 1.22 analog zu Abschnitt 5.1 in simulierter Festkomma-Arithmetik implementiert worden. Um die Operation (2.51) für ein festes $q \in \mathbb{N}$ auf eine Zahl $x \in [-1,1]$ mit der Eigenschaft (2.52) anzuwenden, haben wir jeweils Methode 5.1 angewandt.

Methode 5.1. Sei $x \in [-1,1]$ und $q \in \mathbb{N}$. Dann simulieren wir die Abbildung $\hat{\mathbf{x}} := \mathrm{fix}(\mathbf{x})$ wie folgt:

(1) Die Zahl x wird mit 2^q skaliert, so dass sich das Zwischenergebnis $y := 2^q x$ im Intervall $[-2^q, 2^q]$ befindet.

(2) Mit dem in MATLAB zur Verfügung stehenden Befehl `fix` werden die eventuell noch verbleibenden Nachkommastellen abgeschnitten, so dass als zweites Zwischenergebnis $z :=$`fix`(y) eine ganze Zahl entsteht.

(3) Anschließend wird mit 2^{-q} zurückskaliert, so dass $\hat{x} := 2^{-q} z$ in $\mathbb{M}_q$ liegt. □

Aus dem Basismodell (2.53) für die Festkomma-Arithmetik geht weiter hervor, dass lediglich die Multiplikation zweier Zahlen $\hat{x}, \hat{y} \in \mathbb{M}_q$ gesondert modelliert werden muss. Demnach verfahren wir mit dem Produkt $p := \hat{x}\hat{y}$ genauso wie mit der Abbildung (2.51) nach $\mathbb{M}_q$, d.h., wir skalieren wiederum mit 2^q, schneiden die Nachkommastellen mit dem MATLAB-Befehl `fix` ab und skalieren anschließend mit 2^{-q} zurück. Um die Anzahl der Nachkommastellen variieren zu können, sind die in Festkomma-Arithmetik simulierten Programme und Unterprogramme, welche für die Algorithmen 1.15 – 1.22 benötigt werden, mit einem zusätzlichen Parameter q ausgestattet. Da wir entsprechend der Modellannahmen (2.53) keinen Überlauf erlauben, was bei den Algorithmen 1.15 – 1.22 im Fall (4.15) gewährleistet wird, müssen die Komponenten eines Eingangsvektors $\hat{\mathbf{x}} \in \mathbb{M}_q^n$ mit $2^{-\lceil \log_2(\|\hat{\mathbf{x}}\|_2)\rceil}$ vorskaliert werden. Demzufolge ist die Forderung $q > \lceil \log_2(\|\hat{\mathbf{x}}\|_2)\rceil$ an die Anzahl q der Nachkommastellen sinnvoll, um nach der Abbildung (2.51) nicht nur den Nullvektor zu erhalten.
In einer ersten Simulation untersuchen wir für jeden einzelnen der Algorithmen 1.15 – 1.22 das numerische Verhalten in Abhängigkeit von q, also der Anzahl von Nachkommastellen, wobei die Transformationslänge $n = 2^t$ beibehalten wird. Bei jedem Durchlauf erzeugen wir über den MATLAB-Befehl `rand([n 1])` durch entsprechende Skalierung und Verschiebung $M = 100$ zufällige Vektoren $\boldsymbol{\xi}^{(i)} \in [-1,1]^n$ $(i = 1, \ldots, M)$ mit auf dem Intervall $[-1,1]$ gleichverteilten Komponenten. Die im Fall $\boldsymbol{\xi}^{(i)} = \mathbf{1} \in [-1,1]^n$ für mindestens ein i sogar scharfe Abschätzung

$$\max_{i=1,\ldots,M} \|\boldsymbol{\xi}^{(i)}\|_2 \;\leq\; \sqrt{n} \tag{5.2}$$

erfordert gemäß der Bedingung (4.15) eine Vorskalierung. An dieser Stelle wählen wir

$$2^{-\lceil \frac{t}{2} \rceil} \;\leq\; \frac{1}{\sqrt{n}} \tag{5.3}$$

als Skalierungsfaktor, da somit nur eine Bitverschiebung notwendig ist [40, S. 95]. Nach dem Skalieren wird jeweils noch Methode 5.1 angewandt. Die so entstandenen Vektoren $\mathbf{x}^{(i)} \in \mathbb{M}_q^n$ $(i = 1, \ldots, M)$ erfüllen wie gewünscht Bedingung (4.15). Jeder einzelne Vektor $\mathbf{x}^{(i)}$ wird nun jeweils mit einem der in simulierter Festkomma-Arithmetik für q Nachkommastellen implementierten Algorithmen 1.15 – 1.22 transformiert. Die auf diese Weise erzeugten Vektoren $\tilde{\mathbf{y}}^{(i)} \in \mathbb{M}_q^n$ $(i = 1, \ldots, M)$ werden - da die exakten Ergebnisse $\mathbf{y}^{(i)}$ für einen Vergleich nicht zur Verfügung stehen - anschließend entsprechend Satz 1.4 unter Verwendung eines der Algorithmen 1.15 – 1.22 mit $2q$ Nachkommastellen zurücktransformiert, so dass wir die Vektoren $\tilde{\mathbf{x}}^{(i)} \in \mathbb{M}_q^n$ $(i = 1, \ldots, M)$ erhalten. Dabei können wir die Ausführung der Rücktransformation mit $2q$ Nachkommastellen im Vergleich zur Berechnung mit q Nachkommastellen als näherungsweise exakt ansehen. Die in Satz 1.4 nachgewiesene Orthogonalität der Matrizen (1.2) – (1.7) impliziert daher

$$\|\tilde{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)}\|_2 \;\approx\; \|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|_2 \;, \tag{5.4}$$

so dass wir nun anstelle der Normen der absoluten Fehler $\tilde{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}$ die Größen $\|\tilde{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)}\|_2$ verwenden. Die Abbildungen 5.4 und 5.5 stellen die Ergebnisse exemplarisch für die DCT-II(2^t) mittels Algorithmus 1.19 im Fall $t = 4$ bzw. $t = 11$ dar, indem die Größen (5.4) unter Verwendung einer logarithmischen Skala in Abhängigkeit von q abgetragen sind.

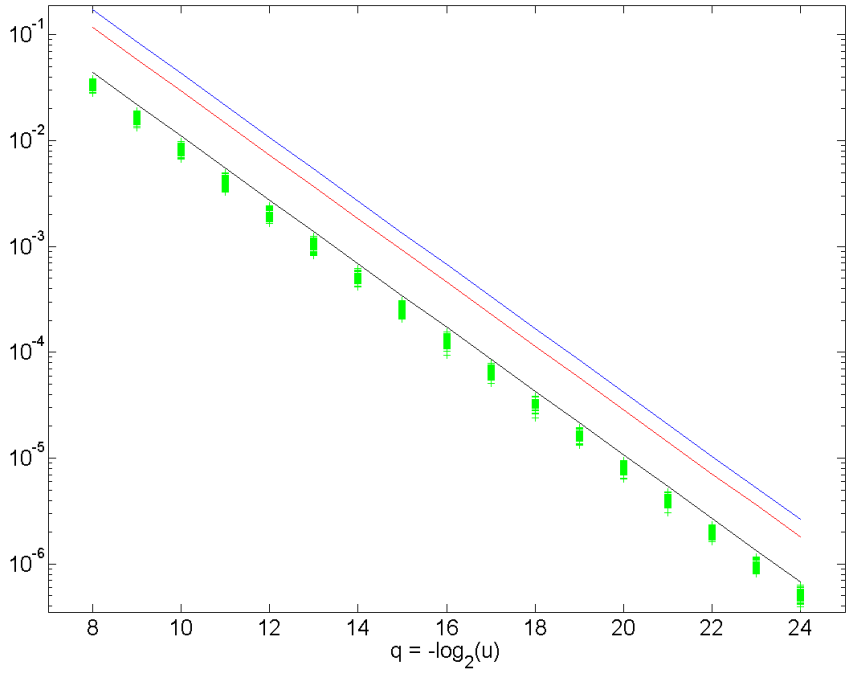


Abbildung 5.4: Absolute Fehler aus (5.4) bei der DCT-II(2^4) mittels Algorithmus 1.19 gefolgt von der DCT-III(2^4) mittels Algorithmus 1.20 in Festkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 4.12 (blau für einfache, rot für doppelte Genauigkeit) und 4.17 (schwarz) mit $\sigma_{\odot}^2 = \frac{1}{12}$ abgetragen.

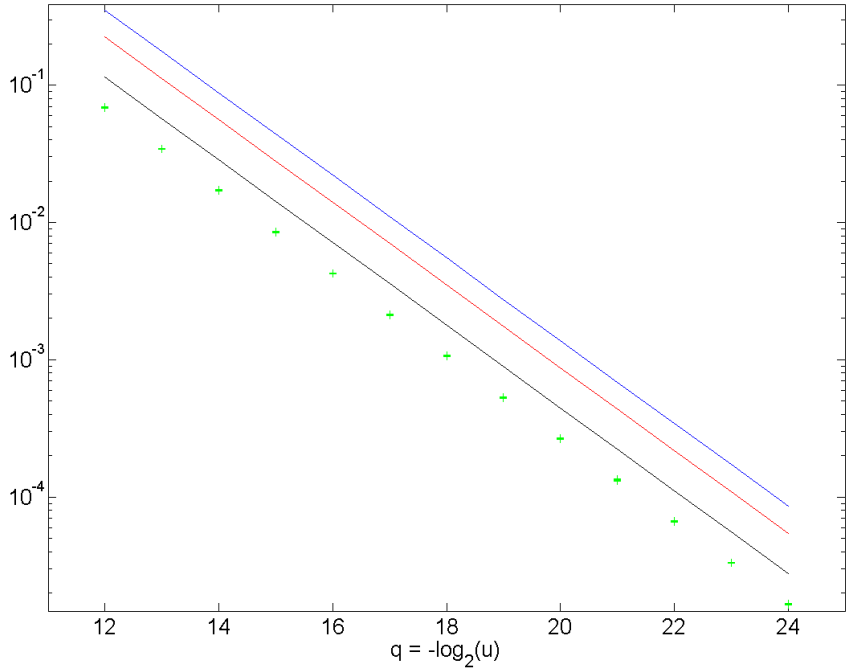


Abbildung 5.5: Absolute Fehler aus (5.4) bei der DCT-II(2^{11}) mittels Algorithmus 1.19 gefolgt von der DCT-III(2^{11}) mittels Algorithmus 1.20 in Festkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 4.12 (blau für einfache, rot für doppelte Genauigkeit) und 4.17 (schwarz) mit $\sigma_{\odot}^2 = \frac{1}{12}$ abgetragen.

Darüber hinaus sind die theoretischen Schranken aus Satz 4.12 (iii) eingezeichnet, wobei anstelle von $\|\mathbf{x}\|_2$ jeweils die Größe

$$\max_{i=1,\dots,M} \|\mathbf{x}^{(i)}\|_2 \tag{5.5}$$

verwendet wird. Zusätzlich sind die aus Satz 4.17 mittels Jensen-Ungleichung (A.35) resultierenden oberen Schranken $2^{-q}\breve{\kappa}_{n,\times}$ und $2^{-q}\breve{\kappa}_{n,\odot}$ für $\mathbb{E}\left(\| \Delta\mathbf{X} \|_2\right)$ angegeben.
In einer zweiten Testreihe haben wir bei konstantem q die Transformationslänge $n = 2^t$ variiert. Wie bereits zuvor begründet wählen wir dazu mindestens $q = \lceil \frac{t_{\max}}{2} \rceil$, wobei $t_{\max}$ den während der Testreihe größten auftretenden Wert für $\log_2(n)$ bezeichnet. Abbildung 5.6 zeigt die Ergebnisse exemplarisch für die DCT-IV(2^t) mittels Algorithmus 1.20 im Fall $q = 24$ und $t = 3, \dots, 12$, wobei wiederum $M = 100$ Zufallsvektoren verwendet werden. Der Übersichtlichkeit enthält Abbildung 5.6 nur die Maxima der Größen (5.4) im Vergleich zu den theoretischen Schranken aus Satz 4.12 (i) für den ungünstigsten Fall bzw. $2^{-q}\breve{\kappa}_{n,\odot}$ aus Satz 4.17 für die durchschnittliche Fehlernorm.

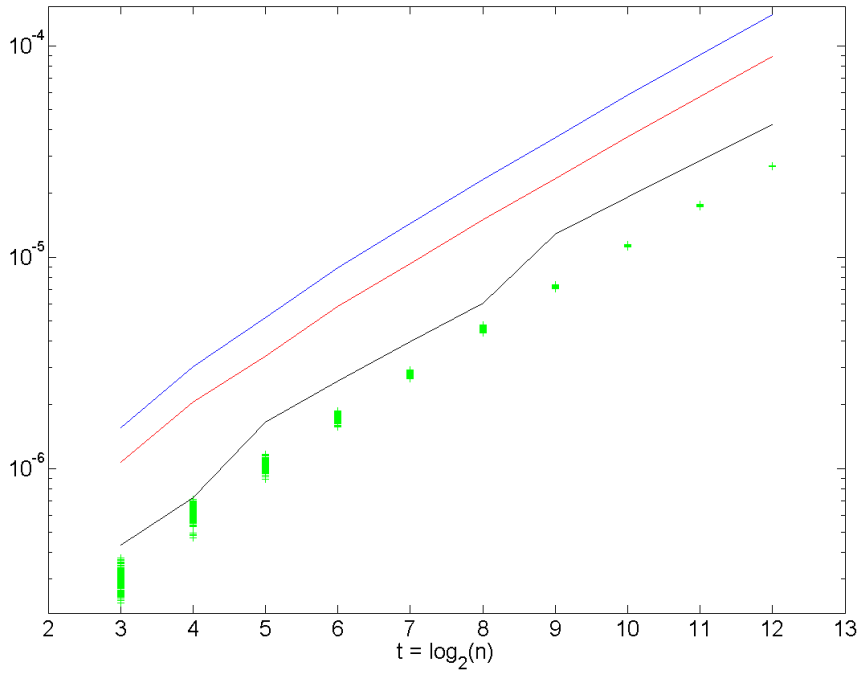


Abbildung 5.6: Absolute Fehler aus (5.4) bei zweifach ausgeführter DCT-IV(n) mittels Algorithmus 1.20 in Festkomma-Arithmetik. Zusätzlich sind die theoretischen Schranken aus den Sätzen 4.12 (blau für einfache, rot für doppelte Genauigkeit) und 4.17 (schwarz) mit $\sigma_\odot^2 = \frac{1}{12}$ abgetragen.

Bemerkung 5.2 (Zusammenfassung von Kapitel 5)**.** Um die theoretischen Abschätzungen aus den Sätzen 3.9, 3.16 und 3.23 bzw. aus den Sätzen 4.12 und 4.17 auf ihre Qualität hin zu überprüfen, sind die Algorithmen 1.15 – 1.22 in MATLAB implementiert worden. Für die Simulationen in Festkomma-Arithmetik hat insbesondere Methode 5.1 bei der jeweiligen Implementation der Algorithmen Berücksichtigung gefunden.

□

A Anhang

A.1 Maß- und Wahrscheinlichkeitstheorie

An dieser Stelle werden die Grundlagen der Maß- und Integrationstheorie sowie der Wahrscheinlichkeitstheorie nach [3] und [4] bereit gestellt, welche für das Verständnis einer allgemeinen Zufallsvariable wesentlich sind.

A.1.1 Grundlagen der Maß- und Integrationstheorie

Zunächst widmen wir uns der Einführung von Maßen, mit deren Hilfe der Integrationsbegriff eingeführt wird. Dazu benötigen wir den Begriff der σ-Algebra.

Definition A.1 (vgl. [3], Definition 1.1)**.** Ein System $\mathfrak{S}$ von Teilmengen einer (nichtleeren) Menge Ω heißt *σ-Algebra* (in Ω), wenn es die folgenden Eigenschaften besitzt:

(i) $\Omega \in \mathfrak{S}$;

(ii) $A \in \mathfrak{S} \;\Rightarrow\; \Omega \setminus A \in \mathfrak{S}$;

(iii) für jede Folge $(A_n)_{n\in\mathbb{N}}$ von Mengen aus $\mathfrak{S}$ liegt $\bigcup\limits_{n\in\mathbb{N}} A_n$ in $\mathfrak{S}$. □

Offenbar werden von jeder σ-Algebra $\mathfrak{S}$ die zu (i) und (iii) „dualen“ Eigenschaften erfüllt:

(i)′ $\emptyset \in \mathfrak{S}$;

(iii)′ für jede Folge $(A_n)_{n\in\mathbb{N}}$ von Mengen aus $\mathfrak{S}$ liegt $\bigcap\limits_{n\in\mathbb{N}} A_n$ in $\mathfrak{S}$.

Wegen $A \cap \Omega = A$ und $A \cup \emptyset = A$ enthält eine σ-Algebra $\mathfrak{S}$ (in Ω) auch mit je endlich vielen Mengen deren Vereinigung und Durchschnitt. Zusammen mit Eigenschaft (ii) folgt weiterhin noch

(ii)’ $A, B \in \mathfrak{S} \Rightarrow A \setminus B = A \cap (\Omega \setminus B) \in \mathfrak{S}$.

Die Menge der σ-Algebren in einer nichtleeren Menge Ω ist nicht leer, da die Potenzmenge $\mathcal{P}(\Omega)$ alle Eigenschaften einer σ-Algebra (in Ω) erfüllt. Darüber hinaus gilt für jedes $\Omega' \subseteq \Omega$, dass $\Omega' \cap \mathfrak{S} := \{\Omega' \cap A \mid A \in \mathfrak{S}\}$ eine σ-Algebra (in Ω') ist, falls $\mathfrak{S}$ eine σ-Algebra (in Ω) war. In diesem Fall wird $\Omega' \cap \mathfrak{S}$ die *Spur* von $\mathfrak{S}$ in Ω' genannt.
Zur Konstruktion von σ-Algebren ist der nachfolgende Satz ein wichtiges Werkzeug.

Satz A.2 (vgl. [3, Satz 1.2])**.** *Sei $I \neq \emptyset$ eine beliebige Indexmenge. Dann ist jeder Durchschnitt $\bigcap\limits_{i\in I} \mathfrak{S}_i$ einer Familie $(\mathfrak{S}_i)_{i\in I}$ von σ-Algebren in einer Menge Ω selbst wiederum eine σ-Algebra in Ω.*

Mit Hilfe von Satz A.2, dessen Beweis im Nachprüfen der Eigenschaften (i)-(iii) aus Definition A.1 besteht, folgt für jedes System $\mathfrak{E}$ von Teilmengen von Ω die Existenz einer kleinsten, $\mathfrak{E}$ enthaltenden σ-Algebra $\sigma(\mathfrak{E})$. Dabei heißt dann $\sigma(\mathfrak{E})$ die von $\mathfrak{E}$ (in Ω) *erzeugte σ-Algebra* und $\mathfrak{E}$ ein *Erzeuger* von $\sigma(\mathfrak{E})$. Falls $\mathfrak{E}$ selbst schon eine σ-Algebra in Ω ist, gilt offenbar $\mathfrak{E} = \sigma(\mathfrak{E})$. Besteht $\mathfrak{E}$ dagegen nur aus Ω, erhalten wir $\sigma(\mathfrak{E}) = \{\emptyset, \Omega\}$.
Im Spezialfall $\Omega = \mathbb{R}$ kann das System $\mathfrak{I}$ betrachtet werden, welches sämtliche Teilintervalle $I \subset \mathbb{R}$ enthält. Die von $\mathfrak{I}$ erzeugte $\sigma-$Algebra $\sigma(\mathfrak{I})$ heißt dann die $\sigma-$Algebra der *Borel-Mengen* von $\mathbb{R}$ und wird mit $\mathcal{B}(\mathbb{R})$ bezeichnet (vgl. [3, Definition 6.1, Satz 6.4]). Für die Erweiterung $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ ist das System

$$\mathcal{B}(\bar{\mathbb{R}}) := \Big\{ B, B \cup \{\infty\}, B \cup \{-\infty\}, B \cup \{\infty\} \cup \{-\infty\} \;\Big|\; B \in \mathcal{B}(\mathbb{R}) \Big\}$$

offenbar eine σ-Algebra in $\bar{\mathbb{R}}$ mit $\mathcal{B}(\mathbb{R})$ als Spur in $\mathbb{R}$, d.h. $\mathbb{R} \cap \mathcal{B}(\bar{\mathbb{R}}) = \mathcal{B}(\mathbb{R})$. Daher wird analog $A \subset \bar{\mathbb{R}}$ als *Borelsch* in $\bar{\mathbb{R}}$ bezeichnet, falls $A \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})$ gilt. Analog konstruieren wir die σ-Algebren $\mathcal{B}(\mathbb{R}^n)$ und $\mathcal{B}(\bar{\mathbb{R}}^n)$, deren Elemente entsprechend als *Borelsch* in $\mathbb{R}^n$ bzw. in $\bar{\mathbb{R}}^n$ bezeichnet werden. Auf allgemeinen σ-Algebren lassen sich wie folgt Maße definieren.

Definition A.3 (vgl. [3], Definitionen 3.1, 3.3)**.** Sei $\mathfrak{S}$ eine σ-Algebra in Ω. Dann heißt $\mu : \mathfrak{S} \to [0, \infty]$ ein *Maß* (auf $\mathfrak{S}$), falls

$$\mu(\emptyset) = 0 \tag{A.1}$$

und für jede Folge $(A_n)_{n\in\mathbb{N}}$ paarweise disjunkter Mengen aus $\mathfrak{S}$ mit in $\mathfrak{S}$ gelegener Vereinigung

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n) \qquad (\sigma\text{-Additivität}) \tag{A.2}$$

gilt. Jeder Funktionswert $\mu(A)$ von μ für ein $A \in \mathfrak{S}$ wird das (μ-)*Maß der Menge* A genannt. Gilt $\mu(\Omega) < \infty$ (und folglich auch $\mu(A) < \infty$ für alle $A \in \mathfrak{S}$), so heißt μ *endlich*. □

Auf der σ-Algebra $\mathcal{B}(\mathbb{R})$ (bzw. $\mathcal{B}(\mathbb{R}^n)$) interessieren uns nun zunächst einmal Maße, welche jedem Intervall $[a, b[\subset \mathbb{R}$ (bzw. $[a_1, b_1[\times \ldots \times [a_n, b_n[\in \mathbb{R}^n$) seinen (*n-dimensionalen*) *Elementarinhalt*

$$b_1 - a_1 \qquad \left(\text{bzw.} \qquad \prod_{i=1}^{n} (b_i - a_i)\right)$$

zuordnen. Es existiert jedoch genau ein solches Maß (vgl. [3], Sätze 6.2, 4.3, 5.4). Dieses Maß heißt *Lebesgue-Borelsches Maß* auf $\mathbb{R}$ (bzw. $\mathbb{R}^n$) und wird mit λ (bzw. λ^n) bezeichnet. Entsprechend wird $\lambda(B)$ (bzw. $\lambda^n(B)$) für eine beliebige Borel-Menge $B \in \mathcal{B}(\mathbb{R})$ (bzw. $B \in \mathcal{B}(\mathbb{R}^n)$) das (*n-dimensionale*) *Lebesgue-Borel-Maß* von B genannt (vgl. [3], Definition 6.3). Darüber hinaus ist die Restriktion λ_C von λ (bzw. λ^n_C von λ^n) für ein $C \in \mathcal{B}(\mathbb{R})$ (bzw. $C \in \mathcal{B}(\mathbb{R}^n)$) ein Maß auf der Spur-$\sigma$-Algebra $C \cap \mathcal{B}(\mathbb{R})$ (bzw. $C \cap \mathcal{B}(\mathbb{R}^n)$), welches als *Lebesgue-Borelsches Maß auf* C bezeichnet wird. Das Lebesgue-Borelsche Maß gehört zur Klasse der *Borel-Maße*. Dies sind genau die Maße auf $\mathcal{B}(\mathbb{R})$ (bzw. $\mathcal{B}(\mathbb{R}^n)$), welche jedem beschränkten $B \in \mathcal{B}(\mathbb{R})$ (bzw. $B \in \mathcal{B}(\mathbb{R}^n)$) ein endliches Maß zuordnen. Wegen $\mu(\emptyset) = 0$ und aufgrund der aus der σ-Additivität folgenden Isotonie-Eigenschaft eines Maßes (d.h. für $A, B \in \mathfrak{S}$ mit $A \subseteq B$ folgt stets $\mu(A) \leq \mu(B)$) ist somit insbesondere jedes endliche Maß auf $\mathcal{B}(\mathbb{R})$ ein Borel-Maß. Zu jedem endlichen Borel-Maß $\mu \neq 0$ lässt sich wegen $0 \leq \mu(B) \leq \mu(\mathbb{R}) < \infty$ für alle $B \in \mathcal{B}(\mathbb{R})$ durch $\nu := \frac{1}{\mu(\mathbb{R})}\mu$ ein weiteres Maß auf $\mathcal{B}(\mathbb{R})$ definieren.

Definition A.4 (vgl. [3], S.38)**.** Ein Maß μ auf einer σ-Algebra $\mathfrak{S}$ in einer Menge Ω wird *Wahrscheinlichkeitsmaß* (*W-Maß*) genannt, wenn $\mu(\Omega) = 1$ gilt. □

Wegen $\mu(\emptyset) = 0$ und wiederum aufgrund der Isotonie gilt für jedes W-Maß

$$0 \leq \mu(A) \leq 1 \qquad (A \in \mathfrak{S}).$$

Betrachten wir nun ein W-Maß auf $\mathcal{B}(\mathbb{R})$. Da jedes offene Intervall $]-\infty, x[$ in $\mathcal{B}(\mathbb{R})$ liegt, wird durch

$$F_\mu(x) := \mu(]-\infty, x[)$$

eine Funktion $F_\mu : \mathbb{R} \to [0, 1]$ definiert, welche *Verteilungsfunktion* von μ heißt. Für eine beliebige solche Verteilungsfunktion $\mathbb{F}_\mu$ gilt wegen $]-\infty, b[\setminus]-\infty, a[\; = \; [a, b[$ für $a \leq b$ somit

$$\mu([a, b[) = F_\mu(b) - F_\mu(a) \; .$$

Es besteht nun der folgende Zusammenhang:

Satz A.5 (vgl. [3], Satz 6.6)**.** *Eine Funktion $F : \mathbb{R} \to [0, 1]$ ist genau dann die Verteilungsfunktion eines – notwendigerweise eindeutig bestimmten – W-Maßes μ auf $\mathcal{B}(\mathbb{R})$, wenn F monoton wachsend und linksseitig stetig ist sowie zusätzlich*

$$\lim_{x\to-\infty} F(x) \; = \; 0 \; , \qquad \lim_{x\to\infty} F(x) \; = \; 1 \tag{A.3}$$

gilt.

Zur Vereinfachung vereinbaren wir die folgenden abkürzenden Sprechweisen. Für eine Menge Ω und eine σ-Algebra $\mathfrak{S}$ in Ω werden das Paar $(\Omega, \mathfrak{S})$ ein *Messraum* und die Mengen aus $\mathfrak{S}$ *messbar* genannt. Ist auf $\mathfrak{S}$ noch ein Maß μ definiert, bezeichnet das Tripel $(\Omega, \mathfrak{S}, \mu)$ einen *Maßraum*. Desweiteren heißen alle Mengen $N \in \mathfrak{S}$, für die $\mu(N) = 0$ gilt, *μ-Nullmengen*. Existiert nun eine μ-Nullmenge N, so dass eine Eigenschaft E für alle $\omega \in \Omega \setminus N$ erfüllt ist, so sagen wir, dass E dann *μ-fast überall* gilt. Falls insbesondere μ ein W-Maß darstellt, nennen wir $(\Omega, \mathfrak{S}, \mu)$ einen *Wahrscheinlichkeitsraum* (*W-Raum*) und jedes Element $A \in \mathfrak{S}$ ein *Ereignis*. In Analogie zum Stetigkeitsbegriff für topologische Räume führen wir den Begriff einer messbaren Abbildung ein.

Definition A.6 (vgl. [3], Definition 7.1)**.** Es seien $(\Omega, \mathfrak{S})$ und $(\Omega', \mathfrak{S}')$ zwei Messräume. Eine Abbildung $T : \Omega \to \Omega'$ heißt $(\mathfrak{S}, \mathfrak{S}')$-*messbar*, wenn für alle $A' \in \mathfrak{S}'$ die Bedingung

$$T^{-1}(A') := \{\omega \in \Omega \ : \ T(\omega) \in A'\} \in \mathfrak{S} \tag{A.4}$$

erfüllt ist, d.h. also, wenn das Urbild einer $\mathfrak{S}'$-messbaren Menge $\mathfrak{S}$-messbar ist. Mit der Bezeichung $T^{-1}(\mathfrak{S}') := \{T^{-1}(A') : A' \in \mathfrak{S}'\}$ ist $T^{-1}(\mathfrak{S}') \subseteq \mathfrak{S}$ eine äquivalente Formulierung von (A.4). Symbolisch drücken wir die $(\mathfrak{S}, \mathfrak{S}')$-Messbarkeit von T auch durch $T : (\Omega, \mathfrak{S}) \to (\Omega', \mathfrak{S}')$ aus.

Die Komposition $T_2 \circ T_1$ messbarer Abbildungen $T_1 : (\Omega_1, \mathfrak{S}_1) \to (\Omega_2, \mathfrak{S}_2)$ und $T_2 : (\Omega_2, \mathfrak{S}_2) \to (\Omega_3, \mathfrak{S}_3)$ ist wegen der für alle $A \subseteq \Omega_3$ gültigen Gleichheit $(T_2 \circ T_1)^{-1}(A) = T_1^{-1}(T_2^{-1}(A))$ dann $(\mathfrak{S}_1, \mathfrak{S}_3)$-messbar. Desweiteren kann mit Hilfe einer messbaren Abbildung $T : (\Omega, \mathfrak{S}) \to (\Omega', \mathfrak{S}')$ für jedes Maß μ auf $\mathfrak{S}$ durch

$$A' \mapsto \mu\left(T^{-1}(A')\right) \qquad\qquad (A' \in \mathfrak{S}') \tag{A.5}$$

ein Maß μ' auf $\mathfrak{S}'$ definiert werden, welches das *Bild von μ unter der Abbildung T* heißt und mit $T(\mu)$ bezeichnet wird. Die Bildung derartiger Bildmaße ist im Fall der Wohldefiniertheit transitiv.
Als Nächstes betrachten wir jetzt auf einem festen Messraum $(\Omega, \mathfrak{S})$ Funktionen $f : \Omega \to \bar{\mathbb{R}}$, die – im Unterschied zu *reellen* Funktionen $f : \Omega \to \mathbb{R}$ – als *numerische* Funktionen bezeichnet werden. Da es sich bei dem Teilmengensystem $\{[\alpha, \infty] \ : \ \alpha \in \mathbb{R}\}$ um einen Erzeuger von $\mathcal{B}(\bar{\mathbb{R}})$ handelt, ist eine numerische Funktion f auf Ω genau dann $(\mathfrak{S}, \mathcal{B}(\bar{\mathbb{R}}))$-messbar (oder kurz $\mathfrak{S}$-messbar), wenn für jedes $\alpha \in \mathbb{R}$ die Urbildmenge $\{\omega \in \Omega \ : \ f(\omega) \geq \alpha\}$ Element der σ-Algebra $\mathfrak{S}$ ist (vgl. [3, Satz 7.2 und 9.1]). Ist die $\mathfrak{S}$-Messbarkeit von numerischen Funktionen f, g auf Ω bekannt und sind darüber hinaus μ-fast überall die Funktionen $f \pm g$ definiert, so sind die Funktionen $f \pm g$ und

$$fg = \frac{(f+g)^2 - (f-g)^2}{4}$$

$\mathfrak{S}$-messbar (vgl. [3, Satz 9.4]). Genauso folgt mit der $\mathfrak{S}$-Messbarkeit einer Folge $(f_k)_{k \in \mathbb{N}}$ numerischer Funktionen auf Ω auch die $\mathfrak{S}$-Messbarkeit von

$$\sup_{k \in \mathbb{N}} f_k \ , \quad \inf_{k \in \mathbb{N}} f_k \ , \quad \limsup_{k \to \infty} f_k \ , \quad \liminf_{k \to \infty} f_k$$

(vgl. [3, Satz 9.5]) sowie die $\mathfrak{S}$-Messbarkeit – falls $(f_k)_{k \in \mathbb{N}}$ punktweise auf Ω konvergiert – von der Grenzfunktion $\lim\limits_{k \to \infty} f_k$. Als Spezialfall des Infimums und Supremums folgt ebenso die $\mathfrak{S}$-Messbarkeit des Minimums und Maximums endlich vieler $\mathfrak{S}$-messbarer numerischer Funktionen. Insbesondere sind der *Positivteil*

$$f^+ \ := \ \max\{f, 0\}$$

und der *Negativteil*

$$f^- \ := \ -\min\{f, 0\}$$

(und damit auch der *Absolutbetrag* $|f| = f^+ + f^-$) einer numerischen Funktion f auf Ω genau dann $\mathfrak{S}$-messbar, wenn f selbst $\mathfrak{S}$-messbar ist (vgl. [3, Satz 9.8]). Die einfachste numerische Funktion ist die *charakteristische Funktion*

$$1_A(\omega) \ := \ \begin{cases} 1, & \omega \in A, \\ 0, & \omega \in \Omega \setminus A \end{cases} \tag{A.6}$$

von A, welche genau dann $\mathfrak{S}$-messbar ist, wenn $A \in \mathfrak{S}$ erfüllt ist. Mit Hilfe von $\mathfrak{S}$-messbaren charakteristischen Funktionen können wir die folgenden Funktionen günstig darstellen, über welche schließlich das Integral nichtnegativer $\mathfrak{S}$-messbarer numerischer Funktionen definiert wird.

Definition A.7 (vgl. [3], Definitionen 10.1, 10.3)**.** Sei $(\Omega, \mathfrak{S}, \mu)$ ein Maßraum. Eine nichtnegative reelle Funktion e auf Ω, welche $\mathfrak{S}$-messbar ist und nur endlich viele Werte annimmt, heißt $\mathfrak{S}$-*Elementarfunktion*. Sind $\{\alpha_1, \ldots, \alpha_k\}$ die paarweise verschiedenen Werte von e, so sind die Urbilder $A_k := \{\omega \in \Omega \ : \ e(\omega) = \alpha_k\}$ paarweise disjunkt und $\mathfrak{S}$-messbar aufgrund der $\mathfrak{S}$-Messbarkeit von e. Die Darstellung

$$e \ = \ \sum_{i=1}^{k} \alpha_i \, 1_{A_i}$$

bezeichnen wir als *Normaldarstellung* und die Zahl

$$\int\limits_{\Omega} e \; d\mu \ := \ \sum_{i=1}^{k} \alpha_i \, \mu(A_i)$$

als das μ-*Integral* von e (über Ω).

Da zu jeder $\mathfrak{S}$-messbaren nichtnegativen numerischen Funktion f auf Ω mindestens eine monotone Folge $(e_m)_{m\in\mathbb{N}}$ von Elementarfunktionen mit $\sup\limits_{m\in\mathbb{N}} e_m = f$ existiert (vgl. [3, Satz 11.6]), wird durch die Zahl

$$\int\limits_{\Omega} f \; d\mu \ := \sup_{\substack{e \le f \\ e \text{ Elementarfunktion}}} \int\limits_{\Omega} e \; d\mu$$

das μ-*Integral* einer $\mathfrak{S}$-messbaren nichtnegativen numerischen Funktion f (über Ω) definiert (vgl. [3, Definition 11.3]). Weiterhin gilt der

Satz A.8 (Monotone Konvergenz (Beppo-Levi) – vgl. [3], Satz 11.4)**.** *Sei $(\Omega, \mathfrak{S}, \mu)$ ein Maßraum. Für jede monotone Folge $(f_n)_{n\in\mathbb{N}}$ von $\mathfrak{S}$-messbaren nichtnegativen numerischen Funktionen existiert der Grenzwert*

$$f \ := \ \sup_{n\in\mathbb{N}} f_n \ = \ \lim_{n\to\infty} f_n \ .$$

Weiterhin ist f eine $\mathfrak{S}$-messbare nichtnegative numerische Funktion mit

$$\int\limits_{\Omega} f \; d\mu \ = \ \sup_{n\in\mathbb{N}} \int\limits_{\Omega} f_n \; d\mu \ = \ \lim_{n\to\infty} \int\limits_{\Omega} f_n \; d\mu \ .$$

Das μ-Integral einer beliebigen $\mathfrak{S}$-messbaren nichtnegativen numerischen Funktion f kann nun durchaus den Wert unendlich annehmen. Um jedoch mit μ-Integralen vernünftig rechnen zu können, definieren wir μ-integrierbare Funktionen wie folgt.

Definition A.9 (vgl. [3], Definitionen 12.1, 12.4)**.** Sei $(\Omega, \mathfrak{S}, \mu)$ ein Maßraum. Eine numerische Funktion f auf Ω heißt μ-*integrierbar*, wenn sie $\mathfrak{S}$-messbar ist und sowohl $\int_\Omega f^+ d\mu < \infty$ als auch $\int_\Omega f^- d\mu < \infty$ gelten. Dann heißt

$$\int\limits_{\Omega} f \; d\mu \ := \ \int\limits_{\Omega} f^+ \; d\mu - \int\limits_{\Omega} f^- \; d\mu \tag{A.7}$$

das μ-*Integral* von f (über Ω). Für $\mu = \lambda$ (bzw. $\mu = \lambda^n$) heißt f dann *Lebesgue-integrierbar*. Im Spezialfall, dass μ ein W-Maß auf $\mathcal{B}(\mathbb{R})$ mit zugehöriger Verteilungsfunktion F ist, schreiben wir anstatt $\int f d\mu$ auch kurz $\int f dF$. Für eine messbare Teilmenge $A \subset \Omega$ definieren wir weiterhin durch

$$\int\limits_{A} f \; d\mu \ := \ \int\limits_{\Omega} 1_A f \; d\mu \tag{A.8}$$

das μ-*Integral von f über A*.

Anhand der Definition A.9 lässt sich leicht überprüfen, dass das μ-Integral linear und monoton ist (vgl. [3, Sätze 12.2, 12.3, 12.4]) sowie

$$\int\limits_{A\cup B} f \; d\mu \ + \ \int\limits_{A\cap B} f \; d\mu \ = \ \int\limits_{A} f \; d\mu \ + \ \int\limits_{B} f \; d\mu \tag{A.9}$$

für eine beliebige μ-integrierbare Funktion f und beliebige messbare Teilmengen $A, B \subset \Omega$ gilt. Insbesondere folgt aus der Äquivalenz

$$\int\limits_{\Omega} f \, d\mu \;=\; 0 \qquad \Longleftrightarrow \qquad f \;=\; 0 \;\; \mu\text{-fast überall} \tag{A.10}$$

für beliebige $\mathfrak{S}$-messbare nichtnegative numerische Funktionen f auf Ω auch die μ-Integrierbarkeit über eine beliebige μ-Nullmenge N für jede $\mathfrak{S}$-messbare numerische Funktion g auf Ω. In diesem Fall wird

$$\int\limits_{N} g \, d\mu \;=\; 0 \tag{A.11}$$

erfüllt (vgl. [3, Satz 13.2, Korollar 13.3]). Desweiteren können zwei verschiedene Maße auf einem Messraum in folgender Weise verbunden sein.

Satz A.10 (vgl. [3], Satz 17.1)**.** *Sei $(\Omega, \mathfrak{S}, \mu)$ ein Maßraum. Ist f eine beliebige $\mathfrak{S}$-messbare nichtnegative numerische Funktion (über Ω), dann ist durch*

$$\nu(A) \;:=\; \int\limits_{A} f \, d\mu \tag{A.12}$$

ein weiteres Maß ν auf $\mathfrak{S}$ gegeben.

Das oben definierte Maß ν heißt das *Maß mit der Dichte f* bezüglich μ und wird auch mit

$$\nu \;=\; f\mu$$

bezeichnet (vgl. [3, Definition 17.2]). Darüber hinaus ist eine $\mathfrak{S}$-messbare numerische Funktion φ genau dann ν-integrierbar, wenn φf bereits μ-integrierbar ist. In diesem Fall gilt

$$\int\limits_{\Omega} \varphi \, d\nu \;=\; \int\limits_{\Omega} \varphi f \, d\mu \tag{A.13}$$

(vgl. [3, Satz 17.3]). Desweiteren ist die Bildung von Maßen mit Dichten transitiv (vgl. [3, Korollar 17.4]) und für μ-integrierbare Funktionen f, g die Äquivalenz

$$f \;=\; g \;\; \mu\text{-fast überall} \qquad \Longleftrightarrow \qquad f\mu \;=\; g\mu \tag{A.14}$$

erfüllt (vgl. [3, Satz 17.5]). Interessant ist nun die Fragestellung, wann ein Maß ν auf einem Messraum $(\Omega, \mathfrak{S})$ eine Dichte bezüglich eines weiteren Maßes μ auf demselben Messraum besitzt.

Definition A.11 (vgl. [3], Definitionen 17.7, 17.12)**.** Auf der σ-Algebra $\mathfrak{S}$ eines Messraumes $(\Omega, \mathfrak{S})$ seien zwei Maße ν und μ gegeben. Dann heißt das Maß ν *stetig bezüglich eines Maßes μ* (abkürzend *μ-stetig*), wenn jede μ-Nullmenge aus $\mathfrak{S}$ auch eine ν-Nullmenge ist. In diesem Fall schreiben wir $\nu \ll \mu$. Man nennt das Maß ν *singulär bezüglich μ* (abkürzend *μ-singulär*) und schreibt $\nu \perp \mu$, falls eine Menge $N \in \mathfrak{S}$ mit $\mu(N) = 0$ und $\nu(\Omega \setminus N) = 0$ existiert.

Im Fall endlicher Maße ist die μ-Stetigkeit von ν nach dem Satz von Radon-Nikodym (vgl. [3, Satz 17.10 – Beweis 1. Fall]) dazu äquivalent, dass ν eine Dichte f bezüglich μ besitzt, welche μ-fast überall eindeutig bestimmt ist (vgl. [3, Satz 17.11 – Beweis 1. Teil]). Allgemeiner gilt der

Satz A.12 (Lebesguescher Zerlegungssatz – vgl. [3], Beweis zu Satz 17.13)**.** *Sind ν und μ zwei endliche Maße auf der σ-Algebra $\mathfrak{S}$ eines Messraumes $(\Omega, \mathfrak{S})$, so lässt sich ν in eindeutiger Weise in der Form $\nu = \nu_1 + \nu_2$ mit Maßen ν_1, ν_2 auf $\mathfrak{S}$ darstellen, für die $\nu_1 \ll \mu$ und $\nu_2 \perp \mu$ erfüllt ist.*

Dabei heißt ν_1 der *reguläre* und ν_2 der *singuläre Teil* von ν *bezüglich* μ. Insbesondere besitzt ν_1 nach dem Satz von Radon-Nikodym eine Dichte bezüglich μ. Als Nächstes benötigen wir noch den Zusammenhang zwischen dem Integral bezüglich eines Maßes μ und dem Integral bezüglich eines mittels μ gebildeten Bildmaßes.

Satz A.13 (Transformationssatz für Integrale – vgl. [3], Satz 19.1, Korollare 19.2, 19.3)**.** *Gegeben seien ein Maßraum $(\Omega, \mathfrak{S}, \mu)$, ein Messraum $(\Omega', \mathfrak{S}')$ und eine $(\mathfrak{S}, \mathfrak{S}')$-messbare Abbildung $T : (\Omega', \mathfrak{S}') \to (\Omega, \mathfrak{S})$. Weiterhin bezeichne $\mu' := T(\mu)$ das Bildmaß wie in* (A.5) *definiert. Dann gelten die folgenden Aussagen:*

(i) *Jede $\mathfrak{S}'$-messbare nichtnegative numerische Funktion f' auf Ω' erfüllt*

$$\int\limits_{\Omega'} f' \; dT(\mu) \;=\; \int\limits_{\Omega} f' \circ T \; d\mu \; . \tag{A.15}$$

(ii) *Jede $\mathfrak{S}'$-messbare numerische Funktion f' auf Ω' ist genau dann bezüglich des Bildmaßes $T(\mu)$ integrierbar, wenn $f' \circ T$ bezüglich μ integrierbar ist. In diesem Fall gilt* (A.15).

(iii) *Ist $T : \Omega \to \Omega'$ zusätzlich bijektiv und besitzt eine $(\mathfrak{S}', \mathfrak{S})$-messbare Umkehrabbildung T^{-1}, dann ist für jede numerische Funktion f' auf Ω' die $T(\mu)$-Integrierbarkeit äquivalent zur μ-Integrierbarkeit von $f' \circ T$. In diesem Fall gilt wiederum* (A.15).

Eine wichtige Anwendung von Satz A.13 ist der in der Analysis oft verwendete

Satz A.14 (Transformationssatz für Lebesgue-Integrale – vgl. [3], Satz 19.4)**.** *Seien $U, V \subset \mathbb{R}^n$ zwei offene Teilmengen und $\Phi : U \to V$ ein C^1-Diffeomorphismus (d.h. Φ sei bijektiv und sowohl Φ als auch Φ^{-1} (mindestens) einmal stetig differenzierbar). Weiter bezeichne $D\Phi$ die Ableitung von Φ (d.h. für jeden Punkt $\mathbf{u} \in U$ diejenige lineare Abbildung $A_{\mathbf{u}} : \mathbb{R}^n \to \mathbb{R}^n$, für die*

$$\lim_{\mathbb{R}^n \setminus \{\mathbf{0}\} \ni \mathbf{h} \to \mathbf{0}} \frac{\Phi(\mathbf{u}+\mathbf{h}) - \Phi(\mathbf{u}) - A_{\mathbf{u}}\mathbf{h}}{\|h\|_2} \;=\; \mathbf{0}$$

gilt). Dann ist $f : V \to \bar{\mathbb{R}}$ genau dann über V bezüglich des n-dimensionalen Lebesgue-Maßes integrierbar, wenn $(f \circ \Phi)|\det D\Phi|$ über U bezüglich des n-dimensionalen Lebesgue-Maßes integrierbar ist. In diesem Fall gilt

$$\int\limits_{V} f \; d\lambda^n \;=\; \int\limits_{U} (f \circ \Phi)|\det D\Phi| \; d\lambda^n \; . \tag{A.16}$$

Desweiteren benötigen wir noch Produkte von σ-Algebren und Maßen sowie die Integration bezüglich eines Produktmaßes. Dazu betrachten wir für endlich viele Messräume $(\Omega_i, \mathfrak{S}_i),\ i = 1, \ldots, n$, mit $n \in \mathbb{N}$ die Produktmenge

$$\Omega \;:=\; \prod_{i=1}^{n} \Omega_i \;:=\; \Omega_1 \times \ldots \times \Omega_n \tag{A.17}$$

und für jedes i die Projektionsabbildung $p_i : \Omega \to \Omega_i$, welche das geordnete Tupel $(\omega_1, \ldots, \omega_n) \in \Omega$ auf seine i-te Koordinate ω_i abbildet. Entsprechend Definition A.6 bezeichnen wir die von den Abbildungen $p_1, \ldots, p_n$ erzeugte σ-Algebra

$$\mathfrak{S} \;:=\; \bigotimes_{i=1}^{n} \mathfrak{S}_i \;:=\; \mathfrak{S}_1 \otimes \ldots \otimes \mathfrak{S}_n \;:=\; \sigma(p_1, \ldots, p_n) \;:=\; \sigma\left(\bigcup_{i=1}^{n} p_i^{-1}(\mathfrak{S}_i) \right) \tag{A.18}$$

als das *Produkt der σ-Algebren* $\mathfrak{S}_1, \ldots, \mathfrak{S}_n$ (vgl. [3, §22]). Insbesondere ist jedes p_i dann $(\mathfrak{S}, \mathfrak{S}_i)$-messbar. Weiterhin ist eine Abbildung

$$f \;:\; \Omega_0 \to \Omega_1 \times \ldots \times \Omega_n$$

eines Messraumes $(\Omega_0, \mathfrak{S}_0)$ in ein Produkt von Messräumen $(\Omega_i, \mathfrak{S}_i)$ genau dann messbar bezüglich der σ-Algebra $\mathfrak{S}$ aus (A.18), wenn jede der Komponenten $f_i := p_i \circ f$ von f eine $(\mathfrak{S}_0, \mathfrak{S}_i)$-messbare Abbildung ist (vgl. [3, §22, Satz 7.4]). Haben wir nun zwei endliche Maßräume $(\Omega_i, \mathfrak{S}_i, \mu_i),\ i = 1, 2$, so existiert genau ein Maß μ auf der Produkt-σ-Algebra $\mathfrak{S}_1 \otimes \mathfrak{S}_2$ mit

$$\mu(A_1 \times A_2) \;=\; \mu_1(A_1) \cdot \mu_2(A_2)$$

für alle $A_i \in \mathfrak{S}_i$ (vgl. [3, Satz 23.3]). Dieses eindeutig auf $\mathfrak{S}_1 \otimes \mathfrak{S}_2$ festgelegte Maß μ heißt das *Produkt der Maße* μ_1 und μ_2 und wird mit $\mu_1 \otimes \mu_2$ bezeichnet (vgl. [3, Definition 23.4]). Es gilt nun der bekannte

Satz A.15 (Satz von Tonelli/Fubini - vgl. [3], Satz 23.6, Korollar 23.7)**.** *Seien* $(\Omega_i, \mathfrak{S}_i, \mu_i)$, $i = 1, 2$, *endliche Maßräume.*

(i) *Für eine* $(\mathfrak{S}_1 \otimes \mathfrak{S}_2)$*-messbare nichtnegative numerische Funktion* f *über* $\Omega_1 \times \Omega_2$ *sind die Funktionen*

$$F_1 : \omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)\ d\mu_1 \qquad \text{bzw.} \qquad F_2 : \omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\ d\mu_2$$

$\mathfrak{S}_2$*- bzw.* $\mathfrak{S}_1$*-messbar und es gilt*

$$\left.\begin{aligned} \int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d(\mu_1 \otimes \mu_2) &= \int_{\Omega_2} F_1(\omega_2)\ d\mu_2 = \int_{\Omega_2} \left(\int_{\Omega_1} f(\omega_1, \omega_2)\ d\mu_1 \right)\ d\mu_2 \\ &= \int_{\Omega_1} F_2(\omega_1)\ d\mu_1 = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2)\ d\mu_2 \right)\ d\mu_1 \end{aligned}\right\} \quad (A.19)$$

(ii) *Sei* f *eine* $(\mu_1 \otimes \mu_2)$*-integrierbare numerische Funktion auf* $\Omega_1 \times \Omega_2$*. Die Funktion* $f_{\omega_1} : \omega_2 \mapsto f(\omega_1, \omega_2)$ *ist dann* μ_2*-integrierbar für* μ_1*-fast alle* ω_1 *und die Funktion* $f_{\omega_2} : \omega_1 \mapsto f(\omega_1, \omega_2)$ *ist dann* μ_1*-integrierbar für* μ_2*-fast alle* ω_2*. Die somit* μ_1*- bzw.* μ_2*-fast überall definierte Funktion*

$$\omega_1 \mapsto \int f_{\omega_1} d\mu_2 \qquad \text{bzw.} \qquad \omega_2 \mapsto \int f_{\omega_2} d\mu_1$$

ist μ_1*- bzw.* μ_2*-integrierbar, und es gilt die Gleichung* (A.19)*.*

Eine wichtige Anwendung der beiden vorangegangenen Sätze ist beispielsweise die Berechnung des Lebesgue-Integrals

$$\int_{\mathbb{R}^2} \mathrm{e}^{-x^2-y^2} d\lambda^2 \ .$$

Aufgrund der Radialsymmetrie des Integranden gehen wir zu Polarkoordinaten über. Verwenden wir die Transformation $\Phi : (r, \varphi) \mapsto (r\cos(\varphi), r\sin(\varphi))$, welche die stetige Ableitung

$$D\Phi(r, \varphi) = \begin{pmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{pmatrix}$$

mit $\det(D\Phi(r, \varphi)) = r$ besitzt und welche die offene Menge $U := \ ]0, \infty[\times \left]0, \frac{\pi}{2}\right[\ \subset \ \mathbb{R}^2$ diffeomorph auf die offene Menge $V := \ ]0, \infty[\times]0, \infty[\subset \mathbb{R}^2$ abbildet, so erhalten wir

$$\int_{\mathbb{R}^2} \mathrm{e}^{-x^2-y^2} d\lambda^2 = 4 \int_V \mathrm{e}^{-x^2-y^2} d\lambda^2 = 4 \int_U r\mathrm{e}^{-r^2} d\lambda^2 = 4 \int_{\left]0, \frac{\pi}{2}\right[} \left(\int_{]0,\infty[} r\mathrm{e}^{-r^2} d\lambda \right) d\lambda = \pi \ .$$

und daraus schließlich die für die Standard-Normalverteilung wichtige Beziehung

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\frac{x^2}{2}} d\lambda = \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \mathrm{e}^{-\zeta^2} d\lambda = \frac{1}{\sqrt{\pi}} \sqrt{\left(\int_{\mathbb{R}} \mathrm{e}^{-\zeta^2} d\lambda \right)^2} = \frac{1}{\sqrt{\pi}} \sqrt{\int_{\mathbb{R}} \mathrm{e}^{-\zeta^2-\eta^2} d\lambda^2} = 1 \ . \quad (A.20)$$

Mit Hilfe der Assoziativität der Produktbildung von σ-Algebren (vgl. [3], S. 162) folgt für endliche Maße $\mu_1, \ldots, \mu_n$ auf σ-Algebren $\mathfrak{S}_1, \ldots, \mathfrak{S}_n$ nach Induktion ebenso die Existenz eines eindeutig bestimmten endlichen Maßes μ auf $\mathfrak{S}_1 \otimes \ldots \otimes \mathfrak{S}_n$ mit

$$\mu(A_1 \times \ldots \times A_n) = \mu_1(A_1) \cdot \ldots \cdot \mu_n(A_n)$$

für alle $A_i \in \mathfrak{S}_i$ $(i = 1, \ldots, n)$, welches das *Produkt der Maße* $\mu_1, \ldots, \mu_n$ genannt und mit

$$\bigotimes_{i=1}^{n} \mu_i = \mu_1 \otimes \ldots \otimes \mu_n \quad (A.21)$$

bezeichnet wird (vgl. [3, Satz 23.9]). Für endliche Maßräume $(\Omega_i, \mathfrak{S}_i, \mu_i)$, $i = 1, \ldots, n$, mit $n \in \mathbb{N}$ heißt

$$\bigotimes_{i=1}^{n} (\Omega_i, \mathfrak{S}_i, \mu_i) := \left(\prod_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} \mathfrak{S}_i, \bigotimes_{i=1}^{n} \mu_i \right) \quad (A.22)$$

entsprechend das *Produkt dieser Maßräume* (vgl. [3, Definition 23.10]). Analog können nun auch die Aussagen aus Satz A.15 für den Produktraum (A.22) formuliert werden. Zum Abschluss betrachten wir noch den Fall, dass jedes endliche Maß μ_i mit einer reellen Dichte versehen wird.

Satz A.16 (vgl. [3], Satz 23.11)**.** *Gegeben seien die endlichen Maßräume* $(\Omega_i, \mathfrak{S}_i, \mu_i)$ *und* $\mathfrak{S}_i$*-messbare nichtnegative reellwertige Funktionen* f_i *auf* Ω_i*. Dann ist das Produkt der endlichen Maße* $\nu_i := f_i\mu_i$ $(i = 1, \ldots, n)$ *definiert, und es gilt*

$$\bigotimes_{i=1}^{n} \nu_i \;=\; F \cdot \left(\bigotimes_{i=1}^{n} \mu_i \right) \tag{A.23}$$

mit der Dichtefunktion

$$F(\omega_1, \ldots, \omega_n) \;:=\; \prod_{i=1}^{n} f_i(\omega_i) \;. \tag{A.24}$$

Die Funktion F aus (A.24) wird *Tensorprodukt* der Dichten $f_1, \ldots, f_n$ genannt.

Bemerkung A.17. Viele der obigen Aussagen gelten auch noch im Fall *σ-endlicher* Maßräume $(\Omega, \mathfrak{S}, \mu)$, d.h. falls Ω durch eine abzählbare Vereinigung von $\mathfrak{S}$-messbaren Mengen $(A_n)_{n\in\mathbb{N}}$ mit $\mu(A_n) < \infty$ für alle $n \in \mathbb{N}$ darstellbar ist. Das Maß μ wird in diesem Fall *σ-endliches Maß auf* Ω genannt. Da Wahrscheinlichkeitsmaße jedoch endliche Maße sind, haben wir an dieser Stelle auf entsprechend allgemeiner formulierte Aussagen bewusst verzichtet. □

Von den maßtheoretischen Grundlagen haben wir nun im Wesentlichen alle diejenigen zusammengestellt, welche für die Wahrscheinlichkeitstheorie von Bedeutung sind.

A.1.2 Grundlagen der Wahrscheinlichkeitstheorie

Mit Hilfe der im Unterabschnitt A.1.1 eingeführten Begriffe aus der Maßtheorie können wir nun die wesentlichen Grundbegriffe und Aussagen der Wahrscheinlichkeitstheorie angeben.

Definition A.18 (vgl. [4], §1)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum. Dann heißen die Elemente der σ-Algebra $\mathfrak{S}$ (*beobachtbare*) *Ereignisse*. Die jedem Ereignis $A \in \mathfrak{S}$ zugeordnete Zahl $P(A)$ wird *Wahrscheinlichkeit* von A genannt. Darüber hinaus spricht man bei $\emptyset$ und Ω vom *unmöglichen* bzw. *sicheren Ereignis*. Im Fall $P(A) = 0$ bzw. $P(A) = 1$ wird das Ereignis A *fast unmöglich* bzw. *fast sicher* genannt. Weiterhin heißen zwei Ereignisse A und B *disjunkt*, wenn $A \cap B = \emptyset$ erfüllt ist.

Ist bereits etwas vom eingetretenen Ereignis bekannt, so gelangen wir zum Begriff der bedingten Wahrscheinlichkeit.

Definition A.19 (vgl. [4], §2)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum und $B \in \mathfrak{S}$ ein Ereignis mit $P(B) > 0$. Dann ist durch

$$P_B \;:\; A \mapsto \frac{P(A \cap B)}{P(B)} \tag{A.25}$$

wiederum ein Wahrscheinlichkeitsmaß P_B auf $\mathfrak{S}$ mit $P_B(B) = 1$ gegeben und $(\Omega, \mathfrak{S}, P_B)$ ein Wahrscheinlichkeitsraum. Die Zahl $P(A|B) := P_B(A)$ wird als *bedingte Wahrscheinlichkeit von A unter der Hypothese B* bezeichnet.

Aufgrund der σ-Additivität von Maßen erhalten wir nun auch

Folgerung A.20 (vgl. [4], §2)**.** *Sei* $(\Omega, \mathfrak{S}, P)$ *ein Wahrscheinlichkeitsraum. Für die Indexmenge* $I = \mathbb{N}$ *oder* $I = \{1, \ldots, n_0\}$ *mit einem* $n_0 \in \mathbb{N}$ *sei eine Familie* $(B_n)_{n\in I}$ *paarweise disjunkter Ereignisse* $B_n \in \mathfrak{S}$ *mit* $P(B_n) > 0$ *für alle* $n \in I$ *gegeben. Dann sind die folgenden Aussagen wahr:*

(i) *Jedes* $A \in \mathfrak{S}$ *erfüllt die* Formel von der totalen Wahrscheinlichkeit

$$P(A) \;=\; \sum_{n\in I} P(B_n)P(A|B_n) \;. \tag{A.26}$$

(ii) *Für jedes* $A \in \mathfrak{S}$ *mit* $P(A) > 0$ *und* $n \in I$ *gilt die* Bayessche Formel

$$P(B_n|A) \;=\; \frac{P(B_n)P(A|B_n)}{\sum\limits_{k\in I} P(B_k)P(A|B_k)} \;. \tag{A.27}$$

Desweiteren betrachten wir messbare Abbildungen von Wahrscheinlichkeiträumen.

Definition A.21 (vgl. [4], Definitionen 3.1, 3.2)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $(\Omega', \mathfrak{S}')$ ein Messraum. Dann heißt jede Abbildung $X : (\Omega, \mathfrak{S}) \to (\Omega', \mathfrak{S}')$ *Zufallsvariable* (*mit Werten in* Ω') und das Bildmaß $X(P)$ (vgl. (A.5)) die *Verteilung* von X (bezüglich des W-Maßes P). In den Fällen $(\Omega', \mathfrak{S}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ bzw. $(\Omega', \mathfrak{S}') = (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ sprechen wir von *reellen* bzw. *numerischen Zufallsvariablen*. Ist $(\Omega', \mathfrak{S}') = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, so wird X *Zufallsvektor der Länge* n oder n*-dimensionale reelle Zufallsvariable* genannt.

Wegen $P\left(X^{-1}(\Omega')\right) = P(\Omega) = 1$ wird durch das Bildmaß $X(P)$ offenbar ein W-Maß auf $(\Omega', \mathfrak{S}')$ definiert und $(\Omega', \mathfrak{S}', X(P))$ somit zu einem Wahrscheinlichkeitsraum. Das Urbild $X^{-1}(A')$ von A' bezeichnet dabei das *Ereignis* „X liegt in A'“ und $P\left(X^{-1}(A')\right)$ die *Wahrscheinlichkeit* dieses Ereignisses. Das einfachste nicht-triviale Beispiel für eine reelle Zufallsvariable auf einem W-Raum $(\Omega, \mathfrak{S}, P)$ ist die charakteristische Funktion (A.6) einer messbaren Menge $A \in \mathfrak{S}$. Desweiteren ist eine Abbildung $X : \Omega \to \mathbb{R}^n$ genau dann eine n-dimensionale reelle Zufallsvariable, wenn jede ihrer Komponenten eine reelle Zufallsvariable ist (vgl. [3], §22).
Ist die Verteilung einer Zufallsvariable bekannt, so lassen sich die Wahrscheinlichkeiten $P\left(X^{-1}(A')\right)$ gegebenenfalls auch ohne explizite Kenntnis des zugrunde liegenden W-Raums $(\Omega, \mathfrak{S}, P)$ bestimmen. Ebenso können wir zu einer Zufallsvariable den Erwartungswert wie folgt definieren.

Definition A.22 (vgl. [4], Definition 3.3)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $X : (\Omega, \mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ eine Zufallsvariable. Gilt dann $X \geq 0$ oder ist X eine P-integrierbare Zufallsvariable, so heißt

$$\mathbb{E}(X) \;:=\; \mathbb{E}_P(X) \;:=\; \int\limits_{\Omega} X \; dP$$

der *Erwartungswert* von X.

Die Eigenschaften des Erwartungswertes sind somit die des P-Integrals. Insbesondere gilt nach dem Transformationssatz A.13 für jede $\mathcal{B}(\mathbb{R})$-messbare Funktion $f : \mathbb{R} \to \mathbb{R}$, welche nicht-negativ oder $X(P)$-integrierbar ist,

$$\mathbb{E}(f \circ X) \;=\; \int\limits_{\mathbb{R}} f \; dX(P) \; . \tag{A.28}$$

Insbesondere ergibt sich für den Erwartungswert $\mathbb{E}(X) = \int_{\mathbb{R}} f \; dX(P)$, wenn wir die Identität $f : x \mapsto x$ auf $\mathbb{R}$ (d.h. $f(x) = x$ für alle $x \in \mathbb{R}$) verwenden. Weiterhin gilt

$$\mathbb{E}(X) \;=\; \sum_{n \in \mathbb{N}} P\left(X^{-1}([n, \infty[)\right) \;, \tag{A.29}$$

falls X nur Werte in $\mathbb{N}$ annimmt (vgl. [4, Satz 3.4]). Mit Hilfe von (A.28) lassen sich nun auch höhere Momente definieren.

Definition A.23 (vgl. [4], Definition 3.5)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und X eine P-integrierbare reelle Zufallsvariable. Weiter seien $m \in \mathbb{N}$ und $\alpha \in \mathbb{R}$ gegeben.

(i) Ist X^m sogar P-integrierbar, so heißen $\mathbb{E}(X^m)$ und $\mathbb{E}(|X|^m)$ das *zentrale* bzw. das *absolute* m*-te Moment* von X. Entsprechend nennen wir $\mathbb{E}((X - \alpha)^m)$ und $\mathbb{E}(|X - \alpha|^m)$ die *in* α *zentrierten* (bzw. *zentrierten absoluten*) m*-ten Momente* von X.

(ii) Die *Varianz* von X ist definiert durch

$$\mathbb{V}(X) \;:=\; \mathbb{E}\left([X - \mathbb{E}(X)]^2\right) \;\leq\; +\infty \; . \tag{A.30}$$

Desweiteren heißt

$$\sigma(X) \;:=\; +\sqrt{\mathbb{V}(X)} \;\leq\; +\infty \tag{A.31}$$

die *Streuung* (oder *Standardabweichung*) von X.

Aus der Linearität des P-Integrals und der aus der Monotonie des P-Integrals resultierenden Nichtnegativität der Varianz ergibt sich dann sofort die

Folgerung A.24 (vgl. [4], Satz 3.6)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum und X eine reelle Zufallsvariable*

(i) *Genau dann ist X^2 eine P-integrierbare Zufallsvariable, wenn X bereits P-integrierbar und $\mathbb{V}(X) < +\infty$ erfüllt ist. In diesem Fall gilt*

$$\mathbb{V}(X) \;=\; \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 \;. \tag{A.32}$$

Insbesondere besitzen die Zufallsgrößen X und $X - \mathbb{E}(X)$ dieselbe Varianz.

(ii) *Für eine P-integrierbare Zufallsvariable X gelten stets*

$$\big|\mathbb{E}(X)\big| \;\leq\; \mathbb{E}(|X|) \tag{A.33}$$

sowie

$$\big(\mathbb{E}(X)\big)^2 \;\leq\; \mathbb{E}(X^2) \;. \tag{A.34}$$

Die in Folgerung A.24 (ii) angegebenen Ungleichungen (A.33) und (A.34) sind lediglich Spezialfälle der *Jensen-Ungleichung*.

Satz A.25 (Jensen-Ungleichung – vgl. [4], Satz 3.9)**.** *Sei X eine P-integrierbare reelle Zufallsvariable auf einem W-Raum $(\Omega, \mathfrak{S}, P)$ mit Werten in einem offenen Intervall $I \subset \mathbb{R}$. Dann liegt der Erwartungswert ebenfalls in I und für jede auf I definierte konvexe Funktion q ist $q \circ X$ wiederum eine Zufallsvariable. Ist diese P-integrierbar, dann gilt*

$$q\left(\mathbb{E}(X)\right) \;\leq\; \mathbb{E}(q \circ X) \;. \tag{A.35}$$

Oftmals ist die Fragestellung interessant, mit welcher Wahrscheinlichkeit eine reelle Zufallsvariable Werte oberhalb einer positiven Konstante a annimmt. Im Falle der P-Integrierbarkeit existieren zahlreiche Ungleichungen, die sich häufig auf die folgende Situation zurückführen lassen.

Lemma A.26 (allgemeine Markov-Ungleichung)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum und $X : (\Omega, \mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ eine P-integrierbare Zufallsvariable. Ferner sei ein $a > 0$ und eine messbare nichtnegative Funktion $h : \mathbb{R} \to \mathbb{R}$ mit $h(a) > 0$ gegeben, welche für $t \geq a$ monoton nichtfallend ist. Dann gilt*

$$P(X^{-1}([a, \infty[)) \;\leq\; \frac{\mathbb{E}(h(X))}{h(a)} \;. \tag{A.36}$$

Beweis: Wegen $[a, \infty[\in \mathcal{B}(\bar{\mathbb{R}})$ und der Messbarkeit von X, ist auch das Urbild $X^{-1}([a, \infty[)$ messbar, so dass wir die charakteristische Funktion $1_{X^{-1}([a,\infty[)}$ bezüglich des Wahrscheinlichkeitsmaßes P integrieren können. Mit der Nichtnegativität und Monotonie von h folgt nun nacheinander.

$$\mathbb{E}(h(X)) \;=\; \int\limits_{\Omega} h(X) dP \;\geq\; \int\limits_{\Omega} h(X) 1_{X^{-1}([a,\infty[)} dP \;\geq\; \int\limits_{\Omega} h(a) 1_{X^{-1}([a,\infty[)} dP \;=\; h(a) P(X^{-1}([a, +\infty[)).$$

Nach Umstellen erhalten wir nun die Behauptung. ■

Mit Hilfe der Varianz bzw. den höheren Momenten lässt sich nun für P-integrierbare reelle Zufallsvariablen X unter Anderem die *Chebyshev-Ungleichung* (A.41) formulieren, welche für $m = 2$ einen Spezialfall der *Chebyshev-Markovschen Ungleichung* (A.40) darstellt (vgl. [3, Lemma 20.1]).

Folgerung A.27. *Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum und $X : (\Omega, \mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ eine P-integrierbare Zufallsvariable. Ferner sei ein $a > 0$ gegeben. Dann gelten*

(i) *die* Markov-Ungleichungen

$$P(X^{-1}([a, +\infty[)) \;\leq\; \frac{\mathbb{E}(\max\{X, 0\})}{a} \tag{A.37}$$

und

$$P(|X|^{-1}([a, +\infty[)) \;\leq\; \frac{\mathbb{E}(|X|)}{a} \tag{A.38}$$

bzw. – falls der Erwartungswert von $|X|$ größer Null ist – äquivalent dazu

$$P\left(|X|^{-1}([c\mu, +\infty[)\right) \;\leq\; \frac{1}{c} \tag{A.39}$$

mit $\mu := \mathbb{E}(|X|)$ und $c = \frac{a}{\mu} > 0$,

(ii) *die* Chebyshev-Markovsche Ungleichung

$$P\left(|X|^{-1}([a,+\infty[)\right) \;\leq\; \frac{1}{a^m}\int\limits_{\Omega}|X|^m dP \qquad (m>0), \tag{A.40}$$

(iii) *die* Chebyshev-Ungleichung

$$P\left(|X-\mathbb{E}(X)|^{-1}([a,+\infty[)\right) \;\leq\; \frac{\mathbb{V}(X)}{a^2} \tag{A.41}$$

bzw. – falls X ein endliches zweites Moment ungleich Null besitzt – äquivalent dazu

$$P\left(|X-\mathbb{E}(X)|^{-1}([c\sigma,+\infty[)\right) \;\leq\; \frac{1}{c^2} \tag{A.42}$$

mit $\sigma := \sqrt{\mathbb{V}(X)}$ und $c = \frac{a}{\sigma} > 0$,

(iv) *die* Chernoff-Ungleichung

$$P(X^{-1}([a,+\infty[)) \;\leq\; \inf_{\varphi\geq 0}\left(\mathrm{e}^{-\varphi a}\mathbb{E}\left(\mathrm{e}^{\varphi X}\right)\right)\;. \tag{A.43}$$

Beweis: Alle Ungleichungen lassen sich mit Folgerung A.26 beweisen:
(i) Sowohl $h := \max\{\cdot,0\} : \mathbb{R}\to\mathbb{R}$ als auch $h := |\cdot| : \mathbb{R}\to\mathbb{R}$ ist nichtnegativ mit $h(a) = a > 0$ und für $t \geq a$ sogar streng monoton wachsend.
(ii) Ebenso ist $h := |\cdot|^m : \mathbb{R}\to\mathbb{R}$ für $m > 0$ nichtnegativ mit $h(a) = a^m > 0$ und auf jedem Intervall $[a,\infty[$ mit $a > 0$ streng monoton wachsend. Darüber hinaus gilt

$$\int\limits_{\Omega}|X|^m dP \;=\; \mathbb{E}\left(|X|^m\right)\;.$$

(iii) Betrachten wir anstelle von X die Zufallsvariable $\tilde{X} := X - \mathbb{E}(X)$, so folgt die Behauptung aus (ii) mit $m = 2$.
(iv) Da die Exponentialfunktion nichtnegativ und streng monoton wachsend ist, gilt die allgemeine Markov-Ungleichung (A.36) mit $h(t) = \mathrm{e}^{\varphi t}$ und beliebigem $\varphi \geq 0$. Da jeweils auf der linken Seite φ nicht auftritt, folgt somit die Behauptung. ■

Zu einem gegebenen Messraum $(\Omega',\mathfrak{S}')$ können alle W-Maße P' als Verteilungen von $(\Omega',\mathfrak{S}')$-Zufallsvariablen auftreten, sofern wir beliebige W-Räume $(\Omega,\mathfrak{S},P)$ zulassen. Es genügt lediglich, dafür $\Omega = \Omega',\ \mathfrak{S} = \mathfrak{S}',\ P = P'$ und für X die identische Abbildung von Ω auf sich zu wählen, um für das Bildmaß $X(P) = P$ zu erhalten. Im Folgenden betrachten wir einige für uns interessante Verteilungen.

Definition A.28 (diskrete und (Lebesgue)-stetige Verteilungen – vgl. [4], §4)**.** Sei $n \in \mathbb{N}$ und $(\Omega,\mathfrak{S},P)$ ein W-Raum sowie $X : (\Omega,\mathfrak{S}) \to (\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n))$ eine Zufallsvariable.

(i) Das für ein festes $\mathbf{x} \in \mathbb{R}^n$ durch

$$\eth_{\mathbf{x}}(B) \;:=\; \begin{cases} 1 & \mathbf{x}\in B\;, \\ 0 & \mathbf{x}\notin B \end{cases} \tag{A.44}$$

für $B \in \mathcal{B}(\mathbb{R}^n)$ gegebene Wahrscheinlichkeitsmaß auf $\mathcal{B}(\mathbb{R}^n)$ heißt *Dirac-Maß*. Besitzt X als Verteilung ein solches Dirac-Maß, so wird X *singulär verteilt* oder *ausgeartet* genannt.

(ii) Sei $(\mathbf{x}_k)_{k\in\mathbb{N}}$ eine Folge im $\mathbb{R}^n$ und $(\alpha_k)_{k\in\mathbb{N}}$ eine Folge nichtnegativer reeller Zahlen mit

$$\sum_{k=1}^{\infty}\alpha_k \;=\; 1\;.$$

Dann heißt das durch

$$\nu \;=\; \sum_{k=1}^{\infty}\alpha_k\,\eth_{\mathbf{x}_k} \tag{A.45}$$

auf $\mathcal{B}(\mathbb{R}^n)$ gegebene Wahrscheinlichkeitsmaß *diskret*. Ist X eine Zufallsvariable mit einer solchen Verteilung ν, so wird X *diskret verteilt* genannt.

(iii) Ein bezüglich des n-dimensionalen Lebesgue-Maßes λ^n stetiges Wahrscheinlichkeitsmaß μ heißt (*Lebesgue*)-*stetig*. Die nach dem Satz von Radon-Nykodym (vgl. [3, Satz 17.10]) existierende fast überall eindeutig bestimmte Borel-messbare Dichte $f \geq 0$ mit $\mu = f\lambda^n$ und folglich $\int f d\lambda^n = 1$ wird die *Wahrscheinlichkeitsdichte* von μ genannt.

Offenbar ist X genau dann singulär verteilt, wenn X fast sicher konstant ist, also $X = \mathbb{E}(X)$ gilt. Dies ist gleichbedeutend mit $\mathbb{V}(X) = 0$. Insbesondere ist jede singuläre Verteilung $\delta_{\mathbf{x}}$ diskret und jedes diskrete Maß singulär bezüglich des n-dimensionalen Lebesgue-Maßes λ^n (vgl. Definition A.11).

Kommen wir nun zum Begriff der stochastischen Unabhängkeit. Diesen wollen wir zunächst für Ereignisse und σ-Algebren einführen, um dann stochastisch unabhängige Zufallsvariablen definieren zu können.

Definition A.29 (stochastische Unabhängigkeit – vgl. [4], Definitionen 6.1, 6.2, 7.1)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $I \neq \emptyset$ eine beliebige Indexmenge.*

(i) *Eine Familie $(A_k)_{k\in I}$ von $\mathfrak{S}$-messbaren Mengen A_k heißt* (stochastisch) unabhängig*, falls*

$$P\left(\bigcap_{j=1}^{n} A_{k_j}\right) = \prod_{j=1}^{n} P\left(A_{k_j}\right) \tag{A.46}$$

für jede nichtleere endliche Teilmenge $\{k_1, \ldots, k_n\} \subset I$.

(ii) *Eine Familie $(\mathfrak{E}_k)_{k\in I}$ von Teilmengen $\mathfrak{E}_k \subset \mathfrak{S}$ heißt* (stochastisch) unabhängig*, falls* (A.46) *für jede nichtleere endliche Teilmenge $\{k_1, \ldots, k_n\} \subset I$ und jede mögliche Wahl von*

$$A_{k_j} \in \mathfrak{E}_{k_j} \qquad (j = 1, \ldots, n) \tag{A.47}$$

erfüllt bleibt.

(iii) *Sei nun $(\Omega_k, \mathfrak{S}_k)_{k\in I}$ eine Familie von Messräumen. Eine Familie $(X_k)_{k\in I}$ von Zufallsvariablen $X_k : (\Omega, \mathfrak{S}) \to (\Omega_k, \mathfrak{S}_k)$ heißt* (stochastisch) unabhängig*, falls die Familie $(\sigma(X_k))_{k\in I}$ der von ihnen erzeugten σ-Algebren $\sigma(X_k) := \sigma\left(X^{-1}(\mathfrak{S}_k)\right)$ unabhängig ist.*

Betrachten wir nun zu einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{S}, P)$ und Messräumen $(\Omega_k, \mathfrak{S}_k), k = 1, \ldots, n$, Zufallsvariablen $X_k : (\Omega, \mathfrak{S}) \to (\Omega_k, \mathfrak{S}_k), k = 1, \ldots, n$, dann ist die durch

$$X_1 \otimes \ldots \otimes X_n : \omega \mapsto (X_1(\omega), \ldots, X_n(\omega)),$$

gegebene *Produktabbildung* eine Zufallsvariable $X_1 \otimes \ldots \otimes X_n : (\Omega, \mathfrak{S}) \to \left(\Omega_1 \times \ldots \times \Omega_n, \bigotimes_{k=1}^{n} \mathfrak{S}_k\right)$, da sie $\left(\mathfrak{S}, \bigotimes_{k=1}^{n} \mathfrak{S}_k\right)$-messbar ist (vgl. Definition (A.18)) und das Bildmaß

$$(X_1 \otimes \ldots \otimes X_n)(P) \tag{A.48}$$

ein Wahrscheinlichkeitsmaß auf $\bigotimes_{k=1}^{n} \mathfrak{S}_k$ darstellt, welches die *gemeinsame Verteilung* der Zufallsvariablen $X_1, \ldots, X_n$ genannt wird. Damit lässt sich die stochastische Unabhängigkeit von Zufallsvariablen auch folgenderweise charakterisieren.

Satz A.30 (vgl. [4], Satz 7.5)**.** *Sei $n \in \mathbb{N}$ und $(\Omega, \mathfrak{S}, P)$ ein W-Raum sowie $(\Omega_k, \mathfrak{S}_k), k = 1, \ldots, n$, Messräume. Weiterhin seien Zufallsvariablen $X_k : (\Omega, \mathfrak{S}) \to (\Omega_k, \mathfrak{S}_k), k = 1, \ldots, n$, gegeben. Dann sind die Zufallsvariablen $X_1, \ldots, X_n$ genau dann stochastisch unabhängig, wenn ihre gemeinsame Verteilung das Produkt ihrer einzelnen Verteilungen ist, d.h., falls für die entsprechenden Bildmaße (vgl.* (A.5))

$$(X_1 \otimes \ldots \otimes X_n)(P) = \bigotimes_{k=1}^{n} X_k(P) \tag{A.49}$$

gilt.

Dabei wird verwendet, dass sich das Bildmaß für beliebige $\left(\bigotimes_{k=1}^{n} \mathfrak{S}_k\right)$-messbare Teilmengen der Gestalt $A_1 \times \ldots \times A_n$ mit der stochastischen Unabhängigkeit als

$$\begin{aligned}
\left((X_1 \otimes \ldots \otimes X_n)(P)\right)(A_1 \times \ldots \times A_n) &= P\left((X_1 \otimes \ldots \otimes X_n)^{-1}(A_1 \times \ldots \times A_n)\right) \\
&= P\left(X_1^{-1}(A_1) \cap \ldots \cap X_n^{-1}(A_n)\right) \\
&= P\left(X_1^{-1}(A_1)\right) \cdot \ldots \cdot P\left(X_n^{-1}(A_n)\right) \\
&= \left(\bigotimes_{k=1}^{n} P\right)\left(X_1^{-1}(A_1) \times \ldots \times X_n^{-1}(A_n)\right) \\
&= \left(\bigotimes_{k=1}^{n} X_k(P)\right)(A_1 \times \ldots \times A_n)
\end{aligned}$$

schreiben lässt. In ähnlicher Weise werden die folgenden Eigenschaften stochastischer Unabhängigkeit bewiesen.

Folgerung A.31 (Vererbung stochastischer Unabhängigkeit - vgl. [5], Satz 20.9)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum. Für $k = 1, \ldots, n$ seien $X_k : (\Omega, \mathfrak{S}) \to (\Omega_k, \mathfrak{S}_k)$ unabhängige Zufallsvariablen. Mit Hilfe der disjunkten Zerlegung $I_1 \dot\cup \ldots \dot\cup I_s = \{1, \ldots, n\}$ mit $s \geq 2$ und $I_j \neq \emptyset$ für alle $j = 1, \ldots, s$, seien die Zufallsvektoren $Y_j : (\Omega, \mathfrak{S}) \to \left(\underset{k \in I_j}{\times} \Omega_k, \bigotimes_{k \in I_j} \mathfrak{S}_k\right)$, $j = 1, \ldots, s$, definiert. Dann gilt:*

(i) *Die Zufallsvariablen $Y_1, \ldots, Y_s$ sind stochastisch unabhängig.*

(ii) *Sind weitere Messräume $(\tilde\Omega_j, \tilde{\mathfrak{S}}_j)$ und messbare Abbildungen $f_j : \left(\underset{k \in I_j}{\times} \Omega_k, \bigotimes_{k \in I_j} \mathfrak{S}_k\right) \to (\tilde\Omega_j, \tilde{\mathfrak{S}}_j)$ für $j = 1, \ldots, s$ gegeben, so sind die Zufallsvariablen $f_1 \circ Y_1, \ldots, f_s \circ Y_s$ stochastisch unabhängig.*

Unter Verwendung des Transformationssatzes A.13, des Satzes von Fubini A.15 und mit Hilfe von Satz A.30 können wir nun den Erwartungswert für das Produkt stochastisch unabhängiger reeller Zufallsvariablen bestimmen.

Satz A.32 (Multiplikationssatz - vgl. [4], Satz 8.1)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum. Für $k = 1, \ldots, n$ seien $X_k : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ unabhängige Zufallsvariablen. Ist $X_k \geq 0$ für alle $k = 1, \ldots, n$ oder ist X_k eine P-integrierbare Zufallsvariable für alle $k = 1, \ldots, n$, dann gilt*

$$\mathbb{E}\left(\prod_{k=1}^{n} X_k\right) = \prod_{k=1}^{n} \mathbb{E}(X_k) \; . \tag{A.50}$$

Im zweiten Fall ist das Produkt $\prod_{k=1}^{n} X_k$ ebenfalls P-integrierbar.

Wenden wir nun Satz A.32 in Kombination mit Folgerung A.31 (ii) an, so erhalten wir (A.50) auch für $f_k \circ X_k$ anstelle von X_k, $k = 1, \ldots, n$, und somit auch eine Verallgemeinerung der Chernoff-Ungleichung (A.43).

Satz A.33 (Cramér-Chernoff)**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $(X_k)_{k \in \mathbb{N}}$ eine unabhängige Folge identisch verteilter P-integrierbarer Zufallsvariablen $X_k : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Für beliebiges $a > 0$ und alle $n \in \mathbb{N}$ gelten dann die Abschätzungen*

$$P\left(\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^{-1}([a, \infty[)\right) \leq \left(\inf_{\varphi \geq 0} \mathrm{e}^{-\varphi a} \mathbb{E}\left(\mathrm{e}^{\varphi X_1}\right)\right)^n \tag{A.51}$$

sowie

$$P\left(\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^{-1}(]-\infty, -a])\right) \leq \left(\inf_{\varphi \leq 0} \mathrm{e}^{\varphi a} \mathbb{E}\left(\mathrm{e}^{\varphi X_1}\right)\right)^n \; . \tag{A.52}$$

Beweis: Aufgrund der Linearität des P-Integrals sowie der P-Integrierbarkeit der Zufallsvariablen X_k ist

$$S_n \;:=\; \sum_{k=1}^{n} X_k$$

ebenfalls P-integrierbar. Desweiteren erhalten wir

$$\mathbb{E}\left(\mathrm{e}^{\varphi S_n}\right) \;=\; \mathbb{E}\left(\prod_{k=1}^{n} \mathrm{e}^{\varphi X_k}\right) \;=\; \prod_{k=1}^{n} \mathbb{E}\left(\mathrm{e}^{\varphi X_k}\right) \;=\; \left(\mathbb{E}\left(\mathrm{e}^{\varphi X_1}\right)\right)^n$$

aufgrund der identischen Verteilung und stochastischen Unabhängigkeit der X_k sowie der sich nach Folgerung A.31 (ii) ergebenden stochastischen Unabhängigkeit der $\mathrm{e}^{\varphi X_k}$. Die Chernoff-Ungleichung (A.43) liefert nun insgesamt

$$P\left(\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^{-1}([a,\infty[)\right) \;=\; P\left(S_n^{-1}([na,\infty[)\right) \;\leq\; \inf_{\varphi\geq 0}\left(\mathrm{e}^{-\varphi na}\left(\mathbb{E}\left(\mathrm{e}^{\varphi X_1}\right)\right)^n\right)$$

und wegen der strengen Monotonie der n-ten Potenz auf der rechten Halbachse somit (A.51). Beim Übergang von S_n zu $-S_n$ ergibt sich mit

$$P\left(\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^{-1}(]-\infty,-a])\right) \;=\; P\left(S_n^{-1}(]-\infty,-na]))\right) \;=\; P\left((-S_n)^{-1}([na,\infty[)\right)$$

nun analog (A.52). ■

Wegen $\mathrm{e}^0 = 1$ und $\mathbb{E}(1) = 1$ sowie der Nichtnegativität der Exponentialfunktion liegen die in den Ungleichungen (A.51) und (A.52) vorkommenden Infima jeweils im Intervall $[0,1]$.

Die Gültigkeit von (A.50) ist nur eine notwendige, jedoch nicht hinreichende Bedingung für stochastische Unabhängigkeit. Ist (A.50) im Fall $n = 2$ gegeben, so gelangen wir zum wichtigen Begriff unkorrelierter Zufallsvariablen. In diesem Zusammenhang definieren wir den Begriff der Kovarianz.

Definition A.34 (Kovarianz - vgl. [4], Definition 8.2)**.** Sei $(\Omega,\mathfrak{S},P)$ ein Wahrscheinlichkeitsraum und seien $X,Y:(\Omega,\mathfrak{S})\to(\bar{\mathbb{R}},\mathcal{B}(\bar{\mathbb{R}}))$ zwei P-integrierbare Zufallsvariablen, deren Produkt P-integrierbar ist. Dann heißt

$$\mathrm{Cov}(X,Y) \;:=\; \mathbb{E}\left((X-\mathbb{E}(X))\,(Y-\mathbb{E}(Y))\right) \;=\; \mathbb{E}\,(XY)-\mathbb{E}(X)\,\mathbb{E}(Y) \tag{A.53}$$

die *Kovarianz* von X und Y. Im Fall $\mathrm{Cov}(X,Y)=0$ heißen die Zufallsvariablen X und Y *unkorreliert.*

Direkt aus der Definition lassen sich die nachstehenden Eigenschaften ablesen.

Folgerung A.35. *Sei $(\Omega,\mathfrak{S},P)$ ein Wahrscheinlichkeitsraum und seien $X,Y,Z:(\Omega,\mathfrak{S})\to(\bar{\mathbb{R}},\mathcal{B}(\bar{\mathbb{R}}))$ jeweils P-integrierbare Zufallsvariablen mit endlichem zweiten Moment und $a\in\mathbb{R}$. Dann gelten:*

(i) $\mathrm{Cov}(X,X)=\mathbb{V}(X)\geq 0$*;*

(ii) $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$*;*

(iii) $\mathrm{Cov}(aX+Y,Z)=a\,\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)$.

Für Summen von paarweise unkorrelierten Zufallsvariablen lässt sich die Varianz einfach berechnen.

Satz A.36 (Gleichheit von Bienaymé - vgl. [4, Satz 8.3])**.** *Sei $(\Omega,\mathfrak{S},P)$ ein W-Raum und $n\in\mathbb{N}$. Weiter seien $X_k:(\Omega,\mathfrak{S})\to(\bar{\mathbb{R}},\mathcal{B}(\bar{\mathbb{R}}))$ für $k=1,\ldots,n$ paarweise unkorrelierte P-integrierbare Zufallsvariablen. Dann gilt*

$$\mathbb{V}\left(\sum_{k=1}^{n} X_k\right) \;=\; \sum_{k=1}^{n}\mathbb{V}\,(X_k)\;. \tag{A.54}$$

Wird die Voraussetzung der paarweisen Unkorreliertheit weggelassen, so gilt

$$\mathbb{V}\left(\sum_{k=1}^{n} X_k\right) \;=\; \sum_{k=1}^{n}\mathbb{V}\,(X_k) \;+\; \sum_{\substack{j,k=1\\ j\neq k}}^{n}\mathrm{Cov}\,(X_j,X_k) \;=\; \sum_{k=1}^{n}\mathbb{V}\,(X_k) \;+\; 2\sum_{k=2}^{n}\sum_{j=1}^{k-1}\mathrm{Cov}\,(X_j,X_k)\;. \tag{A.55}$$

Bemerkung A.37. Eine Verallgemeinerung der Kovarianz ist im Falle der Existenz die sogenannte *Kovarianzmatrix*

$$C \;:=\; \mathrm{Cov}(\mathbf{X}) \;:=\; \mathbb{E}\left((\mathbf{X}-\mathbb{E}(\mathbf{X}))(\mathbf{X}-\mathbb{E}(\mathbf{X}))^{\mathrm{T}}\right) \;:=\; \Big(\mathrm{Cov}(X_j,X_k)\Big)_{j,k=0}^{n-1}\,, \tag{A.56}$$

welche symmetrisch und positiv semidefinit ist. □

Wir betrachten nun einige spezielle Verteilungen.

Definition A.38 (stetige Gleichverteilung)**.** Gegeben seien $a,b\in\mathbb{R}$ mit $a<b$ und ein Wahrscheinlichkeitsraum $(\Omega,\mathfrak{S},P)$. Eine Zufallsvariable $X:(\Omega,\mathfrak{S})\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ heißt *gleichverteilt* auf dem Intervall $[a,b]$, wenn das Bildmaß $X(P)$ mit dem über die Verteilungsfunktion $F:\mathbb{R}\to[0,1]$, definiert durch

$$F(x) \;:=\; \begin{cases} 0\,, & \text{falls } x<a\,,\\ \frac{x-a}{b-a}\,, & \text{falls } a\le x\le b\,,\\ 1\,, & \text{falls } b<x\,, \end{cases} \tag{A.57}$$

eindeutig festgelegten Wahrscheinlichkeitsmaß (vgl. Satz A.5) übereinstimmt. Entsprechend nennen wir die Verteilung von X *Gleichverteilung* auf dem Intervall $[a,b]$.

Die in Definition A.38 auftretende Verteilung der Zufallsgröße X ist (Lebesgue)-stetig mit der Wahrscheinlichkeitsdichte $\frac{1}{b-a}1_{[a,b]}$ (vgl. (A.6)). Folglich besitzt X den Erwartungswert

$$\mathbb{E}(X) \;=\; \int\limits_{\mathbb{R}} x\;dX(P) \;=\; \frac{1}{b-a}\int\limits_{[a,b]} x\;d\lambda_x \;=\; \frac{a+b}{2} \tag{A.58}$$

und die Varianz

$$\mathbb{V}(X) \;=\; \int\limits_{\mathbb{R}} \left(x-\frac{a+b}{2}\right)^2 dX(P) \;=\; \frac{1}{b-a}\int\limits_{[a,b]} \left(x-\frac{a+b}{2}\right)^2 d\lambda_x \;=\; \frac{(b-a)^2}{12}\,. \tag{A.59}$$

Definition A.39 (diskrete Gleichverteilung)**.** Zu einem $k\in\mathbb{N}$ sei eine Menge $A\subset\mathbb{R}$ mit $|A|=k$ gegeben. Die Elemente von A seien mit $a_1<a_2<\ldots<a_{k-1}<a_k$ benannt. Weiter sei $(\Omega,\mathfrak{S},P)$ ein Wahrscheinlichkeitsraum. Eine Zufallsvariable $X:(\Omega,\mathfrak{S})\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ heißt *gleichverteilt* auf der Menge A, wenn das Bildmaß $X(P)$ mit dem über die Verteilungsfunktion $F:\mathbb{R}\to[0,1]$, definiert durch

$$F(x) \;:=\; \begin{cases} 0\,, & \text{falls } x<a_1\,,\\ \frac{r}{k}\,, & \text{falls } a_r\le x<a_{r+1} \qquad (r=1,\ldots,k-1)\,,\\ 1\,, & \text{falls } a_k\le x\,, \end{cases} \tag{A.60}$$

eindeutig festgelegten Wahrscheinlichkeitsmaß (vgl. Satz A.5) übereinstimmt. Entsprechend nennen wir die Verteilung von X *Gleichverteilung* auf der Menge A.

Die in Definition A.39 auftretende Verteilung der Zufallsgröße X ist singulär bezüglich des Lebesgue-Maßes, da die Menge A aufgrund ihrer Abzählbarkeit zu den Lebesgue-Nullmengen zählt (vgl. Definition A.11). Es handelt sich bei X somit um eine diskrete Zufallsvariable, dessen Verteilung sich mit Hilfe des in (A.44) vorgestellten Dirac-Maßes als

$$X(P) \;=\; \sum_{r=1}^{k}\frac{1}{k}\;\eth_{a_r}$$

schreiben lässt. Entsprechend erhalten wir nun den Erwartungswert

$$\mathbb{E}(X) \;=\; \int\limits_{\mathbb{R}} x\;dX(P) \;=\; \frac{1}{k}\sum_{r=1}^{k}\int\limits_{\mathbb{R}} x\;d\eth_{a_r}(x) \;=\; \frac{1}{k}\sum_{r=1}^{k} a_r$$

und nach Folgerung A.24 (i) die Varianz

$$\mathbb{V}(X) \;=\; \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \;=\; \int\limits_{\mathbb{R}} x^2 \; dX(P) - \left(\frac{1}{k}\sum_{r=1}^{k} a_r\right)^2 \;=\; \frac{1}{k}\sum_{r=1}^{k} a_r^2 - \left(\frac{1}{k}\sum_{r=1}^{k} a_r\right)^2 .$$

Sowohl für das über (A.57) als auch für das durch (A.60) festgelegte Wahrscheinlichkeitsmaß gibt es jeweils ein abgeschlossenes beschränktes Intervall $I \subset \mathbb{R}$, so dass $\mathbb{R} \setminus I$ eine entsprechende Nullmenge ist. Die Existenz eines solchen Intervalls ist nicht zwingend notwendig, wie die nachfolgend vorgestellte Verteilung belegt.

Definition A.40 (Normalverteilung)**.** Sei $(\Omega, \mathfrak{S}, P)$ ein Wahrscheinlichkeitsraum. Eine Zufallsvariable $X : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ heißt *normalverteilt* mit Parametern $\mu \in \mathbb{R}$ und $\sigma > 0$, wenn ihr Bildmaß $X(P)$ Lebesgue-stetig ist und die Wahrscheinlichkeitsdichte

$$f_{\mu,\sigma}(x) \;:=\; \frac{1}{\sqrt{2\pi\sigma^2}} \; \mathrm{e}^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} \tag{A.61}$$

besitzt. Im Fall $(\mu, \sigma) = (0, 1)$ nennen wir $X(P)$ *Standard-Normalverteilung.*

Das in Definition A.40 auftretende Bildmaß ist ein Wahrscheinlichkeitsmaß, denn mit der Substitution $s = \frac{x-\mu}{\sigma}$ ergibt sich aus (A.20) sofort

$$\int\limits_{\mathbb{R}} 1 \; dX(P) \;=\; \int\limits_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \; \mathrm{e}^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} d\lambda_x \;=\; \int\limits_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \; \mathrm{e}^{-\frac{s^2}{2}} d\lambda_s \;=\; 1 \; .$$

Desweiteren existiert aufgrund der Nichtnegativität der Dichte (A.61) kein Intervall $[a, b] \subset \mathbb{R}$ mit $a < b$, welches bezüglich des in Definition A.40 auftretenden Wahrscheinlichkeitsmaßes eine Nullmenge ist. Für den Erwartungswert und die Varianz folgen

$$\mathbb{E}(X) \;=\; \int\limits_{\mathbb{R}} x \; dX(P) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{\mathbb{R}} x \; \mathrm{e}^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} d\lambda_x \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{\mathbb{R}} (\mu + \sigma y) \; \mathrm{e}^{-\frac{y^2}{2}} d\lambda_y \;=\; \mu$$

sowie

$$\mathbb{V}(X) \;=\; \int\limits_{\mathbb{R}} (x-\mu)^2 \; dX(P) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{\mathbb{R}} (x-\mu)^2 \; \mathrm{e}^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2} d\lambda_x \;=\; \frac{1}{\sqrt{2\pi}} \int\limits_{\mathbb{R}} (\sigma y)^2 \; \mathrm{e}^{-\frac{y^2}{2}} d\lambda_y \;=\; \sigma^2 \; ,$$

wobei wir jeweils verwendet haben, dass $-\mathrm{e}^{-\frac{y^2}{2}}$ eine Stammfunktion zu $y\mathrm{e}^{-\frac{y^2}{2}}$ ist.

Darüber hinaus interessieren uns auch die Verteilungen von Summen verschiedener Zufallsgrößen.

Lemma A.41. *Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathfrak{S}, P)$ sowie Zufallsvariablen $X : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und $Y : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Dann ist durch $Z := X + Y$ ebenso eine Zufallsvariable $Z : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gegeben. Weiterhin gilt:*

(i) *Besitzen X und Y (Lebesgue-)stetige Verteilungen und ebenso eine gemeinsame (Lebesgue-)Dichte $f_{(X,Y)}(x, y)$, so besitzt auch Z eine (Lebesgue-)stetige Verteilung mit Wahrscheinlichkeitsdichte*

$$f_Z(z) \;=\; \int_{\mathbb{R}} f_{(X,Y)}(x, z-x) d\lambda_x \; .$$

(ii) *Besitzt X eine (Lebesgue)-stetige Verteilung und Y die diskrete Verteilung*

$$Y(P) \;=\; \sum_{s=1}^{m} q_s \eth_{b_s} \qquad \left(q_s > 0, s = 1, \ldots, m, \quad \sum_{s=1}^{m} q_s = 1\right) \tag{A.62}$$

und existiert eine gemeinsame Dichte $f_{(X,Y)}(x, y)$ bezüglich des Produktmaßes $\lambda \otimes \zeta$ mit $\zeta := Y(P)$, dann besitzt Z eine (Lebesgue)-stetige Verteilung mit Wahrscheinlichkeitsdichte

$$f_Z(z) \;=\; \sum_{s=1}^{m} q_s f_{(X,Y)}(z - b_s, b_s) \; .$$

(iii) *Besitzen X und Y die diskreten Verteilungen*

$$X(P) \;=\; \sum_{r=1}^{n} p_r \eth_{a_r} \qquad \left(p_r > 0, r = 1, \ldots, n, \quad \sum_{r=1}^{n} p_r = 1 \right)$$

und (A.62)*, dann ist auch Z eine diskret verteilte Zufallsvariable mit*

$$Z(P) \;=\; \sum_{r=1}^{n} \sum_{s=1}^{m} p_r q_s \eth_{a_r + b_s}.$$

Beweis: (i) Bezeichnet $S : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \times (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ den Summenoperator $(x, y) \mapsto S(x, y) := x + y$, so lässt sich Z als die Komposition $Z = S \circ (X, Y)$ darstellen. Nach Definition und mittels des Satzes A.15 (Satz von Fubini) ergibt sich wegen $S^{-1}(]-\infty, v[) = \{(x, y) \;:\; x + y < v\}$ für die Verteilungsfunktion

$$\begin{aligned} F(v) \;=\; P((X,Y)^{-1}(\{(x,y) \;:\; x+y<v\})) \;&=\; \int_{\{(x,y)\,:\,x+y<v\}} f_{(X,Y)}(x,y)\; d\lambda^2_{(x,y)} \\ &=\; \int_{\mathbb{R}} \left(\int_{]-\infty, v-x[} f_{(X,Y)}(x,y)\; d\lambda_y \right) d\lambda_x \\ &=\; \int_{\mathbb{R}} \left(\int_{]-\infty, v[} f_{(X,Y)}(x, z-x)\; d\lambda_z \right) d\lambda_x \\ &=\; \int_{]-\infty, v[} \left(\int_{\mathbb{R}} f_{(X,Y)}(x, z-x)\; d\lambda_x \right) d\lambda_z \end{aligned}$$

und daher die Behauptung.

(ii) Analog zu (i) erhalten wir mit dem Summenoperator $(x, y) \mapsto S(x, y) := x + y$ und mittels Satz A.15 (Satz von Fubini) für die Verteilungsfunktion zunächst

$$\begin{aligned} F(v) \;=\; P((X,Y)^{-1}(\{(x,y) \;:\; x+y<v\})) \;&=\; \int_{\{(x,y)\,:\,x+y<v\}} f_{(X,Y)}(x,y)\; d\,(\lambda \otimes \zeta)_{(x,y)} \\ &=\; \int_{\mathbb{R}} \left(\int_{]-\infty, v-y[} f_{(X,Y)}(x,y)\; d\lambda_x \right) d\zeta_y \\ &=\; \int_{\mathbb{R}} \left(\int_{]-\infty, v[} f_{(X,Y)}(z-y,y)\; d\lambda_z \right) d\zeta_y \\ &=\; \int_{]-\infty, v[} \left(\int_{\mathbb{R}} f_{(X,Y)}(z-y,y)\; d\zeta_y \right) d\lambda_z \;. \end{aligned}$$

Setzen wir jetzt für ζ_y die Verteilung (A.62) ein, so folgt

$$f_Z(z) \;=\; \int_{\mathbb{R}} f_{(X,Y)}(z-y,y)\; d\zeta_y \;=\; \sum_{s=1}^{m} q_s \int_{\mathbb{R}} f_{(X,Y)}(z-y,y)\; d\eth_{b_s} \;=\; \sum_{s=1}^{m} q_s f_{(X,Y)}(z-b_s, b_s)$$

wie behauptet.

(iii) Wir betrachten zunächst den Fall $X(P) = \eth_a$ und $Y(P) = \eth_b$. Dann besitzt der Zufallsvektor $(X, Y) : (\Omega, \mathfrak{S}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ die Verteilung $(X, Y)(P) \;=\; \eth_{(a,b)} = \eth_a \otimes \eth_b$ (vgl. Produktmaß (A.21)), denn für jedes $A \in \mathcal{B}(\mathbb{R}^2)$ gilt

$$P((X,Y)^{-1}(A)) \;=\; \begin{cases} 1\;, & \text{falls } (a,b) \in A\;, \\ 0\;, & \text{falls } (a,b) \notin A\;. \end{cases}$$

Wegen $(a, b) \in S^{-1}(\{a + b\})$ besitzt die Menge $S^{-1}(\{a + b\})$ bezüglich $(X, Y)(P)$ das Maß 1. Mit den Eigenschaften des Urbildes folgt $S^{-1}(\{a + b\}) \cap S^{-1}(\mathbb{R} \setminus \{a + b\}) = \emptyset$, so dass $S^{-1}(\mathbb{R} \setminus \{a + b\})$ eine Nullmenge bezüglich des Maßes $(X, Y)(P)$ ist. Schließlich erhalten wir wiederum wegen $Z =$

$S \circ (X,Y)$ und der Definition des Bildmaßes wie gewünscht $Z(P) = \eth_{a+b}$. Im allgemeinen Fall besitzt der Zufallsvektor $(X,Y) : (\Omega, \mathfrak{S}) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ die Verteilung

$$(X,Y)(P) \;=\; \left(\sum_{r=1}^{n} p_r \eth_{a_r}\right) \otimes \left(\sum_{s=1}^{m} q_s \eth_{b_s}\right) \;=\; \sum_{r=1}^{n}\sum_{s=1}^{m} p_r q_s \left(\eth_{a_r} \otimes \eth_{b_s}\right) \;.$$

Wie zuvor erhalten wir, dass das Urbild

$$S^{-1}\left(\mathbb{R}^2 \setminus \{a_r + b_s \mid r = 1,\ldots,n,\ s = 1,\ldots,m\}\right)$$

bezüglich $(X,Y)(P)$ eine Nullmenge und die Zufallsvariable $Z = S \circ (X,Y)$ demnach diskret ist, wobei ihre Verteilung die Gestalt

$$Z(P) \;=\; \sum_{r=1}^{n}\sum_{s=1}^{m} \alpha_{rs} \eth_{a_r+b_s} \qquad \left(\alpha_{rs} \geq 0, r = 1,\ldots,n,\ s = 1,\ldots,m, \quad \sum_{r=1}^{n}\sum_{s=1}^{m} \alpha_{rs} = 1\right)$$

besitzt. Offenbar leistet die Wahl $\alpha_{rs} = p_r q_s$ $(r = 1,\ldots,n,\ s = 1,\ldots,m)$ das Gewünschte. ■

Als eine Anwendung erhalten wir insbesondere, dass die Summe gleichverteilter Zufallsvariablen nicht mehr gleichverteilt ist.

Folgerung A.42. *Auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{S}, P)$ seien die Zufallsvariablen $X : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ und $Y : (\Omega, \mathfrak{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gegeben. Desweiteren gelte $a < b$ und $c < d$.*

(i) *Sind X und Y (stetig) gleichverteilt mit Wahrscheinlichkeitsdichten $f_X(x) = \frac{1}{b-a} 1_{[a,b]}(x)$ und $f_Y(y) = \frac{1}{d-c} 1_{[c,d]}(y)$, dann besitzt $Z := X + Y$ die Wahrscheinlichkeitsdichte*

$$f_Z(z) \;=\; \frac{1}{(b-a)(d-c)} \times \begin{cases} 0 & \text{für } z < a+c, \\ z - (a+c) & \text{für } a+c \leq z \leq \min\{a+d, b+c\}, \\ \min\{b-a, d-c\} & \text{für } \min\{a+d, b+c\} < z < \max\{a+d, b+c\}, \\ (b+d) - z & \text{für } \max\{a+d, b+c\} \leq z \leq b+d, \\ 0 & \text{für } z > b+d, \end{cases}$$

den Erwartungswert

$$\mathbb{E}(Z) \;=\; \mathbb{E}(X) + \mathbb{E}(Y) \;=\; \frac{a+b+c+d}{2}$$

sowie die Varianz

$$\mathbb{V}(Z) \;=\; \mathbb{V}(X) + \mathbb{V}(Y) \;=\; \frac{(b-a)^2 + (d-c)^2}{12}.$$

(ii) *Sind X und Y (diskret) gleichverteilt auf den Mengen $A = \{a + k \mid k = 0, 1, \ldots, m-1\}$ und $B = \{b + l \mid l = 0, 1, \ldots, n-1\}$, wobei $m, n \in \mathbb{N}$ mit $m \geq n \geq 2$ gelte, so besitzt $Z := X + Y$ die Verteilung*

$$Z(P) \;=\; \frac{1}{mn} \sum_{j=2}^{n+m} \alpha_j \eth_{a+b+j-2}$$

mit

$$\alpha_j \;:=\; \begin{cases} j-1 & \text{für } j \in \{2,\ldots,n\}, \\ n & \text{für } j \in \{n+1,\ldots,m\}, \\ n+m-(j-1) & \text{für } j \in \{m+1,\ldots,m+n\}. \end{cases} \tag{A.63}$$

Beweis: (i) Die gemeinsame Dichte von (X,Y) ist ganz offenbar $\frac{1}{(b-a)(d-c)} 1_{[a,b]\times[c,d]}(x,y)$. Wegen

$$\frac{1}{(b-a)(d-c)} 1_{[a,b]\times[c,d]}(x,y) \;=\; \frac{1}{b-a} 1_{[a,b]}(x) \cdot \frac{1}{d-c} 1_{[c,d]}(y)$$

erhalten wir $f_{(X,Y)}(x,y) = f_X(x)f_Y(y)$, was hier genau der stochastischen Unabhängigkeit von X und Y entspricht. Zusammen mit der Linearität folgen damit die Behauptungen für den Erwartungswert und die Varianz. Desweiteren ergibt sich über Lemma A.41 (i) die Dichte

$$\begin{aligned} f_Z(z) = \int_{\mathbb{R}} \frac{1}{b-a} 1_{[a,b]}(x) \cdot \frac{1}{d-c} 1_{[c,d]}(z-x)\, d\lambda_x &= \frac{1}{(b-a)(d-c)} \int_{[a,b]} 1_{[c,d]}(z-x)\, d\lambda_x \\ = \frac{1}{(b-a)(d-c)} \int_{[\max\{a,z-d\},\min\{b,z-c\}]} d\lambda_x &= \frac{\max\{\min\{b,z-c\} - \max\{a,z-d\},0\}}{(b-a)(d-c)} \,. \end{aligned}$$

Mittels Durchführung einer Fallunterscheidung schließen wir nun auf die Behauptung für $f_Z(z)$.
(ii) Nach Voraussetzung gelten

$$X(P) \;=\; \frac{1}{m}\sum_{k=1}^{m} \eth_{a+k-1} \qquad \text{bzw.} \qquad Y(P) \;=\; \frac{1}{n}\sum_{l=1}^{n} \eth_{b+l-1}.$$

Nach Lemma A.41 (iii) ergibt sich für die Verteilung von Z dann

$$Z(P) \;=\; \frac{1}{nm}\sum_{k=1}^{m}\sum_{l=1}^{n} \eth_{a+b+l+k-2} \;=\; \frac{1}{nm}\sum_{j=2}^{n+m} \alpha_j \eth_{a+b+j-2}$$

mit α_j aus (A.63), wobei es sich wegen

$$\frac{1}{mn}\sum_{j=2}^{n+m} \alpha_j \;=\; \frac{1}{mn}\left(\sum_{j=1}^{n-1} j \;+\sum_{j=1}^{m-n} n \;+\sum_{k=n}^{1} k\right) \;=\; \frac{1}{mn}\left(\frac{n(n-1)}{2} + n(n-m) + \frac{n(n+1)}{2}\right) \;=\; 1$$

in der Tat um eine (diskrete) Wahrscheinlichkeitsverteilung handelt. ■

In der Wahrscheinlichkeitstheorie treten verschieden starke Konvergenzbegriffe auf, welche in der folgenden Definition zusammengefasst sind.

Definition A.43 (Konvergenzbegriffe für Zufallsvariablen und Verteilungen - [3, 4, 5])**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $(X_n)_{n\in\mathbb{N}}$ eine Folge P-integrierbarer Zufallsvariablen $X_n : (\Omega,\mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$. Desweiteren sei $X : (\Omega,\mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ eine P-integrierbare Zufallsvariable. Dann definieren wir:*

(i) *Die Folge $(X_n)_{n\in\mathbb{N}}$ heißt* fast sicher konvergent *gegen X* (vgl. [3, Definition 13.1]) *genau dann, wenn eine P-Nullmenge $N \in \mathfrak{S}$ existiert, so dass*

$$\lim_{n\to\infty} |X_n(\omega) - X(\omega)| \;=\; 0 \tag{A.64}$$

für alle $\omega \in \Omega \setminus N$ gilt.

(ii) *Für $1 \le p < \infty$ heißt die Folge $(X_n)_{n\in\mathbb{N}}$* im p-ten Mittel konvergent *gegen X* (vgl. [4, §5]) *genau dann, wenn gilt*

$$\lim_{n\to\infty} \mathbb{E}\left(|X_n - X|^p\right) \;=\; \lim_{n\to\infty} \int_{\Omega} |X_n - X|^p \, dP \;=\; 0 \,. \tag{A.65}$$

(iii) *Die Folge $(X_n)_{n\in\mathbb{N}}$ heißt* stochastisch konvergent *gegen X – oder* konvergent *gegen X* in Wahrscheinlichkeit – (vgl. [3, Definition 20.2]) *genau dann, wenn für jedes $\alpha > 0$ gilt*

$$\lim_{n\to\infty} P\left(|X_n - X|^{-1}([\alpha,\infty[)\right) \;=\; 0 \,. \tag{A.66}$$

(iv) *Die Folge $(X_n)_{n\in\mathbb{N}}$ heißt* schwach konvergent *gegen X* (vgl. [3, Definition 30.7], [4, Definition 5.2]) *genau dann, wenn*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f \, dX_n(P) \;=\; \int_{\mathbb{R}} f \, dX(P) \tag{A.67}$$

für alle beschränkten, stetigen, reellen Funktionen $f : \Omega \to \mathbb{R}$ gilt. Dabei bezeichnet $X_n(P)$ bzw. $X(P)$ das Bildmaß (vgl. (A.5)) *von X_n bzw. X.*

(v) *Seien nun F_n bzw. F die Verteilungsfunktionen der Bildmaße $X_n(P)$ bzw. $X(P)$ und $C(F) := \{y \in \mathbb{R} \,:\, F \text{ stetig in } y\}$ bezeichne die Stetigkeitsmenge von F. Die Folge $(X_n)_{n\in\mathbb{N}}$ heißt* verteilungskonvergent *gegen X* (vgl. [5, Definition 26.1]) *genau dann, wenn*

$$\forall x \in C(F) \,:\, \lim_{n\to\infty} F_n(x) \;=\; F(x) \tag{A.68}$$

gilt. In diesem Fall schreiben wir $X_n \xrightarrow{\mathcal{L}} X$.

Bemerkung A.44 (vgl. [3, 4, 5])**.** (i) Die Bedingung für die fast sichere Konvergenz ist (vgl. [3, Lemma 20.6], [4, §5]) äquivalent zu

$$\forall \alpha > 0 \,:\, \lim_{n\to\infty} P\left(\left(\sup_{m\geq n} |X_m - X|\right)^{-1} (]\alpha, \infty[)\right) \;=\; 0\;, \tag{A.69}$$

wobei $\left(\sup\limits_{m\geq n} |X_m - X|\right)^{-1} (]\alpha, \infty[)$ das Urbild des Intervalls $]\alpha, \infty[\in \mathcal{B}(\bar{\mathbb{R}})$ unter der (nach [3, Satz 9.5]) messbaren Funktion $\sup\limits_{m\geq n} |X_m - X|$ bezeichnet.

(ii) Als unmittelbare Folge der Jensen-Ungleichung (A.35) ergibt sich aus der Konvergenz im p-ten Mittel (aufgrund der Endlichkeit von W-Maßen) sofort die Konvergenz im Mittel, d.h.,

$$\lim_{n\to\infty} \mathbb{E}\,(X_n - X) \;=\; 0\;. \tag{A.70}$$

(iii) Sowohl die fast sichere Konvergenz als auch die Konvergenz im p-ten Mittel implizieren die stochastische Konvergenz, wobei ersteres aus einem Vergleich von (A.69) und (A.66) ersichtlich ist und letzteres direkt aus der Chebyshev-Markovschen Ungleichung (A.40) folgt.

(iv) Nach dem Transformationssatz A.13 ist (A.67) äquivalent zu

$$\lim_{n\to\infty} \mathbb{E}\,(f \circ X_n) \;=\; \mathbb{E}\,(f \circ X)\;. \tag{A.71}$$

(v) Aus der schwachen Konvergenz ergibt sich (nach [3, Satz 30.13] oder [5, Satz 26.6]) die Verteilungskonvergenz. Insbesondere ist F der gleichmäßige Grenzwert der F_n, falls F auf $\mathbb{R}$ stetig ist [3, Satz 30.13].

(vi) Stochastische Konvergenz impliziert schwache Konvergenz (nach [4, Satz 5.1]) und nach (v) somit auch die Verteilungskonvergenz (vgl. [5, Lemma 26.2]). □

Mittels oben eingeführte Grenzwertbegriffe können wir nun den zentralen Grenzwertsatz formulieren.

Satz A.45 (Zentraler Grenzwertsatz von Lindeberg-Lévy - [5, Satz 26.7])**.** *Sei $(\Omega, \mathfrak{S}, P)$ ein W-Raum und $(X_n)_{n\in\mathbb{N}}$ eine Folge stochastisch unabhängiger identisch verteilter Zufallsvariablen $X_n : (\Omega, \mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ mit endlichem Erwartungswert $\mathbb{E}(X_n) = \theta$ und endlicher Varianz $\mathbb{V}(X_n) = \sigma^2$. Desweiteren sei*

$$\bar{X}_n \;:=\; \frac{1}{n}\sum_{k=1}^{n} X_k \tag{A.72}$$

die Folge der arithmetischen Mittel und $Y : (\Omega, \mathfrak{S}) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ standard-normalverteilt. Dann ist die Folge

$$Y_n \;:=\; \frac{\sqrt{n}\,(\bar{X}_n - \theta)}{\sigma} \tag{A.73}$$

verteilungskonvergent gegen Y.

In [4] wird Satz A.45 als Satz von de Moivre-Laplace bezeichnet und zusätzlich die Gültigkeit des zentralen Grenzwertsatzes unter schwächeren Voraussetzungen diskutiert.

A.2 Ausgelagerte Beweise

Beweis von Lemma 2.14

Wir betrachten zunächst die sequentielle Summation. Wie im Beispiel 2.4 gilt

$$\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; \mathbf{G}^{\mathrm{T}}\mathbf{1} + \sum_{j=1}^{n-1}\varepsilon_j^+ \sum_{k=0}^{j} G_k \prod_{l=j+1}^{n-1}(1+\varepsilon_l^+)$$

und aufgrund von $\mathbb{E}(G_k) = 0$ sowie mit der Unabhängigkeit dann auch

$$\mathbb{E}(\Delta) \;=\; \mathbb{E}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; \sum_{j=1}^{n-1}\mathbb{E}(\varepsilon_j^+)\sum_{k=0}^{j}\mathbb{E}(G_k)\prod_{l=j+1}^{n-1}\mathbb{E}((1+\varepsilon_l^+)) \;=\; 0\;.$$

Weiterhin folgt für die Varianz

$$\begin{aligned}
\mathbb{V}(\Delta) \;&=\; \mathbb{V}\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;=\; \mathbb{E}\left(\left(\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{1}\right)^2\right) \;-\; 0^2 \\
&=\; \mathbb{E}\left(\left(\sum_{j=1}^{n-1}\varepsilon_j^+\sum_{k=0}^{j}G_k\prod_{l=j+1}^{n-1}(1+\varepsilon_l^+)\right)\left(\sum_{r=1}^{n-1}\varepsilon_r^+\sum_{s=0}^{r}G_s\prod_{t=r+1}^{n-1}(1+\varepsilon_t^+)\right)\right) \\
&=\; \sum_{j,r=1}^{n-1}\mathbb{E}\left(\varepsilon_j^+\varepsilon_r^+\right)\sum_{k=0}^{j}\sum_{s=0}^{r}\mathbb{E}\left(G_kG_s\right)\left(1+\mathcal{O}(u)\right)\;.
\end{aligned}$$

Aufgrund der Unkorreliertheit bzw. Unabhängigkeit gilt einerseits

$$\mathbb{E}\left(G_kG_s\right) \;=\; \delta(s-k)\cdot\mathbb{E}\left((G_k)^2\right) \;=\; \delta(s-k)\cdot\mathbb{V}(G_k)$$

und andererseits

$$\mathbb{E}\left(\varepsilon_j^+\varepsilon_r^+\right) \;=\; \left(\mu_+^2 \;+\; \delta(j-r)\sigma_+^2\right)u^2\;,$$

weil alle relativen Fehler nach Modellannahme dieselbe Varianz und denselben Erwartungswert besitzen. Somit folgt – unter Vernachlässigung von Termen der Ordnung $\mathcal{O}(u^3)$ – weiter

$$\begin{aligned}
\mathbb{V}\left(\Delta\right) \;&=\; \mu_+^2u^2\sum_{j=1}^{n-1}\left(\sum_{r=1}^{j-1}\sum_{k=0}^{r}\mathbb{V}(G_k) \;+\; \sum_{r=j}^{n-1}\sum_{k=0}^{j}\mathbb{V}(G_k)\right) \;+\; \sigma_+^2u^2\sum_{j=1}^{n-1}\sum_{k=0}^{j}\mathbb{V}(G_k) \\
&=\; \mu_+^2u^2\sum_{j=1}^{n-1}\left((j-1)\mathbb{V}(G_0) + \sum_{k=1}^{j-1}(j-k)\mathbb{V}(G_k) \;+\; (n-j)\sum_{k=0}^{j}\mathbb{V}(G_k)\right) \;+\; \sigma_+^2u^2\sum_{j=1}^{n-1}\sum_{k=0}^{j}\mathbb{V}(G_k) \\
&=\; \mu_+^2u^2\sum_{j=1}^{n-1}\left((n-1)\mathbb{V}(G_0) \;+\; \sum_{k=1}^{j}(n-k)\mathbb{V}(G_k)\right) \;+\; \sigma_+^2u^2\left((n-1)\mathbb{V}(G_0) + \sum_{j=1}^{n-1}(n-j)\mathbb{V}(G_j)\right)
\end{aligned}$$

und schließlich

$$\mathbb{V}\left(\Delta\right) \;=\left((n-1)\mathbb{V}(G_0) + \sum_{k=1}^{n-1}(n-k)\mathbb{V}(G_k)\right)\sigma_+^2u^2 \;+\left((n-1)^2\mathbb{V}(G_0) + \sum_{k=1}^{n-1}(n-k)^2\mathbb{V}(G_k)\right)\mu_+^2u^2\;.$$

Jetzt betrachten wir den Fall der Kaskaden-Summation. Dabei sei ε_{kl}^+ der relative Fehler, welcher bei der l-ten Addition im Summationslevel k auftritt. Abkürzend bezeichnen wir mit $\Delta_n^{k_0,\ldots,k_{n-1}}$ den absoluten Rundungsfehler bei der Kaskaden-Summation der Komponenten $G_{k_0},\ldots,G_{k_{n-1}}$.
Im Fall $n = 2$ ist der absolute Fehler offenbar $\Delta_2^{0,1} = \varepsilon_{11}^+(G_0+G_1)$. Für eine allgemeine Zweierpotenz $n = 2^t,\ t \geq 2$, erhalten wir entsprechend die rekursive Darstellung

$$\left.\begin{aligned}
\Delta \;:=\; \Delta_n^{0,\ldots,n-1} \;&:=\; \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{1}\right) \;-\; \mathbf{G}^{\mathrm{T}}\mathbf{1} \\
&=\; \varepsilon_{\log_2(n),1}^+\sum_{k=0}^{n-1}G_k \;+\; (1+\varepsilon_{\log_2(n),1}^+)\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1} \;+\; \Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)
\end{aligned}\right\} \qquad (A.74)$$

mit Rekursionsanfang

$$\Delta_2^{2l-2,2l-1} \;=\; \varepsilon_{1l}^{+}\left(G_{2l-2}+G_{2l-1}\right)\;, \qquad l=1,\ldots,\frac{n}{2}\;.$$

Zu den weiteren interessanten Fällen $n\in\{3,5,6,7\}$ ergeben einfache Rechnungen

$$\Delta \;=\; \Delta_n^{0,\ldots,n-1} = \begin{cases} \varepsilon_{21}^{+}(G_0+G_1+G_2)+(1+\varepsilon_{21}^{+})\Delta_2^{0,1}\;, & n=3,\\ \varepsilon_{31}^{+}(G_0+G_1+G_2+G_3+G_4)+(1+\varepsilon_{31}^{+})\Delta_4^{0,1,2,3}\;, & n=5,\\ \varepsilon_{31}^{+}(G_0+G_1+G_2+G_3+G_4+G_5)+(1+\varepsilon_{31}^{+})\left(\Delta_4^{0,1,2,3}+\Delta_2^{4,5}\right), & n=6,\\ \varepsilon_{31}^{+}(G_0+G_1+G_2+G_3+G_4+G_5+G_6)+(1+\varepsilon_{31}^{+})\left(\Delta_4^{0,1,2,3}+\Delta_2^{4,5}\right), & n=7, \end{cases}$$

für den absoluten Rundungsfehler. Aufgrund der Unabhängigkeit, der Unkorreliertheit der G_k und $\mathbb{E}(G_k)=0$ folgt zunächst

$$\mathbb{E}\left(\Delta_2^{2l-2,2l-1}\right) \;=\; \mathbb{E}(\varepsilon_{1l}^{+})\left(\mathbb{E}\left(G_{2l-2}\right)+\mathbb{E}\left(G_{2l-1}\right)\right)=0\;, \qquad l=1,\ldots,\frac{n}{2},$$

und demnach für den absoluten Rundungsfehler aus (A.74) dann ebenso per Induktion

$$\begin{aligned}\mathbb{E}\left(\Delta\right) \;&=\; \mathbb{E}\left(\Delta_n^{0,\ldots,n-1}\right)\\ &=\; \mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\right)\sum_{k=0}^{n-1}\mathbb{E}\left(G_k\right) \;+\; \left(1+\mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\right)\right)\left(\mathbb{E}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \mathbb{E}\left(\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right) \;=\; 0\;.\end{aligned}$$

Bevor wir aus (A.74) eine Rekursionsformel für die Varianz herleiten, stellen wir fest, dass

$$\mathrm{Cov}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1},\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; 0$$

für beliebige Zweierpotenz $n\geq 4$ erfüllt ist, da die beiden absoluten Rundungsfehler keine Zufallsgröße gemeinsam haben. Mit dem gleichen Argument und der Linearität der Kovarianz folgen auch

$$\mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{n-1}G_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;=\; \mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{\frac{n}{2}-1}G_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;=\; \mu_+ u\sum_{k=0}^{\frac{n}{2}-1}\mathbb{E}\left(G_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right)$$

und

$$\mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{n-1}G_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; \mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=\frac{n}{2}}^{n-1}G_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; \mu_+ u\sum_{k=\frac{n}{2}}^{n-1}\mathbb{E}\left(G_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right).$$

Somit ergibt sich mit (A.74) nach den Rechenregeln für die Varianz die Rekursion

$$\begin{aligned}\mathbb{V}\left(\Delta\right) \;&=\; \mathbb{V}\left(\Delta_n^{0,\ldots,n-1}\right)\\ &=\; \left(\sigma_+^2+\mu_+^2\right)u^2\sum_{k=0}^{n-1}\mathbb{V}\left(G_k\right) \;+\; \left(\mathbb{V}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \mathbb{V}\left(\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right)(1+\mathcal{O}(u))\\ &\quad+\; 2\mu_+ u\cdot\left(\sum_{k=0}^{\frac{n}{2}-1}\mathbb{E}\left(G_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \sum_{k=\frac{n}{2}}^{n-1}\mathbb{E}\left(G_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right)(1+\mathcal{O}(u))\;.\end{aligned}$$

Aufgrund der Unkorreliertheit der $G_r,\ r=0,\ldots,n-1$, ergibt sich aus (A.74) für eine Zweierpotenz $m\geq 4$ zunächst die Rekursion

$$\mathbb{E}\left(G_r\cdot\Delta_m^{k_0,\ldots,k_{m-1}}\right) \;=\; \begin{cases} 0, & r\notin\{k_0,\ldots,k_{m-1}\},\\ \mu_+ u\mathbb{V}(G_r) \;+\; \mathbb{E}\left(G_r\cdot\Delta_{\frac{m}{2}}^{k_0,\ldots,k_{\frac{m}{2}-1}}\right)(1+\mathcal{O}(u)), & r\in\{k_0,\ldots,k_{\frac{m}{2}-1}\},\\ \mu_+ u\mathbb{V}(G_r) \;+\; \mathbb{E}\left(G_r\cdot\Delta_{\frac{m}{2}}^{k_{\frac{m}{2}},\ldots,k_{m-1}}\right)(1+\mathcal{O}(u)), & r\in\{k_{\frac{m}{2}},\ldots,k_{m-1}\}, \end{cases}$$

mit Rekursionsanfang $\mathbb{E}\left(G_r \cdot \Delta_2^{r,\ldots,r+1}\right) \;=\; \mathbb{E}\left(G_r \cdot \Delta_2^{r-1,\ldots,r}\right) \;=\; \mu_+ u \mathbb{V}(G_r)$ und demnach

$$\sum_{k=0}^{\frac{n}{2}-1} \mathbb{E}\left(G_k \cdot \Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) + \sum_{k=\frac{n}{2}}^{n-1} \mathbb{E}\left(G_k \cdot \Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; \mu_+ u \sum_{k=0}^{n-1} \mathbb{V}(G_k) \cdot (\log_2(n)-1)\,(1+\mathcal{O}(u)) \;.$$

Durch Induktion gelangen wir dann zu

$$\mathbb{V}(\Delta) \;=\; \sum_{s=1}^{\log_2(n)} \left(\sigma_+^2 + \left(1+2(\log_2(n)-s)\right)\cdot\mu_+^2\right) \sum_{k=0}^{n-1} \mathbb{V}\left(G_k\right)\left(u^2+\mathcal{O}(u^3)\right) \;.$$

Mittels Gaußscher Summenformel und der daraus resultierenden Gleichung

$$\sum_{s=1}^{t}(1+2(t-s)) \;=\; t^2 \tag{A.75}$$

für $t=\log_2(n)$ folgt nun die letzte Behauptung. ■

Beweis von Lemma 2.16

Wir betrachten zunächst wieder den Fall der sequentiellen Summation. Ähnlich wie im Beispiel 2.4 ergibt sich für das berechnete Skalarprodukt

$$\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right) \;=\; \Big(\ldots\big(G_0X_0(1+\varepsilon_0^{\times})+G_1X_1(1+\varepsilon_1^{\times})\big)(1+\varepsilon_1^{+})+\ldots+G_{n-1}X_{n-1}(1+\varepsilon_{n-1}^{\times})\Big)(1+\varepsilon_{n-1}^{+}) \;.$$

In den Fällen $n\in\{2,3,4,5\}$ erhalten wir für $\Delta=\mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right)-\mathbf{G}^{\mathrm{T}}\mathbf{X}$ nach einfachen Rechnungen

$$\Delta \;=\; \begin{cases} \sum\limits_{k=0}^{1} G_kX_k\varepsilon_k^{\times} \;+\; \varepsilon_1^{+}\sum\limits_{k=0}^{1} G_kX_k(1+\varepsilon_k^{\times}) \;, & n=2 \\ \sum\limits_{k=0}^{2} G_kX_k\varepsilon_k^{\times} \;+\; (1+\varepsilon_2^{+})\varepsilon_1^{+}\sum\limits_{k=0}^{1} G_kX_k(1+\varepsilon_k^{\times}) \;+\; \varepsilon_2^{+}\sum\limits_{k=0}^{2} G_kX_k(1+\varepsilon_k^{\times}) \;, & n=3 \\ \sum\limits_{k=0}^{3} G_kX_k\varepsilon_k^{\times} \;+\; (1+\varepsilon_3^{+})(1+\varepsilon_2^{+})\varepsilon_1^{+}\sum\limits_{k=0}^{1} G_kX_k(1+\varepsilon_k^{\times}) & \\ \qquad + (1+\varepsilon_3^{+})\varepsilon_2^{+}\sum\limits_{k=0}^{2} G_kX_k(1+\varepsilon_k^{\times}) \;+\; \varepsilon_3^{+}\sum\limits_{k=0}^{3} G_kX_k(1+\varepsilon_k^{\times}) \;, & n=4 \\ \sum\limits_{k=0}^{4} G_kX_k\varepsilon_k^{\times} \;+\; \sum\limits_{j=1}^{4}\varepsilon_j^{+}\sum\limits_{k=0}^{j} G_kX_k(1+\varepsilon_k^{\times})\left(\prod\limits_{l=j+1}^{4}(1+\varepsilon_l^{+})\right) \;, & n=5 \end{cases}$$

und nach Induktion für allgemeines $n\geq 2$ somit

$$\Delta \;=\; \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right)-\mathbf{G}^{\mathrm{T}}\mathbf{X} \;=\; \sum_{k=0}^{n-1} G_kX_k\varepsilon_k^{\times} \;+\; \sum_{j=1}^{n-1}\varepsilon_j^{+}\sum_{k=0}^{j} G_kX_k(1+\varepsilon_k^{\times})\left(\prod_{l=j+1}^{n-1}(1+\varepsilon_l^{+})\right).$$

Mit der Unabhängigkeit und $\mathbb{E}(X_k)=0,\ k=0,\ldots,n-1$, folgt dann sofort

$$\mathbb{E}(\Delta) \;=\; \sum_{k=0}^{n-1}\mathbb{E}\left(G_k\right)\mathbb{E}\left(X_k\right)\mathbb{E}\left(\varepsilon_k^{\times}\right) \;+\; \sum_{j=1}^{n-1}\mathbb{E}\left(\varepsilon_j^{+}\right)\sum_{k=0}^{j}\mathbb{E}\left(G_k\right)\mathbb{E}\left(X_k\right)\mathbb{E}(1+\varepsilon_k^{\times})\left(\prod_{l=j+1}^{n-1}\mathbb{E}(1+\varepsilon_l^{+})\right) \;=\; 0\,,$$

da jeder einzelne Summand den Erwartungswert Null besitzt. Mit dem gleichen Argument sowie mit der Unkorreliertheit der X_k ergeben sich auch

$$\mathbb{V}\left(\sum_{k=0}^{n-1} G_kX_k\varepsilon_k^{\times}\right) \;=\; \mathbb{E}\left(\left(\sum_{k=0}^{n-1} G_kX_k\varepsilon_k^{\times}\right)^2\right) \;=\; \sum_{k,l=0}^{n-1}\mathbb{E}\left(G_kG_l\right)\mathbb{E}\left(X_kX_l\right)\mathbb{E}\left(\varepsilon_k^{\times}\varepsilon_l^{\times}\right)$$

$$\;=\; \sum_{k=0}^{n-1}\mathbb{E}\left((G_k)^2\right)\mathbb{V}(X_k)\left(\sigma_{\times}^2+\mu_{\times}^2\right)u^2 \tag{A.76}$$

sowie

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{k=0}^{n-1} G_k X_k \varepsilon_k^{\times}\ ,\ \sum_{r=1}^{n-1}\varepsilon_r^{+}\sum_{s=0}^{r} G_s X_s(1+\varepsilon_s^{\times})\left(\prod_{t=r+1}^{n-1}(1+\varepsilon_t^{+})\right)\right)\\
&\quad = \ \mu_{\times}\mu_{+}\left(\sum_{r=1}^{n-1}\sum_{s=0}^{r}\sum_{k=0}^{n-1}\mathbb{E}(G_kG_s)\mathbb{E}(X_kX_s)\right)\left(u^2+\mathcal{O}(u^3)\right)\\
&\quad = \ \mu_{\times}\mu_{+}\left((n-1)\ \mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0)\ +\ \sum_{k=1}^{n-1}(n-k)\ \mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\right)\left(u^2+\mathcal{O}(u^3)\right) \quad \text{(A.77)}
\end{aligned}$$

und ebenso

$$\begin{aligned}
&\mathbb{V}\left(\sum_{j=1}^{n-1}\varepsilon_j^{+}\sum_{k=0}^{j} G_kX_k(1+\varepsilon_k^{\times})\left(\prod_{l=j+1}^{n-1}(1+\varepsilon_l^{+})\right)\right)\\
&= \ \mathbb{E}\left(\left(\sum_{j=1}^{n-1}\varepsilon_j^{+}\sum_{k=0}^{j} G_kX_k(1+\varepsilon_k^{\times})\left(\prod_{l=j+1}^{n-1}(1+\varepsilon_l^{+})\right)\right)\left(\sum_{r=1}^{n-1}\varepsilon_r^{+}\sum_{s=0}^{r} G_sX_s(1+\varepsilon_s^{\times})\left(\prod_{t=r+1}^{n-1}(1+\varepsilon_t^{+})\right)\right)\right)\\
&= \ \sum_{j,r=1}^{n-1}\mathbb{E}\left(\varepsilon_j^{+}\varepsilon_r^{+}\right)\sum_{k=0}^{j}\sum_{s=0}^{r}\mathbb{E}\left(G_kG_s\right)\mathbb{E}\left(X_kX_s\right)\left(1+\mathcal{O}(u)\right)\ .
\end{aligned}$$

Analog zum Beweis von Lemma 2.14 erhalten wir wegen

$$\mathbb{E}\left(\varepsilon_j^{+}\varepsilon_r^{+}\right) \ = \ \left(\mu_+^2\ +\ \delta(j-r)\sigma_+^2\right)u^2$$

weiter

$$\begin{aligned}
&\sum_{j,r=1}^{n-1}\mathbb{E}\left(\varepsilon_j^{+}\varepsilon_r^{+}\right)\sum_{k=0}^{j}\sum_{s=0}^{r}\mathbb{E}\left(G_kG_s\right)\mathbb{E}\left(X_kX_s\right)\\
&= \ \mu_+^2u^2\sum_{j=1}^{n-1}\left(\sum_{r=1}^{j-1}\sum_{k=0}^{r}\mathbb{E}\big((G_s)^2\big)\mathbb{V}(X_k)+\sum_{r=j}^{n-1}\sum_{k=0}^{j}\mathbb{E}\big((G_s)^2\big)\mathbb{V}(X_k)\right)+\ \sigma_+^2u^2\sum_{j=1}^{n-1}\sum_{k=0}^{j}\mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\\
&= \ \left((n-1)\ \mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0)+\sum_{k=1}^{n-1}(n-k)\ \mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\right)\sigma_+^2u^2\\
&\quad +\ \left((n-1)^2\ \mathbb{E}\big((G_0)^2\big)\mathbb{V}(X_0)+\sum_{k=1}^{n-1}(n-k)^2\ \mathbb{E}\big((G_k)^2\big)\mathbb{V}(X_k)\right)\mu_+^2u^2\ . \qquad \text{(A.78)}
\end{aligned}$$

Aus (A.76) bis (A.78) resultiert nun für die Varianz von $\Delta := \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right) - \mathbf{G}^{\mathrm{T}}\mathbf{X}$ nach den üblichen Rechenregeln und unter Zusammenfassung von Termen der Größe $\mathcal{O}(u^3)$ die Behauptung.

Jetzt betrachten wir den Fall der Kaskaden-Summation. Dabei sei einerseits ε_{kl}^{+} der relative Fehler, welcher bei der l-ten Addition im Level k auftritt, und andererseits $\varepsilon_r^{\times}$ der relative Fehler, welcher bei der Multiplikation von G_rX_r entsteht. Abkürzend bezeichnen wir mit $\Delta_n^{k_0,\ldots,k_{n-1}}$ den absoluten Rundungsfehler des Zwischenergebnisses, welches bei der Kaskaden-Summation der Produkte $G_{k_0}X_{k_0},\ldots,G_{k_{n-1}}X_{k_{n-1}}$ auftritt. Analog zum Beweis von Lemma 2.14 erhalten wir im Fall $n=2$ für den absoluten Rundungsfehler $\Delta_2^{0,1} \ = \ \varepsilon_{11}^{+}\left(G_0X_0+G_1X_1\right)\ +\ (1+\varepsilon_{11}^{+})\left(G_0X_0\varepsilon_0^{\times}+G_1X_1\varepsilon_1^{\times}\right)$ und für eine allgemeine Zweierpotenz $n=2^t,\ t\geq 2$, die rekursive Darstellung

$$\left.\begin{aligned}
\Delta \ := \ \Delta_n^{0,\ldots,n-1} \ &:= \ \mathrm{fl}\left(\mathbf{G}^{\mathrm{T}}\mathbf{X}\right)\ -\ \mathbf{G}^{\mathrm{T}}\mathbf{X}\\
&= \ \varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{n-1}G_kX_k\ +\ (1+\varepsilon_{\log_2(n),1}^{+})\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\ +\ \Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)
\end{aligned}\right\} \qquad \text{(A.79)}$$

mit Rekursionsanfang

$$\Delta_2^{2l-2,2l-1} \;=\; \varepsilon_{1l}^{+}\left(G_{2l-2}X_{2l-2}+G_{2l-1}X_{2l-1}\right) \;+\; (1+\varepsilon_{1l}^{+})\left(G_{2l-2}X_{2l-2}\varepsilon_{2l-2}^{\times}+G_{2l-1}X_{2l-1}\varepsilon_{2l-1}^{\times}\right)$$

für $l=1,\ldots,\frac{n}{2}$ oder – falls wir den Fall $n=2$ in der Rekursion zulassen wollen und dementsprechend die entartete Summe mit nur einem Summanden betrachten – mit Rekursionsanfang

$$\Delta_1^k \;:=\; G_kX_k\varepsilon_k^{\times}\;, \qquad k=0,\ldots,n-1.$$

Aufgrund der Unabhängigkeit und mit $\mathbb{E}(X_k)=0,\ k=0,\ldots,n-1$, folgt zunächst

$$\mathbb{E}\left(\Delta_1^k\right) \;=\; \mathbb{E}(G_k)\mathbb{E}(X_k)\mu_{\times}u \;=\; 0\;, \qquad k=0,\ldots,n-1,$$

und induktiv somit auch

$$\mathbb{E}\left(\Delta_n^{0,\ldots,n-1}\right) \;=\; \mu_+u\sum_{k=0}^{n-1}\mathbb{E}(G_k)\mathbb{E}(X_k) \;+\; \left(1+\mu_+u\right)\left(\mathbb{E}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \mathbb{E}\left(\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right) \;=\; 0\;.$$

Mit demselben Argument wie im Beweis von Lemma 2.14 ergeben sich wiederum

$$\mathrm{Cov}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1},\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; 0$$

sowie mit der Linearität des Erwartungswertes dann auch analog

$$\mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{n-1}G_kX_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;=\; \mu_+u\sum_{k=0}^{\frac{n}{2}-1}\mathbb{E}\left(G_kX_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right)$$

und

$$\mathbb{E}\left(\varepsilon_{\log_2(n),1}^{+}\sum_{k=0}^{n-1}G_kX_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right) \;=\; \mu_+u\sum_{k=\frac{n}{2}}^{n-1}\mathbb{E}\left(G_kX_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)$$

für beliebige Zweierpotenz $n\geq 2$. Somit resultiert aus (A.79) nach den Rechenregeln für die Varianz

$$\begin{aligned}\mathbb{V}\left(\Delta\right) \;&=\; \mathbb{V}\left(\Delta_n^{0,\ldots,n-1}\right)\\ &=\; \left(\sigma_+^2+\mu_+^2\right)u^2\sum_{k=0}^{n-1}\mathbb{E}\left(G_k^2\right)\mathbb{V}\left(X_k\right) \;+\; \left(\mathbb{V}\left(\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \mathbb{V}\left(\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right)\left(1+\mathcal{O}(u)\right)\\ &\quad+\; 2\mu_+u\cdot\left(\sum_{k=0}^{\frac{n}{2}-1}\mathbb{E}\left(G_kX_k\cdot\Delta_{\frac{n}{2}}^{0,\ldots,\frac{n}{2}-1}\right) \;+\; \sum_{k=\frac{n}{2}}^{n-1}\mathbb{E}\left(G_kX_k\cdot\Delta_{\frac{n}{2}}^{\frac{n}{2},\ldots,n-1}\right)\right)\left(1+\mathcal{O}(u)\right)\end{aligned}$$

mit Rekursionsanfang $\mathbb{V}\left(\Delta_1^r\right) \;=\; \mathbb{E}\left(G_r^2\right)\mathbb{V}(X_r)\left(\sigma_{\times}^2+\mu_{\times}^2\right)u^2$. Ähnlich wie im Beweis zu Lemma 2.14 erhalten wir aus (A.79) für eine Zweierpotenz $m\geq 2$ die Rekursion

$$\mathbb{E}\left(G_rX_r\Delta_m^{k_0,\ldots,k_{m-1}}\right)=\begin{cases}0, & r\notin\{k_0,\ldots,k_{m-1}\},\\ \mu_+u\mathbb{V}(X_r)\mathbb{E}\left(G_r^2\right) \;+\; \mathbb{E}\left(G_rX_r\Delta_{\frac{m}{2}}^{k_0,\ldots,k_{\frac{m}{2}-1}}\right)\left(1+\mathcal{O}(u)\right), & r\in\{k_0,\ldots,k_{\frac{m}{2}-1}\},\\ \mu_+u\mathbb{V}(X_r)\mathbb{E}\left(G_r^2\right) \;+\; \mathbb{E}\left(G_rX_r\Delta_{\frac{m}{2}}^{k_{\frac{m}{2}},\ldots,k_{m-1}}\right)\left(1+\mathcal{O}(u)\right), & r\in\{k_{\frac{m}{2}},\ldots,k_{m-1}\},\end{cases}$$

mit Rekursionsanfang $\mathbb{E}\left(G_rX_r\Delta_1^r\right) \;=\; \mu_{\times}u\mathbb{V}(X_r)\mathbb{E}\left(G_r^2\right)$ und demnach

$$\mathbb{E}\left(G_rX_r\Delta_m^{k_0,\ldots,k_{m-1}}\right)=\begin{cases}0, & r\notin\{k_0,\ldots,k_{m-1}\},\\ \left(\mu_+\log_2(m)+\mu_{\times}\right)u\mathbb{V}(X_r)\mathbb{E}\left(G_r^2\right), & r\in\{k_0,\ldots,k_{m-1}\}.\end{cases}$$

Durch Induktion gelangen wir dann zu

$$\mathbb{V}\left(\Delta\right)=\left(\sigma_{\times}^2+\mu_{\times}^2+\sum_{s=1}^{\log_2(n)}\left(\sigma_+^2+\left(1+2(\log_2(n)-s)\right)\cdot\mu_+^2+2\mu_+\mu_{\times}\right)\right)\sum_{r=0}^{n-1}\mathbb{E}\left(G_r^2\right)\mathbb{V}(X_r)\left(u^2+\mathcal{O}(u^3)\right)$$

und mittels (A.75) für $t=\log_2(n)$ dann zur Behauptung. ■

Beweis von Lemma 2.32

(i) Zunächst halten wir fest, dass wegen $\hat{X} = \mathrm{fix} \circ X$ und der Transitivität der Bildung von Bildmaßen durch $\hat{X}(P)$ ein Wahrscheinlichkeitsmaß auf dem Messraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ gegeben ist. Da X nach Voraussetzung eine auf dem Intervall $[-1,1]$ gleichverteilte Zufallsgröße ist, wird $X(P)$ von der Verteilungsfunktion (A.57) mit $a = -1$ und $b = 1$ eindeutig bestimmt. Daraus folgt, dass $]-1,1[$ bezüglich des Bildmaßes $X(P)$ das Maß 1 besitzt und jedes $B \in \mathcal{B}(\mathbb{R})$ mit $B \cap]-1,1[= \emptyset$ bezüglich $X(P)$ eine Nullmenge ist. Aufgrund der Definition des Bildmaßes folgt nun ebenso $P(X^{-1}(]-1,1[)) = 1$ und – mit den Eigenschaften des Urbildes – ebenso $P(A) = 0$ für jedes $A \in \mathfrak{S}$ mit $A \cap X^{-1}(]-1,1[) = \emptyset$. Da nun $]-1,1[$ eine Teilmenge von $\mathrm{fix}^{-1}(A_q)$ ist, hat $\mathrm{fix}^{-1}(\mathbb{R} \setminus A_q)$ leeren Schnitt mit $]-1,1[$ und ist demnach eine Nullmenge bezüglich $X(P)$. Nach Definition des Bildmaßes bedeutet dies

$$0 \;=\; P(X^{-1}(\mathrm{fix}^{-1}(\mathbb{R} \setminus A_q))) \;=\; P(\hat{X}^{-1}(\mathbb{R} \setminus A_q)) \;,$$

wonach $\mathbb{R} \setminus A_q$ eine Nullmenge bezüglich des Bildmaßes $\hat{X}(P)$ ist. Da es sich bei $\hat{X}(P)$ um ein Wahrscheinlichkeitsmaß handelt, kann es insbesondere nicht mit dem Nullmaß übereinstimmen. Somit muss $\hat{X}(P)$ diskret sein und die Gestalt

$$\hat{X}(P) \;=\; \sum_{k=-2^q+1}^{2^q-1} p_k \; \eth_{k \cdot 2^{-q}}$$

besitzen. Da X eine auf dem Intervall $[-1,1]$ gleichverteilte Zufallsgröße ist und somit das Maß $X(P)$ auf $[-1,1]$ mit $\frac{1}{2}\lambda$ übereinstimmt und da nach Definition

$$\mathrm{fix}^{-1}(\{k \cdot 2^{-q}\}) \cap]-1,1[\;=\; \begin{cases} [k \cdot 2^{-q}, (k+1) \cdot 2^{-q}[& \text{für } k = 1, \ldots, 2^q - 1 \;, \\ ]-2^{-q}, 2^{-q}[& \text{für } k = 0, \\ ](k-1) \cdot 2^{-q}, k \cdot 2^{-q}] & \text{für } k = -2^q + 1, \ldots, -1 \;, \end{cases}$$

gilt, ergeben sich nun wie behauptet $p_0 = 2^{-q}$ und $p_k = 2^{-(q+1)}$ für $k \neq 0$.

(ii) Wiederum aufgrund der Transitivität der Bildung von Bildmaßen wird wegen $\mathfrak{d} = (\mathrm{Id} - \mathrm{fix}) \circ X$ durch $\mathfrak{d}(P)$ ein Wahrscheinlichkeitsmaß auf dem Messraum $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ definiert. Da aus der Definition von fix hervorgeht, dass $[-1,1]$ Teilmenge von $(\mathrm{Id} - \mathrm{fix})^{-1}([-2^{-q}, 2^{-q}])$ ist, können wir analog zu (i) schließen, dass

$$0 \;=\; P(X^{-1}((\mathrm{Id} - \mathrm{fix})^{-1}(\mathbb{R} \setminus [-2^{-q}, 2^{-q}]))) \;=\; P(\mathfrak{d}^{-1}(\mathbb{R} \setminus [-2^{-q}, 2^{-q}]))$$

und $\mathbb{R} \setminus [-2^{-q}, 2^{-q}]$ bezüglich des Bildmaßes $\mathfrak{d}(P)$ eine Nullmenge ist. Somit wissen wir bereits, dass die $\mathfrak{d}(P)$ eindeutig bestimmende Verteilungsfunktion auf den Intervallen $]-\infty, -2^{-q}[$ sowie $]2^{-q}, \infty[$ konstant Null bzw. Eins ist. Es bleibt zu zeigen, dass sie auf dem Intervall $[-2^{-q}, 2^{-q}]$ linear mit Anstieg 2^{q-1} wächst. Da X eine auf dem Intervall $[-1,1]$ gleichverteilte Zufallsgröße ist und nach Definition

$$(\mathrm{Id} - \mathrm{fix})^{-1}(]-\infty, v[) \cap [-1,1] \;=\; \begin{cases} \emptyset & \text{für } v \leq -2^{-q} \;, \\ \bigcup_{k=0}^{2^q-1} [-(k+1) \cdot 2^{-q}, v - k \cdot 2^{-q}[& \text{für } -2^{-q} < v \leq 0, \\ [-1,0] \cup \bigcup_{k=0}^{2^q-1} [k \cdot 2^{-q}, k \cdot 2^{-q} + v[& \text{für } 0 < v \leq 2^{-q} \;, \\ [-1,1] & \text{für } 2^{-q} < v \;, \end{cases}$$

gilt, ergibt sich wiederum aufgrund der Übereinstimmung von $X(P)$ auf $[-1,1]$ mit $\frac{1}{2}\lambda$ schließlich

$$F(v) \;=\; \begin{cases} 0 & \text{für } v \leq -2^{-q} \;, \\ 2^{q-1}(v + 2^{-q}) & \text{für } -2^{-q} < v \leq 2^{-q} \;, \\ 1 & \text{für } 2^{-q} < v \;, \end{cases}$$

als entsprechende Verteilungsfunktion wie behauptet. Analog sehen wir, dass $\mathbb{R} \setminus [0, 2^{-q}]$ eine Nullmenge bezüglich des Wahrscheinlichkeitsmaßes $|\mathfrak{d}|(P)$ ist und wegen $\mathfrak{d}^{-1}(]-v, v[) = |\mathfrak{d}|^{-1}([0, v[)$ für beliebiges $v > 0$ das Maß $|\mathfrak{d}|(P)$ auf $[0, 2^{-q}]$ mit $2\mathfrak{d}(P)$ übereinstimmt. ■

A.3 Statistische Testrechnungen mit MATLAB

In Kapitel 2 treten beim Modell (2.20)-(2.21) für den relativen Fehler $\varepsilon^\bullet$ die Größen $\mu_\bullet$ und $\sigma_\bullet$ auf, welche dort nicht weiter spezifiziert worden sind. Um die Ergebnisse aus 2.2 auf Testreihen anwenden zu können, benötigen wir nun entsprechende Schätzwerte für $\mu_\bullet$ und $\sigma_\bullet$. In [75], Remark 1.8, sind nicht weiter beschriebene Simulationen in MATLAB sowohl mit standardnormalverteilten als auch mit auf $[0,1]$ und $[-1,1]$ gleichverteilten Daten durchgeführt worden. Jedoch bleibt die Anzahl der durchgeführten Versuche sowie die Testmethode unerwähnt. Erstaunlicherweise werden hierbei für die drei angegebenen Verteilungen bei der Addition unterschiedliche Schätzungen angegeben. Dieser Unterschied ist bei unseren Experimenten in MATLAB nicht zu beobachten gewesen. Dabei sind wir wie folgt vorgegangen.

Experiment A.46. Für eine vorgegebene Anzahl $n < 2^{24}$ werden jeweils zwei zufällige Vektoren $\mathbf{X}$ und $\mathbf{Y}$ erzeugt, deren Komponenten X_k und Y_k sämtlich

(a) standardnormalverteilt (d.h. normalverteilt mit Erwartungswert Null und Varianz Eins),

(b) gleichverteilt auf dem Intervall $[-1,1]$ oder

(c) gleichverteilt auf dem Intervall $[0,1]$

sind. Dabei erzwingen wir (in MATLAB) über den Befehl `double(...)`, dass alle Komponenten mit 52 Nachkommastellen, also mit doppelter Genauigkeit abgespeichert werden. Die beiden Vektoren werden nun jeweils in einfacher Genauigkeit, d.h. (in MATLAB) über den Befehl `single(...)` mit maximal 24 Nachkommastellen, als auch in doppelter Genauigkeit addiert, subtrahiert und multipliziert. Für $\bullet \in \{+,-,\times\}$ bezeichnen wir dabei mit $\mathrm{fl}(X_k \bullet Y_k)$ jeweils das in einfacher Genauigkeit erhaltene Ergebnis in der k-ten Komponente. Da wir die exakten Ergebnisse nicht zur Verfügung haben, wollen wir die in doppelter Genauigkeit berechneten Ergebnisse als ausreichend genau annehmen. Anschließend bestimmen wir für jede der drei Operationen den Vektor $\mathfrak{E}^\bullet$ der Quotienten

$$\varepsilon_k^\bullet \;:=\; \begin{cases} \frac{\mathrm{fl}(X_k \bullet Y_k) - X_k \bullet Y_k}{X_k \bullet Y_k} \;, & \text{falls} \;\; X_k \bullet Y_k \neq 0 \;, \\ 0 \;, & \text{falls} \;\; X_k \bullet Y_k = 0 \;, \end{cases} \tag{A.80}$$

für die jeweils aufgetretenen relativen Fehler. Über ein Histogramm sowie ein kummulatives Histogramm erhalten wir dann Aufschluss über die Verteilungseigenschaften der entsprechenden Approximationen für den als Zufallsgröße angenommenen relativen Fehler. Um Näherungen für den Erwartungswert und die Varianz zu erhalten, verwenden wir die folgenden erwartungstreuen Schätzer

$$\hat{\mu} \;:=\; \frac{1}{n}\sum_{k=1}^{n} X_k \;, \qquad \hat{\sigma}^2 \;:=\; \frac{1}{n-1}\sum_{k=1}^{n} (X_k - \hat{\mu})^2 \;. \tag{A.81}$$

Sind nämlich alle $X_k,\ k = 1,\ldots,n$, unabhängig identisch verteilt mit Erwartungswert μ und Varianz σ^2, gilt für die neuen Zufallsgrößen $\hat{\mu}$ und $\hat{\sigma}^2$ aufgund der Linearität des Erwartungswertes sowie mit der aus der stochastischen Unabhängigkeit resultierenden Beziehung $\mathbb{E}((X_k-\mu)(X_j-\mu)) = \delta(j-k)\sigma^2$ für $j,k = 1,\ldots,n$ wie gewünscht $\mathbb{E}\,(\hat{\mu}) \;=\; \mu$ und

$$\begin{aligned} \mathbb{E}\left(\hat{\sigma}^2\right) \;&=\; \frac{1}{n-1}\sum_{k=1}^{n} \mathbb{E}\,(X_k - \hat{\mu})^2 \;=\; \frac{1}{n-1}\sum_{k=1}^{n} \mathbb{E}\left((X_k-\mu) - \frac{1}{n}\sum_{j=1}^{n}(X_j-\mu)\right)^2 \\ &=\; \frac{1}{n-1}\sum_{k=1}^{n} \mathbb{E}\left(\frac{n-1}{n}(X_k-\mu) - \frac{1}{n}\sum_{\substack{j=1\\ j\neq k}}^{n}(X_j-\mu)\right)^2 \;=\; \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{(n-1)^2}{n^2}\sigma^2 + \frac{n-1}{n^2}\sigma^2\right) \\ &=\; \frac{n}{n-1}\cdot\frac{(n-1)(n-1+1)}{n^2}\,\sigma^2 \;=\; \sigma^2 \;. \end{aligned}$$

(a) **Test für standardnormalverteilte Eingangsgrößen**
Zunächst erzeugen wir mit dem Befehl `normrnd`$(0,1,[1\ n])$ die benötigten Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$ der Länge n, deren Komponenten als normalverteilt mit Erwartungswert Null und Varianz

Eins generiert werden. Zur Überprüfung der Korrektheit des Zufallsgenerators lassen wir uns über den Befehl `hist`(...) jeweils über die Komponenten von $\mathbf{X}$ und $\mathbf{Y}$ Histogramme mit Klassenbreite 0.01 erstellen (s. Abb. A.1). Verwenden wir für den Erwartungswert und die Varianz einer Eingangsgröße x die erwartungstreuen Schätzer (A.81), so erhalten wir für die beiden Vektoren $\mathbf{X}$ und $\mathbf{Y}$ die in Abb. A.1 angegebenen Schätzer, so dass wir die Komponenten X_k, Y_k tatsächlich als standardnormalverteilt akzeptieren können.

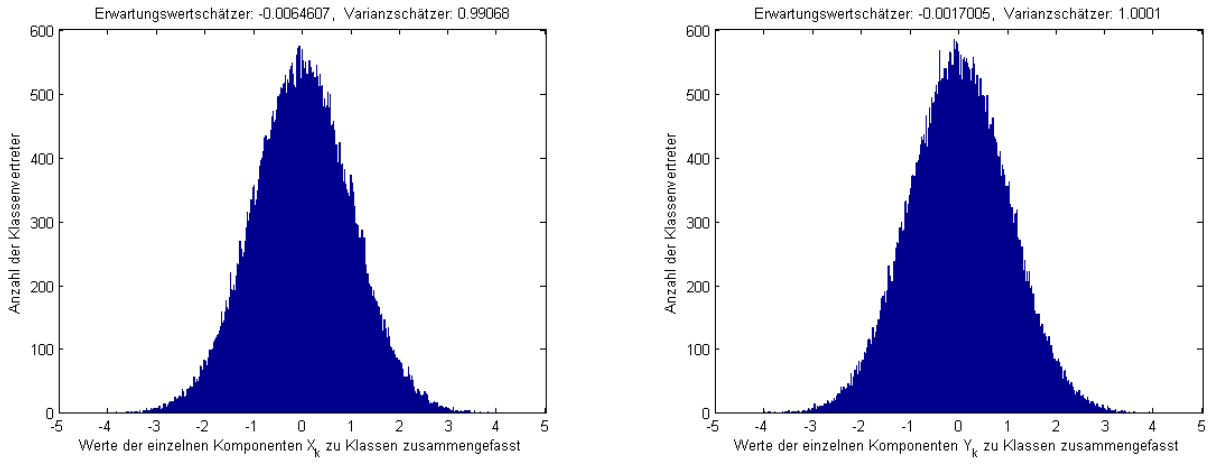


Abbildung A.1: Histogramm für die normalverteilten Komponenten von $\mathbf{X}$ und $\mathbf{Y}$. Auf der Abszissenachse sind jeweils in Klassen der Breite 0.01 die Werte der einzelnen Komponenten abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Wie oben beschrieben addieren und multiplizieren wir die Vektoren $\mathbf{X}$ und $\mathbf{Y}$ nun komponentenweise und erhalten die vier Vektoren

$$\begin{aligned} \mathbf{S}_{\text{single}} &:= \Big(\text{fl}(X_k + Y_k)\Big)_{k=1}^{n}, & \mathbf{S}_{\text{double}} &:= (X_k + Y_k)_{k=1}^{n}, \\ \mathbf{P}_{\text{single}} &:= \Big(\text{fl}(X_k \times Y_k)\Big)_{k=1}^{n}, & \mathbf{P}_{\text{double}} &:= (X_k \times Y_k)_{k=1}^{n}. \end{aligned} \tag{A.82}$$

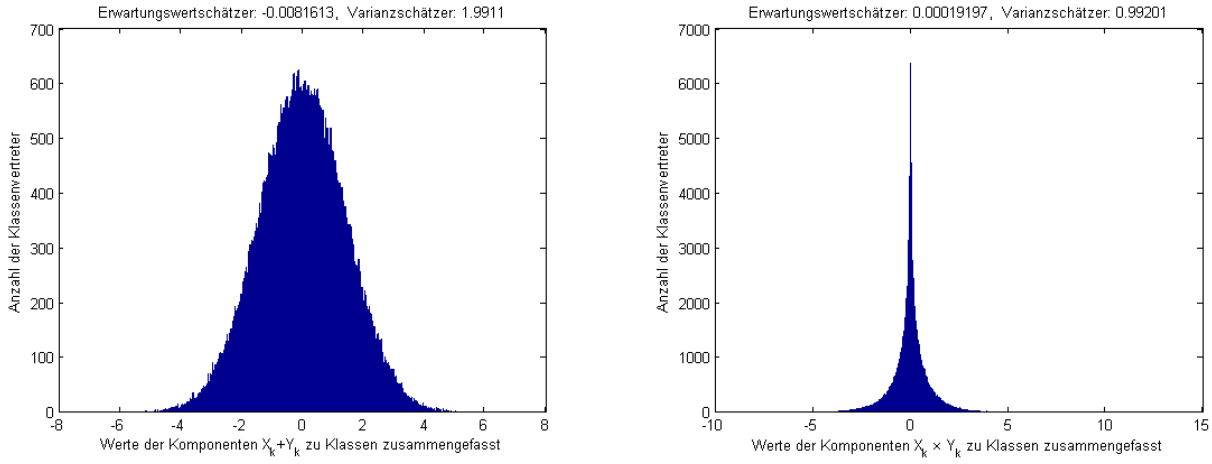


Abbildung A.2: Histogramm für die Komponenten von $\mathbf{S}_{\text{double}}$ (links) und $\mathbf{P}_{\text{double}}$ für normalverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils in Klassen der Breite 0.01 die Werte der einzelnen Komponenten abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Da die Komponenten von $\mathbf{Y}$ symmetrisch zu 0 verteilt sind, können Betrachtungen zur Subtraktion vernachlässigt werden. Wiederum mit dem Befehl `hist`(...) erhalten wir Informationen über

die Verteilungen der als exakt ausgeführt angenommenen Additionen $\mathbf{S}_{\text{double}}$ und Multiplikationen $\mathbf{P}_{\text{double}}$ (s. Abb. A.2).

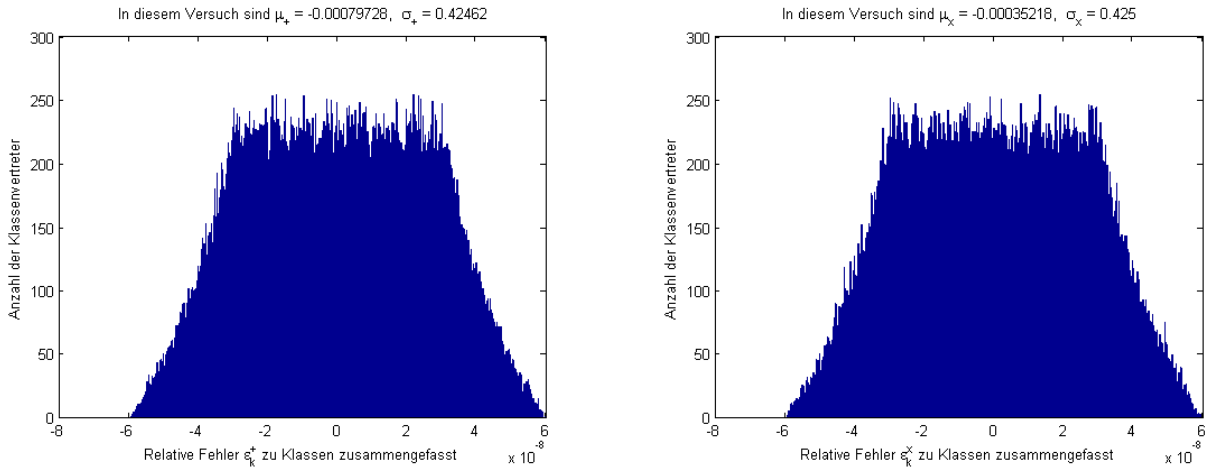


Abbildung A.3: Histogramm für die Komponenten von $\mathfrak{E}^+$ (links) und $\mathfrak{E}^\times$ für normalverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils in Klassen der Breite $0.01u$ die Werte der einzelnen Komponenten abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Mit Hilfe der in (A.82) berechneten Ergebnisvektoren erhalten wir Approximationen für die Vektoren $\mathfrak{E}^+$ und $\mathfrak{E}^\times$, deren Komponenten genau die in (A.80) definierten Größen sind. Die gewünschten Schätzwerte für σ_+, μ_+ und $\sigma_\times$, $\mu_\times$ erhalten wir nun wiederum über (A.81), die in Abb. A.3 zu finden sind.

(b) **Test für auf $[-1, 1]$ gleichverteilte Eingangsgrößen**

Um die benötigten Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$ der Länge n mit auf $[-1, 1]$ gleichverteilten Komponenten zu erhalten, verwenden wir jeweils den Befehl `unifrnd`$(-1, 1, [1\ n])$.

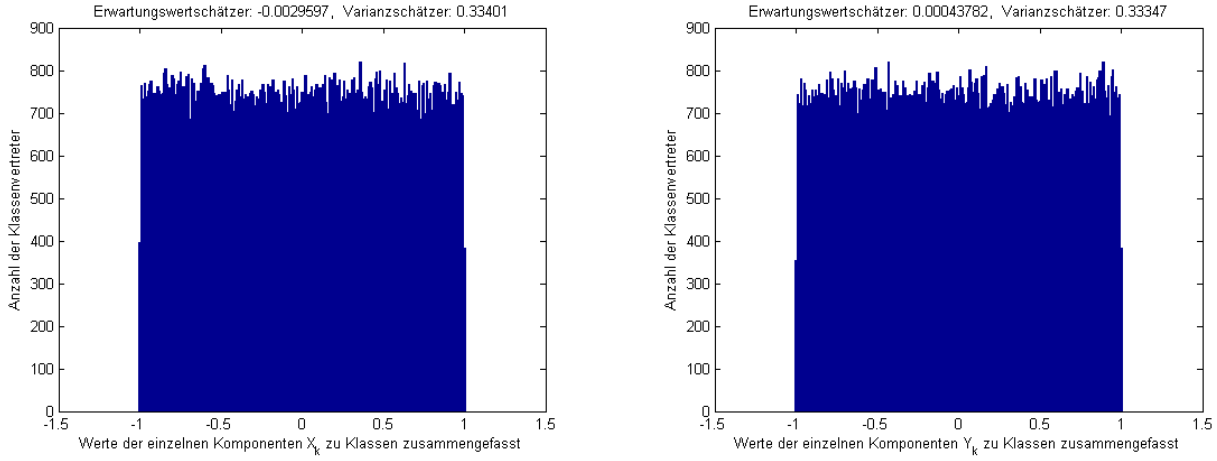


Abbildung A.4: Histogramm für die auf $[-1, 1]$ gleichverteilten Komponenten von $\mathbf{X}$ und $\mathbf{Y}$. Auf der Abszissenachse sind jeweils in Klassen der Breite 0.01 die Werte der einzelnen Komponenten abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Wie zuvor lassen wir uns über den Befehl `hist`$(\ldots)$ Histogramme über die Komponenten von $\mathbf{X}$ und $\mathbf{Y}$ mit Klassenbreite 0.01 erstellen (s. Abb. A.4) und geben wiederum die erwartungstreuen Schätzer (A.81) an. Da bei auf $[-1, 1]$ gleichverteilten Daten der theoretische Erwartungswert nach (A.58) bei 0 und die Varianz nach (A.59) bei $\frac{1}{3}$ liegen, können wir auch hier die Daten

akzeptieren. Wiederum aufgrund der zu 0 symmetrischen Verteilung der Komponenten Y_k, lassen sich für Addition und Subtraktion gleiche Aussagen ableiten. Für die Vektoren $\mathbf{S}_{\text{double}}$ und $\mathbf{P}_{\text{double}}$ – wie in (A.82) definiert – ergeben sich hier die Histogramme aus Abb. A.5.

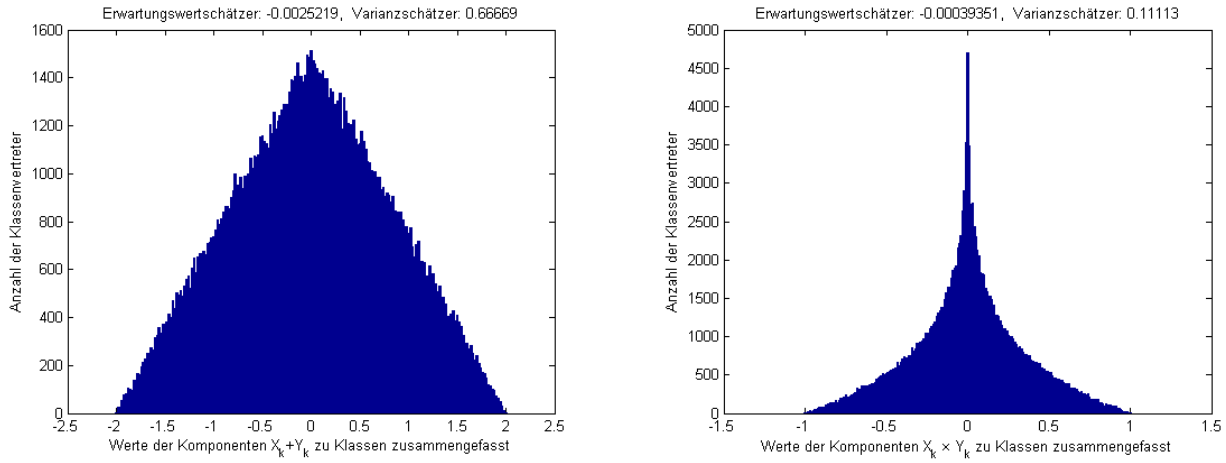


Abbildung A.5: Histogramm für die Komponenten von $\mathbf{S}_{\text{double}}$ (links) und $\mathbf{P}_{\text{double}}$ für auf $[-1, 1]$ gleichverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils die Werte der einzelnen Komponenten in 1000 Klassen zusammengefasst abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Die entsprechenden Approximationen für die Vektoren $\mathfrak{E}^+$ und $\mathfrak{E}^\times$ sowie die sich über (A.81) ergebenden gewünschten Schätzwerte für $\sigma_+,\ \mu_+$ und $\sigma_\times,\ \mu_\times$ finden wir in Abb. A.6.

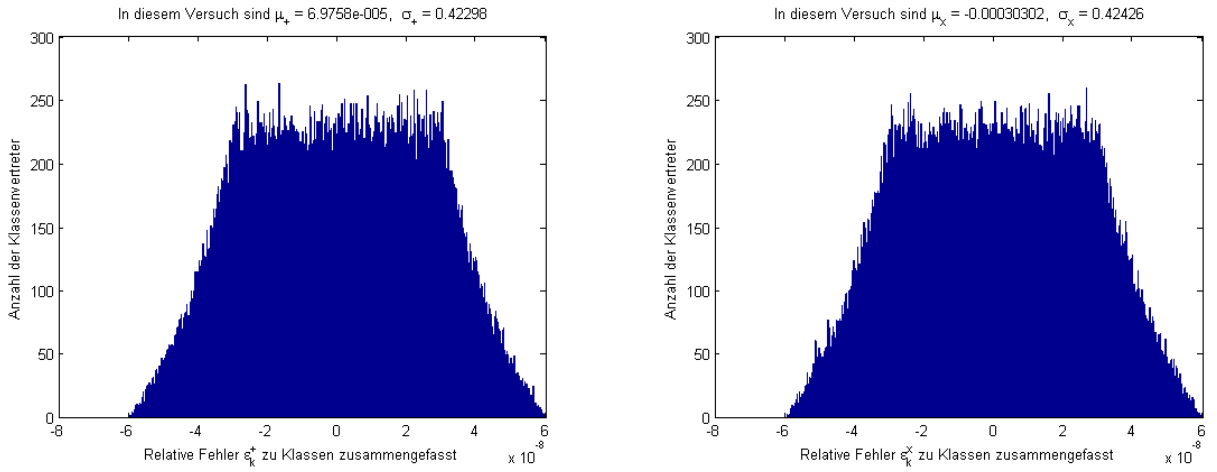
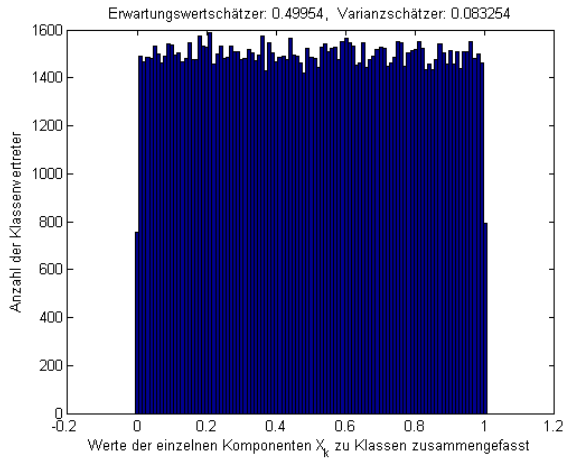


Abbildung A.6: Histogramm für die Komponenten von $\mathfrak{E}^+$ (links) und $\mathfrak{E}^\times$ für auf $[-1, 1]$ gleichverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils in Klassen der Breite $0.01u$ die Werte der einzelnen Komponenten abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

(c) **Test für auf $[0, 1]$ gleichverteilte Eingangsgrößen**

Analog zum vorangegangenen Test verwenden wir den Befehl `unifrnd`$(0, 1, [1\ n])$, um die benötigten Zufallsvektoren $\mathbf{X}$ und $\mathbf{Y}$ der Länge n mit auf $[0, 1]$ gleichverteilten Komponenten zu erhalten. Die entsprechenden mit dem Befehl `hist`$(\ldots)$ erzeugten Histogramme über die Komponenten von $\mathbf{X}$ und $\mathbf{Y}$ mit Klassenbreite 0.01 (s. Abb. A.7) ähneln denen aus Abb. A.4. Ebenso finden wir in Abb. A.7 die erwartungstreuen Schätzer (A.81) für den Erwartungswert und die Varianz. Da bei auf $[0, 1]$ gleichverteilten Daten der theoretische Erwartungswert nach (A.58) bei $\frac{1}{2}$ und die Varianz nach (A.59) bei $\frac{1}{12}$ liegen, können wir auch hier die Daten akzeptieren.



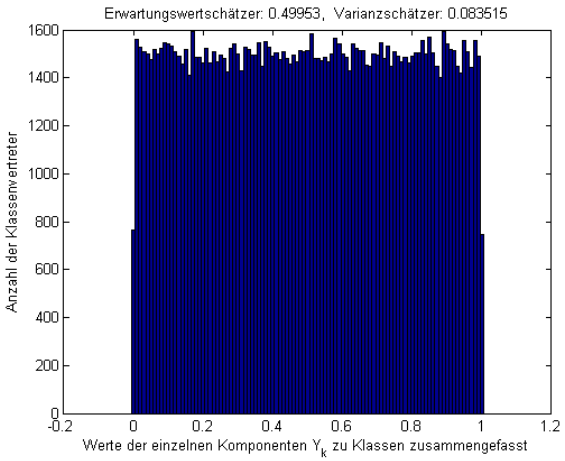
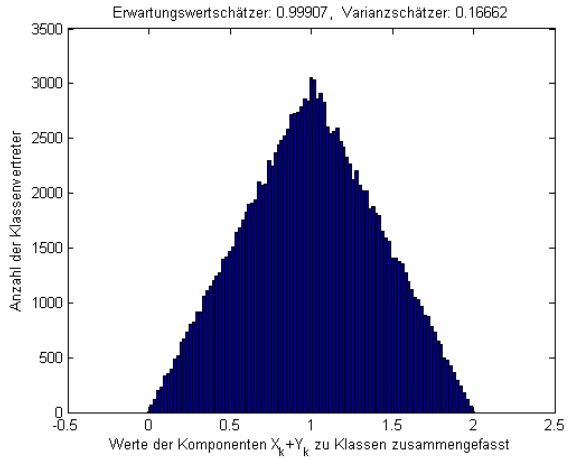


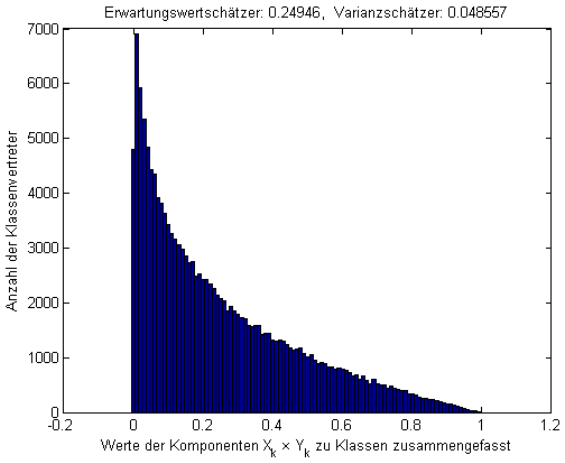
Abbildung A.7: Histogramm für die auf $[0,1]$ gleichverteilten Komponenten von $\mathbf{X}$ und $\mathbf{Y}$. Auf der Abszissenachse sind die Werte der einzelnen Komponenten jeweils in Klassen der Breite 0.01 abgetragen, während auf der Ordinatenachse die Klassenstärke abzulesen ist.

Da die Komponenten von $\mathbf{Y}$ in diesem Fall nicht symmetrisch zu 0 verteilt sind, betrachten wir neben den Vektoren aus (A.82), deren zugehörige wiederum mit dem Befehl `hist`(...) erzeugten Histogramme in Abb. A.8 zu finden sind, weiterhin noch die Differenzvektoren

$$\mathbf{D}_{\text{single}} \;:=\; \Big(\,\mathrm{fl}(X_k - Y_k)\Big)_{k=1}^{n}\,, \qquad\qquad \mathbf{D}_{\text{double}} \;:=\; (X_k - Y_k)_{k=1}^{n}\,. \tag{A.83}$$

Das entsprechende Histogramm ist ebenfalls in Abb. A.8 enthalten.





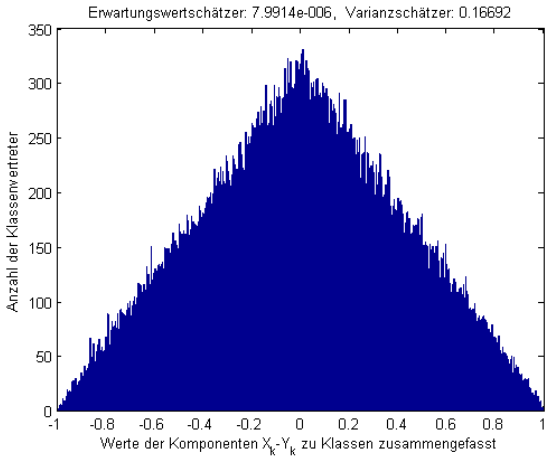


Abbildung A.8: Histogramm für die Komponenten von $\mathbf{S}_{\text{double}}$ (links oben), $\mathbf{P}_{\text{double}}$ (rechts oben) und $\mathbf{D}_{\text{double}}$ für auf $[0,1]$ gleichverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils die Werte der einzelnen Komponenten in 1000 Klassen zusammengefasst abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

Analog zuvor erhalten wir entsprechende Approximationen für die Vektoren $\mathfrak{E}^{+}$, $\mathfrak{E}^{-}$ und $\mathfrak{E}^{\times}$

sowie die sich über (A.81) ergebenden gewünschten Schätzwerte für $\sigma_{\pm}$, $\mu_{\pm}$ und $\sigma_{\times}$, $\mu_{\times}$, welche wir in Abb. A.9 ablesen können.

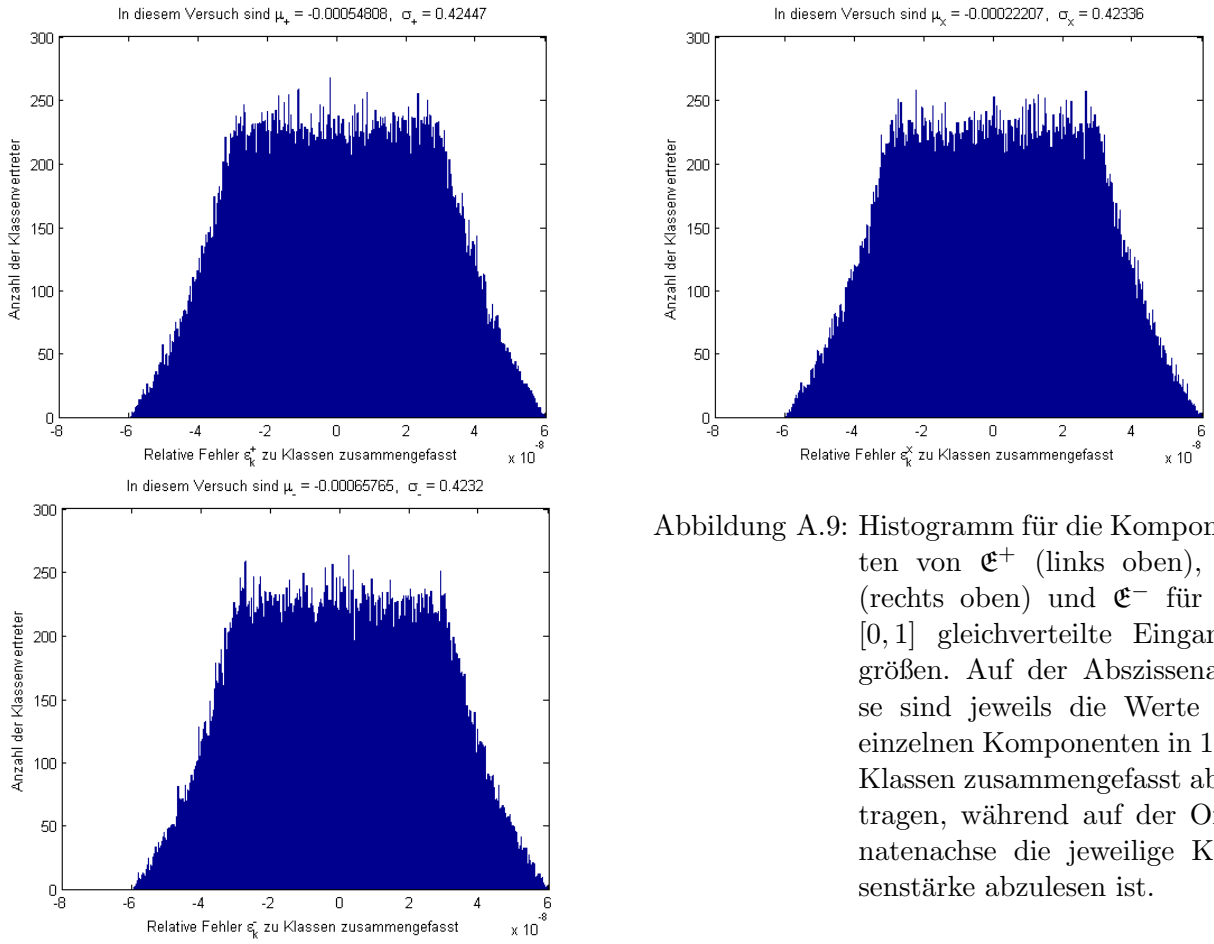


Abbildung A.9: Histogramm für die Komponenten von $\mathfrak{E}^{+}$ (links oben), $\mathfrak{E}^{\times}$ (rechts oben) und $\mathfrak{E}^{-}$ für auf $[0,1]$ gleichverteilte Eingangsgrößen. Auf der Abszissenachse sind jeweils die Werte der einzelnen Komponenten in 1000 Klassen zusammengefasst abgetragen, während auf der Ordinatenachse die jeweilige Klassenstärke abzulesen ist.

□

Zu beobachten ist, dass sich die Histogramme aus dem Experiment A.46 für die jeweiligen relativen Fehler in den Abbildungen A.3, A.6 und A.9 sehr stark ähneln und daher die Vermutung zulassen, dass der relative Fehler – zumindest bei standardnormalverteilten oder gleichverteilten Eingangsdaten – weder von den Eingangsdaten noch von der gewählten Operation $\bullet \in \{+,-,\times\}$ abhängt. Desweiteren können wir verifizieren, dass sich der relative Fehler offenbar jeweils im Intervall $[-u,u]$ mit dem hier modellierten $u = 2^{-24} \approx 6 \cdot 10^{-8}$ befindet. In einem weiteren Experiment untersuchen wir den bei einer skalierten Drehung auftretenden relativen Fehler.

Experiment A.47. Für eine vorgegebene Anzahl $n < 2^{24}$ werden jeweils zwei zufällige Vektoren $\mathbf{X}$ und $\mathbf{Y}$ erzeugt, deren Komponenten X_k und Y_k sämtlich

(a) standardnormalverteilt (d.h. normalverteilt mit Erwartungswert Null und Varianz Eins),

(b) gleichverteilt auf dem Intervall $[-1,1]$ oder

(c) gleichverteilt auf dem Intervall $[0,1]$

sind. Dabei erzwingen wir wiederum über den Befehl `double`(...), dass alle Komponenten mit 52 Nachkommastellen, also mit doppelter Genauigkeit abgespeichert werden. Die beiden Vektoren werden als Zeilen einer $2 \times n$ Matrix aufgefasst, deren sämtliche Spalten

$$\mathbf{V}^{(k)} := \begin{pmatrix} X_k \\ Y_k \end{pmatrix}, \qquad k = 1, \ldots, n,$$

mit einer konstanten Drehmatrix

$$A := \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix}, \qquad \varphi \in \left]0, \frac{\pi}{4}\right[$$

jeweils sowohl in einfacher Genauigkeit, d.h. über den Befehl `single`(...) mit maximal 24 Nachkommastellen, als auch in doppelter Genauigkeit multipliziert werden. Dabei wollen wir die in doppelter Genauigkeit vorhandenen Einträge der Drehmatrix A als (vergleichsweise) exakt annehmen. Analog zuvor bezeichnen wir mit $\mathrm{fl}(A\mathbf{V}^{(k)})$ jeweils die in einfacher Genauigkeit gedrehte k-te Spalte. Da wir die exakten Ergebnisse nicht zur Verfügung haben, wollen wir die in doppelter Genauigkeit gedrehten Spalten als ausreichend genau annehmen. Anschließend bestimmen wir die Vektoren

$$\mathfrak{E}^k := \begin{cases} \frac{\mathrm{fl}(A\mathbf{V}^{(k)}) - A\mathbf{V}^{(k)}}{\|A\mathbf{V}^{(k)}\|_2} , & \text{falls } \mathbf{V}^{(k)} \neq \mathbf{0} , \\ \mathbf{0} , & \text{falls } \mathbf{V}^{(k)} = \mathbf{0} , \end{cases} \tag{A.84}$$

der jeweils aufgetretenen relativen Fehler sowie die entsprechenden euklidischen Normen dieser Vektoren. Über ein Histogramm erhalten wir dann Aufschluss über die Verteilungseigenschaften der entsprechenden Approximationen für den in diesem Fall als zweidimensionalen Zufallsvektor angenommenen relativen Fehler. Abbildung A.10 zeigt den Vektor $\mathfrak{E}^k$ aus (A.84) bei einer Drehung um den Winkel $\frac{3\pi}{64}$ exemplarisch für den Fall standardnormalverteilter Zufallsgrößen.

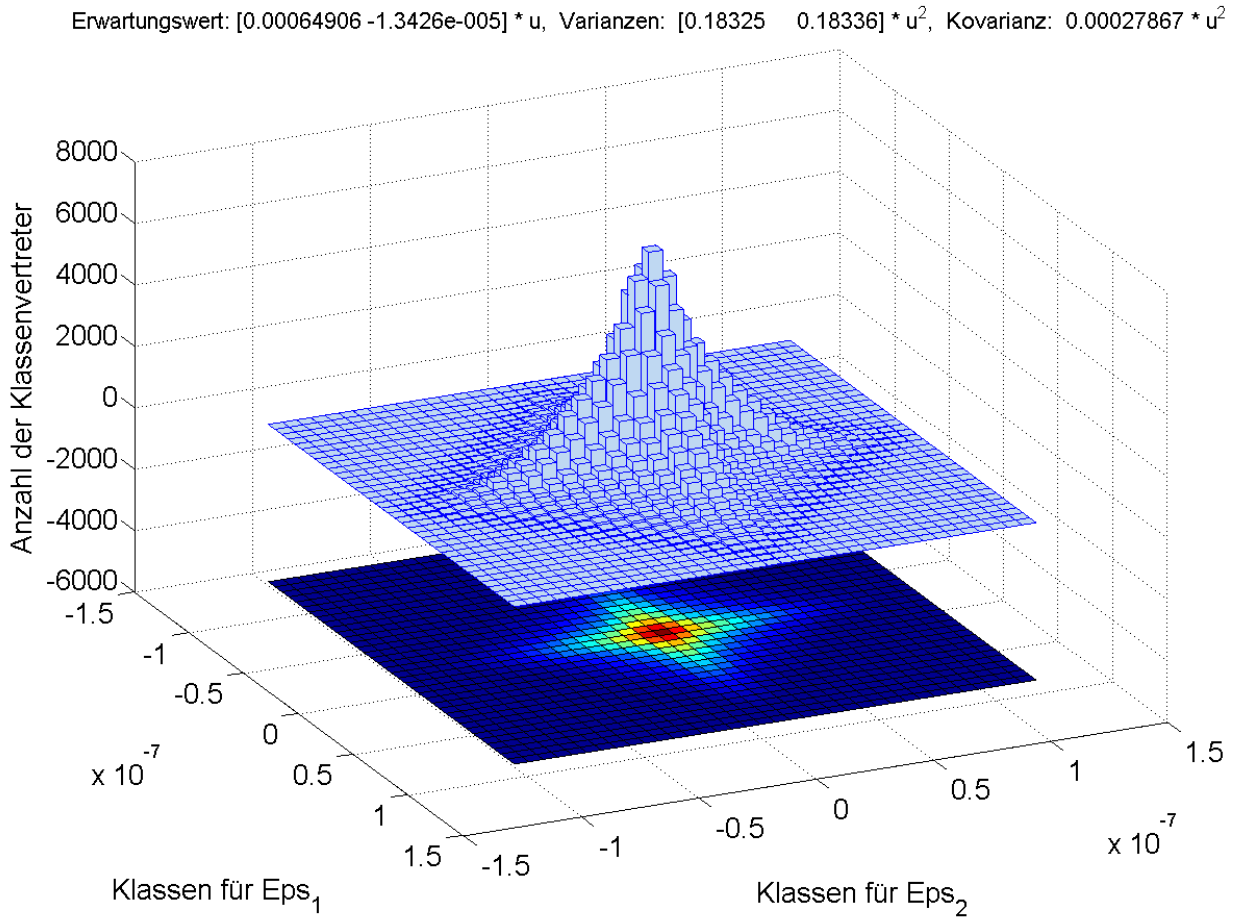


Abbildung A.10: Relativer Fehler aus (A.84) für den Drehwinkel $\frac{3\pi}{64}$.

Sachwortverzeichnis

Literaturverzeichnis

[1] N. Ahmed, T. Natarajan, K.R. Rao. *Discrete cosine transform.* IEEE Trans. Comput. **23** (1974), 90 – 93.

[2] R. Ansari, C. Guillemot, N. Memon. *Lossy Image Compression: JPEG and JPEG*2000 *Standards*, in: Handbook of Image and Video Processing (A. Bovik, ed.). Second Edition, Academic Press, New York, 2000, 709 – 731.

[3] H. Bauer. *Maß- und Integrationstheorie.* 2. Auflage, de Gruyter, Berlin, 1992.

[4] H. Bauer. *Wahrscheinlichkeitstheorie.* 5. Auflage, de Gruyter, Berlin, 2002.

[5] K. Behnen & G. Neuhaus. *Grundkurs Stochastik.* 3. Auflage, B.G. Teubner, Stuttgart, 1995.

[6] E.O. Brigham. *The Fast Fourier Transform and its Applications.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[7] W.S. Brown. *A simple but realistic model of floating-point arithmetic.* ACM Trans. Math. Software **7** (1981), 445 – 480.

[8] J. Bustoz, A. Feldstein, R. Goodman, S. Linnainmaa. *Improved trailing digits estimates applied to optimal computer arithmetic.* J.ACM **26** (1979), 716 – 730.

[9] D. Calvetti. *A stochastic roundoff error analysis for the fast Fourier transform.* Math. Comput. **194** (1991), 755 – 774.

[10] J.W. Carr III. *Error analysis in floating point arithmetic.* Comm. ACM **5** (1959), 10 – 15.

[11] R.J. Clarke. *Relation between the Karhunen Loève and cosine transforms.* IEE Proc. **128** (1981), 359 – 360.

[12] J.W. Cooley & J.W. Tukey *An algorithm for the machine calculation of complex Fourier series.* Math. Comp. **19** (1965), 297 – 301.

[13] J.W. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia, 1997.

[14] S. Didas, S. Setzer, G. Steidl. *Combines l_2 data and gradient fitting in conjunction with l_1 regularization.* Adv. Comput. Math **30** (2009), 79 – 99.

[15] P.S.R. Diniz, E.A.B. da Silva, S.L. Netto. *Digital Signal Processing: System Analysis and Design.* Second Edition, Cambridge University Press, Cambridge, 2010.

[16] C.H. Dommergues, J.-L. Dommergues & E.P. Verrecchia. *The discrete cosine transform, a Fourier-related method for morphometric analysis of open contours.* Math. Geol. **39** (2007), 749 – 763.

[17] J. Elstrodt. *Maß- und Integrationstheorie.* 2. Auflage, Springer, Berlin, 1999.

[18] E. Feig & M. Ben-Or. *On algebras related to the discrete cosine transform.* Linear Algebra Appl. **266** (1997), 81 – 106.

[19] J.W. Goodman. *Introduction to Fourier Optics.* Third edition, Roberts & Company, Englewood, CO, 2005.

[20] J.-I. Guo & J.-C. Yen. *An efficient* IDCT *processor design for* HDTV *applications.* J. VLSI Signal Process. **33** (2003), 147 – 155.

[21] Z.M. Hafed & M.D. Levine. *Face recognition using the discrete cosine transform.* Int. J. Comput. Vision **43** (2001), 167 – 188.

[22] M.T. Heidemann, D.H. Johnson, C.S. Burrus. *Gauss and history of the fast Fourier transform.* Arch. Hist. Exact. Sci. **34** (1985), 265 – 277.

[23] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations.* Wiley, New York, 1962.

[24] N.J. Higham. *Accuracy and Stability of Numerical Algorithms.* Second edition, SIAM, Philadelphia, 2002.

[25] H.S. Hou. *A fast recursive algorithm for computing the discrete cosine transform.* IEEE Trans. Acoust. Speech Signal Process. **35** (1987), 1455 – 1461.

[26] C.-Y. Hsu & Y.C. Yao. *Comparative performance of fast cosine transform with fixed-point roundoff error analysis.* IEEE Trans. Signal Process. **42** (1994), 1256 – 1259.

[27] K. Ihsberner. *Roundoff error analysis of fast DCT algorithms in fixed point arithmetic.* Numer. Algorithms **46** (2007), 1 – 22.

[28] S. Jirachaweng & V. Areekul. *Fingerprint enhancement based on discrete cosine transform.* LNCS **4642** (ICB 2007), 96 – 105.

[29] S.G. Johnson & M. Frigo. *A modified split-radix FFT with fewer arithmetic operations.* IEEE Trans. Signal Process. **55** (2007), 111 – 119.

[30] T. Kailath & V. Olshevsky. *Displacement structure approach to discrete-trigonometric-transform based preconditioners of G. Strang type and of T. Chan type.* SIAM J. Matrix Anal. Appl. **26** (2005), 706 – 734.

[31] D.W. Kammler. *A First Course in Fourier Analysis.* Second edition, Cambridge University Press, New York, 2007.

[32] T. Kaneko & B. Liu. *Accumulation of roundoff error in fast Fourier transform.* J. Assoc. Comput. Mach. **17** (1970), 637 – 654.

[33] T. Kaneko & B. Liu. *On local roundoff errors in floating-point arithmetic.* J. Assoc. Comput. Mach. **20** (1973), 391 – 398.

[34] W.R. Knight & R. Kaiser. *A simple fixed-point error bound for the fast Fourier transform.* IEEE Trans. Acoust. Speech Signal Process. **27** (1979), 615 – 620.

[35] V. Kober & J.G. Agis. *Space-variant restoration with sliding discrete cosine transform.* in: Lecture Notes in Computer Science (LNCS **4673**), Computer Analysis of Images and Patterns (W.G. Kropatsch, M. Kampel, A. Hanbury, eds.). Springer, Berlin, 2007, 903 – 911.

[36] J. Liang & T.D. Tran. *Fast multiplierless approximations of the DCT with the lifting scheme.* IEEE Trans. Signal Process. **49** (2001), 3032 – 3044.

[37] H.S. Malvar. *Lapped transforms for efficient transform/subband coding.* IEEE Trans. Acoust. Speech Signal Process. **38** (1990), 969 – 978.

[38] A.G. Marshall & F.R. Verdun. *Fourier Transforms in NMR, Optical and Mass Spectroscopy.* Elsevier, New York, 1990.

[39] S.A. Martucci. *Symmetric convolution and the discrete sine and cosine transforms.* IEEE Trans. Signal Process. **42** (1994), 1038 – 1051.

[40] U. Meyer-Bäse. *Schnelle digitale Signalverarbeitung.* Springer, Berlin, 2000.

[41] J. von Neumann & H.H. Goldstine. *Numerical inverting of matrices of high order.* Bull. Amer. Math. Soc. **53** (1947), 1021 – 1099.

[42] G. Plonka & M. Tasche. *Invertible integer DCT algorithms.* Appl. Comput. Harmon. Anal. **15** (2003), 70 – 88.

[43] G. Plonka & M. Tasche. *Fast and numerically stable algorithms for discrete cosine transforms.* Linear Algebra Appl. **394** (2005), 309 – 345.

[44] M. Primbs. *Worst-case error analysis of lifting-based fast* DCT*-algorithms.* IEEE Trans. Signal Process. **53** (2005), 3211 – 3218.

[45] M. Püschel & J.M.F. Moura. *The algebraic approach to the discrete cosine and sine transform and their fast algorithms.* SIAM J. Comput. **32** (2003), 1280 – 1316.

[46] K. R. Rao & P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, Boston, 1990.

[47] V. G. Reju, S. N. Koh, I. Y. Soon. *Convolution using discrete sine and cosine transforms.* IEEE Signal Process. Letters **14** (2007), 445 – 448.

[48] K. Samelson & F.L. Bauer *Optimale Rechengenauigkeit bei Rechenanlagen mit gleitendem Komma.* Z. Angew. Math. Phys. **4** (1953), 312 – 316.

[49] V. Sáchez, P. García, A.M. Peinado, J.C. Segura, A.J. Rubio. *Diagonalizing properties of the discrete cosine transforms.* IEEE Trans. Signal Process. **43** (1995), 2631 – 2641.

[50] U. Schreiber, *Schnelle und numerisch stabile trigonometrische Transformationen.* Diss., Univ. Rostock, 1999.

[51] G. Steidl & M. Tasche. *A polynomial approach to fast algorithms for discrete Fourier-cosine and Fourier-sine transforms.* Math. Comput. **56** (1991), 281 – 296.

[52] G. Strang. *The discrete cosine transform.* SIAM Rev. **41** (1999), 135 – 147.

[53] G. Strang. *Introduction to Linear Algebra.* Third Edition, Wellesley-Cambridge Press, Wellesley, 2005

[54] M. Tasche & H. Zeuner. *Roundoff error analysis for fast trigonometric transforms*, in: Handbook of Analytic-Computational Methods in Applied Mathematics (G. Anastassiou, ed.). Chapman & Hall / CRC Press, Boca Raton, 2000, 357 – 406.

[55] M. Tasche & H. Zeuner. *Worst and average case roundoff error analysis for FFT.* BIT **41** (2001), 563 – 581.

[56] M. Tasche & H. Zeuner. *Improved roundoff error analysis for precomputed twiddle factors.* J. Comput. Anal. Appl. **4** (2002), 1 – 18.

[57] P.P. Vaidyanathan. *Multirate Systems and Filterbanks.* Prentice Hall, 1993.

[58] C. F. Van Loan. *Computational Framework for the Fast Fourier Transform.* SIAM, Philadelphia, 1992.

[59] Z. Wang. *Reconsideration of "A fast computational algorithm for the discrete cosine transform".* IEEE Trans. Commun. **31** (1983), 121 – 123.

[60] Z. Wang. *Fast algorithms for the discrete W transform and the discrete Fourier transform.* IEEE Trans. Acoust. Speech Signal Process. **32** (1984), 803 – 816.

[61] Z. Wang. *On computing the discrete Fourier and cosine transforms.* IEEE Trans. Acoust. Speech Signal Process. **33** (1985), 1341 – 1344.

[62] Z. Wang, Z. He, C. Zou, J.D.Z. Chen. *A generalized fast algorithm for* n*-*D *discrete cosine transform and its application to motion picture coding.* IEEE Trans. Circuits Systems **46** (1999), 617 – 627.

[63] Z. Wang & B. Hunt. *The discrete W transform* Appl. Math. Comput. **16** (1985), 19 – 48.

[64] C.J. Weinstein. *Roundoff noise in floating point fast Fourier transform computation.* IEEE Trans. Audio Electroacoust. **17-3** (1969), 209 – 215.

[65] P.D. Welch. *A fixed-point fast Fourier transform error analysis.* IEEE Trans. Audio Electroacoust. **17-2** (1969), 151 – 157.

[66] J.H. Wilkinson. *Error analysis of floating-point computation.* Numer. Math. **2** (1960), 319 – 340.

[67] J.H. Wilkinson. *Rounding Errors in Algebraic Processes.* Prentice Hall, London, 1963.

[68] J.H. Wilkinson. *Rundungsfehler.* Springer, Berlin, 1969. Deutsche Übersetzung von [67].

[69] P. Yip & K.R. Rao. *A fast computational algorithm for the discrete sine transform.* IEEE Trans. Commun. **28** (1980), 304 – 307.

[70] P. Yip & K.R. Rao. *Fast decimation-in-time algorithms for a family of discrete sine and cosine transforms.* Circuits Systems Signal Process. **3** (1984), 387 – 408.

[71] P. Yip & K.R. Rao. *Fast decimation-in-frequency algorithms for a family of discrete sine and cosine transforms.* Circuits Systems Signal Process. **7** (1988), 3 – 19.

[72] I.D. Yun & S.U. Lee. *On the fixed-point-error analysis of several fast DCT algorithms.* IEEE Trans. Circuits Syst. Video Technol. **3** (1993), 27 – 41.

[73] I.D. Yun & S.U. Lee. *On the fixed-point-error analysis of several fast IDCT algorithms.* IEEE Trans. Circuits Systems **42** (1995), 685 – 693.

[74] Y. Zeng, Z. Lin, G. Bi, L. Cheng. *Fast computation of* MD-DCT-IV/MD-DST-IV *by* MD-DWT *or* MD-DCT-II. SIAM J. Sci. Comput. **24** (2003), 1903 – 1918.

[75] H. Zeuner. *A general theory of stochastic roundoff error analysis with applications to* DFT *and* DCT. J. Concr. Appl. Math. **3** (2005), 283 – 311 .

Zusammenfassung

Diskrete Kosinus- und Sinustransformationen (DCT, DST) zählen zu den wichtigen und häufig gebrauchten Werkzeugen der digitalen Signalverarbeitung. In dieser Dissertation wird eine umfassende und einheitliche Stabilitätsanalyse sowohl in Festkomma- als auch in Gleitkomma-Arithmetik für eine Klasse von schnellen DCT- und DST-Algorithmen durchgeführt, welche auf Faktorisierungen der orthogonalen Transformationsmatrizen in Produkte von dünnbesetzten orthogonalen Matrizen beruhen. Neben Untersuchungen für den ungünstigsten Fall (worst case) wird auch jeweils eine stochastische Rundungsfehleranalyse (average case) durchgeführt, die das Verhalten des durchschnittlichen relativen bzw. absoluten Rundungsfehlers abschätzt. Insbesondere für die Festkomma-Arithmetik sind bisher keine Aussagen über den Rundungsfehler bei den DCT und DST bekannt. Neu ist hierbei auch, dass das stochastische Modell für den Rundungsfehler ohne die in der Bildverarbeitung selten gegebene Unkorreliertheit der Eingangsdaten auskommt und nur wenige Informationen über die Verteilung der beteiligten Größen benötigt.

Abstract

In this thesis, a comprehensive and unified stability analysis for a class of fast DCT (discrete cosine transform) and DST (discrete sine transform) algorithms is performed, both for fixed-point and floating-point arithmetic. These algorithms belong to important and often used tools within digital signal processing. Each of them is based on a factorization of the underlying orthogonal transform matrix into a product of sparse orthogonal matrices. Additionally to worst case analysis, also the average case is considered using stochastic models for the relative and absolute roundoff errors. Especially for fixed-point arithmetic, results concerning the roundoff errors for DCT- and DST-algorithms are rarely known. Particularly with regard to applications in digital image processing, the stochastic analysis of roundoff error is done without assuming the data to be uncorrelated or independent.

Selbständigkeitserklärung nach §4 Abs. 1 der Promotionsordnung

Ich versichere hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig angefertigt und ohne fremde Hilfe verfasst habe, keine außer den von mir angegebenen Hilfsmitteln und Quellen dazu verwendet habe und die den benutzten Werken inhaltlich und wörtlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, den 31. Januar 2011

Katja Ihsberner
Ulmenstraße 69
18057 Rostock
Tel.: 0381-498 6582

Curriculum Vitae

5. September 1979	Geboren in Rostock
1986 – 1988	54. POS Alexander Fadejew, Rostock-Evershagen
1988 – 1991	35. POS Albin Köbis, Rostock-Lütten Klein
1991 – 1992	Integrierte Gesamtschule, Rostock-Lütten Klein
1992 – 1998	Gymnasium CJD Rostock
1998	Abitur mit Gesamtprädikat „sehr gut“ (1.2)
1998 – 2004	Studium der Mathematik an der Universität Rostock,
	Nebenfach Informatik
2001	Vordiplom mit Gesamtprädikat „gut“ (1.7)
2004	Diplomarbeit zum Thema:
	„Numerisch stabile DCT- und DST-Algorithmen“
	Betreuer: Prof. Dr. M. Tasche
Dezember 2004	Diplom mit Gesamtprädikat „sehr gut“ (1.2)
seit 2005	Promotionsstudium an der Universität Rostock bei Prof. Dr. M. Tasche
	Tätigkeit als wissenschaftliche Mitarbeiterin an der
	Universität Rostock bei Prof. Dr. K.P. Rybakowski

Thesen zur Dissertation

Deterministische und stochastische Rundungsfehleranalysen von schnellen trigonometrischen Algorithmen in Gleitkomma- bzw. Festkomma-Arithmetik

von Dipl. Math. Katja Ihsberner

- Die Menge $\mathbb{M}$ der in einem Computer darstellbaren Zahlen ist endlich. In der elektronischen Datenverarbeitung treten daher neben Eingangsfehlern ebenso bei arithmetischen Operationen kaum vermeidbare Rundungsfehler auf, welche von der implementierten Arithmetik abhängig sind. Erste Untersuchungen zum Rundungsfehlerverhalten bei Algorithmen zum Lösen linearer Gleichungssysteme gehen auf J. von Neumann & H.H. Goldstine [41, (1947)] zurück, die überwiegend von J.H. Wilkinson [67, (1963)] fortgeführt wurden.

- Aufgrund der höheren Flexibilität verwenden die meisten Computer die sogenannte Gleitkomma-Arithmetik, für welche daher in erster Linie Rundungsfehleranalysen entwickelt wurden (vgl. Higham [24, (2002)]).

- Bei der deterministischen Rundungsfehleranalyse wird der Rundungsfehler im ungünstigsten Fall (worst-case) bei jedem Schritt nach oben abgeschätzt. Die auf diese Weise ermittelten oberen Schranken sind im Allgemeinen jedoch viel zu pessimistisch. Daher wird im Rahmen einer stochastischen Rundungsfehleranalyse zusätzlich das durchschnittliche Verhalten (average-case) des Rundungsfehlers abgeschätzt.

- In der digitalen Signalverarbeitung ist der Bedarf an anwendungsspezifisch konzipierten Signalprozessoren in den letzten Jahren enorm angestiegen. Nicht selten wird dabei auf die im Vergleich zur konventionellen Gleitkomma-Arithmetik *einfachere* und bezüglich der Hardware *weniger kostenintensive* Festkomma-Arithmetik zurückgegriffen (vgl. Diniz [15, (2010)]). Somit besteht großes Interesse an deterministischen bzw. stochastischen Rundungsfehleranalysen in Festkomma-Arithmetik, was das Thema dieser Dissertation ist.

- Zu den bekanntesten schnellen und numerisch stabilen Algorithmen gehört die schnelle Fourier-Transformation (FFT). Diese Methode zur Berechnung der diskreten Fourier-Transformation (DFT) wurde von Cooley und Tukey [12, (1965)] vorgeschlagen und beruht auf der Teile-und-Herrsche-Strategie. Bei einer Transformationslänge $n = 2^t$ reduziert die FFT gegenüber einer naiven Implementierung der DFT die Anzahl der arithmetischen Operationen von $\mathcal{O}(n^2)$ auf $\mathcal{O}(n \log_2 n)$.

- Bei zahlreichen Aufgabenstellungen sind reelle Datenvektoren gegeben. Um die komplexe Arithmetik bei der DFT zu umgehen, werden beispielsweise die *diskreten Kosinus-Transformationen* (DCT) und die *diskreten Sinus-Transformationen* (DST) angewandt.

- Aufgrund ihrer Linearität, leichten Invertierbarkeit und Nähe zur DFT besitzen die DCT und DST zahlreiche Anwendungen, beispielsweise in der digitalen Signalverarbeitung. Eine herausgehobene Rolle spielen dabei die DCT-II in der Datenkompression (vgl. Ansari [2, (2000)]) und die DCT-IV beim Entwurf von effizienten Filterbänken (vgl. Vaidyanathan [57, (1993)] und Diniz [15, (2010)]).

- Für die DCT und DST lassen sich reelle, schnelle und numerisch stabile Algorithmen angeben, welche ebenso auf der Teile-und-Herrsche-Strategie beruhen. Wie bei der DFT gibt es für die DCT und DST verschiedene Möglichkeiten einer schnellen Realisierung (vgl. C. van Loan [58, (1992)]). Im Vordergrund stehen hier Algorithmen, welche auf Matrixfaktorisierungen

$$A \;=\; \prod_{m=1}^{\nu} A^{(m)} \;:=\; A^{(\nu)} \dots A^{(2)} A^{(1)} \tag{T.1}$$

der entsprechenden orthogonalen Transformationsmatrix A basieren, deren Faktoren $A^{(m)}$ *dünnbesetzt* und orthogonal sind. Dabei bedeutet hier die Dünnbesetztheit von $A^{(m)}$, dass **in jeder Zeile und in jeder Spalte maximal zwei Elemente ungleich Null stehen.**

- Die Effizienz eines schnellen Algorithmus lässt sich im Vergleich zu seinen Stabilitätseigenschaften wesentlich einfacher vorhersagen (vgl. Higham [24, (2002)]). Letztere erfordert eine sorgfältige Analyse (vgl. Abbildung T.1).

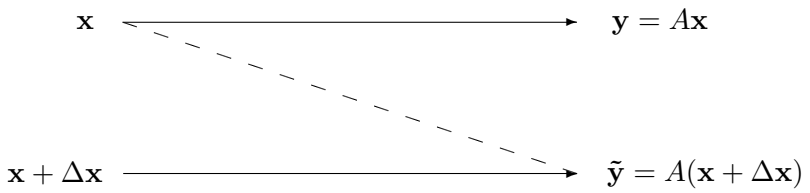


Abbildung T.1: Modellierung des Vorwärts- und Rückwärtsfehlers.

- In **Kapitel 1** werden verschiedene bekannte Algorithmen zur effizienten Realisierung diskreter Kosinus- und Sinustransformationen vom Typ II – IV vorgestellt.
  - Bei einer solchen Transformation wird für einen Vektor $\mathbf{x}$ fester Länge $n \in \mathbb{N}$ jeweils eine Matrix-Vektor-Multiplikation $A\mathbf{x}$ mit einer quadratischen vollbesetzten, jedoch orthogonalen Matrix, z.B.,

$$C_n^{\mathrm{II}} \;:=\; \sqrt{\frac{2}{n}} \left(\varepsilon_n(j) \cos \frac{j(2k+1)\pi}{2n} \right)_{j,k=0}^{n-1}, \qquad C_n^{\mathrm{IV}} \;:=\; \sqrt{\frac{2}{n}} \left(\cos \frac{(2j+1)(2k+1)\pi}{4n} \right)_{j,k=0}^{n-1},$$

    ($\varepsilon_n(0) = \frac{\sqrt{2}}{2}$ und $\varepsilon_n(j) = 1$ sonst) ausgeführt. Insbesondere interessiert der Fall $n = 2^t$.
  - Zu jeder dieser vollbesetzten Kosinus- und Sinusmatrizen werden jeweils Matrixfaktorisierungen angegeben, welche als Faktoren zwei dünnbesetzte orthogonale Matrizen und eine Blockdiagonalmatrix besitzen. Darüber hinaus enthält die jeweilige Blockdiagonalmatrix als Blöcke wiederum Kosinus- bzw. Sinusmatrizen der Größe $\frac{n}{2} \times \frac{n}{2}$ – jedoch nicht notwendigerweise vom gleichen Typ.
  - Induktiv lassen sich somit Faktorisierungen herleiten, die nur noch aus dünnbesetzten orthogonalen Matrizen bestehen, wobei sich jeder Faktor durch Permutationen und zeilen- bzw. spaltenweise Vorzeichenskalierungen (nette Matrizen) in eine Blockdiagonalmatrix mit Blöcken der Gestalt

$$Q_2(\varphi) \;:=\; \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix} \qquad \left(\varphi \in \left[0, \tfrac{\pi}{4}\right]\right) \tag{T.2}$$

    überführen lässt.
  - Die Multiplikation eines Vektors $\mathbf{x} \in \mathbb{R}^2$ mit einer Matrix $Q_2(\varphi)$ bedeutet geometrisch genau eine Drehung um den Winkel $-\varphi$, so dass es sich bei $Q_2(\varphi)$ wegen $Q_2(\varphi)^{-1} = Q_2(-\varphi) = Q_2(\varphi)^{\mathrm{T}}$ um eine orthogonale Matrix handelt.
  - Aus den hergeleiteten Faktorisierungen werden schnelle Algorithmen entworfen. Dabei bestechen insbesondere die *rekursiven Algorithmen* 1.19 – 1.22 durch ihre einfache Struktur.

- In **Kapitel 2** werden sowohl deterministische als auch stochastische Modelle für die Gleitkomma- und Festkomma-Arithmetik untersucht bzw. weiterentwickelt.
  - Ausgehend von dem auf Wilkinson [67, (1963)] zurückgehenden Modell zur Gleitkomma-Arithmetik werden jeweils Abschätzungen für den entstehenden relativen Rundungsfehler bei elementaren Operationen (Addition, Subtraktion und Multiplikation) und ebenso beim Skalarprodukt von Vektoren sowie bei der Matrix-Vektor-Multiplikation hergeleitet, welche sich für Drehmatrizen (T.2) verschärfen.
  - Im Rahmen eines stochastischen Modells und unter zusätzlichen Annahmen werden Vorhersagen für die euklidische Norm des Vektors der bei Matrix-Vektor-Multiplikationen mit den Drehmatrizen (T.2) auftretenden relativen Rundungsfehler bereitgestellt.
    * Die zunächst geforderte Annahme der stochastischen Unabhängigkeit an die Eingangsdaten erweist sich als ungeeignet, da sie im Allgemeinen bereits nach einer Matrix-Vektor-Multiplikation nicht mehr gewährleistet werden kann.
    * Aufgrund der speziellen Blockdiagonalgestalt gelingt nach geringfügiger Modifikation die Übertragung der Idee von Zeuner [75, (2005)], welcher ein stochastisches Modell für die Multiplikation komplexer Zahlen entwickelt hat, auf den Fall von Drehmatrizen (vgl. Satz 2.22). **Neu ist hierbei, dass die Eingangsdaten korreliert sein können.**
    * Unter Verwendung der sinnvollen Annahmen $G_1^2 + G_2^2 = 1 + \varepsilon_{\|\cdot\|_2^2}$, $|\varepsilon_{\|\cdot\|_2^2}| \leq 2u + u^2$, und $\mathbb{E}\big(\varepsilon_{\|\cdot\|_2^2}\big) = 0$ bzw. $\varepsilon_{\|\cdot\|_2^2} \leq 0$ sind die Hauptergebnisse in Satz 2.26 zusammengefasst.
  - Das bereits auf J. von Neumann und Goldstine [41, (1947)] zurückgehende Modell zur Festkomma-Arithmetik, welches die Vorzeichenbetragsdarstellung (sign-magnitude representation) verwendet, wird betrachtet.
    * Im Vergleich zur Gleitkomma-Arithmetik ist die Addition innerhalb der Festkomma-Arithmetik fehlerfrei ausführbar, solange kein Überlauf auftritt.
    * Obere Schranken für den bei elementaren Operationen, Skalarprodukt und Matrix-Vektor-Multiplikation entstehenden absoluten Rundungsfehler werden ausgehend von

$$\boxed{\begin{array}{lcll} \mathrm{fix}(m+\tilde{m}) & = & m+\tilde{m}\ , & \text{falls} \quad |m+\tilde{m}| \leq 1; \\ m \times \tilde{m} & = & m\tilde{m} - \delta_{m\tilde{m}}\ , & (|\delta_{m\tilde{m}}| \leq u, \quad m\tilde{m}\,\delta_{m\tilde{m}} \geq 0, \quad |m \times \tilde{m}| \leq |m\tilde{m}|)\ . \end{array}}$$

      für $m, \tilde{m} \in \mathbb{M}_q$ hergeleitet. Dabei ist $\mathrm{fix} : [-1,1] \to \mathbb{M}_q$ das Abschneiden und

$$\mathbb{M}_q := \left\{ m \in [-1,1] : m = \mathrm{sign}(m) \sum_{k=1}^{q} \mu_k 2^{-k}, \quad \mu_k \in \{0,1\} \right\} \cup \{-1,1\}$$

      für ein fest gewähltes $q \in \mathbb{N}$ sowie $u := 2^{-q}$ die Rechnergenauigkeit.
    * Skalarprodukte werden ebenso in doppelter Genauigkeit untersucht.
    * Satz 2.31 liefert die wichtigste Aussage von Abschnitt 2.3.
  - Schließlich untersuchen wir ein stochastisches Modell zur Festkomma-Arithmetik:
    * **Auf die Unkorreliertheit der Eingangsdaten kann dabei verzichtet werden.**
    * Die in dieser allgemeinen Form völlig neuen Hauptergebnisse, welche die besondere Gestalt der betrachteten Matrizen essentiell verwendet, finden sich in den Sätzen 2.37, 2.38 und 2.39.

- In **Kapitel 3** untersuchen wir verschiedene Situationen für die Gleitkomma-Arithmetik, bei der Rundungsfehler auftreten können. Eingangsfehler der Daten werden zusätzlich berücksichtigt. Zu den Algorithmen 1.15 – 1.22 werden von der Transformationslänge n abhängige Stabilitätskonstanten für den ungünstigsten bzw. für den durchschnittlich auftretenden Fall hergeleitet.

– Für die deterministische Rundungsfehleranalyse sind die auf den neuen nicht verbesserbaren Ungleichungen

$$(2|ac| + 2|bd| + |ac - bd|)^2 + (2|ad| + 2|bc| + |ad + bc|)^2 \quad \leq \quad \frac{27}{2}(a^2 + b^2)(c^2 + d^2)$$

bzw.

$$(|as| + |bs| + |as - bs|)^2 + (|as| + |bs| + |as + bs|)^2 \quad \leq \quad 2\big(3 + \sqrt{5}\big)\big(a^2 + b^2\big)s^2 \ ,$$

$(a, b, c, d, s \in \mathbb{R})$ basierenden Hauptergebnisse in den Sätzen 3.12 und 3.18 zusammengefasst. Letzterer berücksichtigt, dass einige der Matrixfaktoren höchstens skalierte Butterfly-Matrizen

$$s \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad\qquad (s \in \mathbb{R}, s \neq 0)$$

enthalten, so dass hier pro Block zwei Rundungsfehler erzeugende Multiplikationen eingespart werden können.

– Für die auf den Annahmen aus Modell 3.22 basierende stochastische Rundungsfehleranalyse sind die Hauptergebnisse in Satz 3.23 zusammengefasst. Im Vergleich zu den Arbeiten von Tasche & Zeuner [54, (2000)] bzw. Zeuner [75, (2005)] verzichtet Modell 3.22 darauf, die Eingangsdaten als unkorreliert oder stochastisch unabhängig anzunehmen. Auf diese Weise ist das Anwendungsspektrum erheblich größer.

- In **Kapitel 4** werden die Algorithmen 1.15 – 1.22 auf ihre numerische Stabilität in Festkomma-Arithmetik untersucht.
  – Sowohl in einfacher als auch in doppelter Genauigkeit werden Abschätzungen für den ungünstigsten Fall angegeben. Satz 4.5 fasst dabei die Resultate für eine spezielle Blockdiagonalmatrix und einen Eingangsvektor $\mathbf{x} \in [-1, 1]^n$ mit $\|\mathbf{x}\|_2 \leq 1$ zusammen, aus denen dann die Stabilitätskonstanten für die Algorithmen 1.15 – 1.22 bei verschiedenen Skalierungsarten in den Sätzen 4.12 – 4.14 gewonnen werden.
  – Basierend auf Modell 4.16 findet analog zur Gleitkomma-Arithmetik eine stochastische Rundungsfehleranalyse statt. Als Hauptergebnis sind die entsprechenden durchschnittlichen Stabilitätskonstanten in Satz 4.17 angegeben.

- In **Kapitel 5** werden die theoretischen Schranken aus den Kapiteln 3 und 4 auf ihre Qualität überprüft, indem die Algorithmen aus Kapitel 1 für verschiedene Transformationslängen $n = 2^t$ und zufällig erzeugte Testvektoren ausgeführt werden. Dazu sind die Algorithmen 1.15 – 1.22 jeweils in MATLAB implementiert worden.

Die wichtigsten neuen Ergebnisse lassen sich wie folgt zusammenfassen:

(1) Die teilweise bekannten (vgl. Plonka & Tasche [43, (2005)]) Matrixfaktorisierungen für Kosinus- und Sinusmatrizen vom Typ II – IV werden systematisch ausgearbeitet und führen zu einer Klasse von einfachen, rekursiven Algorithmen für die DCT und DST vom Typ II – IV, falls die Transformationslänge n eine Zweierpotenz ist.

(2) Es werden deterministische und stochastische Modelle zur Gleit- und Festkomma-Arithmetik bereitgestellt. Für Matrix-Vektor-Multiplikationen erfolgt eine deterministische und stochastische Rundungsfehleranalyse. Besitzen die Matrizen spezielle Blockdiagonalgestalt, so bleiben die Ergebnisse auch im Fall korrelierter Eingangsdaten gültig.

(3) Die entwickelten Methoden der deterministischen und stochastischen Rundungsfehleranalysen werden auf die schnellen DCT- und DST-Algorithmen 1.15 – 1.22 angewandt. Für die numerische Stabilität dieser Algorithmen werden im Fall der Gleitkomma-Arithmetik verbesserte Konstanten angegeben und im Fall der Festkomma-Arithmetik erstmals Konstanten bestimmt. Die numerischen Testrechnungen illustrieren die theoretischen Ergebnisse und ermöglichen einen sehr guten Vergleich des Rundungsfehler-Verhaltens in beiden Rechnerarithmetiken.