KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT COMPUTERWETENSCHAPPEN
AFDELING NUMERIEKE ANALYSE EN
TOEGEPASTE WISKUNDE
Celestijnenlaan 200A – B-3001 Leuven

MULTISCALE AND HYBRID METHODS FOR THE SOLUTION OF OSCILLATORY INTEGRAL EQUATIONS

Promotor:
Prof. Dr. ir. S. Vandewalle

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Daan HUYBRECHS

Mei 2006

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT COMPUTERWETENSCHAPPEN
AFDELING NUMERIEKE ANALYSE EN
TOEGEPASTE WISKUNDE
Celestijnenlaan 200A – B-3001 Leuven

MULTISCALE AND HYBRID METHODS FOR THE SOLUTION OF OSCILLATORY INTEGRAL EQUATIONS

Jury:
Prof. Dr. ir. L. Froyen, voorzitter
Prof. Dr. ir. S. Vandewalle, promotor
Prof. Dr. A. Bultheel
Prof. Dr. ir. G. Vandenbosch
Prof. Dr. ir. R. Cools
Prof. Dr. ir. D. Roose
Prof. Dr. A. Iserles
(University of Cambridge)
Prof. Dr. R. Stevenson
(Universiteit Utrecht)

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Daan HUYBRECHS

U.D.C. 519.64

Mei 2006

© Katholieke Universiteit Leuven — Faculteit Ingenieurswetenschappen
Arenbergkasteel, B-3001 Leuven, Belgium

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2006/7515/48

ISBN 90-5682-714-6

Rev. 1.1

Multiscale and hybrid methods for the solution of oscillatory integral equations

Daan Huybrechs
Departement Computerwetenschappen, K.U.Leuven
Celestijnenlaan 200A, B-3001 Leuven, België

Abstract

Waves and oscillatory phenomena abound in many disciplines of science and engineering. Prime examples are electromagnetic and acoustic waves that permeate the atmosphere. In this thesis, we analyse and develop algorithms for the efficient numerical simulation of the scattering of such waves.

Time-harmonic scattering problems are modelled by an integral equation formulation. We consider three multiscale methods for the efficient solution of the resulting oscillatory integral equation: methods based on wavelets, methods based on hierarchical matrices and fast multipole methods. Although the discretisation matrix for integral equations is a dense matrix, each of these methods yields a fast matrix-vector product, where the number of operations scales approximately linearly in the number of unknowns. The solution can then be obtained efficiently in combination with an iterative Krylov subspace solver.

We show that wavelet based methods are not suitable for high frequency problems, where the number of oscillations is large with respect to the size of the scattering obstacle. We quantify the behaviour of the method in the oscillatory setting, and propose an improvement based on wavelet packets. Quadrature techniques are constructed for the efficient implementation of wavelet Galerkin discretisations. Methods based on hierarchical matrices and fast multipole methods are discussed for low frequency and high frequency scattering problems, and their applicability is compared.

Due to their ubiquitous nature in wave phenomena, oscillatory integrals are studied. A new method is proposed for the evaluation of univariate and multivariate oscillatory integrals, based on an extension of the method of steepest descent. Contrary to traditional methods, the accuracy of the new method increases rapidly with increasing frequency of the integrand, and it is shown that its computational cost is very low.

Finally, the insights in the behaviour of oscillatory integrals lead to the formulation of a novel method for highly oscillatory integral equations. We propose a hybrid method that combines asymptotic estimates of the solution with a classical boundary element discretisation. The hybrid asymptotic method requires a number of operations that is fixed with respect to the frequency. Results are given for the case of smooth and convex scattering obstacles. We show that the discretisation matrix in this case is small and highly sparse.

Multiscale and hybrid methods for the solution of oscillatory integral equations

Daan Huybrechs
Departement Computerwetenschappen, K.U.Leuven
Celestijnenlaan 200A, B-3001 Leuven, België

Samenvatting

Golven en golfverschijnselen komen voor in verchillende disciplines van de wetenschap en in vele ingenieurstoepassingen. De voornaamste voorbeelden zijn electromagnetische en akoestische golven die ons overal omgeven. In deze doctoraatsthesis analyzeren en ontwikkelen we algoritmes voor de efficiënte numerieke simulatie van de weerkaatsing van dergelijke golven.

Tijdsharmonische verstrooiïngsproblemen worden gemodelleerd met een wiskundige formulering in de vorm van een integraalvergelijking. We bekijken drie multischaalmethodes voor het oplossen van de resulterende oscillerende integraalvergelijking: methodes gebaseerd op wavelets, methodes gebaseerd op hiërarchische matrices en snelle multipoolmethodes. Hoewel de discretizatiematrix voor integraalvergelijkingen een volle matrix is, maakt elk van die methodes een snel matrix-vectorproduct mogelijk waarbij het aantal bewerkingen bij benadering linear is in het aantal onbekenden. De oplossing kan dan snel gevonden worden in combinatie met een iteratieve Krylov deelruimte oplossingsmethode.

We tonen aan dat waveletgebaseerde methodes niet geschikt zijn voor problemen met hoge frequenties, waarbij het aantal oscillaties groot is ten opzichte van de grootte van het weerkaatsende obstakel. We analyzeren het gedrag van de methode in een sterk oscillerend regime, en stellen een verbetering voor op basis van wavelet pakketten. Kwadratuurtechnieken worden

opgesteld voor de efficiënte implementatie van wavelet-Galerkin discretizaties. Methodes gebaseerd op hiërarchische matrices en snelle multipoolmethodes worden onderzocht voor lage- en hoge frequentieregimes, en de toepasbaarheid van de methodes wordt vergeleken.

Omwille van hun verschijnen in de beschrijving van tal van golfproblemen, worden vervolgens integralen bestudeerd met een sterk oscillerende integrand. Er wordt een nieuwe methode voorgesteld voor de evaluatie van dergelijke integralen in één of meerdere dimensies, gebaseerd op een uitbreiding van de methode van de steilste helling. In tegenstelling tot traditionele methodes verhoogt de nauwkeurigheid van de nieuwe methode sterk bij stijgende frequenties, en we tonen aan dat de berekeningstijd klein blijft.

Tenslotte worden de verworven inzichten in het gedrag van oscillerende integralen aangewend in een originele oplossingsmethode voor sterk oscillerende integraalvergelijkingen. We stellen een hybride methode voor die een asymptotische schatting van de oplossing combineert met een klassieke randelementendiscretizatie. Resultaten worden gegeven voor het geval van convexe obstakels met een zachtverlopende rand. We tonen aan dat de discretisatiematrix in dit geval klein is en in hoge mate ijl.

Preface

This thesis is the result of four years of research in numerical analysis, of analysing and developing algorithms for the simulation of engineering problems. You could say that I have spent these years taking things apart to see what makes them work. This is much like I did when I was young, although, admittedly, wavelets are somewhat less tangible than clock radio's.

Several people have helped me to get to this point, and I'd like to mention a few that have made a difference. Foremost, I would like to thank Stefan Vandewalle, my supervisor. He has originally given me the chance to pursue the study of integral equations. Since then, he has actively supported me in various ways, and introduced me to many people. Also, it is no coincidence that almost every referee report that we received on one of our papers stated that the paper was well written. This is certainly due to Stefan and his strong desire for clarity and perfection.

A number of people from abroad have also played an important role. I thank Stephen Langdon and Simon Chandler-Wilde (University of Reading) for inviting me to Reading and for the great experience I've had there. Likewise, I thank Arieh Iserles (University of Cambridge) for a wonderful visit to Cambridge, but also for his support in my research and for the opportunity to participate in the HOP programme at the Isaac Newton Institute in Cambridge next year. I am very much looking forward to doing research in the United Kingdom.

I thank all the members of the jury for agreeing to take part in my defence, and especially my assessors Adhemar Bultheel and Guy Vandenbosch for their careful reading of the text and their suggestions for improvements.

I have learned many aspects of numerical integration from Ronald Cools. There are numerous other people at the department who had an influence on this thesis, in particular the members of the TWR research group and the other groups, with whom I've had many interesting discussions. I'd like to thank Bert, Dominik and Bart for making our office an enjoyable place to work, and Pieter for the many snooker games. It's funny to notice how the level of our game accurately reflected the deadlines we faced. Judging from past experiences, we should be breaking some records after our PhD.

Finally, I would like to thank my family and friends for their support over the years, especially my fiancée Elise. The times I was faced with a choice of different research paths to pursue, her advice has invariably been to work harder and do it all. She then made sure that I could do so. It is impossible to say how much her motivation and encouragement have helped me, except that it must be a lot.

Four years ago, in the preface of my master's thesis, I wrote that research is an ongoing process without end. It's an obvious truth, that I can easily repeat here. It is perhaps a telling fact that the list of suggestions for future research at the end of this thesis is longer than the list of achievements. One can not possibly hope to find all the answers. Instead, in the years to come, I hope to find many more interesting questions.

Daan Huybrechs
May 2006

Acknowledgement

The research in this thesis was financially supported by the Institute for Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Nederlandse samenvatting

Meerschalige en hybride oplossingsmethodes voor oscillatorische integraalvergelijkingen

Inhoudsopgave

1 Inleiding **xii**
  1.1 Integraalvergelijkingen xii
  1.2 Randelementenmethode xiii
  1.3 Zwak en sterk oscillerende regimes xiii
  1.4 Overzicht van de thesis xiv

2 Waveletgebaseerde methodes **xiv**
  2.1 Een snel matrix-vector product xiv
  2.2 Afhankelijkheid van het golfgetal xv
  2.3 Verbeterde compressie met waveletpakketten xv
  2.4 Nauwkeurige kwadratuurformules xvi

3 Snelle multipoolmethodes **xvii**

4 Hiërarchische matrix methodes **xviii**

5 Oscillerende integralen **xviii**
  5.1 Modelvorm en eigenschappen xviii
  5.2 De numerieke methode van de steilste helling xix

5.3 Hogerdimensionale integralen xx

6 Een asymptotische hybride methode xx
6.1 Formulering xx
6.2 Een ijle discretisatiematrix xxi

7 Slotbemerkingen xxi
7.1 Eigen bijdragen xxi
7.2 Conclusies xxii
7.3 Suggesties voor verder onderzoek xxii

1 Inleiding

1.1 Integraalvergelijkingen

Het doel van deze thesis is de analyse en ontwikkeling van snelle oplossingsmethodes voor integraalvergelijkingen met een oscillerend karakter. Deze vergelijkingen duiken op bij de wiskundige modellering van de voortplanting en de weerkaatsing van, bijvoorbeeld, electromagnetische of akoestische golven. We behandelen zogenaamde *integraalvergelijkingen van de eerste soort*

$$(Av)(x) = \int_\Gamma G(x,y)v(y)\,\mathrm{d}s_y = f(x), \qquad x \in \Gamma, \tag{1}$$

en *integraalvergelijkingen van de tweede soort*

$$\lambda v(x) + \int_\Gamma G(x,y)v(y)\,\mathrm{d}s_y = f(x), \qquad x \in \Gamma, \lambda \neq 0. \tag{2}$$

Hierin is Γ de rand van een obstakel waardoor een invallende golf verstrooid wordt. De invallende golf wordt beschreven door de randvoorwaarde $f(x)$. De onbekende in de vergelijkingen is de *dichtheidsfunctie* $v(x)$. De functie $G(x,y)$ noemt men de functie van Green, of ook de kernfunctie van de integraaloperator A. Voor twee-dimensionale verstrooiïngsproblemen met tijdsharmonische golven is de kernfunctie gekend, $G(x,y) = \frac{i}{4}H_0^{(1)}(k|x-y|)$. De integraalvergelijking (1) komt in dat geval wiskundig overeen met het oplossen van de Helmholtzvergelijking

$$\Delta u + k^2 u = 0, \tag{3}$$

met de randvoorwaarde $u(x) = f(x)$ op Γ.

1.2 Randelementenmethode

De randelementenmethode is een eindige-elementenmethode waarvan de basisfuncties ϕ_i, $i = 1, \ldots, N$, gedefinieerd zijn op de rand Γ van het obstakel. De basisfuncties worden daarom ook *randelementen* genoemd. De Galerkindiscretisatie van integraalvergelijking (1) leidt tot het lineaire stelsel $Mx = b$. De elementen van de discretisatiematrix M en van het rechterlid b worden gegeven door

$$M_{ij} = \langle A\phi_j, \phi_i \rangle \quad \text{en} \quad b_i = \langle f, \phi_i \rangle. \tag{4}$$

Hierin stelt $\langle \cdot, \cdot \rangle$ het L_2 inwendig product voor. De matrixelementen worden expliciet gegeven door een dubbelintegraal met de vorm (2.61).

De discretisatiematrix M is een volle matrix. Het oplossen van het stelsel $Mx = b$ met directe methodes vereist daarom $O(N^3)$ bewerkingen. De snelle methodes die we bestuderen leiden tot een snel matrix-vectorproduct met een complexiteit van $O(N)$ of $O(N \log^p N)$ bewerkingen, $p > 0$. Het snelle matrix-vectorproduct kan aangewend worden in combinatie met een iteratieve Krylov-deelruimte oplossingsmethode, zoals GMRES, wat leidt tot een efficiënte oplossingsmethode voor de integraalvergelijking. Het conditiegetal van de discretisatiematrix hangt af van de *orde* r van de operator A. Voor een integraalvergelijking met een ééndimensionale rand Γ geldt dat

$$\kappa(M) = O(N^{|r|}). \tag{5}$$

De operator die overeenkomt met het Helmholtzprobleem heeft orde $r = -1$; het conditiegetal stijgt dus linear met het aantal basisfuncties.

1.3 Zwak en sterk oscillerende regimes

De frequentie van het golfprobleem wordt uitgedrukt door het golfgetal k, dat zowel in de Helmholtzvergelijking (3) als in de kernfunctie verschijnt. De grootte van het golfgetal is slechts relatief. Van belang om te spreken van een hoogfrequent problem is dat het getal groot is in vergelijking met de omvang van de rand Γ.

We maken in de analyse van de complexiteit van snelle oplossingsmethodes in functie van het aantal basisfuncties N een onderscheid tussen twee oscillerende regimes. In het *zwak oscillerende regime* blijft het golfgetal k constant, terwijl N stijgt. De oplossing wordt daardoor nauwkeuriger berekend. In het *sterk oscillerende regime* stijgt het golfgetal evenredig met N. De nauwkeurigheid van de oplossing blijft daarbij ongeveer dezelfde, maar de frequentie van het probleem stijgt.

1.4 Overzicht van de thesis

We bespreken drie verschillende multischaalmethodes voor het oplossen van integraalvergelijkingen. We starten in Hoofdstuk 2 van deze samenvatting met methodes gebaseerd op wavelets. Eerst wordt de afhankelijkheid van het golfgetal bestudeerd. De analyse toont aan dat waveletgebaseerde methodes enkel efficiënt werken in het zwak oscillerend regime. Een verbetering voor het sterk oscillerend regime wordt voorgesteld met behulp van waveletpakketten. Er worden ook kwadratuurformules voorgesteld om integralen met wavelets in de integrand nauwkeurig te benaderen. Vervolgens bespreken we snelle multipoolmethodes en methodes gebaseerd op hiërarchische matrices in Hoofdstuk 3 en Hoofdstuk 4. Beide multischaalmethodes vertonen een asymptotische complexiteit van $O(N \log N)$ in het sterk oscillerende regime. De principes waarop de methodes gebaseerd zijn worden geïllustreerd met numerieke experimenten.

In Hoofdstuk 5 stellen we enkele methodes voor om de waarde van sterk oscillerende bepaalde integralen zeer nauwkeurig te benaderen. Integralen met een sterk oscillerende integrand duiken op in verschillende toepassingen, maar ook de integraalvergelijkingen (1)-(2) zelf bevatten een sterk oscillerende integraal indien het golfgetal groot is. De methodes hebben de eigenschap dat ze nauwkeuriger worden bij hogere frequenties. We bespreken ééndimensionale en hogerdimensionale integralen. In Hoofdstuk 6 worden de methodes toegepast op de integraalvergelijking zelf. Dit leidt tot een methode die slechts een constant aantal bewerkingen vereist voor stijgende waarden van het golfgetal. De methode wordt een *hybride* methode genoemd omdat ze een klassieke eindige-elementendiscretisatie combineert met een asymptotische methode.

2 Waveletgebaseerde methodes

2.1 Een snel matrix-vector product

De waveletmethode is een eindige-elementenmethode waarbij waveletfuncties ψ_{jk} gebruikt worden als basisfuncties. Wavelets worden gedefinieerd op verschillende schalen j en posities k in termen van de *moederwavelet* ψ,

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k). \tag{6}$$

De moederwavelet ψ en de zogenaamde *schaalfunctie* ϕ worden gekarakteriseerd door de tweeschaalvergelijkingen

$$\phi(t) = \sum_{k\in\mathbb{Z}} h_k \phi(2t-k), \quad \text{en} \quad \psi(t) = \sum_{k\in\mathbb{Z}} g_k \phi(2t-k). \tag{7}$$

Wavelets hebben een aantal nulmomenten $\tilde{d}$,

$$\int_{-\infty}^{\infty} \psi(x)x^i \, \mathrm{d}x = 0, \qquad i = 0, \ldots, \tilde{d} - 1, \tag{8}$$

waardoor zij geschikt zijn voor het benaderen van functies. Deze eigenschap zorgt er namelijk voor dat de matrixelementen (4) in de waveletbasis,

$$W_{(j,k),(j',k')} = \langle A\psi_{j'k'}, \psi_{jk} \rangle, \tag{9}$$

veelal klein zijn. De discretisatiematrix W in de waveletbasis kan bijgevolg sterk gecomprimeerd worden tot een ijle matrix. Men toont aan dat het aantal significante elementen grootteorde $O(N)$ heeft, wat rechtstreeks aanleiding geeft tot een matrix-vectorproduct in $O(N)$ bewerkingen. Daarnaast kan het conditiegetal van de matrix uniform begrensd worden in N met een eenvoudige diagonale preconditionering.

2.2 Afhankelijkheid van het golfgetal

De asymptotische lineaire complexiteit $O(N)$ geldt enkel in het zwak oscillerende regime. We analyseerden ook het gedrag van de waveletmethode voor het sterk oscillerende regime, waarbij het golfgetal k evenredig stijgt met N. Het resultaat wordt samengevat in de volgende stelling.

Stelling 1 (Theorem 3.5.6). *Het aantal significante elementen in de gecomprimeerde discretisatiematrix W stijgt asymptotisch lineair in N, met een evenredigheidsconstante die zich gedraagt als $O(k^{1+1/(2\tilde{d}-2)})$.*

De evenredigheidsconstante in de uitdrukking $O(N)$ is ongeveer linear in het golfgetal k. In het sterk oscillerende regime verloopt het aantal significante elementen na compressie dus kwadratisch in N. De compressie gaat uiteindelijk verloren bij stijgende frequenties. De afhankelijkheid van het golfgetal wordt geïllustreerd in Figuur 3.6.

2.3 Verbeterde compressie met waveletpakketten

Naast een controleerbare subdivisie van schaal en positie, veroorzaken wavelets ook een oncontroleerbare subdivisie van het frequentiespectrum. Bij de overgang van een fijne naar een ruwere schaal, wordt bij benadering telkens enkel het lage deel van het frequentiespectrum verder opgedeeld. Er treedt geen compressie meer op indien het frequentiespectrum van een functie voornamelijk in het hogere deel gelegen is.

Dit fenomeen treedt op in het sterk oscillerende regime, omdat het grote golfgetal aanleiding geeft tot hoge frequenties. Om dit te vermijden onderzochten we het gebruik van waveletpakketten. Waveletpakketten w_n worden gedefinieerd in analogie met (7) door

$$w_{2n}(t) = \sqrt{2} \sum_{k \in \mathbb{Z}} h_k w_n(2t - k), \tag{10}$$

$$w_{2n+1}(t) = \sqrt{2} \sum_{k \in \mathbb{Z}} g_k w_n(2t - k). \tag{11}$$

Ze worden verder gedefinieerd op alle schalen en posities door $w_{njk} = 2^{j/2} w_n(2^j t - k)$. De index n is een aanduiding van de frequentie van de bijhorende waveletpakketfunctie.

Er is een groot aantal mogelijkheden om uit de waveletpakketten een volledig stel basisfuncties te kiezen. We vergeleken verschillende methodes om een geschikte basis op te stellen in functie van de waarde van het golfgetal. De beste resultaten werden bekomen door toepassing van het tweedimensionale *beste-basisalgoritme* op de discretisatiematrix. Op basis van numerieke experimenten en een heuristische schatting, concluderen we dat het aantal significante elementen na compressie zich gedraagt als $O(N^{1.4})$. De totale complexiteit van de methode blijft echter hoog, omdat de volle discretisatiematrix aanvankelijk moet berekend worden. De methode wordt wel interessant indien hetzelfde stelsel opgelost wordt voor verschillende randvoorwaarden. Andere basiskeuzes met een lagere complexiteit leidden eveneens tot sterkere compressie in vergelijking met de klassieke waveletmethode. De resultaten worden vergeleken in Figuur 3.9 en Figuur 3.10.

2.4 Nauwkeurige kwadratuurformules

De wavelettransformatie van een functie vereist de evaluatie van een groot aantal integralen van de vorm

$$c_{jk} = \int_{-\infty}^{\infty} f(x)\phi_{jk}(x)\,\mathrm{d}x, \quad \text{of} \quad d_{jk} = \int_{-\infty}^{\infty} f(x)\psi_{jk}(x)\,\mathrm{d}x. \tag{12}$$

De convergentie van kwadatuurformules voor deze integralen hangt af van de regulariteit van de schaalfunctie ϕ. De integratie kan verder bemoeilijkt worden door singulariteiten of discontinuïteiten van de functie f. We onderzoeken kwadratuurformules van de vorm

$$\int_a^b f(x)\phi(x)\,\mathrm{d}x \approx Q[f] = \sum_{i=1}^{r} w_i f(x_i), \tag{13}$$

waarvan de convergentie onafhankelijk is van de regulariteit van ϕ. De gewichten kunnen automatisch berekend worden op basis van de coëfficiënten h_k in (7). Door een geschikte keuze van het integratie-interval $[a, b]$ kunnen discontinuïteiten van f steeds op de rand gelegd worden, zodat ze de convergentie niet negatief beïnvloeden. De kwadratuurregel (13) kan ook uitgebreid worden naar singuliere functies f. Ook in dat geval kunnen de gewichten automatisch en efficiënt berekend worden.

Tenslotte worden de kwadratuurpunten x_i zodanig gekozen dat zij op een regelmatig rooster liggen. De functie-evaluaties $f(x_i)$ kunnen dan herbruikt worden voor de evaluatie van integralen met naburige basisfuncties. In vele gevallen is de evaluatie van de kernfunctie tijdrovend in vergelijking met eenvoudige bewerkingen zoals optellen en het vermenigvuldigen met gewichten. De tijdsbesparing door het herbruiken van de functie-evaluaties is dan bijzonder groot.

3 Snelle multipoolmethodes

Snelle multipoolmethodes leiden tot een snel matrix-vectorproduct op een andere manier. Er worden separabele expansies van de kernfunctie $G(x, y)$ opgesteld, met de algemene vorm

$$G(x, y) \approx \sum_{l=1}^{L} u_l(x) v_l(y). \tag{14}$$

Deze *multipoolexpansie* is slechts geldig in bepaalde gebieden $x \in \Omega_x$ en $y \in \Omega_y$. De integraal van de integraaloperator A over het gebied Ω_y kan geschreven worden als

$$\int_{\Omega_y} G(x, y) q(y) \, \mathrm{d}s_y \approx \sum_{l=1}^{L} u_l(x) \int_{\Omega_y} v_l q(y) \, \mathrm{d}s_y. \tag{15}$$

Aangezien de integralen in het rechterlid van (15) nu onafhankelijk zijn van x, kunnen zij berekend worden als tussenresultaat, en herbruikt worden voor verschillende waarden van x. Het efficiënt opstellen van expansies van de vorm (14) in gebieden die heel de rand Γ bestrijken, en het delen van de tussenresultaten, maken een matrix-vector product mogelijk in $O(N)$ bewerkingen voor een vaste discretisatiefout ϵ.

In tegenstelling tot de waveletmethode, kan dezelfde techniek gebruikt worden in het sterk oscillerende regime. Het nodige aantal termen in de expansie stijgt echter ongeveer lineair met het golfgetal. Een matrix-vectorproduct met complexiteit $O(N \log N)$ blijft mogelijk indien de expansies hiërarchisch opgesteld worden. De multipoolcoëfficiënten voor een bepaald gebied worden dan berekend uit de coëfficiënten van de deelgebieden. De convergentie van de expansie die gebruikt wordt voor de kernfunctie van het Helmholtzprobleem wordt geïllustreerd in Figuur 4.4.

4 Hiërarchische matrix methodes

Hiërarchische matrices zijn matrices met een blokstructuur, waarvan de deelblokken bestaan uit matrices van lage rang. De vermenigvuldiging met een matrix van lage rang kan efficiënt uitgevoerd worden. Een volle matrix van rang L kan geschreven worden als $M = AB^T$, met $M \in \mathbb{C}^{N\times N}$, en $A, B \in \mathbb{C}^{N\times L}$. De vermenigvuldiging met een vector $x \in \mathbb{C}^N$,

$$Mx = AB^Tx = A(B^Tx), \tag{16}$$

vergt slechts $O(NL)$ bewerkingen, in plaats van $O(N^2)$ voor een matrix van volle rang. De vermenigvuldiging met een hiërarchische matrix wordt versneld door het toepassen van (16) voor elke deelmatrix.

De lage-rangbenadering van een deelmatrix van de discretisatiematrix wordt gevonden door een separabele expansie van de kernfunctie te gebruiken, gelijkaardig aan (14). In dat opzicht is de hiërarchische matrix techniek vergelijkbaar met de snelle multipoolmethode. De aanpak is echter eerder algebraïsch en algemeen, terwijl de snelle multipoolmethode typisch opgesteld wordt voor een specifieke integraalvergelijking met bijhorende kernfunctie. Een algemene separabele expansie kan bijvoorbeeld bereikt worden door veelterminterpolatie.

5 Oscillerende integralen

5.1 Modelvorm en eigenschappen

Een weerkerend probleem in de numerieke simulatie van golfverschijnselen is de evaluatie van oscillerende integralen. Als modelvorm beschouwen we de integraal

$$I = \int_a^b f(x)e^{i\omega g(x)}\,\mathrm{d}x. \tag{17}$$

Hierin zijn f en g zachtverlopende functies, respectievelijk de *amplitude* en de *oscillator* van (17) genoemd. De parameter ω bepaalt de frequentie van de oscillerende integrand. De waarde van I wordt bepaald door het gedrag van f and g in de buurt van de eindpunten a en b, en van de *stationaire punten*. Deze punten worden gevonden als oplossing van de vergelijking

$$g'(\xi) = 0, \qquad \xi \in [a, b]. \tag{18}$$

Men zegt dat een stationair punt ξ orde r heeft indien $g^{(i)}(\xi) = 0$, $i = 1, \ldots, r$. Hun belang volgt uit het feit dat de oscillerende factor in (17) rond een stationar punt lokaal quasi constant is. Het deel van de integraal rond een stationair punt heeft daardoor een belangrijk bijdrage tot de waarde van I. In de andere delen van het interval $[a, b]$ heffen de oscillaties elkaar in toenemende mate op bij stijgende waarden van ω.

De eigenschappen van de integraal blijken uit de asymptotische expansie

$$I \sim \sum_{i=1}^{\infty} a_i \omega^{-n_i}, \qquad \forall i : n_i > 0. \tag{19}$$

De coëfficiënten a_i zijn volledig bepaald door een eindig aantal functiewaarden en afgeleiden van f en g in de punten a, b en alle stationaire punten ξ. Deze eigenschap kan numeriek benut worden om nauwkeurige benaderingen voor I te vinden die sterk verbeteren met toenemende frequentie.

5.2 De numerieke methode van de steilste helling

De klassieke methode van de steilste helling leidt tot een asymptotische reeks van de vorm (19). We stellen een efficiënte numerieke implementatie van de methode van de steilste helling voor. De methode is gebaseerd op de volgende observaties:

- de oscillerende factor $e^{i\omega g(x)}$ daalt exponentieel voor complexe waarden van $g(x)$ met een groeiend positief imaginair deel;
- de oscillerende factor $e^{i\omega g(x)}$ oscilleert *niet* voor waarden van $g(x)$ met een constant reëel deel.

Op basis van de stelling van Cauchy mag het integratiepad $[a, b]$ verlegd worden in het complexe vlak zonder de waarde van de integraal te veranderen, indien f en g analytische functies zijn. We kiezen vanop het reële punt x het pad met parameterisatie $h_x(p)$ dat voldoet aan de vergelijking

$$g(h_x(p)) = g(x) + ip, \qquad p > 0. \tag{20}$$

Dit is het pad van de steilste helling. Deze keuze leidt, in de afwezigheid van stationaire punten, tot de decompositie $I = F(a) - F(b)$, met

$$\begin{aligned} F(x) &= \int_0^{\infty} f(h_x(p)) e^{i\omega g(h_x(p))} h_x'(p) \, \mathrm{d}p \\ &= e^{i\omega g(x)} \int_0^{\infty} f(h_x(p)) h_x'(p) e^{-\omega p} \, \mathrm{d}p. \end{aligned}$$

De integrand van $F(x)$ oscilleert niet en daalt exponentieel snel. De evaluatie van $F(x)$ via Gauss-Laguerre kwadratuur met n punten leidt tot een fout die zich gedraagt als $O(\omega^{-2n})$. De convergentie in functie van stijgende frequentie is bijzonder hoog.

We bekijken enkele uitbreidingen. Het pad van de steilste helling kan efficiënt bepaald worden met een iteratief algoritme. Een stationair punt geeft aanleiding tot twee bijkomende contributies, $F_1(\xi)$ en $F_2(\xi)$, die elk afzonderlijk met dezelfde hoge nauwkeurigheid kunnen bepaald worden. Tenslotte kunnen de resultaten uitgebreid worden naar functies die niet analytisch zijn, door de afgeleiden van f en g in de kritieke punten te interpoleren.

5.3 Hogerdimensionale integralen

De numerieke methode van de steilste helling kan uitgebreid worden naar hogerdimensionale integralen. De modelintegraal heeft de vorm

$$I_n = \int_S f(\mathbf{x}) e^{i\omega g(\mathbf{x})} \, \mathrm{d}\mathbf{x}. \tag{21}$$

De waarde van I_n wordt opnieuw bepaald door een aantal speciale punten. Dit zijn vooreerst de hoekpunten van het integratiedomein S, als tegenhanger van de eindpunten a en b in het ééndimensionale geval. Verder zijn er de *kritieke punten* van de oscillator g. Dat zijn punten waar de gradiënt van g nul wordt,

$$\nabla g(\xi) = 0, \qquad \xi \in S. \tag{22}$$

Tenslotte zijn er *resonantiepunten*. Dat zijn punten waar de gradiënt van de oscillator orthogonaal staat op de rand van het integratiedomein, $\nabla g \perp \partial S$.

De integraal wordt behandeld door herhaalde enkelvoudige integratie. We tonen aan dat resonantiepunten overeenkomen met stationaire punten van een lagerdimensionale integraal. De kritieke punten ξ zijn een stationair punt in elk van de integratieveranderlijken. Efficiënte cubatuurformules worden opgesteld voor enkele voorbeeldintegralen, zoals de Fouriertransformatie van een functie die gedefinieerd is op de driedimensionale eenheidsbol.

6 Een asymptotische hybride methode

6.1 Formulering

De integraal Av in integraalvergelijking (1) is sterk oscillerend wanneer het golfgetal k groot is. Aangezien de rand Γ van een gesloten obstakel periodiek is, zijn er geen bijdragen van eindpunten. De waarde van Av wordt volledig bepaald door het gedrag van de integrand nabij de stationaire punten van de oscillator. Indien de oscillator gekend is, kan de integraaloperator bijgevolg snel geëvalueerd worden met de numerieke methode van de steilste helling, en kan de integraalvergelijking zelf ook snel opgelost worden.

We veronderstellen een invallende golf van de vorm $u^i(x) = u_s(x)e^{ikg(x)}$, met gekende zachtverlopende functies u_s en g. Het is geweten dat, voor een convex obstakel, de dichtheidsfunctie zich gedraagt als

$$q(x) = q_s(x)e^{ikg(x)}. \tag{23}$$

Met andere woorden, de oplossing van de integraalvergelijking heeft hetzelfde oscillerende gedrag als de invallende golf. Met deze kennis kan de integraalvergelijking $Av = f$ geschreven worden in de algemene vorm

$$\int_0^1 H(x, y; k)e^{ik\tilde{g}(x,y)}q_s(y; k)e^{ikg(y)}\,\mathrm{d}y = f(x). \tag{24}$$

De oscillator $\tilde{g}(x, y)$ is afkomstig van het gekende oscillerende gedrag van de kernfunctie $G(x, y)$. Functie $q_s(y; k)$ is een ongekende, zachtverlopende functie die afhankelijk is van de waarde van k. Functie $H(x, y; k)$ is een gekende, zachtverlopende functie. Integraal (24) is een oscillerende integraal die zeer vergelijkbaar is met de modelvorm (17).

6.2 Een ijle discretisatiematrix

De zachtverlopende onbekende functie $q_s(y)$ wordt geschreven in een basis van kubische spline-functies,

$$q_s(y) = \sum_{i=1}^{N} c_i\phi_i(x). \tag{25}$$

Collocatie van vergelijking (24) leidt tot een oscillerende integraal voor elk collocatiepunt x_i. De waarde van de integraal wordt bepaald door het gedrag van $q_s(y)$ rond de stationaire punten van de oscillator $g(x) = g_1(x) + g^i(x)$. De collocatie-integraal wordt dus enkel bepaald door de coëfficiënten c_j bij basisfuncties c_j waarvan de drager overlapt met een stationair punt. Een rechtstreeks gevolg is dat de discretisatiematrix een ijle matrix wordt. Deze matrix wordt geïllustreerd in Figuur 7.4. De matrix is klein en kan opgesteld worden met een aantal bewerkingen dat onafhankelijk is van de waarde van het golfgetal. Numerieke experimenten illustreren dat de oplossing van het stelsel nauwkeuriger wordt bij stijgende frequenties in Figuur 7.8.

7 Slotbemerkingen

7.1 Eigen bijdragen

We vermelden eerst de belangrijkste eigen bijdragen van deze thesis tot het onderzoeksdomein:

- de analyse van de waveletmethode in het sterk oscillerende regime;
- verbeterde matrixcompressie in het sterk oscillerende regime met een nieuwe methode op basis van waveletpakketten;
- de constructie van efficiënte kwadratuurformules voor integralen met wavelets in de integrand;
- een zeer nauwkeurige numerieke methode om sterk oscillerende integralen te evalueren;
- de uitbreiding van deze methode naar algemene hogerdimensionale integralen;
- de ontwikkeling van een hybride numeriek-asymptotische randelementenmethode voor hoogfrequente integraalvergelijkingen.

7.2 Conclusies

Elke multischaalbenadering die in deze thesis onderzocht werd, leidt tot een robuuste en snelle oplossingsmethode in het zwak oscillerende regime. De waveletmethode is de enige die theoretisch (asymptotisch) optimaal is. De snelle multipoolmethode en methodes gebaseerd op hiërarchische matrices vereisen een aparte preconditionering indien de orde van de operator niet nul is. Daartegenover staat dat het gebruik van separabele expansies robuuster kan zijn dan het gebruik van wavelets wanneer het obstakel een grillige vorm heeft. De juiste methode hangt dus af van de toepassing.

Problemen in het sterk oscillerend regime zijn beduidend moeilijker te simuleren. Van de multischaalmethodes die in deze thesis onderzocht werden, geven enkel de snelle multipoolmethode en een specifieke gerelateerde implementatie van de hiërarchische matrix methode aanleiding tot een snel matrix-vectorproduct. Ondanks de asymptotische voordelige complexiteit $O(N \log N)$ blijven de methodes rekenintensief.

Een andere benadering werd in deze thesis voorgesteld door de combinatie van de randelementemethode met een asymptotische methode. De efficiënte methodes voor oscillerende integralen maken een efficiënte implementatie van deze aanpak mogelijk. Problemen werden opgelost met zeer grote waarden van het golfgetal, voor obstakels met een convexe en zachtverlopende vorm.

7.3 Suggesties voor verder onderzoek

Het verdere onderzoek na deze thesis kan zich toespitsen op methodes voor de evaluatie van oscillerende integralen, op verbeteringen van de randelementenmethode en op uitbreidingen van de hybride methode voor hoogfrequente integraalvergelijkingen. Enkele toepassingen worden verder uitgediept in Appendix C. Mogelijke onderzoeksrichtingen voor de evaluatie van oscillerende integralen zijn:

- efficiënte algoritmes voor een aantal varianten van de modelvorm met een meer algemene oscillator, zoals een cosinus of een Besselfunctie;
- een robuuste implementatie van de numerieke methode van de steilste helling in de aanwezigheid van complexe stationaire punten, polen of singulariteiten;
- de asymptotische analyse van de methode voor meerdimensionale integralen met een gedegenereerd stationair punt in het integratiedomein;
- de toepassing van de ontwikkelde methodes in bestaande methodes voor wetenschappelijke berekeningen en ingenieursproblemen, zoals de J-matrixmethode voor quantum verstrooiïng [194], en de golfgebaseerde methode voor akoestische berekeningen [76].

Verdere onderzoeksrichtingen in het domein van de randelementenmethode en de hybride methode zijn:

- het onderzoeken van het gedrag van het conditiegetal van de discretisatiematrix in de randelemenmethode bij hoge frequenties. Het conditiegetal heeft een invloed heeft op de rekentijd bij het gebruik van iteratieve oplossingsmethodes;
- de separabele benadering van de kernfunctie in de snelle multipoolfunctie is gebaseerd op de discretisatie van een oscillerende integraal. Met de nieuwe technieken kan de benadering mogelijk verbeterd worden, waardoor de snelle multipoolmethode efficiënter wordt;
- de uitbreiding van de $O(1)$ hybride methode voor hoogfrequente integraalvergelijkingen naar meer algemene verstrooiïngsproblemen. Problemen met meervoudige verstrooiïng en niet-convexe obstakels kunnen mogelijk met een iterative benadering gesimuleerd worden;
- de methode kan verder uitgebreid worden naar drie-dimensionale problemen, de Maxwell-vergelijkingen en obstakels met hoeken;
- het is een open vraag of hybride methodes geformuleerd kunnen worden voor algemene problemen met industriele relevantie, zoals planaire antennes bestaande uit verschillende elektronische componenten. De asymptotische vorm van de oplossing kan in dat geval bijzonder complex worden. De ontwikkeling van $O(1)$ methodes voor realistische problemen bij hoge frequenties blijft een fascinerende uitdaging.

Contents

Preface vii

Acknowledgement ix

Nederlandse samenvatting xi

Contents xxv

Notations xxxi

Abbreviations xxxiii

1 Introduction 1
- 1.1 Problem statement 1
- 1.2 Motivation and goals 3
- 1.3 Scope of the research project 4
- 1.4 Outline of the thesis 5

2 Integral equations 7
- 2.1 Introduction 7
- 2.2 The Helmholtz equation and applications 8
  - 2.2.1 Time-harmonic wave scattering 8
  - 2.2.2 Acoustic scattering 9
  - 2.2.3 Electromagnetic waves and Maxwell's equations 10
  - 2.2.4 Radar cross section 11
- 2.3 Types of integral equations 12
  - 2.3.1 Volterra integral equations 13
  - 2.3.2 Fredholm integral equations 14
- 2.4 Mathematical preliminaries 15
  - 2.4.1 Function spaces 15
  - 2.4.2 Linear compact operators 17
  - 2.4.3 Riesz-Fredholm theory 18

2.5 Boundary value problems 21
2.6 Boundary integral equations 22
2.6.1 Green's identities 23
2.6.2 An integral representation 23
2.6.3 Jump relations 26
2.6.4 Boundary integral equations 28
2.6.5 Mapping properties of the integral operators 32
2.6.6 Uniqueness and resonance frequencies 34
2.7 The boundary element method 36
2.7.1 The boundary element method 36
2.7.2 Convergence 37
2.7.3 Conditioning 38

3 Wavelet based methods 39
3.1 Introduction 39
3.2 Overview of the method 40
3.3 Wavelets 41
3.4 Theory of wavelet based methods 43
3.4.1 Element size estimates 43
3.4.2 Compression of the discretisation matrix 44
3.4.3 Diagonal preconditioning 45
3.4.4 Convergence 46
3.5 Dependence on the wavenumber 47
3.5.1 Estimates for the derivatives of the kernel 47
3.5.2 Estimates for the size of the elements 49
3.5.3 Density of the compressed stiffness matrix 52
3.5.4 Wavenumber dependence of the wavelet compression . 53
3.6 Wavelet-packet based methods 54
3.6.1 Motivation 54
3.6.2 Wavelet packets 55
3.6.3 Application in boundary element methods 58
3.6.4 Computational complexity 61
3.7 Wavelet quadrature 64
3.7.1 Integration techniques 64
3.7.2 An integration rule for smooth functions 65
3.7.3 Improving accuracy by composite quadrature 67
3.7.4 An integration rule for singular functions 73
3.8 Numerical results 77
3.8.1 Wavenumber dependence 78
3.8.2 Wavelet-packet based methods 79
3.8.3 Wavelet quadrature 84
3.9 Three-dimensional problems 85
3.10 Conclusions 86

4 Fast multipole methods 87
4.1 Introduction 87
4.2 Overview of the method 88
4.3 Multilevel fast multipole method 89
4.3.1 Multipole expansions and local expansions 89
4.3.2 Single level fast multipole method 90
4.3.3 Multilevel fast multipole method 92
4.3.4 Application to integral equations 94
4.4 Expansions and translation operators 95
4.5 High frequency fast multipole method 96
4.6 Numerical results 97
4.6.1 Low frequency fast multipole method 98
4.6.2 High frequency fast multipole method 99
4.7 Three-dimensional problems 101
4.8 Conclusions 101

5 Hierarchical matrix methods 103
5.1 Introduction 103
5.2 Overview of the method 104
5.3 $\mathcal{H}$-matrices 105
5.3.1 Definition of hierarchical matrices 105
5.3.2 $\mathcal{H}$-matrix arithmetic 106
5.4 The $\mathcal{H}$-matrix method 107
5.4.1 Construction of the $\mathcal{H}$-matrix 107
5.4.2 $\mathcal{H}^2$-matrices 109
5.4.3 Black-box separable approximations 110
5.5 High frequency $\mathcal{H}$-matrix methods 111
5.6 Numerical results 112
5.6.1 Convergence of the boundary element method 112
5.6.2 $\mathcal{H}$-matrix method 113
5.6.3 The choice of separable approximations 114
5.7 Three-dimensional problems 117
5.8 Conclusions 117

6 Oscillatory integrals 119
6.1 Introduction 119
6.2 Methods with high asymptotic order 121
6.2.1 The asymptotic method 121
6.2.2 Filon-type methods 123
6.2.3 Levin-type methods 124
6.3 Numerical steepest descent method 125
6.3.1 Overview of the method 125
6.3.2 Some motivating examples 126

6.3.3 The ideal case: analytic integrand and no stationary points . . . 128
6.3.4 The case of stationary points . . . 136
6.3.5 The case where the oscillator is not easily invertible . 142
6.3.6 Filon-type methods for a non-analytic function f . . . 144
6.3.7 Generalisation to a non-analytic function g . . . 149
6.4 Multivariate numerical steepest descent . . . 151
6.4.1 Overview . . . 151
6.4.2 Extension to two-dimensional integrals . . . 153
6.4.3 A decomposition of multivariate highly oscillatory integrals . . . 159
6.4.4 The construction of cubature rules using derivatives . 167
6.4.5 Numerical results . . . 169
6.5 Conclusions . . . 176

7 A hybrid asymptotic method 177
7.1 Introduction . . . 177
7.2 Overview of the method . . . 178
7.3 Specialised quadrature rules . . . 180
7.3.1 A generalised model form . . . 180
7.3.2 Construction of the quadrature rule . . . 181
7.3.3 Convergence properties of the quadrature rule . . . 183
7.4 High frequency scattering problems . . . 186
7.4.1 High frequency integral equation formulation . . . 186
7.4.2 Asymptotic behaviour of the solution . . . 187
7.5 A high frequency boundary element method . . . 189
7.5.1 Collocation approach for the discretisation . . . 190
7.5.2 Convergence of the specialised quadrature rule as a function of k . . . 193
7.5.3 A sparse discretisation . . . 195
7.6 Numerical results . . . 197
7.6.1 Convergence and total solution time . . . 197
7.6.2 Error of the solution . . . 199
7.7 Three-dimensional problems . . . 202
7.8 Conclusions . . . 202

8 Conclusions and future research 203
8.1 Multiscale methods . . . 203
8.2 Hybrid methods . . . 205
8.3 Future research directions . . . 205
8.3.1 Oscillatory integrals . . . 205
8.3.2 Boundary element methods . . . 206

A Scattering by the circle 209
A.1 Solution by separation of variables 209
A.2 An analytical solution 211
A.3 Eigenvalues 212

B The method of steepest descent 213
B.1 Watson's Lemma 213
B.2 An asymptotic expansion for $I[f]$ 214
  B.2.1 An asymptotic expansion for $F_j(x)$ 215
  B.2.2 Computational issues 216

C Engineering applications 219
C.1 The Wave Based method in acoustics 219
C.2 Maxwell's equations 221
C.3 Quantum scattering 222
C.4 Scattering by convex polygons 223
C.5 Multiple scattering 225

Bibliography 227

Index 245

Notations

Functional analysis

D	finite dimensional open or closed region $D \subset \mathbb{R}^d$
$C^m(D)$	m-fold continuously differentiable functions on D
$C_0^m(D)$	m-fold continuously differentiable functions on D with compact support
$L_2(D)$	square integrable functions on D
$H^s(D)$	Sobolev space with index s on D
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	linear normed spaces (unless mentioned otherwise)
$\mathcal{X}^*$	dual of $\mathcal{X}$
$L(\mathcal{X}, \mathcal{Y})$	space of linear and bounded operators from $\mathcal{X}$ to $\mathcal{Y}$
$\mathcal{K}(\mathcal{X}, \mathcal{Y})$	space of compact operators from $\mathcal{X}$ to $\mathcal{Y}$
$(\cdot, \cdot)$	inner product on H or sesquilinear form on $H^* \times H$
γ	trace operator
δ	Dirac delta-function
f.p.	Hadamard finite part integral

Scattering problems

k	wavenumber of the Helmholtz equation
Ω	bounded and open subset of $\mathbb{R}^d$, $d = 2, 3$
Γ	boundary of the domain Ω
Ω^-, Ω^+	interior and exterior domains with respect to Γ
κ	parameterisation of Γ

Boundary element method

S	single-layer potential operator
D	double-layer potential operator
N	hypersingular integral operator
$G(x, y)$	kernel function
$\tilde{G}(t, \tau)$	kernel function in the parameter domain

$H_\nu^{(1)}(z)$	Hankel function of the first kind and order ν
$u(x)$	density function
r	order of an operator
α	$r/2$
$\phi(x)$	basis function
Ω_i	support of basis function ϕ_i
γ	regularity of the basis functions
d	approximation order of the basis functions
η	coupling parameter in the combined field integral equation

Multiscale methods

τ, σ	row cluster and column cluster
U^τ, V^σ	row cluster basis and column cluster basis
T_I	cluster tree
$T_{I \times I}$	block cluster tree
$\mathcal{L}^+, \mathcal{L}^-$	set of admissible leaves, set of inadmissible leaves
$\phi(x)$	scaling function
$\psi(x)$	wavelet
d	approximation order of the basis functions
$\tilde{d}$	number of vanishing moments
γ	regularity of the scaling function
$\tilde{\gamma}$	regularity of the dual scaling function

Oscillatory integrals

ω	frequency parameter
f	smooth amplitude function
g	smooth oscillator
a, b	endpoints of the integration variable
ξ	stationary point
r	order of a stationary point
$I[f]$	the oscillatory integral from a to b
$h_x(p)$	steepest descent path at x

Other symbols

$\mathrm{d}\sigma_x$,	infinitesimal element in volume or area integrals
$\mathrm{d}s_x$,	infinitesimal surface or line element in integrals over a boundary
$\Re, \Im$	real and imaginary part of a complex number

Abbreviations

BEM	boundary element method
CDF	Cohen-Daubechies-Feauveau (wavelet)
CFIE	combined field integral equation
DB	Daubechies (wavelet)
FFT	fast Fourier transform
FMM	fast multipole method
GMRES	generalised minimal residual
NSD	numerical steepest descent
NZE	non-zero entries
ODE	ordinary differential equation
PDE	partial differential equation
SVD	singular value decomposition

Chapter 1

Introduction

1.1 Problem statement

The main focus of this thesis is the numerical simulation of wave phenomena. Everywhere we go, we are surrounded by waves. Acoustic waves enable sound and music, we use microwave ovens, wireless communication devices and computer chips every day. Our hospitals make use of x-ray computed tomography (CT), magnetic resonance imaging (MRI) and ultrasonography. The smallest currently known particles exhibit wave-like properties on very small time scales. On a much larger scale, cosmic background radiation in the universe reveals a possible Big Bang billions of years ago. Wave problems arise in many disciplines of science.

Both from a physical and a mathematical point of view, there is an important difference between static or non-oscillatory behaviour, such as in gravitation, and dynamic or oscillatory behaviour, such as in electromagnetic radiation. The difference lies in the way that information is conveyed to the far field. For static problems, information is lost with increasing distance from a source. For example, one can easily infer the existence and the weight of the moon from its gravitational field, but one can not determine its shape. The moon, and any other mass of any shape, behaves as a single point mass from a sufficiently large distance. In contrast, one can easily observe different craters on the surface of the moon, even with the naked eye. The highly oscillatory nature of light permits very detailed information to travel the long distance to earth. Numerically, this means that static or low frequency behaviour in the far field can be well approximated, opening up possibilities for efficient algorithms in numerical simulations. On the other hand, it can be expected that the simulation of high frequency problems will be more computationally intensive.

We are interested in this thesis in the scattering of time-harmonic waves by a scattering obstacle $\Omega \subset \mathbb{R}^d$ with boundary $\Gamma := \partial\Omega$. This can be modelled by the Helmholtz equation

$$\Delta u + k^2 u = 0,$$

with $\Delta = \nabla^2$ being the Laplacian, and where the *wavenumber* k determines the frequency of the waves. The unknown $u(x)$ is defined on the infinitely large domain surrounding the obstacle Ω. A boundary condition $u(x) = f(x)$ is imposed on Γ. This elliptic boundary value problem can be reformulated as an integral equation over the boundary, of the form

$$(Av)(x) = \int_\Gamma G(x,y)v(y)\,\mathrm{d}s_y = f(x), \qquad x \in \Gamma, \tag{1.1}$$

or

$$\lambda v(x) + \int_\Gamma G(x,y)v(y)\,\mathrm{d}s_y = f(x), \qquad x \in \Gamma. \tag{1.2}$$

Contrary to the Helmholtz equation, the unknown function $v(y)$ in (1.1) and (1.2) is defined on a finite domain Γ. Moreover, the dimension of Γ is lower than that of $\mathbb{R}^d$. Numerically, these are two important advantages.

The integral equation formulation comes at a cost however. The discretisation matrix, corresponding to a finite element or *boundary element* approach with N basis functions, is a dense matrix with N^2 elements. Direct solution methods for the corresponding system of equations require $O(N^3)$ operations. Efficient iterative Krylov subspace methods still require $O(N^2)$ operations per matrix-vector product. For large problems, or complicated boundaries, the numerical simulation becomes computationally intractable.

Another issue is the efficient solution for increasing values of the wavenumber. Specifically, at larger frequencies, the solution $v(y)$ itself will be more oscillatory. The number of basis functions N needs to grow with k, in order to represent the solution with a fixed accuracy. Usually, one chooses a fixed number of basis functions per wavelength per dimension. This means that, for three-dimensional problems involving a two-dimensional boundary, the number of unknowns increases quadratically with k.

Similar to the difference between static and dynamic problems, we make a distinction between the *low frequency regime* and the *high frequency regime*. In the low frequency regime, the wavenumber k is fixed, and the number of basis functions N is increased to improve the accuracy. In the high frequency regime, the wavenumber k and N increase proportionally to solve the problem at a higher frequency with a fixed accuracy. The problem considered in this thesis is the study of efficient solution methods for equations (1.1) and (1.2) in the low frequency regime and in the high frequency regime.

1.2 Motivation and goals

The motivation for this research project lies in a collaboration of the research group Scientific Computing, at the Department of Computer Science, with research groups from other departments in the Engineering Faculty of K.U.Leuven. The research group Telemic, at the Department of Electrical Engineering, employs integral equations for the modelling of antennas. Their research focuses on antennas at millimetre wave frequencies, electronic beam steering, and small integrated antennas. The software package MAGMAS (Model for the Analysis of General Multilayered Antenna Structures) that was developed relies heavily on integral equation formulations [73, 192, 198]. Currently, the dense nature of the discretisation matrix and the high frequency nature of the problem are the bottlenecks that preclude the numerical simulation of larger antenna structures.

The boundary element method has also been considered in the package OLYMPOS, developed by research group Electa at the Department of Electrical Engineering [71, 139]. Their research involves the analysis, design and optimisation of the steady state and dynamic behaviour of electromagnetic energy transducers at low frequency, such as permanent magnets, actuators and motors. The large scale numerical simulation using OLYMPOS is restricted due to the computational cost of the boundary element method.

Finally, methods similar to the boundary element method are being developed by the research group Noise and Vibration, at the Department of Mechanical Engineering, for the modelling of noise and vibration propagation inside vehicles [76, 77, 190]. The new methods developed in that group require the efficient evaluation of a large number of oscillatory integrals.

The purpose of this thesis is the study of the theoretical and numerical properties of fast solution methods for integral equations and evaluation methods for oscillatory integrals. Several fast solution methods for integral equations have been developed in the past decades. We focus on fast multipole methods [98], hierarchical matrix methods [101] and wavelet based methods [19]. We analyse the properties and the computational complexity of the methods, both in the low frequency regime and in the high frequency regime. We aim at suggesting and developing improvements, and at closing some of the gaps in the theoretical analysis.

We already briefly summarise the main results. Of the methods considered, we found that the fast multipole method is the only viable efficient solution method for high frequency problems (together with an intimately related implemention of hierarchical matrices). The wavelet method is optimal for low frequency problems, but is less robust for scattering obstacles with irregular shapes. A novel approach is developed in this thesis for high frequency problems, based on a thorough analysis of the properties of oscillatory integrals. A hybrid method is proposed that combines the standard boundary element method with asymptotic methods for high frequencies. For a restricted class of model problems with a smooth and convex scattering obstacle problems are solved with wavenumbers that are several times larger than what is currently computationally feasible with an efficient implementation of multiscale methods.

1.3 Scope of the research project

The field of integral equations, scattering problems and efficient solution methods is rich and varied. It is therefore necessary to limit the scope of the research project. Specifically, we restrict the scattering problems in this thesis to two-dimensional integral equation formulations of the Helmholtz equation, leading to linear Fredholm integral equations of the first kind or the second kind. The issues that arise in three-dimensional problems will be briefly discussed in each chapter.

In the study of fast solution methods, we emphasise the construction of a fast matrix-vector product. This matrix-vector product can be used in combination with an iterative linear solver for the efficient overall solution of the problem. With the exception of wavelet based methods, a study of preconditioning strategies is not included in this thesis. This choice is motivated by a similar subdivision that is observed in literature: preconditioning techniques can be combined with different matrix-vector products, independently of each other. Pointers are given to the relevant literature.

The discussion of fast multipole methods and hierarchical matrices is mostly limited to a literature study, with numerical results that are restricted to an illustration of the principles and basic properties. It was found that the theory and properties of these methods have been sufficiently studied and described elsewhere, both for the low frequency regime and the high frequency regime. Still, the numerical results for hierarchical matrices include the largest problem that is considered in this thesis. A problem is solved with half a million unknowns, which corresponds to a dense matrix of approximately 274 billion elements.

Several own contributions are discussed to multiscale methods based on wavelets. A gap was observed in the theory and applications of wavelet based methods. We show that the wavelet method requires $O(N^2)$ operations in the high frequency regime. We propose a new method, based on wavelet packets, that leads to a matrix-vector product with a much reduced computational complexity. Wavelets enable a subdivision of scale and position. Wavelet packets introduce an additional subdivision of the frequency spectrum. We show that the use of wavelet packets allows the choice of oscillatory basis functions, that can be adaptively matched to the frequency of the scattered wave. Finally, we also focus on new quadrature techniques for the implementation, showing that, contrary to a widespread perception, a full Galerkin method can be implemented almost as efficiently as a collocation method.

Next, a relatively new area of research is explored, that has originated in the advent of efficient techniques for the evaluation of highly oscillatory integrals. Own contributions are discussed for univariate and multivariate integrals, based on a numerical implementation of the method of steepest descent. It is shown that oscillatory integrals can be evaluated using a number of operations that is independent of the frequency of the integrand. The accuracy of these methods increases with increasing frequency.

The further analysis of these methods has led to a novel hybrid method for oscillatory integral equations that is presented in Chapter 7 of this thesis. The new method is formulated for the case of a smooth and convex scattering obstacle. In contrast to classical boundary element methods, the hybrid method leads to a discretisation matrix that is small and sparse. The matrix can be constructed with a number of operations that is independent of the frequency. Moreover, the accuracy of a large part of the solution increases with increasing frequency.

1.4 Outline of the thesis

We commence in Chapter 2 with a study of integral equations and their applications. We derive a set of boundary integral equations that are equivalent to the Laplace equation or the Helmholtz equation on a two- or three-dimensional domain, and we introduce the boundary element method.

Next, we consider fast solution methods based on a multiscale representation of the problem. Multiscale methods based on wavelets are discussed in Chapter 3. The results in this chapter have been published in a number of articles. The wavenumber dependence of the wavelet method is analysed in [118]. The efficient computation of integrals involving wavelets is the subject of [120]. The method based on wavelet packets was proposed in [121], with more numerical results for multiple scattering configurations given in [122].

The fast multipole method and methods based on hierarchical matrices are discussed in Chapter 4 and Chapter 5. These chapters represent a literature study, with numerical results that illustrate the principles and basic properties.

Chapter 6 focuses on efficient evaluation methods for oscillatory integrals. The numerical steepest descent method was proposed in [124]. The extension to multivariate integrals was described in [123]. The application of these methods for the evaluation of the matrix entries of the discretisation matrix is explored in [119].

The new hybrid method for high frequency scattering problems is described next in Chapter 7. The contents of this chapter correspond to the article [125].

The conclusions of this research project are formulated in Chapter 8, and directions for future research are indicated. A number of appendices follow this chapter. An analytic solution for the scattering by a circle further illustrates the scattering problem in Appendix A. The steepest descent method is discussed in Appendix B. Finally, the ongoing research in some applications is summarised in Appendix C.

Chapter 2

Integral equations

2.1 Introduction

The main focus of this thesis is the solution of scattering problems using integral equation formulations. The purpose of this chapter is to introduce such scattering problems and their applications, and to relate them to the corresponding mathematical formulations. We also aim to give an overview of the relevant mathematical theory of integral equations. The chapter is meant to be illustrative of the general theory. As such, it is not entirely self-contained; the reader is referred to the given references for a more complete description.

We commence with an introduction to the Helmholtz equation and its applications in §2.2. The Helmholtz equation and related equations appear in many problems involving wave characteristics, including acoustics and electromagnetics. The equation can be derived from the wave equation for *time-harmonic* problems - problems involving one specific frequency ω. The main classification of general integral equations is described in §2.3. In the remainder of the chapter, we will focus on a particular class of integral equations that arises from the reformulation of boundary value problems involving partial differential equations. The resulting integral equations are so-called *Fredholm integral equations of the first kind* or *Fredholm integral equations of the second kind*.

A number of mathematical preliminaries are presented in §2.4. We introduce Sobolev function spaces and the theory of compact operators as necessary tools for the understanding and characterisation of the mapping properties of an integral operator. The main result of this section is the formulation of the *Fredholm Alternative* - the theorem that decides on the solvability of an operator equation.

The relevant boundary value problems are formally defined in §2.5. The construction of boundary integral equations that are equivalent to the Helmholtz or Laplace boundary value problems is described in §2.6. The boundary integral equations are essentially derived from an integral representation of the solution to the boundary value problem. We introduce the boundary element method for the solution of boundary integral equations in §2.7. We discuss the convergence properties of a Galerkin approach, and the conditioning issues of the resulting discretisation matrices.

The theory in this chapter mostly follows the approach of [11, 49, 100, 116, 154, 159]. Boundary integral equations for general elliptic partial differential equations are derived in [154]. A detailed description of the numerical properties of integral equations is given in [100], with special attention for the n-dimensional Laplace equation. The issues associated with integral equations of the first kind are explored in [203]. Solution methods for integral equations of the second kind are discussed in [11]. Integral equation formulations for the three-dimensional Helmholtz problem are developed in [159] for applications in electromagnetics, and in [49, 138] for acoustics.

2.2 The Helmholtz equation and applications

The Helmholtz equation takes its name from Hermann von Helmholtz (1821-1894). The equation is often related to problems involving wave characteristics. For example, we will show in §2.2.1 how the equation can be derived from the wave equation. In that context, it is sometimes called the *reduced wave equation.* The Helmholtz equation is encountered in many different applications, including non-oscillatory problems, such as the eigenvalue problem for the Laplace operator Δ. We present an overview of some of the applications of the Helmholtz equation and related equations, that have motivated the research of this thesis. For more information, the reader is referred to [107, 159] for applications in electromagnetics, to [49, 138] for applications in acoustics, and to [99] for an overview of the Helmholtz equation in general. For more general information on wave problems, that is not specific to integral equations, the reader is referred to [147, 201].

2.2.1 Time-harmonic wave scattering

The first important equation in the modelling of wave propagation is the *wave equation* itself, given by

$$\frac{\partial^2 U}{\partial t^2} + \gamma \frac{\partial U}{\partial t} - c^2 \Delta U = 0. \tag{2.1}$$

The wave equation models a wave with propagation speed c. Parameter γ is a positive damping factor - if it is nonzero, the equation is dissipative.

Equation (2.1) is a linear second order hyperbolic differential equation. Since it is linear, the sum of any two solutions is again a solution of the equation. Using this fact, we can single out one specific frequency. Assume that $U(x,t) = u(x)e^{-i\omega t}$ is a time-harmonic wave with frequency $\omega > 0$.[1] Substitution in (2.1) yields the *Helmholtz equation*,

$$\Delta u + k^2 u = 0, \tag{2.2}$$

where the *wavenumber* $k \neq 0$ is given by $k^2 = \omega(\omega + i\gamma)/c^2$. In the absence of damping, we have $k = \omega/c$. The sign of k is usually chosen such that

$$\Im(k) \geq 0.$$

Henceforth, we will mostly assume that k is real, unless mentioned otherwise. The name wavenumber is related to the case of a propagating plane wave, where the *wavelength* is given by $\lambda = 2\pi/k$. In that case, the wavenumber k is the number of wavelengths per 2π units of length.

The Helmholtz equation is a linear second order elliptic partial differential equation. The description of a wave that is scattered by an obstacle Ω leads to a boundary value problem. Say a time-harmonic incoming wave is given by the position-dependent amplitude function $u^i(x)$. Then the total wave is given by

$$u(x) = u^i(x) + u^s(x),$$

where u^s denotes the scattered wave. The boundary condition should be given on $\Gamma = \partial\Omega$, and depends on the underlying physical problem. One can have Dirichlet boundary conditions of the form $u^s = f$, or Neumann boundary conditions of the form $\frac{\partial u^s}{\partial n} = g$, or combinations of both. In general, different boundary conditions may be required for different parts of Γ. For exterior problems, a suitable boundary condition should also be satisfied at infinity. We discuss these conditions in more detail in §2.5.

2.2.2 Acoustic scattering

Consider the propagation of acoustic waves in a homogeneous, inviscid medium with propagation speed c. The velocity vector $v(x,t)$ of each particle is a function of space and time. It can be derived from a scalar function $U(x,t)$, called the *velocity potential* function, by

$$v(x,t) = \rho^{-1}\nabla U(x,t),$$

---

[1] Some authors use a time-dependence of the form $U(x,t) = u(x)e^{i\omega t}$. The differing sign has implications for the exact form of many relations that follow from this equation.

where ρ is the density of the medium. Based on the conservation of mass and momentum, it can be shown that the velocity potential satisfies the wave equation (2.1) if the perturbations in speed and pressure are sufficiently small. In time-harmonic problems, where $U(x,t) = u(x)e^{-i\omega t}$, the amplitude function $u(x)$ hence satisfies the scalar Helmholtz equation (2.2). The pressure $p(x,t)$ can be obtained from

$$p - p_0 = -\frac{\partial U}{\partial t} - \gamma U.$$

The amplitude of the pressure function in time-harmonic problems also satisfies the Helmholtz equation.

The boundary conditions in the problem of scattering by an obstacle Ω relate to a physical property of the obstacle. The Dirichlet condition $u^s = -u^i$ ensures that $u = 0$ on $\Gamma = \partial\Omega$. Physically, this corresponds to a so-called *sound-soft obstacle.* The Neumann boundary condition ensures $\frac{\partial u}{\partial n} = 0$ on Γ, and corresponds to a *sound-hard obstacle.*

For general media, the wavenumber $k(\omega)$ depends on the frequency in a medium-dependent manner. A consequence is that, in complex media, the speed of sound is also a function of the frequency ω. The quantity

$$c_p = \frac{\omega}{\Re(k)}$$

is called the *phase velocity.* The imaginary part of ω/k characterises the attenuation of waves in the medium. The effect that waves at different frequencies may propagate with different velocities is called *dispersion.*

2.2.3 Electromagnetic waves and Maxwell's equations

The Helmholtz equation also appears in the modelling of electromagnetic waves. Electromagnetic waves are described by the Maxwell equations. For a homogeneous medium with electric permittivity ϵ and magnetic permeability μ, and in the absence of charges and currents, the Maxwell equations are given by

$$\nabla \times E = -\mu\frac{\partial H}{\partial t}, \quad \nabla \times H = \epsilon\frac{\partial E}{\partial t}, \quad \nabla \cdot E = 0, \quad \nabla \cdot H = 0. \tag{2.3}$$

The field vectors E and H are the electric and magnetic field respectively.

The vectorial identity

$$\Delta E = \nabla \cdot \nabla \times E - \nabla \times \nabla \times E$$

can be used to show that E and H satisfy the vectorial wave equations

$$\begin{cases} \Delta E - \dfrac{1}{c^2}\dfrac{\partial^2 E}{\partial t^2} = 0, \\ \Delta H - \dfrac{1}{c^2}\dfrac{\partial^2 H}{\partial t^2} = 0, \end{cases}$$

where $c = 1/\sqrt{\epsilon\mu}$ is the speed of light in the medium. Transformation to the frequency domain yields

$$\begin{cases} \Delta \hat{E} + k^2 \hat{E} = 0, \\ \Delta \hat{H} + k^2 \hat{H} = 0, \end{cases} \tag{2.4}$$

The notation $\hat{E}$ is used to denote the Fourier transform of E, and is called the *phasor* of E. Thus, the phasors of the electric and magnetic field vectors satisfy a vector Helmholtz equation. Equations (2.4) should be augmented with the conditions $\nabla \cdot \hat{E} = 0$ and $\nabla \cdot \hat{H} = 0$ to be equivalent to (2.3).

2.2.4 Radar cross section

Integral equation formulations for boundary value problems will be discussed in depth in §2.6. Nevertheless, as an introduction, we already state the form of the most common integral equation of this thesis in order to describe the *radar cross section* of a scattering obstacle.

Consider a two-dimensional scattering obstacle Ω with boundary $\Gamma = \partial\Omega$. The scattered wave $u^s(x)$ that satisfies the Helmholtz equation in the exterior of Ω, and the Dirichlet boundary condition $u^s(x) = -u^i(x)$ on Γ, can be found by solving the integral equation

$$\int_\Gamma \frac{i}{4} H_0^{(1)}(k|x-y|)u(y)\,\mathrm{d}s_y = -u^i(x), \qquad x \in \Gamma, \tag{2.5}$$

where $H_0^{(1)}(z)$ is the Hankel function of the first kind and order zero. The unknown function in (2.5) is the *density function* $u(y)$, which is defined on Γ. The scattered wave is then given by

$$u^s(x) = \int_\Gamma \frac{i}{4} H_0^{(1)}(k|x-y|)u(y)\,\mathrm{d}s_y, \qquad x \in \mathbb{R}^2 \setminus \Omega.$$

Based on the asymptotic behaviour of the Hankel function for large arguments [4], one obtains the asymptotic behaviour of the scattered wave itself as

$$u^s(x) = \frac{e^{ik|x|}}{\sqrt{|x|}} \left(u^\infty\left(\frac{x}{|x|}\right) + O\left(\frac{1}{|x|}\right)\right), \qquad |x| \to \infty.$$

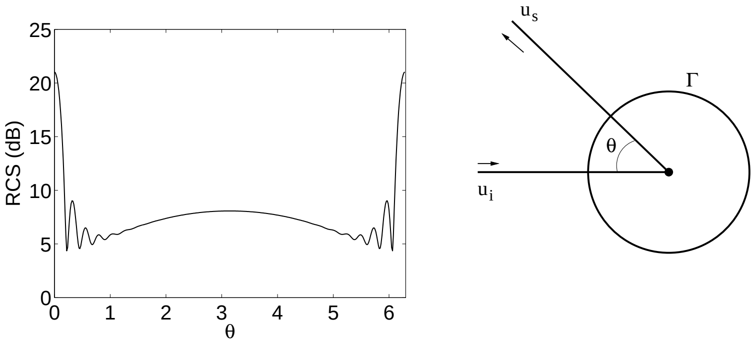


Figure 2.1: Illustration of the radar cross section of a circular obstacle with a radius of 0.5 m at a frequency of 1.3 GHz as a function of the observing angle θ in radians.

The function $u^{\infty}(\frac{x}{|x|})$ is called the *far-field pattern*; it depends only on the direction of the vector x. The far field pattern can be expressed in terms of the original density function,

$$u^{\infty}\left(\frac{x}{|x|}\right) = \frac{e^{i\frac{\pi}{4}}}{8\pi k}\int_{\Gamma} u(y)e^{-ik\frac{x}{|x|}\cdot y}\,\mathrm{d}s_y.$$

The *radar cross section* is defined for $x = (\cos\theta, \sin\theta)$ as

$$\sigma^c(\theta) := 2\pi \lim_{|x|\to\infty} |x| \frac{|u_s(x)|^2}{|u_i(x)|^2}. \tag{2.6}$$

It is the effective surface area of an isotropic antenna that would return the same amount of power to a receiver at the distance $|x|$, as the power reflected by Ω at the backscattering angle θ. The radar cross section can also be found in terms of the far field pattern,

$$\sigma^c(\theta) := 2\pi \left| u^{\infty}\left(\frac{x}{|x|}\right)\right|^2.$$

The radar cross section for a circular obstacle is shown in Figure 2.1. The figure shows $10\log_{10}(\sigma^c/\lambda)$.

2.3 Types of integral equations

The main classification of integral equations into two major groups depends on the integration domain in the equation. If the integration domain depends on the variable x, the equation is called a *Volterra* integral equation.

If the integration domain is fixed, it is called a *Fredholm* integral equation. A second classification is used depending on whether the unknown function appears only inside the integral, or both inside and outside the integral; these are called integral equations of the *first kind* and integral equations of the *second kind* respectively.

The theoretical properties of Volterra and Fredholm integral equations are quite different, as are the numerical treatment and the applications. The most important difference between equations of the first kind and equations of the second kind, both from a theoretical and from a numerical point of view, is the conditioning of the problem. Specifically, integral equations of the first kind can be severely ill-conditioned. The meaning of ill-conditioned in this context is that small changes in the right hand side may correspond to large changes of the solution. We present a brief overview that illustrates the scope of this classification for one-dimensional integral equations.

2.3.1 Volterra integral equations

Volterra integral equations of the second kind have the general form

$$\lambda u(x) + \int_a^x G(x, y, u(y))\,\mathrm{d}y = f(x), \quad x \geq a, \lambda \neq 0. \tag{2.7}$$

The function $u(x)$ is the only unknown in the equation. Depending on the function G, the equation may be nonlinear. This type of problem can be seen as a generalisation of the initial value problem of a non-linear ordinary differential equation,

$$\begin{aligned} &u'(x) = F(x, u(x)), \quad x \geq a, \\ &u(a) = u_0. \end{aligned} \tag{2.8}$$

The initial value problem is equivalent to the integral equation

$$u(x) = u_0 + \int_a^x F(y, u(y))\,\mathrm{d}y, \qquad x \geq a,$$

and indeed, solution methods for (2.7) resemble solution methods for the initial value problem (2.8).

Linear Volterra integral equations of the first kind have the form

$$\int_a^x G(x, y) u(y)\,\mathrm{d}y = f(x), \qquad x \geq a. \tag{2.9}$$

Such equations can be very ill-conditioned, depending on the continuity properties of the kernel function $G(x, y)$. We will investigate this ill-conditioning later in detail for Fredholm integral equations of the first kind.

In some cases, linear Volterra integral equations of the first kind can be reduced to equations of the second kind. If the derivatives $\frac{\partial G}{\partial x}$ and f' exist and are continuous, and if $G(x,x) \neq 0$, $x \in [a,b]$, then differentiating (2.9) with respect to x leads to

$$u(x) + \int_a^x \frac{\frac{\partial G}{\partial x}(x,y)}{K(x,x)} u(y)\,\mathrm{d}y = \frac{f'(x)}{G(x,x)}, \qquad x \in [a,b].$$

One particularly simple example is the Volterra integral equation of the first kind

$$\int_a^x u(y)\,\mathrm{d}y = f(x), \qquad x \geq a.$$

This problem is equivalent to $u(a) = 0$ and $u(x) = f'(x)$, $x \geq a$.

We will not consider Volterra integral equations in the remainder of this thesis. The reader is referred to [148] for an overview of analytical and numerical solution methods, and to [157] for the treatment of nonlinear Volterra integral equations.

2.3.2 Fredholm integral equations

Linear Fredholm integral equations of the second kind have the general form

$$\lambda u(x) + \int_a^b G(x,y)u(y)\,\mathrm{d}y = f(x), \qquad x \in [a,b], \lambda \neq 0. \tag{2.10}$$

The difference with Volterra integral equations is that the product of the unknown function and the kernel function is integrated over a fixed integration domain. Although Volterra equations could be interpreted as a special case of Fredholm equations, using the modified kernel function

$$G_F(x,y) = \begin{cases} G(x,y), & x \leq y, \\ 0, & x > y, \end{cases}$$

they are not usually treated that way.

Linear Fredholm integral equations of the first kind have the form

$$\int_a^b G(x,y)u(y)\,\mathrm{d}y = f(x), \qquad x \in [a,b]. \tag{2.11}$$

The kernel function $G(x,y)$ in (2.10) and (2.11) may be continuous or singular. Typically, in the context of boundary value problems, the kernel function is singular when $x = y$. We call the kernel *weakly singular* if the integral is (improperly) integrable, and *strongly singular* if it is not integrable. A weak singularity in the case of one-dimensional integrals may be a logarithmic singularity, or it may have the form $|x-y|^\alpha$ with $\alpha > -1$. A weakly singular kernel for two-dimensional integrals can be $|x-y|^{-1}$. Strongly singular kernels result in a so-called *hypersingular* integral equation. The efficient numerical solution of Fredholm integral equations of the first and second kind is the main subject of this thesis.

2.4 Mathematical preliminaries

2.4.1 Function spaces

We briefly introduce the main function spaces that will be used in the text. Sobolev spaces will play an important role in the characterisation of integral operators. A detailed discussion of Sobolev spaces is given in [5].

Consider a domain $D \subset \mathbb{R}^d$, that may be open or closed. We denote by $C(D)$ the space of all continuous functions on D, and by $C_0(D)$ the subspace of all continuous functions on D that have compact support. The corresponding spaces of all functions with continuous m-th order derivative on D are denoted by $C^m(D)$ and $C_0^m(D)$ respectively. In particular, we have $C(D) = C^0(D)$. The space $C^\infty(D)$ contains all infinitely differentiable functions that are bounded on D. These spaces are useful in characterising the smoothness of a function; the higher the index of the function space, the smoother the functions. Yet, they are not completely adequate for our purposes, as they are not Hilbert spaces.

The space $L_2(D)$ is the space of all square integrable functions on D,

$$L_2(D) = \left\{ f : \left| \int_D f^2(x) \, \mathrm{d}\sigma_x \right| < \infty \right\}.$$

It is a Hilbert space equipped with the scalar product

$$(f,g) = \int_D f(x)\overline{g(x)} \, \mathrm{d}\sigma_x, \qquad f,g \in L_2(D).$$

The space $L_2(D)$ can be seen as the *completion* of $C_0(D)$ with respect to the norm induced by its scalar product. We define the *Sobolev spaces* $H^m(D)$ with a positive integer index m for an open domain D by

$$H^m(D) = \{ f : \partial^\alpha f \in L_2(D) \text{ with } |\alpha| \le m \},$$

where we have used ∂^α to denote the partial derivatives of f,

$$\partial^\alpha f = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d} f, \qquad \alpha = (\alpha_1, \alpha_2, \ldots).$$

The Sobolev space $H^m(D)$ with index m contains all functions with each m-th order derivative square integrable. The index m hence corresponds to a certain smoothness of the functions. Sobolev spaces are Hilbert spaces with the scalar product

$$(f, g)_{H^m} = \sum_{|\alpha| \leq m} \int_D \partial^\alpha f(s) \overline{\partial^\alpha g(x)} \, \mathrm{d}\sigma_x, \qquad f, g \in H^m(D).$$

Sobolev spaces correspond closely to the C^m spaces, as they both characterise smoothness. For example, we always have the imbedding

$$H^m(\mathbb{R}^d) \subset C_0^k(\mathbb{R}^d), \quad 0 \leq k < m - \frac{d}{2}.$$

We have $H^s \subseteq H^m$ if $s \geq m$, and $H^0 = L_2$. One can also define Sobolev spaces with a fractional or real index. For the particular case when $D = \mathbb{R}^d$, the norm of a Sobolev space with positive real index s is given by

$$\|f\|_{H^s}^2 = \int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\hat{f}(\omega)|^2 \, \mathrm{d}\omega < \infty,$$

where $\hat{f}$ is the Fourier transform of f. For more general domains D, the norms are hard to compute. We will see later that equivalent norms can be computed from the discrete wavelet transform of a function.

We will also require function spaces on the boundary Γ of a bounded open domain Ω. We assume that Ω is a *Lipschitz domain.* A function f is *Lipschitz continuous* if there exists a constant $M > 0$ such that

$$|f(x) - f(y)| \leq M|x - y|, \qquad x, y \in \mathbb{R}^d.$$

A Lipschitz domain is a domain with a boundary that can be locally represented by a Lipschitz continuous function. Lipschitz domains include most domains of interest, including domains with corners and edges. Exceptions are domains with cracks or cusps. Sobolev spaces $H^s(\Gamma)$ on Γ can be defined by relating them to Sobolev spaces in the parameter domain. The precise definition is rather involved, and can be found in [149].

Sobolev spaces with index m on Γ are well defined if Ω is a C^m domain, i.e., there exists a parameterisation for the boundary of Ω that is m times continuously differentiable. It is useful to know the smoothness of a function on $\mathbb{R}^d$ that is restricted to Γ. Define the *trace operator* γ by

$$\gamma f = f|_\Gamma. \tag{2.12}$$

If Ω is a C^m domain, then for $1/2 < s \leq m$ the trace operator is a bounded linear operator

$$\gamma : H^s(\Omega) \to H^{s-1/2}(\Gamma). \tag{2.13}$$

The Sobolev index of the restriction of f to Γ is $1/2$ less than the Sobolev index of f itself on Ω.

We end our discussion of function spaces with the concept of *duality* of function spaces. We denote the space of all bounded linear functionals from a normed space $\mathcal{X}$ to a normed space $\mathcal{Y}$ by $L(\mathcal{X}, \mathcal{Y})$.

Definition 2.4.1. *The* dual space $\mathcal{X}^*$ *of a normed space* $\mathcal{X}$ *is the space of all bounded linear functionals* $g : \mathcal{X} \to \mathbb{C}$*, i.e.,* $\mathcal{X}^* = L(\mathcal{X}, \mathbb{C})$*.*

The dual space $\mathcal{X}^*$ can be equipped with the norm

$$\|g\|_{\mathcal{X}^*} = \sup_{0 \neq x \in \mathcal{X}} \frac{|g(x)|}{\|x\|_{\mathcal{X}}}.$$

For a Hilbert space H, there exists a bijection between H and H^*. This bijection can be used to *identify* H with H^*, i.e., to define the meaning of $H = H^*$.

Theorem 2.4.2 (Riesz' representation theorem). *If* H *is a Hilbert space, then for each* $g \in H^*$ *there exists a unique* $x \in H$ *such that*

$$g(y) = (x, y)_H, \qquad \forall y \in H. \tag{2.14}$$

Furthermore, $\|g\|_{H^*} = \|x\|_H$*.*

Conversely, for each $x \in H$, there exists a bounded linear functional $g \in H^*$ defined by (2.14).

2.4.2 Linear compact operators

In this section we introduce the concept of a *compact operator*. Recall that a set $U \in \mathcal{X}$ is called *compact* if each sequence in U has a convergent subsequence. A set is called *relatively compact* if its closure is compact.

Definition 2.4.3. *A linear operator* $A : \mathcal{X} \to \mathcal{Y}$ *from a normed space* $\mathcal{X}$ *into a normed space* $\mathcal{Y}$ *is called a* compact operator *if it maps each bounded set in* $\mathcal{X}$ *into a relatively compact set in* $\mathcal{Y}$*.*

A compact operator A is necessarily bounded, i.e., $A \in L(\mathcal{X}, \mathcal{Y})$. Any linear combination of two compact operators is again compact, and we denote the space of compact operators by $\mathcal{K}(\mathcal{X}, \mathcal{Y}) \subset L(\mathcal{X}, \mathcal{Y})$. As most common integral operators are compact operators, we describe some relevant properties of compact operators.

A common characteristic of compact operators is that they increase smoothness. The function Ku, with $u \in C([a, b])$, is typically smoother than u itself if K is a compact operator. A similar observation holds for integral operators, because the integration increases smoothness. Indeed, it can be shown that the integral operator

$$(Au)(x) = \int_{\Omega} G(x,y)u(y)\,\mathrm{d}s_y, \quad x \in \Omega,$$

is compact if $G : \Omega \times \Omega \to \mathbb{C}$ is a continuous kernel. The integral operator is also compact if G is a weakly singular kernel. Note however that these statements may no longer hold if the domain Ω has corners or edges.

Another characteristic is the typical spectrum of a compact operator. Recall that λ is an eigenvalue of a linear operator $A : \mathcal{X} \to \mathcal{X}$ if $A\phi = \lambda\phi$, for some non-zero $\phi \in \mathcal{X}$.

Theorem 2.4.4 ([11],Th.1.4.1). *Let $A \in \mathcal{K}(\mathcal{X},\mathcal{X})$ be a compact operator, and let $\mathcal{X}$ be a Banach space. Then the eigenvalues of A form a discrete set in the complex plane $\mathbb{C}$, with 0 as the only possible limit point.*

The fact that the eigenvalues of a compact operator can accumulate near 0 has important implications for the conditioning of an integral equation of the first kind. The discretisation of such an equation represents the discretisation of a compact operator, and hence, the eigenvalues of the discretisation matrix will also tend to zero. This causes the ill-conditioning. The rate at which the eigenvalues tend to zero is investigated in [113]. Generally, the speed of convergence of the eigenvalues to zero increases with increasing differentiability of the kernel function. For analytic kernel functions, the eigenvalues decay exponentially fast [150].

2.4.3 Riesz-Fredholm theory

We discuss the solvability of the operator equation $Au = f$. First, we consider the case where the operator $A : \mathcal{X} \to \mathcal{X}$ is defined as

$$A = I - K,$$

with I the identity operator and with a compact operator $K \in \mathcal{K}(\mathcal{X},\mathcal{X})$. We say that A is the identity up to a *compact perturbation*. For such an operator, it can be shown that injectivity is a sufficient condition for the existence of a bounded inverse.

Theorem 2.4.5 ([49],Th.1.16). *Let $\mathcal{X}$ be a normed space, and $K \in \mathcal{K}(\mathcal{X},\mathcal{X})$ be a compact linear operator. If $I-K$ is injective, then the inverse operator $(I-K)^{-1}$ exists and is bounded.*

The theorem shows that if the homogeneous equation $u - Ku = 0$ has only the trivial solution $u = 0$, then the non-homogeneous equation $u -$

$Ku = f$ has a unique solution for all $f \in \mathcal{X}$, and the solution depends continuously on f. In order to characterise the solvability of $Au = f$ when $A = I - K$ is not injective, we require a stronger result. First, we introduce the concept of a *Fredholm* operator. For the remainder of this section, we will follow the modern theory of [154].

Definition 2.4.6. *An operator* $A \in L(\mathcal{X}, \mathcal{Y})$ *is called* Fredholm *if*

1. *the subspace* $\operatorname{Range}(A)$ *is closed in* $\mathcal{Y}$ *;*
2. *the spaces* $\operatorname{Null}(A)$ *and the quotient space* $\mathcal{Y}/\operatorname{Range}(A)$ *are finite-dimensional.*

The index *of* A *is the integer defined by*

$$\operatorname{index} A = \dim \operatorname{Null}(A) - \dim(\mathcal{Y}/\operatorname{Range}(A)).$$

For example, a matrix operator $A : \mathbb{C}^n \to \mathbb{C}^m$ is Fredholm, with index $n - m$. A matrix with Fredholm index 0 therefore corresponds to a square matrix. The theory of Fredholm operators also covers the previously discussed case where $A = I - K$, because $I - K$ is Fredholm.

Theorem 2.4.7. *If* $A = I - K$*, where* $K \in \mathcal{K}(\mathcal{X}, \mathcal{X})$ *is compact, then* $A : \mathcal{X} \to \mathcal{X}$ *is Fredholm and* $\operatorname{index} A = 0$*. More generally, if* $B : \mathcal{X} \to \mathcal{Y}$ *is Fredholm and if* $B \in \mathcal{K}(\mathcal{X}, \mathcal{Y})$*, then* $A = B - K$ *is Fredholm, and* $\operatorname{index}(A) = \operatorname{index} B$*.*

The invertibility of the general equation $Au = f$ is decided by the so-called *Fredholm Alternative* theorem. In order to state the theorem, we require the concept of a *sesquilinear form* and of an *adjoint operator.*

Definition 2.4.8. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be two normed spaces. We call* $(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{C}$ *a* sesquilinear form *if for any* $x_1, x_2, x \in \mathcal{X}$*,* $y_1, y_2, y \in \mathcal{Y}$*,* $\alpha_1, \alpha_2 \in \mathbb{C}$ *we have*

$$\begin{aligned}
&(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_1, y), \\
&(x, \alpha_1 y_1 + \alpha_2 y_2) = \overline{\alpha_1} (x, y_1) + \overline{\alpha_2} (x, y_2).
\end{aligned}$$

For example, a sesquilinear form corresponding to the function space $\mathcal{X} = \mathcal{Y} = C(\Omega)$ may be

$$(f, g) = \int_\Omega f(x) \overline{g(x)} \, \mathrm{d}s_x, \qquad f, g \in C(\Omega).$$

A useful sesquilinear form on $\mathcal{X}^* \times \mathcal{X}$ can be defined using the concept of a dual space that was defined in Definition 2.4.1. The sesquilinear form $(\cdot, \cdot) : \mathcal{X}^* \times \mathcal{X} \to \mathbb{C}$ for dual spaces is defined by

$$(g, x) = g(\overline{x}), \qquad x \in \mathcal{X},\ g \in \mathcal{X}^*. \tag{2.15}$$

Now consider an operator $A : \mathcal{X} \to \mathcal{Y}$. The *adjoint operator* $A^* : \mathcal{Y}^* \to \mathcal{X}^*$ is defined using the sesquilinear form (2.15) by

$$(A^* y, x) = (y, Ax), \quad x \in \mathcal{X}, y \in \mathcal{Y}^*. \tag{2.16}$$

The Fredholm Alternative reads as follows.

Theorem 2.4.9 (Fredholm Alternative). *Assume* $A : \mathcal{X} \to \mathcal{Y}$ *is Fredholm with* $\operatorname{index} A = 0$. *Then there are two, mutually exclusive possibilities:*

1. *The homogeneous equation* $Au = 0$ *has only the trivial solution* $u = 0$. *In this case,*
   (a) *for each* $f \in \mathcal{Y}$, *the inhomogeneous equation* $Au = f$ *has a unique solution* $u \in \mathcal{X}$;
   (b) *for each* $g \in \mathcal{X}^*$, *the adjoint equation* $A^* v = g$ *has a unique solution* $v \in \mathcal{Y}^*$.
2. *The homogeneous equation* $Au = 0$ *has exactly* p *linearly independent solutions* $u_1, \ldots, u_p$ *for some finite* $p \geq 1$. *In this case,*
   (a) *the homogeneous adjoint equation* $A^* v = 0$ *has exactly* p *linearly independent solutions* $v_1, \ldots, v_p$;
   (b) *the inhomogeneous equation* $Au = f$ *is solvable if and only if the right-hand side* f *satisfies* $(v_j, f) = 0$, $j = 1, \ldots, p$;
   (c) *the inhomogeneous adjoint equation* $A^* v = g$ *is solvable if and only if the right-hand side* g *satisfies* $(g, u_j) = 0$, $j = 1, \ldots, p$.

The proof of the Fredholm Alternative relies on the fact that the range of A *annihilates* the null space of the adjoint operator A^* and vice-versa, in the sense that

$$\begin{aligned}
\operatorname{Range}(A) &= \{y \in \mathcal{Y} : (u, y) = 0, \forall u \in \operatorname{Null}(A^*) \subset \mathcal{Y}^*\}, \\
\operatorname{Range}(A^*) &= \{v \in \mathcal{X}^* : (v, x) = 0, \forall x \in \operatorname{Null}(A) \subset \mathcal{X}\}.
\end{aligned}$$

In the context of Hilbert spaces, one can define orthogonality using the scalar product. The conditions of the Theorem then become orthogonality conditions. Returning to an example from linear algebra: if $A \in \mathbb{C}^{n \times n}$ is a matrix, then the Fredholm Alternative states the familiar condition that the linear system of equations $Ax = b$ is solvable if and only if $b \in \operatorname{Range} A$, or $b \perp \operatorname{Null}(A^*)$.

2.5 Boundary value problems

In this section, we define the boundary value problems that will be considered in this thesis. Integral equations can be formulated for boundary value problems of general elliptic partial differential equations [154]. The resulting integral equations are all highly similar. We therefore restrict our attention to boundary value problems for the Helmholtz equation. The theory for the Laplace equation mostly follows from the theory described here by substituting $k = 0$, with some exceptions (see [100]).

We are interested in the solution of the Helmholtz equation on an open and simply connected bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with boundary $\Gamma = \partial\Omega$, or on the exterior domain $\mathbb{R}^d \setminus \overline{\Omega}$. We denote the exterior of the domain by Ω^+, and the interior by Ω^-, such that $\mathbb{R}^d = \Omega^+ \cup \Gamma \cup \Omega^-$. In the remainder of the text, we will adopt the convention that the normal derivative on Γ points into the exterior domain Ω^+. We assume that Γ is at least C^1 continuous, unless mentioned otherwise.

In order to characterise a unique solution to the exterior Helmholtz problem, the solution should satisfy the so-called *Sommerfeld radiation condition* at infinity. This condition states that the solution can only be an *outgoing wave*, by requiring that the solution vanishes at a certain rate near infinity. The Sommerfeld radiation condition for a d-dimensional problem is

$$\lim_{r\to\infty} r^{(d-1)/2} \left| \frac{\partial u}{\partial r} - iku \right| = 0, \qquad r = \|x\|. \tag{2.17}$$

The limit states that the expression between the absolute value lines should vanish faster than $1/\sqrt{r}$ for two-dimensional problems, i.e.,

$$\left| \frac{\partial u}{\partial r} - iku \right| = o(1/\sqrt{r}), \qquad r \to \infty, \tag{2.18}$$

and faster than $1/r$ for three-dimensional problems,

$$\left| \frac{\partial u}{\partial r} - iku \right| = o(1/r), \qquad r \to \infty. \tag{2.19}$$

The boundary value problems are now formally defined. We postpone the characterisation of the function spaces to which the solution and the boundary condition belong to a later stage in the text.

Problem 2.5.1 (Interior Dirichlet). *Find the function u, such that*

$$\begin{aligned} &\Delta u(x) + k^2 u(x) = 0, \qquad x \in \Omega^-, \\ &u(x) = f(x), \qquad x \in \Gamma. \end{aligned}$$

For the exterior Dirichlet problem, we require that the Sommerfeld radiation condition is satisfied.

Problem 2.5.2 (Exterior Dirichlet). *Find the function u, such that*

$$\begin{aligned} &\Delta u(x) + k^2 u(x) = 0, \qquad x \in \Omega^+, \\ &u(x) = f(x), \qquad x \in \Gamma, \end{aligned}$$

and u satisfies the Sommerfeld radiation condition *(2.17).*

The interior and exterior Neumann problem are defined similarly.

Problem 2.5.3 (Interior Neumann). *Find the function u, such that*

$$\begin{aligned} &\Delta u(x) + k^2 u(x) = 0, \qquad x \in \Omega^-, \\ &\frac{\partial u}{\partial n}(x) = g(x), \qquad x \in \Gamma. \end{aligned}$$

Problem 2.5.4 (Exterior Neumann). *Find the function u, such that*

$$\begin{aligned} &\Delta u(x) + k^2 u(x) = 0, \qquad x \in \Omega^+, \\ &\frac{\partial u}{\partial n}(x) = g(x), \qquad x \in \Gamma, \end{aligned}$$

and u satisfies the Sommerfeld radiation condition *(2.17).*

It is also possible to consider boundary value problems with *mixed* boundary conditions. The conditions can, e.g., have the form

$$u(x) + \alpha \frac{\partial u}{\partial n}(x) = f(x), \quad x \in \Gamma,$$

or

$$\begin{cases} u(x) = f(x), & x \in \Gamma_D, \\ \frac{\partial u}{\partial n}(x) = g(x), & x \in \Gamma_N, \end{cases}$$

with $\Gamma = \Gamma_D \cup \Gamma_N$. Boundary conditions of this type are treated in [154], in the context of integral equations.

2.6 Boundary integral equations

Boundary value problems for elliptic partial differential equations can be transformed into a boundary integral equation through the use of Green's identities. This approach has the advantage that the properties of the integral equations, such as uniqueness of the solution, can be established using existing theory for the corresponding differential equation. Much insight into the problem is gained by relating the unknown of the integral equation to a physical interpretation in the context of the differential equation.

2.6.1 Green's identities

Assume that Ω is an open domain with a smooth boundary. The divergence theorem relates the area or volume integral over Ω of a vector function to the boundary integral of its normal component. The theorem is also referred to as *Gauss's theorem*, or as *Green's theorem* in the two-dimensional case.

Theorem 2.6.1 (Divergence theorem). *Let $F : \overline{\Omega} \to \mathbb{R}^d$, $d = 2, 3$, with each component of F in $C^1(\overline{\Omega})$, then*

$$\int_{\Omega} \nabla \cdot F(x)\, \mathrm{d}\sigma_x = \int_{\partial\Omega} n(x) \cdot F(x)\, \mathrm{d}s_x.$$

Green's identities follow from the divergence theorem. First, assume $u \in C^1(\overline{\Omega})$ and $v \in C^2(\overline{\Omega})$. Taking $F = u\nabla v$ yields *Green's first identity*

$$\int_{\Omega} u\Delta v\, \mathrm{d}\sigma_x = \int_{\partial\Omega} u\frac{\partial v}{\partial n}\, \mathrm{d}s_x - \int_{\Omega} \nabla u \cdot \nabla v\, \mathrm{d}\sigma_x. \tag{2.20}$$

If also $u \in C^2(\overline{\Omega})$, then interchanging the roles of u and v in (2.20) and subtracting the two relations yields *Green's second identity*

$$\int_{\Omega} (u\Delta v - v\Delta u)\, \mathrm{d}\sigma_x = \int_{\partial\Omega} (u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n})\, \mathrm{d}s_s. \tag{2.21}$$

Green's identities will be central in deriving an integral representation of the solution of the Helmholtz equation. The identities remain valid if the domain Ω has corners or edges.

2.6.2 An integral representation

In this section, we will motivate the definition of two relevant functions, called the *single-layer potential* and the *double-layer potential.* To that end, we first define the *fundamental solution* of an elliptic partial differential equation. Consider a second order elliptic partial differential equation $Lu = f$. We define the *fundamental solution* of the PDE as the solution to the equation

$$L_x G(x, y) = -\delta(x - y), \tag{2.22}$$

in the sense of distributions, where δ is Dirac's delta function. The fundamental solution is also called the *Green's function* or, in the context of integral equations, the *kernel function*. The notation L_x indicates that the derivatives in the operator L occur with respect to the variable x. For the

Laplace equation, we have the fundamental solutions

$$G(x,y) = \frac{1}{2\pi}\log(|x-y|), \qquad d=2, \tag{2.23}$$

$$G(x,y) = \frac{1}{4\pi}\frac{1}{|x-y|}, \qquad d=3, \tag{2.24}$$

where $|x-y|$ denotes the Euclidian norm of the vector $x-y \in \mathbb{R}^d$. The fundamental solutions for the Helmholtz equation are given by

$$G(x,y) = \frac{i}{4}H_0^{(1)}(k|x-y|), \qquad d=2, \tag{2.25}$$

$$G(x,y) = \frac{1}{4\pi}\frac{e^{ik|x-y|}}{|x-y|}, \qquad d=3, \tag{2.26}$$

with $H_0^{(1)}(z)$ the Hankel function of the first kind and order zero. The above fundamental solutions are singular when $x=y$. This is not necessarily always the case. For example, the fundamental solution for the biharmonic differential equation $\Delta^2 u = 0$ in $\mathbb{R}^2$ is

$$G(x,y) = \frac{1}{8\pi}|x-y|^2\log(|x-y|).$$

This function is singular when $x=y$ only in a higher order derivative.

Now consider the domain Ω with boundary Γ, and the non-homogeneous differential equation

$$Lu = \Delta u + k^2 u = f, \tag{2.27}$$

with possibly $k=0$, i.e., we consider the non-homogeneous Laplace and Helmholtz equation. Recall the *sifting property* of the δ-function,

$$\int_\Omega u(y)\delta(y-x)\,\mathrm{d}\sigma_y = u(x), \qquad x\in\Omega. \tag{2.28}$$

By the defining property of the fundamental solution (2.22), we can write

$$u(x) = -\int_\Omega u(y)(\Delta_y G(x,y) + k^2 G(x,y))\,\mathrm{d}\sigma_y, \qquad x\in\Omega.$$

Using Green's second identity with $v(y)=G(x,y)$, we arrive at

$$\begin{aligned} u(x) = &-\int_\Gamma \left[u(y)\frac{\partial G}{\partial n_y}(x,y) - G(x,y)\frac{\partial u}{\partial n}(y)\right]\,\mathrm{d}s_y \\ &-\int_\Omega \left[\Delta_y u(y) + k^2 u(y)\right] G(x,y)\,\mathrm{d}\sigma_y. \end{aligned}$$

If u is a solution to the non-homogeneous differential equation (2.27), then

$$
\begin{aligned}
u(x) = &- \int_\Gamma \left[u(y) \frac{\partial G}{\partial n_y}(x,y) - G(x,y) \frac{\partial u}{\partial n}(y) \right] \mathrm{d}s_y \\
&- \int_\Omega f(y) G(x,y) \, \mathrm{d}\sigma_y.
\end{aligned}
$$

In the absence of a boundary, the solution $u(x)$ can thus be written as a convolution-type integral of the right-hand side of the non-homogeneous partial differential equation (2.27) with the fundamental solution,

$$u(x) = - \int_\Omega f(y) G(x,y) \, \mathrm{d}\sigma_y, \qquad x \in \Omega.$$

In this sense, the integral involving the fundamental solution can be seen as the inverse of the partial differential operator.

In general, an integral equation involving only integrals over Γ cannot be derived for non-homogeneous partial differential equations. A useful integral representation for the solution of the homogeneous equation, relating the solution to its values on the boundary of the domain, is given in the following Theorem. The non-homogeneous problem can be solved by adding a solution of the homogeneous problem to a particular solution of the non-homogeneous problem.

Theorem 2.6.2 (Green's formula). *Assume Γ is C^1 continuous, and $u \in C^2(\overline{\Omega})$ solves $\Delta u + k^2 u = 0$. Then*

$$u(x) = - \int_\Gamma \left[u(y) \frac{\partial G}{\partial n_y}(x,y) - G(x,y) \frac{\partial u}{\partial n}(y) \right] \mathrm{d}s_y, \qquad x \in \Omega. \tag{2.29}$$

Note that if $x \notin \Omega$, then the right hand side of (2.28) is zero, and we obtain

$$\int_\Gamma \left[u(y) \frac{\partial G}{\partial n_y}(x,y) - G(x,y) \frac{\partial u}{\partial n}(y) \right] \mathrm{d}s_y = 0, \qquad x \in \mathbb{R}^d \setminus \overline{\Omega}. \tag{2.30}$$

The result of Theorem 2.6.2 is that the solution u of the homogeneous equation is completely determined by its values and its normal derivative along the boundary Γ of the domain Ω. These values are sometimes called the *Cauchy data.* The form (2.29) suggests the definition of two functions; they are the *single-layer potential*

$$u(x) = \int_\Gamma G(x,y) q_1(y) \, \mathrm{d}s_y, \tag{2.31}$$

and the *double-layer potential*

$$u(x) = \int_\Gamma \frac{\partial G}{\partial n_y}(x,y) q_2(y) \,\mathrm{d}s_y. \tag{2.32}$$

Theorem 2.6.2 states that each solution to the homogeneous equation can be written in terms of the single-layer potential and the double-layer potential with certain *density functions* $q_1(y)$ and $q_2(y)$. The names of the layer potentials originate in the following property.

Theorem 2.6.3. *The single-layer potential* (2.31) *and double-layer potential* (2.32) *satisfy* $Lu = 0$, $x \notin \Gamma$.

Proof. Since the integrands of the single-layer and double-layer potential are smooth for $x \notin \Gamma$, we can move the derivative inside the integral, and interchange the derivatives. The result is an immediate consequence of (2.22). ☐

The name 'potential' arises from the fact that for the Laplace equation, the layer potentials are harmonic functions or *potentials*. For the Helmholtz equation, the name 'potential' should be understood in a generalised sense. The term single-layer stems from the fact that (2.31) can be seen as the field due to a continuous distribution of field sources with *density* $q_1(y)$, i.e., there is a layer of sources on Γ. The double-layer potential can equivalently be seen as a distribution of dipoles on Γ.

The solution of $Lu = 0$ by the integral representation (2.29) requires the knowledge of both $u|_\Gamma$ and $\frac{\partial u}{\partial n}|_\Gamma$. In general, only one of these functions, or a linear combination of both functions, is known through the boundary condition. In order to solve the differential equation, we need to investigate the limiting behaviour of the integral representation as x approaches the boundary Γ. There are two difficulties associated with this limit: first, the fundamental solution has a singularity when $x = y$, and second, the representation was derived with the assumption that $x \in \Omega$. The behaviour of the layer potentials as x approaches Γ is our next subject.

2.6.3 Jump relations

In this section, we investigate the regularity properties of the single-layer and double-layer potential. A first and useful observation is that both layer potentials for the Helmholtz equation satisfy the Sommerfeld radiation condition at infinity. This means that a solution to the homogeneous equation, written in terms of the layer potentials according to Theorem 2.6.2, automatically satisfies the Sommerfeld radiation condition.

The behaviour of the layer potentials as x approaches Γ is not straightforward. The continuity properties as x crosses the boundary depend in general on the continuity properties of the fundamental solution or kernel function. In general, the higher the smoothness of the kernel function, the higher the smoothness of the potentials. In particular, for singular kernels the potentials may be discontinuous across the boundary. In the following, we state the continuity results for the Helmholtz equation in two and three dimensions. The proofs are rather lengthy and technical in nature, and are omitted. A full description with similar theorems can be found in [49] for the three-dimensional Helmholtz and Laplace problems, in [100] for the n-dimensional Laplace problem, and in [154] for more general strongly elliptic second order partial differential equations.

As it turns out, the single-layer potential is continuous everywhere.

Theorem 2.6.4. *The single-layer potential u given by* (2.31) *is continuous in $x \in \mathbb{R}^d$.*

Now consider the normal derivative of the single-layer potential at Γ, which can be understood as the limit

$$\frac{\partial u^\pm}{\partial n}(x) = \lim_{h\to 0, h>0} \left[n(x) \cdot \nabla u(x \pm h n(x)) \right], \qquad x \in \Gamma,$$

where $n(x)$ represents the outward normal to Γ at the point x. It can be shown that both limits exist, but they are not equal.

Theorem 2.6.5. *For the single-layer potential u with continuous density q_1, we have*

$$\frac{\partial u^\pm}{\partial n}(x) = \int_\Gamma \frac{\partial G}{\partial n_x}(x,y) q_1(y)\, \mathrm{d}s_y \mp \frac{1}{2} q_1(x), \qquad x \in \Gamma, \tag{2.33}$$

where the integral exists as an improper integral.

The following corollary relates the jump in the normal derivative of the single-layer potential to the density function q_1. It is an immediate consequence of Theorem 2.6.5.

Corollary 2.6.6. *For the single-layer potential u with continuous density q_1, we have*

$$\frac{\partial u^-}{\partial n}(x) - \frac{\partial u^+}{\partial n}(x) = q_1(x), \qquad x \in \Gamma, \tag{2.34}$$

$$\frac{1}{2}\left(\frac{\partial u^-}{\partial n}(x) + \frac{\partial u^+}{\partial n}(x) \right) = \int_\Gamma \frac{\partial G}{\partial n_x}(x,y) q_1(y)\, \mathrm{d}s_y, \qquad x \in \Gamma. \tag{2.35}$$

The double-layer potential has a similar jump as the normal derivative of the single-layer potential, differing only in sign.

Theorem 2.6.7. *The double-layer potential u with continuous density q_2 can be continuously extended from Ω^+ to $\overline{\Omega^+}$ and from Ω^- to $\overline{\Omega^-}$ with limiting values*

$$u_\pm(x) = \int_\Gamma \frac{\partial G}{\partial n_y}(x,y) q_2(y)\,\mathrm{d}s_y \pm \frac{1}{2} q_2(x), \qquad x \in \Gamma, \tag{2.36}$$

where the integral exists as an improper integral.

Corollary 2.6.8. *For the double-layer potential u with continuous density q_2, we have*

$$u^-(x) - u^+(x) = -q_2(x), \qquad x \in \Gamma, \tag{2.37}$$

$$\frac{1}{2}\left(u^-(x) + u^+(x)\right) = \int_\Gamma \frac{\partial G}{\partial n_y}(x,y) q_2(y)\,\mathrm{d}s_y, \qquad x \in \Gamma. \tag{2.38}$$

Finally, we consider the normal derivative of the double-layer potential. The normal derivative of the double-layer potential is continuous, but the kernel function is strongly singular. The corresponding integral can only be defined as a Hadamard finite part integral, denoted by *f.p.* (see [89, 177]).

Theorem 2.6.9. *For the double-layer potential u with density $q_2 \in C^2(\Gamma)$, we have*

$$\frac{\partial u^+}{\partial n}(x) = \frac{\partial u^-}{\partial n}(x), \qquad x \in \Gamma,$$

and

$$\frac{\partial u}{\partial n}(x) = \textit{f.p.} \int_\Gamma \frac{\partial^2 G}{\partial n_x \partial n_y}(x,y) q_2(y)\,\mathrm{d}s_y.$$

2.6.4 Boundary integral equations

Based on the definitions and the properties of the layer-potentials, we can define the following four integral operators:

$$(Sq)(x) = \int_\Gamma G(x,y) q(y)\,\mathrm{d}s_y, \tag{2.39}$$

$$(Dq)(x) = \int_\Gamma \frac{\partial G}{\partial n_y}(x,y) q(y)\,\mathrm{d}s_y, \tag{2.40}$$

$$(D^*q)(x) = \int_\Gamma \frac{\partial G}{\partial n_x}(x,y) q(y)\,\mathrm{d}s_y, \qquad x \in \Gamma, \tag{2.41}$$

$$(Nq)(x) = \text{f.p.} \int_\Gamma \frac{\partial^2 G}{\partial n_x \partial n_y}(x,y) q(y)\,\mathrm{d}s_y, \quad x \in \Gamma. \tag{2.42}$$

It can be verified that, for sufficiently smooth boundaries, all kernel functions are only weakly singular, except the kernel function of the so-called *hypersingular operator* N. The notation D and D^* is justified because the operators are indeed adjoint when $x \in \Gamma$. The single-layer and double-layer potential operators S and D are defined for each $x \in \mathbb{R}^d$. They can be restricted to $x \in \Gamma$ by the trace operators γ^+ and γ^-.

Using the results of the previous section, we can generalise the integral representation of Theorem 2.6.2 as follows.

Theorem 2.6.10. *Assume* Γ *is* C^1 *continuous, and* u *solves* $\Delta u + k^2 u = 0$, $x \notin \Gamma$, *and satisfies the Sommerfeld radiation condition* (2.18) *or* (2.19) *at infinity. Denote the jumps of* u *and its normal derivative on* Γ *by*

$$[u] = u^- - u^+, \quad \textit{and} \quad \left[\frac{\partial u}{\partial n}\right] = \frac{\partial u^-}{\partial n} - \frac{\partial u^+}{\partial n}.$$

Then we can write

$$u = S\left[\frac{\partial u}{\partial n}\right] - D[u], \qquad x \notin \Gamma. \tag{2.43}$$

For $x \in \Gamma$ *we have*

$$\frac{u^-(x) + u^+(x)}{2} = S\left[\frac{\partial u}{\partial n}\right] - D[u]. \tag{2.44}$$

Proof. From Theorem 2.6.2 and (2.30), we know that

$$u^- = S\frac{\partial u^-}{\partial n} - Du^-, \qquad x \in \Omega^-, \tag{2.45}$$

$$S\frac{\partial u^-}{\partial n} - Du^- = 0, \qquad x \in \Omega^+.$$

Using a similar reasoning, and the Sommerfeld radiation condition, one can show that similar relations hold for the converse case, (see, e.g., [49])

$$u^+ = -S\frac{\partial u^+}{\partial n} + Du^+, \qquad x \in \Omega^+, \tag{2.46}$$

$$-S\frac{\partial u^+}{\partial n} + Du^+ = 0, \qquad x \in \Omega^-.$$

Together, these relations establish (2.43) for $x \notin \Gamma$.

We can take the limit $x \to \Gamma$ using the results of the previous section. By Theorem 2.6.4 and Theorem 2.6.7, equation (2.45) leads to

$$u^-(x) = (S\frac{\partial u^-}{\partial n})(x) - (Du^-) + u^-(x)/2, \qquad x \in \Gamma.$$

Taking a similar limit in equation (2.46) and adding both relations proves the second result (2.44). ☐

2.6.4.1 Integral equations for the single-layer potential

The jump relations of the layer potentials can be used to obtain boundary integral equations. Using the continuity of the single-layer potential, we find that it can be used to solve the interior and exterior Dirichlet problem simultaneously.

Theorem 2.6.11. *The single-layer potential Sq with continuous density q is a solution of the Interior Dirichlet Problem 2.5.1, provided q is a solution to the integral equation of the first kind*

$$Sq = f. \tag{2.47}$$

In that case, the single-layer potential also solves the Exterior Dirichlet Problem 2.5.2.

Proof. The single-layer potential satisfies the Helmholtz equation in Ω^- by Theorem 2.6.3. By Theorem 2.6.4, we see that the single-layer potential assumes the prescribed boundary condition on Γ. This finishes the proof for the interior Dirichlet problem. The proof for the exterior Dirichlet problem is similar; suffice to note that the Sommerfeld radiation condition is automatically satisfied. ☐

Integral equation (2.47) is a linear Fredholm integral equation of the first kind with the weakly singular kernel $G(x, y)$. A consequence of the integral representation Theorem 2.6.10 is that we can characterise the solution of the equation in terms of variables of the physical problem: the density function corresponds exactly to the jump in the normal derivative of the solution.

Theorem 2.6.12. *If the single-layer potential Sq with density function q satisfies the Helmholtz equation in $\Omega^+ \cup \Omega^-$, then*

$$q = \left[\frac{\partial u}{\partial n}\right], \qquad x \in \Gamma.$$

Proof. Since the single-layer potential is continuous across the boundary Γ, we have $[u] = 0$. The result follows from (2.43). ☐

An integral equation for the Neumann problem can be constructed using the normal derivative of the single-layer potential. Due to the discontinuity jump across Γ, the formulations for the interior and the exterior problem are different.

Theorem 2.6.13. *The single-layer potential Sq with continuous density q is a solution of the Interior Neumann Problem 2.5.3, provided q is a solution to the integral equation of the second kind*

$$\left(\frac{I}{2} + D^*\right) q = g. \tag{2.48}$$

It is a solution for the Exterior Neumann Problem 2.5.4 if

$$\left(-\frac{I}{2}+D^*\right)q=g. \tag{2.49}$$

Proof. The result follows from the limits of the normal derivative of the single-layer potential at Γ, characterised by Theorem 2.6.5. ☐

2.6.4.2 Integral equations for the double-layer potential

Integral equations can also be found using the double-layer potential. The resulting equations have different properties for the same problem. For example, the double-layer potential leads to an integral equation of the second kind for the Dirichlet problem.

Theorem 2.6.14. *The double-layer potential Dq with continuous density q is a solution of the Interior Dirichlet Problem 2.5.1, provided q is a solution to the integral equation of the second kind*

$$\left(-\frac{I}{2}+D\right)q=f. \tag{2.50}$$

It is a solution for the Exterior Dirichlet Problem 2.5.2 if

$$\left(\frac{I}{2}+D\right)q=f. \tag{2.51}$$

The proof is based on the jump relations of Theorem 2.6.7. We can again characterise the solution of the integral equation in terms of physical variables. Since by Theorem 2.6.9 the normal derivative of the double-layer potential is continuous across Γ, we have that $\left[\frac{\partial u}{\partial n}\right]=0$. The following theorem follows from the integral representation Theorem 2.6.10.

Theorem 2.6.15. *If the double-layer potential Dq with density function q solves the Helmholtz problem in $\Omega^+\cup\Omega^-$, then*

$$q=[u], \quad x\in\Gamma.$$

Finally, the double-layer potential can also be used to solve the interior and exterior Neumann problems. This leads to the hypersingular integral equation.

Theorem 2.6.16. *The double-layer potential Dq with density $q\in C^2(\Gamma)$ is a solution of the Interior Neumann Problem 2.5.3, provided q is a solution to the integral equation of the first kind*

$$Nq=g. \tag{2.52}$$

In that case, it also solves the Exterior Neumann Problem 2.5.4.

Table 2.1: Summary of the constructed boundary integral equations.

	$u = Sq$	$u = Dq$
Dirichlet problem		
interior	$Sq = f$	$(-\frac{I}{2} + D)q = f$
exterior	$Sq = f$	$(\frac{I}{2} + D)q = f$
Neumann problem		
interior	$(\frac{I}{2} + D^*)q = g$	$Nq = g$
exterior	$(-\frac{I}{2} + D^*)q = g$	$Nq = g$

Integral equation (2.52) involves the strongly singular kernel function $\frac{\partial^2 G}{\partial n_x \partial n_y}(x, y)$. The integral operator can only be evaluated in the sense of Hadamard finite part integration. For this reason, equation (2.52) is called the *hypersingular* integral equation. Since a constant density has no influence on the value of the Hadamard integral, a normalisation condition

$$\int_\Gamma q(x)\,\mathrm{d}s_x = 0$$

is imposed. The singularity needs to be regularised before computations can be performed [158].

A summary of the constructed integral equations is given in Table 2.1. All these formulations were based on representing the solution as either a single-layer potential or a double-layer potential. The equations are therefore called *indirect* integral equations. We note that it is also possible to derive so-called *direct* integral equations, by considering the limit case $x \to \Gamma$ directly for Green's formula in Theorem 2.6.2. The resulting equations are called *direct*, since they yield the Dirichlet or Neumann data of the unknown function as the solution, without an intermediate evaluation of a potential operator. The equations themselves however typically contain several applications of the integral operators.

2.6.5 Mapping properties of the integral operators

The motivation for finding the exact domain and range of the integral operators defined in §2.6.4 does not lie solely in purely theoretical considerations. Indeed, we will see that the properties of the function spaces involved relate directly to important numerical properties of the solution methods, such as the condition number of the discretisation matrix. From a theoretical point of view, the exact domain and range of the operators serve to prove existence and uniqueness results. The importance is summarised in the following lemma.

Lemma 2.6.17 ([154],Cor.2.2). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. If $A \in L(\mathcal{X}, \mathcal{Y})$, then the following conditions are equivalent:*

(i) The subspace $\operatorname{Range} A$ *is closed in* $\mathcal{Y}$.

(ii) The induced map $A_/ : \mathcal{X}/\operatorname{Null}(A) \to \operatorname{Range}(A)$ *has a bounded inverse.*

In particular, there exists a bounded inverse $A^{-1} \in L(\mathcal{Y}, \mathcal{X})$ if and only if $\operatorname{Range}(A) = \mathcal{Y}$ *and* $\operatorname{Null}(A) = \{0\}$.

Given an integral or differential operator A, one wants to find spaces $\mathcal{X}$ and $\mathcal{Y}$ such that $A : \mathcal{X} \to \mathcal{Y}$ is bounded and $\operatorname{Range}(A)$ is closed. If $\mathcal{Y}$ is too big relative to $\mathcal{X}$, than the image of A will fail to be closed. If $\mathcal{Y}$ is too small, than restricting $\mathcal{X}$ to a suitable subspace typically yields an unbounded operator. If both spaces match for the particular operator, then the equation $Au = f$ is solvable, up to a an element of the null space of A.

Suitable function spaces are given by the Sobolev spaces on the boundary Γ, as defined in §2.4.1 for general Lipschitz domains. This characterisation allows the definition of the *order* of an operator.

Definition 2.6.18. *The* order *of an operator A is r if $A : H^s \to H^{s-r}$ satisfies the conditions (i) and (ii) in Lemma 2.6.17.*

The following characterisations were derived in [51]. The single-layer potential and double-layer potentials for general Lipschitz domains give rise to bounded linear operators

$$\begin{aligned} S &: H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma), & D &: H^{1/2}(\Gamma) \to H^{1/2}(\Gamma), \\ D^* &: H^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma), & N &: H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma). \end{aligned} \tag{2.53}$$

The order of the operator S is therefore $r = -1$. The negative order means that applying the integral operator results in a function that is smoother than the density function. This corresponds to the smoothing process of integration. The mapping properties (2.53) can be extended to hold for a whole range of Sobolev spaces. If $-1/2 \leq s \leq 1/2$, then we have

$$\begin{aligned} S &: H^{s-1/2}(\Gamma) \to H^{s+1/2}(\Gamma), & D&\colon H^{s+1/2}(\Gamma) \to H^{s+1/2}(\Gamma), \\ D^* &: H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma), & N&\colon H^{s+1/2}(\Gamma) \to H^{s-1/2}(\Gamma). \end{aligned}$$

The range of Sobolev spaces increases with increasing smoothness of the boundary Γ. In particular, if Γ is C^{m+1} for an integer index $m \geq 0$, then the mapping properties hold for the extended range $-m \leq s \leq m$.

2.6.6 Uniqueness and resonance frequencies

The existence and uniqueness of the solutions to the equations in Table 2.1 can be established using the results from the overview of the Riesz-Fredholm theory in §2.4.3. For example, Theorem 2.4.5 applies to the integral equations of the second kind of the form $I - K$, where the integral operator K is a compact operator. The theorem states that showing injectivity of $I - K$ is sufficient for the existence of a bounded inverse, and hence, for the solvability of $(I - K)q = f$ for each function f. If $I - K$ is not injective, necessary and sufficient conditions for the solvability are given by the Fredholm Alternative Theorem 2.4.9.

However, the Fredholm Alternative does not state that there is a unique solution to the equation $Au = f$. Indeed, the null space of the operator A may be non-trivial. For the Helmholtz equation, there are certain *resonance* frequencies for which the interior problem is not uniquely solvable. These resonance frequencies are related to the eigenvalues of the Laplacian Δ. Assume that

$$\begin{aligned} \Delta u(x) &= -k^2 u(x), && x \in \Omega^-, \\ u(x) &= 0, && x \in \Gamma, \end{aligned}$$

i.e., $-k^2$ is an eigenvalue of the Laplacian on Ω^- with eigenfunction u. The eigenfunction u satisfies the Helmholtz equation in Ω^- with homogeneous Dirichlet boundary conditions. Hence, the eigenfunction lies in the null space of the operator S, and S is not invertible for these critical values. The interior Dirichlet problem is not uniquely solvable. Moreover, the exterior Dirichlet problem can not be solved using the integral equation

$$Sq = f, \qquad \text{(interior and exterior problem)}$$

although one can show that the exterior Dirichlet problem has a unique solution for each value of the wavenumber.

For the same eigenvalue of the interior Dirichlet problem for the Laplacian, the exterior Neumann problem can not be solved using the single-layer potential, i.e., the integral equation of the second kind

$$(-\frac{I}{2} + D^*)q = g, \qquad \text{(exterior problem)}$$

is not solvable for these critical values. In order to see why, consider a function u that is zero in Ω^+, and that coincides with the interior Dirichlet eigenfunction w in Ω^-, such that $[u] = 0$ and $\frac{\partial u^+}{\partial n} = 0$. Then, by (2.43) we have

$$S\frac{\partial w}{\partial n} = 0, \qquad x \in \Omega^+,$$

Table 2.2: Values of the wavenumber for which the integral equations in Table 2.1 are not solvable. The notation $\sigma(\Delta^D)$ denotes the eigenvalues of the interior Dirichlet problem of the Laplacian. Similarly, $\sigma(\Delta^N)$ is used to denote the eigenvalues of the interior Neumann problem.

	$u = Sq$	$u = Dq$
Dirichlet problem		
interior	$-k^2 \in \sigma(\Delta^D)$	$-k^2 \in \sigma(\Delta^D)$
exterior	$-k^2 \in \sigma(\Delta^D)$	$-k^2 \in \sigma(\Delta^N)$
Neumann problem		
interior	$-k^2 \in \sigma(\Delta^N)$	$-k^2 \in \sigma(\Delta^N)$
exterior	$-k^2 \in \sigma(\Delta^D)$	$-k^2 \in \sigma(\Delta^N)$

and by the exterior limit in Theorem 2.6.5 we have

$$(-\frac{I}{2} + D^*)\frac{\partial w}{\partial n} = 0. \tag{2.54}$$

It follows that there exists a non-trivial solution to integral equation (2.54). Similar problems occur for the remaining integral equations in Table 2.1, either for the eigenvalues of the interior Dirichlet problem or the interior Neumann problem of the Laplacian. These critical values present a fundamental problem for the solution of the exterior Helmholtz problems, although a solution exists uniquely. The problems are summarised in Table 2.2.

A solution was suggested by Brakhage and Werner in [28], and independently by Leis and by Panich in [143, 165]. They show that the solution to the Helmholtz equation can be formulated as a linear combination of single-layer and double-layer potentials,

$$u(x) = \int_\Gamma \left[i\eta G(x,y) - \frac{\partial G}{\partial n_y}(x,y) \right] q(y) \, \mathrm{d}s_y.$$

Integral equations can be derived from this representation using the jump relations of §2.6.3. For example, an integral equation for the exterior Neumann problem is given by

$$(\frac{I}{2} - D - i\eta S)q = g. \tag{2.55}$$

Integral equation (2.55) is uniquely solvable if the *coupling parameter* η is nonzero and real. A disadvantage of this method is that the density function q is no longer easily associated with physical variables. A different approach is given by the *combined field integral equation*,

$$(i\eta S + \frac{I}{2} + D^*)q = i\eta f + g, \tag{2.56}$$

proposed by Harrington and Mautz [109]. The correct boundary conditions f and g can both be derived from the incoming wave. The solution to the Dirichlet problem is given by the single-layer potential Sq. The density function q is therefore related to the physical variables by Theorem 2.6.12. For numerical purposes, the coupling parameter η in both (2.55) and (2.56) should be chosen proportional to the wavenumber k [7].

Finally, we note that the eigenvalues of the Laplacian are always real, because Δ is self-adjoint. Hence, $-k^2$ can never be an eigenvalue for complex wavenumbers. The integral equations for complex values of k are always uniquely solvable.

2.7 The boundary element method

2.7.1 The boundary element method

The boundary element method is essentially a finite element method, applied to the variational formulation of a boundary integral equation. A general variational formulation for the operator equation $Au = f$, with $A : V \to V^*$ and $f \in V^*$, is the following: *find an element $u \in V$ such that*

$$a(u, v) = l_f(v), \qquad \forall v \in V. \tag{2.57}$$

The sesquilinear form a is defined by $a(u, v) = (Au, v)$, the functional l_f is given by $l_f(v) = (f, v)$, $v \in V$. When $V = H^\alpha$ is a Sobolev space, then one can simply use the L_2 inner product $(\cdot, \cdot)_{L_2} \in L_2 \times L_2 \to \mathbb{C}$, rather than the duality pairing $(\cdot, \cdot) \in H^{-\alpha} \times H^\alpha \to \mathbb{C}$. The two forms agree if both arguments are elements of L_2.

Let the spaces V_h be a family of finite-dimensional subspaces of V. The Galerkin method for solving (2.57) is: *find an element $u_h \in V_h$ such that*

$$a(u_h, v_h) = l_f(v_h), \qquad \forall v_h \in V_h. \tag{2.58}$$

Assume that the space V_h is spanned by N_h basis functions $\phi_{h,i}(x)$, $i = 1, \ldots, N_h$. The term *boundary element* is used to denote a basis function $\phi_{h,i}$. The Galerkin formulation leads to a linear system of equations

$$M_h x = b_h, \tag{2.59}$$

with elements given by

$$M_{h,i,j} = a(\phi_{h,j}, \phi_{h,i}), \tag{2.60}$$

and with the right hand side b_h given by $b_{h,i} = l_f(\phi_{h,i})$.

Consider for example integral equation (2.47) of the first kind $Sq = f$. Using the L_2 inner product, the elements of the discretisation matrix are given explicitly by

$$M_{h,i,j} = \int_\Gamma \int_\Gamma G(x, y)\phi_{h,j}(y)\phi_{h,i}(x)\,\mathrm{d}s_y\,\mathrm{d}s_x, \tag{2.61}$$

and the right hand side elements by

$$b_{h,i} = \int_\Gamma f(x)\phi_{h,i}(x)\,\mathrm{d}s_x. \tag{2.62}$$

In order to construct basis functions on the boundary Γ, we introduce a parameterisation $\kappa : [0, 1]^{d-1} \to \Gamma$. We define basis functions in the parameter domain $\Box = [0, 1]^{d-1}$. Expression (2.61) for the elements becomes

$$M_{h,i,j} = \int_\Box \int_\Box G(\kappa(t), \kappa(\tau))\phi_{h,j}(\tau)\phi_{h,i}(t)|\nabla\kappa(\tau)||\nabla\kappa(t)|\,\mathrm{d}\tau\,\mathrm{d}t. \tag{2.63}$$

The definitions can be extended to the more general case where $\Gamma = \cup\Gamma_i$, and each part Γ_i has a separate parameterisation.

2.7.2 Convergence

The convergence rate of the Galerkin method depends on the order of the operator, on the smoothness of the boundary Γ and on the smoothness of the basis functions of V_h. The latter is expressed by the notion of *regularity.*

Definition 2.7.1. *A function f has* regularity γ *if*

$$\gamma = \sup\{s : f \in H^s\}. \tag{2.64}$$

Assume that γ is the regularity of the basis functions $\phi_{h,i}$, and that the function space V_h contains all polynomials of degree $d-1$. Let the operator $A : H^\alpha \to H^{-\alpha}$ have order $r = 2\alpha$, and let the index h denote the mesh width of an approximately uniform grid for Γ or the parameter domain $\Box$. Then an asymptotic optimal convergence rate is given in the following theorem.

Theorem 2.7.2. *Assume that $u_h \in V_h$ is the solution of the Galerkin formulation* (2.58)*, and u is the exact solution. Then we have*

$$\|u - u_h\|_{H^t(\Gamma)} \le ch^{s-t}\|u\|_{H^s(\Gamma)}, \tag{2.65}$$

for $-d + r \le t < \gamma$, $t \le s$ and $r/2 \le s \le d$.

Consider a discretisation for the integral equation of the first kind $Sq = f$ with piecewise linear basis functions. We have $r = -1$ and $d = 2$. The optimal rate of convergence is given by $t = -d + r = -3$ and $s = d = 2$, which leads to a convergence rate of h^5 in the norms given by Theorem 2.7.2.

2.7.3 Conditioning

The order of an operator has an influence on the condition number of the discretisation matrix M_h. Assume that $N = h^{-(d-1)}$ is the number of basis functions on an approximately uniform grid for $\Gamma \subset \mathbb{R}^{d-1}$ with mesh width h. Using the $L_2(\Gamma)$ inner product for the computation of the matrix elements (2.60), one can show that the condition number of M_h behaves as (see [115, 116])

$$\kappa(M_h) = O(h^{-|r|}) = O(N^{|r|/(d-1)}). \tag{2.66}$$

For $r = -1$ in two-dimensional problems (with a one-dimensional boundary Γ), the condition number increases linearly with N. Without additional measures, the discretisation of the integral equation of the first kind $Sq = f$ is therefore ill-conditioned. This property is related to the decay of eigenvalues of a compact operator, as discussed in Theorem 2.4.4.

The discretised systems may still be solvable with sufficient accuracy for applications. However, the conditioning (2.66) will have an adverse impact on the convergence of iterative methods, since the number of iterations in such methods depends on the size of the condition number. Integral equations of the second kind are well conditioned, since they have order $r = 0$. Increasing the number of unknowns does not have an impact on the condition number of the resulting linear system, i.e., $\kappa(M_h) = O(1)$, uniformly in h and N for integral equations of the second kind. The condition number may still depend on the wavenumber of the Helmholtz equation.

The ill-conditioning of integral equations of the first kind can be avoided by taking the mapping properties of the operator into account. The use of basis functions in L_2 corresponds to a discretisation for the compact operator $S : L_2(\Gamma) \to L_2(\Gamma)$, which leads to the ill-conditioned representation. Using appropriate basis functions of Sobolev spaces, one may regard the operator as an isomorphism $S : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$, which is numerically more stable. We will see that wavelets are basis functions for a whole range of Sobolev spaces, and they can be used to obtain a well-conditioned representation for the isomorphism. This will be described in Chapter 3.

As we will not consider other preconditioning techniques in detail, we include a number of references that cover different approaches. A wavelet preconditioner can be used in combination with fast multipole methods or hierarchical matrix methods for the matrix-vector product [174]. Alternatively, the LU decomposition of a coarse hierarchical matrix of the problem can be used [17]. So-called sparse approximate inverses have been combined with fast multipole methods to solve large problems in [33]. Finally, some techniques to reduce the condition number corresponding to irregular meshes are explored in [6, 94].

Chapter 3

Wavelet based methods

3.1 Introduction

The first multiscale method we consider for the solution of scattering problems is the wavelet method. The approximation properties of wavelets are used to obtain a sparse representation of an integral operator. In this chapter, we describe the wavelet method, and analyse its behaviour for increasing wavenumbers. We describe a new approach for scattering problems in the high frequency regime, and we develop wavelet-specific quadrature rules for an efficient implementation of wavelet based methods.

The wavelet method was originally introduced by Beylkin, Coifman and Rokhlin in [19]. These authors noted that the discretisation matrix can be compressed to a sparse matrix in the wavelet basis. Preconditioning schemes for integral operators of nonzero order were introduced in [55, 59, 60]. The analysis focused on schemes that match the error of the scheme in each step to the discretisation error of the Galerkin method. An implementation with $O(N \log^4 N)$ computational complexity was presented in [197, 196]. The complexity was reduced to $O(N)$ by considering a so-called *second compression* step and adapted quadrature rules in [175]. Recent research is conducted towards adaptive wavelet schemes [41, 53, 179].

In the low frequency regime, the wavelet method enables a matrix-vector product in $O(N)$ operations, where N is the number of basis functions. We show that the sparsity of the wavelet representation is lost in the presence of strong oscillations, leading to an $O(N^2)$ method in the high frequency regime. We propose a new approach based on the use of wavelet packets. Wavelet packets combine the typical subdivision of scale and position in wavelet analysis with a subdivision of frequency. As such, wavelet packet basis functions can be very oscillatory functions themselves. An adaptive algorithm enables the sparse representation of an oscillatory integral operator. We show that the complexity of the matrix-vector product can be reduced to approximately $O(N^{1.4})$.

We start the chapter with an overview of the wavelet method in §3.2. We recall the definition and the main properties of wavelets in §3.3. Subsequently, we give a more detailed overview of the theory of wavelet based solution methods for integral equations in §3.4. We prove in §3.5 that the complexity of the matrix-vector product of the wavelet method is $O(N^2)$ in the high frequency regime. We propose a remedy for this behaviour that is based on wavelet packets in §3.6. Next, we discuss an efficient implementation of quadrature techniques for integrals involving wavelet functions in §3.7. Finally, we end the chapter with references for three-dimensional problems, and with some concluding remarks.

3.2 Overview of the method

The approximation properties of wavelets can be used in the context of integral equations by considering wavelets as basis functions in a boundary element approach. The resulting discretisation is a multiscale representation of the integral operator A, with elements of the form

$$W_{(j,k),(j',k')} = \langle A\psi_{j',k'}, \psi_{j,k} \rangle, \tag{3.1}$$

where $\psi_{j,k}$ represents a wavelet function on scale j. Alternatively, matrix W can be found by applying the wavelet transform to the regular discretisation matrix M, if the regular basis functions correspond to the *scaling function* of the wavelet. Due to the approximation properties of wavelets and the smoothness of the kernel function away from the diagonal, many elements of the form (3.1) are small. They can be discarded, without introducing a significant error. The fast matrix-vector product is then an immediate result of the sparsity of the compressed discretisation matrix. One can show that the number of remaining significant elements is $O(N \log N)$ or even $O(N)$, depending on the wavelets that are used, where N is the number of unknowns in the boundary element method.

Applying the wavelet transform to the regular discretisation matrix and to the right hand side of the linear system $Mx = b$ yields

$$Wy = c, \qquad \text{with} \quad W = TMT^T, \quad c = Tb, \tag{3.2}$$

where T represents the matrix corresponding to the wavelet transform. The solution of the original system $Mx = b$ is given by $x = T^T y$. Application of the matrix T or T^T can be performed by the fast wavelet transform in $O(N)$ operations. An iterative solver can be used to find the solution of (3.2) in, say, L iterations. The total time required to solve (3.2) is then $O(LN)$.

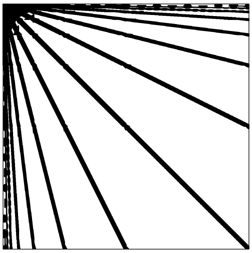


Figure 3.1: Illustration of a discretisation matrix in the wavelet basis. The lines correspond to the singularity of the kernel function on all scales.

In order to have an efficient total solution method, one requires the efficient construction of the wavelet transformed discretisation matrix W, and a bound on the number of iterations L. Obviously, the transformation $W = TMT^T$ is too expensive, even if the fast wavelet transform is used, because constructing the dense matrix M requires at least $O(N^2)$ operations. Instead, the elements of W are computed directly, using the representation (3.1). A priori estimates for the size of the elements are available, so that small elements need not be computed. The efficient construction of W hence requires good integration routines that can compute $O(N)$ significant elements using only $O(N)$ computations. Such routines are available, but they are not trivial, because the integration domain can be large for wavelets on rough scales, and the integrals (3.1) may be singular. The condition number of W can be bounded independently of N with a simple diagonal preconditioner. This result is possible by considering the mapping properties of an integral operator in terms of Sobolev spaces, and using appropriately scaled wavelet bases that are, in fact, stable basis functions for these Sobolev spaces. The number of iterations L required by the iterative solver is then independent of N. Hence, the total solution time for the integral equation is $O(N)$. A typical discretisation matrix is shown in Figure 3.1.

3.3 Wavelets

The theory of wavelets is described extensively in [67]. In this section, we recall the basic properties of multiresolution analysis and wavelets. Specifically, consider a nested series of function spaces

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots ,$$

such that their union is dense in L_2, and with the properties

$$f(t) \in V_j \Rightarrow f(2t) \in V_{j+1}, \quad \text{and}$$
$$f(t) \in V_0 \Rightarrow f(t-k) \in V_0.$$

If the set $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ forms a Riesz basis for V_0, then we call the function spaces V_i a *multiresolution analysis* for L_2. The function ϕ is called the *scaling function*; it satisfies the two-scale relation

$$\phi(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \phi(2t-k), \tag{3.3}$$

with suitable coefficients h_k. The scaling function can be defined on all scales by scaling and translating ϕ,

$$\phi_{jk}(t) := 2^{j/2}\phi(2^j t - k). \tag{3.4}$$

Define W_j as the complement of V_j in V_{j+1},

$$V_{j+1} = V_j \oplus W_j \quad \text{and} \quad V_J = V_0 \oplus_{j=0}^{J-1} W_j. \tag{3.5}$$

There exists a *wavelet* function $\psi(t) \in W_0$ such that $\{\psi(t-k)\}_{k\in\mathbb{Z}}$ is a Riesz basis for W_0. We can define wavelets on all scales similar to (3.4),

$$\psi_{jk}(t) := 2^{j/2}\psi(2^j t - k). \tag{3.6}$$

Since $\psi(t) \in V_1$, there exist coefficients g_k such that

$$\psi(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} g_k \phi(2t-k). \tag{3.7}$$

Any function $f \in V_J$ can be expanded in the basis of scaling functions on scale J, or in the basis suggested by (3.5),

$$f = \sum_k v_{Jk}\phi_{Jk} \quad \text{and} \quad f = \sum_k v_{0k}\phi_{0k} + \sum_{j=0}^{J-1}\sum_k w_{jk}\psi_{jk}. \tag{3.8}$$

One can go from one representation to the other by using the fast wavelet transform and its inverse in $O(N)$ operations, where N is the number of coefficients v_{Jk}, $k = 1, \ldots, N$.

The scaling function ϕ and wavelet ψ have a regularity γ (see (2.64)). The scaling functions are said to have approximation order d if all polynomials of maximal degree $d-1$ are in V_0. Wavelets typically have a certain number $\tilde{d}$ of vanishing moments,

$$\int_{-\infty}^{\infty} \psi(t)t^l \,\mathrm{d}t = 0, \qquad l = 0, \ldots, \tilde{d}-1. \tag{3.9}$$

If the scaling functions $\phi(t-k)$ are not orthogonal, a dual set of basis functions $\tilde{\phi}(t-k)$ exists with approximation order $\tilde{d}$, and dual wavelet functions $\tilde{\psi}(t-k)$ exist with d vanishing moments and regularity $\tilde{\gamma}$. In that case, the following biorthogonality relations hold,

$$\begin{aligned}
&(\phi_{jk}, \tilde{\phi}_{jk'}) = \delta_{k-k'}, && (\psi_{jk}, \tilde{\psi}_{j'k'}) = \delta_{k-k'}\delta_{j-j'}, \\
&(\phi_{jk}, \tilde{\psi}_{jk'}) = 0, && (\tilde{\phi}_{jk}, \psi_{jk'}) = 0.
\end{aligned}$$

A popular choice of basis functions in boundary element methods are CDF-wavelets [43]. They are biorthogonal wavelets, and the scaling function corresponds to a B-spline. The wavelets are therefore piecewise polynomial. Wavelets on a periodic one-dimensional boundary Γ can be constructed using periodic wavelets on $[0,1]$, and a parameterisation $\kappa : [0,1] \to \Gamma$,

$$\hat{\psi}_{jk}(x) = \psi_{jk}(\kappa^{-1}(x)), \qquad x \in \Gamma. \tag{3.10}$$

3.4 Theory of wavelet based methods

We describe the theory of wavelet based methods for the discretisation and solution of integral operator equations. A detailed discussion can be found in [52, 54]. For the purposes of subsequent sections, we will write some constants in the form $C(k)$, to denote explicitly that they depend on the wavenumber of the Helmholtz equation. The exact dependence will be quantified in §3.5. We restrict the discussion to two-dimensional problems.

3.4.1 Element size estimates

The elements of the form (3.1) are given explicitly by a double integral of the form (2.63), with basis functions $\psi_{jk}(t)$ and $\psi_{j'k'}(\tau)$. In order to avoid having to compute matrix entries that will later be discarded, a priori estimates are derived for the size of an entry. Elements are not computed if they are predicted to be smaller than a prescribed threshold value. Suitable estimates can be developed using a property that expresses the smoothness of the kernel function away from the diagonal,

$$\left| \frac{\partial^{|\alpha|+|\beta|} G}{\partial x^\alpha \partial y^\beta}(x,y) \right| \leq C_1(k) |x-y|^{-(n+r+|\alpha|+|\beta|)}, \qquad x \neq y, \tag{3.11}$$

which is developed in the theory of pseudodifferential operators [52, 178]. The orders α and β of the derivative are written in multi-index form, i.e., $\partial x^\alpha = \partial x_1^{\alpha_1} \partial x_2^{\alpha_2}$ with $|\alpha| = \alpha_1 + \alpha_2$. The integer n is the dimension of the boundary manifold, and r is the order of the operator. For two-dimensional obstacles, we have $n = 1$.

Define the modified kernel function $\tilde{G}(t,\tau) := G(\kappa(t),\kappa(\tau))$. An estimate similar to (3.11) can be found for $\tilde{G}(t,\tau)$,

$$\left|\frac{\partial^{\alpha+\beta}\tilde{G}}{\partial t^{\alpha}\partial\tau^{\beta}}(t,\tau)\right| \leq C_2(k)|\kappa(t)-\kappa(\tau)|^{-(n+r+\alpha+\beta)}, \qquad t\neq\tau. \tag{3.12}$$

The parameters α and β in (3.12) are scalars. An estimate for the size of an element can now be found by developing $\tilde{G}$ into a Taylor series. The first terms are cancelled by the vanishing moment properties of the basis functions. If the wavelets in t and τ have disjunct support, then the size of the corresponding element in the stiffness matrix can be bounded by

$$|\langle A\hat{\psi}_{j'k'},\hat{\psi}_{jk}\rangle| \leq C_3(k)\frac{2^{-(j+j')(\frac{n}{2}+\tilde{d})}}{\mathrm{dist}(\mathrm{supp}\ \hat{\psi}_{jk},\mathrm{supp}\ \hat{\psi}_{j'k'})^{n+2\tilde{d}+r}}. \tag{3.13}$$

The support of the wavelet $\hat{\psi}_{jk}$ is denoted by $\mathrm{supp}\,\hat{\psi}_{jk} \subset \Gamma$. The bound (3.13) decreases very rapidly with increasing distance between the supports of the basis functions. Wavelets with a higher number of vanishing moments $\tilde{d}$ yield smaller elements. Such wavelets typically also have a larger support however.

When the supports of the wavelets ψ_{jk} and $\psi_{j'k'}$ overlap, the size of the corresponding element can still be small if the difference in scale is large enough, and if the smaller wavelet is contained entirely in an interval defined by two successive singular points of the larger wavelet. The singular points are those points in the support where the basis function or its derivatives are discontinuous, and are denoted by $\mathrm{sing\,supp}\,\hat{\psi}_{jk}$. For $j>j'$, one has

$$|\langle A\hat{\psi}_{j'k'},\hat{\psi}_{jk}\rangle| \leq C_4(k)\frac{2^{-j(\frac{n}{2}+\tilde{d})}2^{j'\frac{n}{2}}}{\mathrm{dist}(\mathrm{sing\,supp}\ \hat{\psi}_{j'k'},\mathrm{supp}\ \hat{\psi}_{jk})^{\tilde{d}+r}}. \tag{3.14}$$

A similar estimate can be derived for the case $j'>j$.

3.4.2 Compression of the discretisation matrix

Based on estimates (3.13) and (3.14), a compression scheme can be devised to approximate the stiffness matrix in the wavelet basis by a sparse matrix. In an optimal scheme, the error introduced by the compression is matched to the discretisation error of the entire method. To that end, one defines two scale-dependent thresholds,

$$\delta_{j,j'} = \max\left\{a2^{-\min\{j,j'\}}, a2^{\frac{J(2d'-r)-(j+j')(\tilde{d}+d')}{2\tilde{d}+r}}\right\}, \quad \text{and} \tag{3.15}$$

$$\delta^S_{j,j'} = \max\left\{a'2^{-\max\{j,j'\}}, a'2^{\frac{J(2d'-r)-\max\{j,j'\}\tilde{d}-(j+j')d'}{\tilde{d}+r}}\right\}. \tag{3.16}$$

The two constants a and a' determine the amount of compression, and have to be selected carefully, see [175]. Too large a value for a and a' will lead to a denser matrix; not all of its elements may be needed for the required accuracy of the solution. Too small a value results in loss of convergence.

The first threshold (3.15), based on estimate (3.13), gives rise to a sparse matrix M_J^ϵ with elements

$$m^\epsilon_{(j,k),(j',k')} := \begin{cases} m_{(j,k),(j',k')} & \text{if } \operatorname{dist}(\operatorname{supp}\hat{\psi}_{jk}, \operatorname{supp}\hat{\psi}_{j'k'}) \le \delta_{j,j'}, \\ 0 & \text{otherwise.} \end{cases}$$

The number of remaining elements is linear in $N_J = 2^J$, up to a logarithmic factor. The second estimate (3.12) enables a second compression step, by making a sparse matrix $\hat{M}_J$ with elements $\hat{m}_{(j,k),(j',k')}$ defined by

$$\begin{cases} m^\epsilon_{(j,k),(j',k')} & \text{if } j' < j \text{ and } \operatorname{dist}(\operatorname{supp}\hat{\psi}_{jk}, \operatorname{sing\,supp}\hat{\psi}_{j'k'}) \le \delta^S_{j,j'}, \\ m^\epsilon_{(j,k),(j',k')} & \text{if } j < j' \text{ and } \operatorname{dist}(\operatorname{sing\,supp}\hat{\psi}_{jk}, \operatorname{supp}\hat{\psi}_{j'k'}) \le \delta^S_{j,j'}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.17}$$

This second compression leads to an order of $O(N_J)$ remaining significant entries, without a logarithmic term [175]. For this optimal case, it is required that $\tilde{d} > d - r$. For operators with negative order, this means that $\tilde{d} > d$: the primal wavelet needs to have more vanishing moments than the dual wavelet. This condition can only be satisfied by biorthogonal wavelets.

3.4.3 Diagonal preconditioning

Recall from the discussion in §2.7.3 that an operator $A : H^\alpha \to H^{-\alpha}$ with nonzero order $r = 2\alpha$ has a condition number that grows as $O(N_J^{|r|}) = O(2^{-J|r|})$. This is due to the shift in Sobolev spaces. As it turns out, wavelets are stable basis functions for a whole range of Sobolev spaces ([54]),

$$\begin{aligned} \|v\|^2_{H^t} &\sim \textstyle\sum_j \sum_k 2^{2jt} |\langle v, \psi_{jk}\rangle|^2, \qquad t \in (-\tilde{\gamma}, \gamma), \\ \|v\|^2_{H^t} &\sim \textstyle\sum_j \sum_k 2^{2jt} |\langle v, \tilde{\psi}_{jk}\rangle|^2, \qquad t \in (-\gamma, \tilde{\gamma}), \end{aligned} \tag{3.18}$$

where the notation $a \sim b$ means the existence of constants $c, C > 0$ such that $a \le cb$ and $b \le Ca$. This fact can be used to represent the bijective operator $A : H^\alpha \to H^{-\alpha}$, rather than the compact operator $A : L_2 \to L_2$. To that end, the wavelets need to be scaled with a level-dependent scaling factor. Define the diagonal matrix D_J^s as

$$D^s_{J,(j,k),(j',k')} = 2^{sj} \delta_{j,j'} \delta_{k,k'}. \tag{3.19}$$

Table 3.1: Conditions for optimality in the wavelet method.

$\gamma > r/2$	conformity of the basis functions
$\tilde{\gamma} > -r/2$	preconditioning
$\tilde{d} > d - r$	optimal compression
d	convergence rate $2^{-J(2d-r)}$

Multiplication with the diagonal matrix D_J^s corresponds to a scaling of a wavelet coefficient with the scale-dependent factor 2^{sj}. The following result was found in various levels of generality in [18, 55, 57, 60, 133, 164].

Theorem 3.4.1 ([60]). *Assume that the Galerkin discretisation is stable,*

$$b\|v\|_{H^\alpha} \leq \|\sum_j \sum_k \langle Av, \psi_{jk}\rangle \tilde{\psi}_{jk}\|_{H^{-\alpha}} \leq B\|v\|_{H^\alpha}, \quad \forall v \in V_J,$$

with $b, B > 0$, and that the wavelet basis is stable as in (3.18) with $\gamma > r/2$ and $\tilde{\gamma} > -r/2$. Then the preconditioned matrices

$$P_J = D_J^{-\alpha} W_J D_J^{-\alpha} \tag{3.20}$$

have a uniformly bounded spectral condition number.

3.4.4 Convergence

The compression strategy was devised such that the compression error has the same order as the discretisation error of the entire scheme. As such, it has no influence on the accuracy of the overall solution method. Likewise, the compression has no influence on the convergence rate. It can be shown that the optimal convergence rate of Theorem 2.7.2 is preserved: one has

$$\|u - u_h\|_{H^t(\Gamma)} \leq c 2^{-J(s-t)} \|u\|_{H^s(\Gamma)},$$

for $-d + r \leq t < \gamma$, $t \leq s$ and $r/2 \leq s \leq d$. The approximation order d determines the optimal convergence rate $2^{-J(2d-r)}$. Recall that the number of vanishing moments $\tilde{d}$ determines the compression, and we should have $\tilde{d} > d - r$. We also imposed conditions on the regularities γ and $\tilde{\gamma}$ of the primal and dual wavelet. To summarise, we have grouped together the main conditions on the properties of the wavelets for an optimal implementation in Table 3.1. *Optimal* in this context means that an implementation is possible with a number of operations that is linear in N.

3.5 Dependence on the wavenumber

It is known that the wavelet-based matrix compression becomes less effective for higher frequencies in Helmholtz problems [39, 199]. More precisely, by means of intuitive arguments, given in the previous references, it has been demonstrated that there is a linear relation between the matrix-fill and the frequency. In this section, we analyse this effect rigorously for the Helmholtz equation in two dimensions. We quantify the achievable compression by mathematical deduction and are able to establish an upper bound that is close to, but not entirely, linear.

We will proceed by taking the effect of the wavenumber into account in every step of the derivation of the wavelet method. First, we derive in §3.5.1 an estimate for the derivatives of the kernel function

$$G(x,y) = \frac{i}{4} H_0^{(1)}(|x-y|). \tag{3.21}$$

This corresponds to quantifying the behaviour of the constant $C_1(k)$ in (3.11). Likewise, we determine the behaviour of the constant $C_2(k)$ in (3.12). Then, in §3.5.2, we derive the estimates for the size of the elements in the stiffness matrix. Finally, in §3.5.3, the density of the stiffness matrix after compression is analysed. The result of this derivation is formulated in Theorem 3.5.6 of §3.5.4. An upper bound is found that is shown to be sharp by means of some numerical results presented in §3.8.1.

3.5.1 Estimates for the derivatives of the kernel

It is well-known that the derivatives of the 2D Helmholtz kernel (3.21) satisfy (3.11) with $n = 1$ and $r = -1$, for some constant $C_1(k)$ [175]. In order to obtain the dependence of C_1 on the wavenumber, we must explicitly calculate these derivatives. The result is summarised in the following lemma.

Lemma 3.5.1. *The function* $G(x,y) := \frac{i}{4} H_0^{(1)}(k|x-y|)$ *satisfies* (3.11) *with* $n = 1$, $r = -1$, *and*

$$C_1(k) = \mathcal{O}(k^{|\alpha|+|\beta|-\frac{1}{2}}), \quad k \to \infty. \tag{3.22}$$

Proof. In order to estimate the left hand side of (3.11), we define $z(x,y) := |x-y|$ and $f(z) := H_0^{(1)}(kz)$. We then apply the chain rule and product rule for derivatives. The resulting sum contains contributions that are derivatives of f w.r.t. z, and partial derivatives of z. The latter are independent of k; the former are k-dependent. A recursive argument shows that the p-th order derivative of f with respect to z contains a term

$$(-1)^{\frac{p}{2}} k^p H_0^{(1)}(kz) \quad \text{or} \quad (-1)^{\frac{p-1}{2}} k^p H_1^{(1)}(kz)$$

for p even, resp. odd. The term with the highest order derivative of f is

$$\frac{\partial^{|\alpha|+|\beta|} f}{\partial z^{|\alpha|+|\beta|}} \left(\frac{\partial z}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial z}{\partial x_2}\right)^{\alpha_2} \left(\frac{\partial z}{\partial y_1}\right)^{\beta_1} \left(\frac{\partial z}{\partial y_2}\right)^{\beta_2}.$$

First, assume $p := |\alpha| + |\beta|$ to be even. The sum then contains the term

$$T := (-1)^{\frac{p}{2}} k^p H_0^{(1)}(kz) \frac{(x_1 - y_1)^{\alpha_1}}{z^{\alpha_1}} \frac{(x_2 - y_2)^{\alpha_2}}{z^{\alpha_2}} \frac{(x_1 - y_1)^{\beta_1}}{z^{\beta_1}} \frac{(x_2 - y_2)^{\beta_2}}{z^{\beta_2}}.$$

We know from (3.11) that T is bounded on Γ by Dz^{-p} with $D > 0$ a constant that depends on k. We have

$$|T| z^p \leq k^p |H_0^{(1)}(kz)| z^p,$$

so that

$$|T| z^p \leq k^p |H_0^{(1)}(kL)| L^p \quad \text{with} \quad L = \max |x - y|, \ \forall x, y \in \Gamma.$$

The dependence (3.22) now follows from the asymptotic expression [4]

$$H_\nu^{(1)}(x) \sim \sqrt{\frac{2}{\pi x}} e^{i(x - \frac{\pi}{4} - \frac{\nu \pi}{2})}, \quad x \to \infty, \tag{3.23}$$

and the fact that the term T has the highest exponent of k. The argument for p odd is completely analogous. □

Recall the definition of the modified kernel function,

$$\tilde{G}(t, \tau) := G(\kappa(t), \kappa(\tau)). \tag{3.24}$$

The constant $C_2(k)$ in estimate (3.12) for the derivatives of $\tilde{G}(t, \tau)$ in the parameter domain can be found by applying the chain rule for derivatives, and by using the previous estimate for every term in the sum.

Defining $\kappa(t) = (x_1(t), x_2(t))$ and $\kappa(\tau) = (y_1(\tau), y_2(\tau))$, we apply the product and chain rule to (3.24). An upper bound for each partial derivative of G is known through Lemma 3.5.1. We assume the parameterisation sufficiently smooth, so that the derivatives of κ are bounded. They are, of course, independent of k. It is clear that the highest order derivative of G determines the asymptotic behaviour around the diagonal $\kappa(t) = \kappa(\tau)$, where the term $|\kappa(t) - \kappa(\tau)|^{-(n+r+\alpha+\beta)}$ grows to infinity. We arrive at

$$\frac{\partial^{\alpha+\beta} \tilde{G}}{\partial t^\alpha \partial \tau^\beta} = \mathcal{O}(|\kappa(t) - \kappa(\tau)|^{-(n+r+\alpha+\beta)}), \quad t - \tau \to 0.$$

The asymptotic behaviour of $C_2(k)$ is determined by the largest exponent of k in the constants of the upper bounds for every term. This means that the exponent is again $\alpha + \beta - 1/2$. Thus, we have proved the following lemma.

Lemma 3.5.2. *The function* $\tilde{G}(t,\tau) := \frac{i}{4}H_0^{(1)}(k|\kappa(t) - \kappa(\tau)|)$ *with* $\kappa : [0,1] \to \Gamma$ *satisfies* (3.12) *with*

$$C_2(k) = \mathcal{O}(k^{\alpha+\beta-\frac{1}{2}}), \quad k \to \infty.$$

3.5.2 Estimates for the size of the elements

The analysis of the matrix compression is based on the size estimates (3.13) and (3.14). The constants $C_3(k)$ and $C_4(k)$ in these expressions depend on the wavenumber. Using the results of the last section, we can now quantify that dependence. To that end, we will first repeat the derivation of (3.13).

3.5.2.1 First estimate

The modified kernel function is developed into a Taylor expansion around a point in the support of $\psi_{j'k'}(\tau)$. For a wavelet $\psi_{j'k'}$ with $\tilde{d}$ vanishing moments, the first $\tilde{d}$ terms of the expansion will vanish,

$$\begin{aligned}\langle A\hat{\psi}_{j'k'}, \hat{\psi}_{jk}\rangle &= \langle \tilde{A}\psi_{j'k'}, \psi_{jk}\rangle \\ &= \int_0^1\int_0^1 \frac{\partial^{\tilde{d}}\tilde{G}}{\partial t^{\tilde{d}}}(t,\tau')\frac{(\tau'-\tau)^{\tilde{d}}}{\tilde{d}!}\,\psi_{jk}(t)\psi_{j'k'}(\tau)|\kappa'(t)||\kappa'(\tau)|\,\mathrm{d}\tau\,\mathrm{d}t,\end{aligned}$$

with $\tau' \in \operatorname{supp}\psi_{j'k'}$, and where $\tilde{A}f = A(f\circ\kappa^{-1})$ is an integral operator in the parameter domain with the modified kernel function $\tilde{G}$. Doing the same for t leads to

$$\begin{aligned}\langle \tilde{A}\psi_{j'k'}, \psi_{jk}\rangle &= \int_0^1\int_0^1 \frac{\partial^{2\tilde{d}}\tilde{G}}{\partial t^{\tilde{d}}\partial\tau^{\tilde{d}}}(t',\tau')\frac{(t'-t)^{\tilde{d}}}{\tilde{d}!}\frac{(\tau'-\tau)^{\tilde{d}}}{\tilde{d}!} \\ &\qquad \psi_{jk}(t)\psi_{j'k'}(\tau)|\kappa'(t)||\kappa'(\tau)|\,\mathrm{d}\tau\,\mathrm{d}t \\ &\le \frac{C_2(k)}{\operatorname{dist}(\operatorname{supp}\hat{\psi}_{jk}, \operatorname{supp}\hat{\psi}_{j'k'})^{2\tilde{d}}}\int_0^1\int_0^1 \frac{(t'-t)^{\tilde{d}}}{\tilde{d}!}\frac{(\tau'-\tau)^{\tilde{d}}}{\tilde{d}!} \\ &\qquad \psi_{jk}(t)\psi_{j'k'}(\tau)|\kappa'(t)||\kappa'(\tau)|\,\mathrm{d}t\,\mathrm{d}\tau,\end{aligned}$$

with $t' \in \operatorname{supp}\psi_{jk}$. Knowing that $\operatorname{supp}\psi_{jk} \sim 2^{-j}$ and $\int|\psi_{jk}(t)|dt \sim 2^{-j/2}$, and assuming a sufficiently smooth parameterisation, we arrive at

$$\langle A\hat{\psi}_{j'k'}, \hat{\psi}_{jk}\rangle \le \frac{C_2(k)}{\operatorname{dist}(\operatorname{supp}\hat{\psi}_{jk}, \operatorname{supp}\hat{\psi}_{j'k'})^{2\tilde{d}}}\,B\,2^{-j\tilde{d}}\,2^{-j'\tilde{d}}\,2^{-j/2}\,2^{-j'/2}.$$

with B a constant, independent of k. The result has the same form as (3.13) with $n=1$ and $r=-1$. The dependence on the wavenumber k is similar to that of $C_2(k)$, with $\alpha = \beta = \tilde{d}$. We have

$$C_3(k) = \mathcal{O}(k^{2\tilde{d}-1/2}). \qquad (3.25)$$

3.5.2.2 Second estimate

The derivation of estimate (3.14) for wavelets with overlapping support is somewhat more involved, and was first established in [175]. The property of the vanishing moments can be used only once, for the smaller wavelet that is fully contained within the singular points of the other wavelet. The use of this property as was done above in the double integral $\langle \tilde{A}\psi_{j'k'}, \psi_{jk}\rangle$, is not immediately possible due to the singularity of the kernel function. We note, however, that the result of the application of the integral operator $\tilde{A}$ to a smooth function $f \in C_0^\infty$ is also smooth. The restriction of $\psi_{j'k'}$ to the interval in the parameter domain that contains ψ_{jk} can be extended to a smooth function $f \in C_0^\infty$, with $\operatorname{supp}(f) \sim 2^{-j'}$ and $\|f\|_{H^s(\mathbb{R})} \leq c\, 2^{j's}$ [54]. After applying operator $\tilde{A}$ to f, we can again use the property of vanishing moments to establish the estimate (3.14). A concise mathematical proof is given in [54].

Define $\psi_{j'k'}(\tau) = f(\tau) + \tilde{f}(\tau)$, such that the support of $\tilde{f}(\tau)$ does not overlap with the support of ψ_{jk}. We analyse the wavenumber dependence of the estimate (3.14) for the functions f and $\tilde{f}$. To this end, we need to derive the dependence on k of the derivatives of $\tilde{A}f$, since

$$\begin{aligned}\langle \tilde{A}f, \psi_{jk}\rangle &= \int_0^1 \frac{\partial^{\tilde{d}}\tilde{G}f}{\partial t^{\tilde{d}}}(t')\frac{(t'-t)^{\tilde{d}}}{\tilde{d}!}\,\psi_{jk}(t)|\kappa'(t)|\,\mathrm{d}t \\ &\leq A\,2^{-j(\tilde{d}+1/2)} \sup_{t\in\operatorname{supp}(\psi_{jk})}\left|\frac{\partial^{\tilde{d}}\tilde{A}f}{\partial t^{\tilde{d}}}(t)\right| \end{aligned} \tag{3.26}$$

with $t' \in \operatorname{supp}(\psi_{jk})$. An explicit formula for the derivative of the function $\tilde{A}f$, is given in [100, Chapter 3 (3.4.5)] for the special case where the kernel $\tilde{G}(t,\tau)$ is only a function of $(t-\tau)$. An expression can also be found for the more general case where the kernel depends on $(\kappa(t) - \kappa(\tau))$. We prove it here specifically for the kernel (3.24).

Theorem 3.5.3. *Define* $g(t) := \int_a^b \tilde{G}(t,\tau)v(\tau)\,\mathrm{d}\tau$ *with* $v \in C^1$ *and* $\tilde{G}(t,\tau)$ *as* (3.24). *Then* $\forall t \in (a,b)$,

$$\begin{aligned} g'(t) = \int_a^b \tilde{G}(t,\tau)v'(\tau)\,\mathrm{d}\tau &+ \int_a^b \left(\frac{\partial \tilde{G}}{\partial t} + \frac{\partial \tilde{G}}{\partial \tau}\right) v(\tau)\,\mathrm{d}\tau \\ &+ \tilde{G}(t,a)v(a) - \tilde{G}(t,b)v(b). \end{aligned} \tag{3.27}$$

Proof. We first show that both integrals in the right hand side of (3.27) exist. The first integrand is improperly integrable. To show the existence of the second integral, define $r(t,\tau) = |\kappa(t) - \kappa(\tau)|$. Note that $r(t,t) = 0$ and $\frac{\partial r}{\partial t}(t,t) + \frac{\partial r}{\partial \tau}(t,t) = 0$, and also $\frac{\partial r}{\partial t}(t,t+\delta) + \frac{\partial r}{\partial \tau}(t,t+\delta) = \mathcal{O}(\delta)$. Hence, the function

$$\frac{\partial \tilde{G}}{\partial t} + \frac{\partial \tilde{G}}{\partial \tau} = \frac{i}{4}\frac{\partial H_0^{(1)}}{\partial r}\left(\frac{\partial r}{\partial t} + \frac{\partial r}{\partial \tau}\right)$$

is continuous in $t = \tau$, and therefore the second integral exists.

To prove the expression for the derivative, we note that

$$\begin{aligned} g(t+\delta) &= \int_a^b \tilde{G}(t+\delta,\tau)v(\tau)\,\mathrm{d}\tau = \int_{a-\delta}^{b-\delta} \tilde{G}(t+\delta,\tau+\delta)v(\tau+\delta)\,\mathrm{d}\tau \\ &= \int_{a-\delta}^{a} \tilde{G}(t+\delta,\tau+\delta)v(\tau+\delta)\,\mathrm{d}\tau + \int_{a}^{b-\delta} \tilde{G}(t+\delta,\tau+\delta)v(\tau+\delta)\,\mathrm{d}\tau, \end{aligned}$$

and

$$\tilde{G}(t+\delta,\tau+\delta) = \tilde{G}(t,\tau) + \delta\left(\frac{\partial \tilde{G}}{\partial t} + \frac{\partial \tilde{G}}{\partial \tau}\right) + \mathcal{O}(\delta^2).$$

We can also write $g(t) = \int_a^{b-\delta} \tilde{G}(t,\tau)v(\tau)\,\mathrm{d}\tau + \int_{b-\delta}^{b} \tilde{G}(t,\tau)v(\tau)\,\mathrm{d}\tau$. Now, by the definition of the derivative, $g'(t) = \lim_{\delta\to 0}\frac{g(t+\delta)-g(t)}{\delta}$ yields the result (3.27). □

Higher order derivatives of $\tilde{A}f$ can be found by applying Theorem 3.5.3 recursively, with $v(\tau) := f(\tau)|\kappa'(\tau)|$. The factor $|\kappa'(\tau)|$ is independent of j and k, and therefore will not influence the estimate. Assuming $f \in C_0^\infty$ with support contained in $[a,b]$, we have $v^{(i)}(a) = v^{(i)}(b) = 0$. The higher order derivatives of $g = \tilde{A}f$ are then given by

$$g^{(n)}(t) = \sum_{i=0}^{n}\binom{n}{i}\int_a^b\left(\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial \tau}\right)^i \tilde{G}(t,\tau)\right)\frac{d^{n-i}}{dt^{n-i}}(f(\tau)|\kappa'(\tau)|)\,\mathrm{d}\tau. \quad (3.28)$$

Again, each integral on the right hand side exists, since $\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial \tau}\right)^i r(t,t+\delta) = \mathcal{O}(\delta)$, $i >= 0$. We combine (3.28) with (3.12) to find a wavenumber dependence for $\left|\frac{\partial^{\tilde{d}}\tilde{A}f}{\partial t^{\tilde{d}}}(t)\right|$ of $\mathcal{O}(k^{\tilde{d}-1/2})$.

In order to establish a bound for the right hand side of (3.26), we first note that by the Sobolev embedding theorem $\sup|f^{(i)}(t)| = \|f\|_{W^{\infty,i}} \leq c_1\|f\|_{H^{i+1/2}} = c_2\,2^{j'(i+1/2)}$. Now, using the fact that $\mathrm{dist}(\mathrm{sing\,supp}\ \hat{\psi}_{j'k'}, \mathrm{supp}\ \hat{\psi}_{jk}) \leq c2^{-j'}$, $g^{(\tilde{d})}(t)$ can be bounded by

$$\begin{aligned} B\int_a^b 2^{j'(\tilde{d}+1/2)}\,\mathrm{d}\tau &\leq C\,2^{j'(\tilde{d}+1/2-1)} \\ &\leq D\,2^{j'/2}\,\mathrm{dist}(\mathrm{sing\,supp}\ \hat{\psi}_{j'k'}, \mathrm{supp}\ \hat{\psi}_{jk})^{-(\tilde{d}-1)}. \end{aligned}$$

Combined with (3.26), this concludes an estimate of the form (3.14),

$$|\langle \tilde{A}f, \psi_{jk}\rangle| \leq C_4(k) \frac{2^{-j(\tilde{d}+1/2)} 2^{j'/2}}{\mathrm{dist}(\mathrm{sing\,supp}\ \hat{\psi}_{j'k'}, \mathrm{supp}\ \hat{\psi}_{jk})^{\tilde{d}-1}}.$$

It remains to bound the part $|\langle \tilde{A}\tilde{f}, \psi_{jk}\rangle|$. This is more straightforward, as the integrand is not singular. Using the fact that $|\tilde{f}(t)| \leq c 2^{j'/2}$ and applying (3.12) once, we can proceed like in (3.26) (with f replaced by $\tilde{f}$),

$$\begin{aligned}
|\langle \tilde{A}\tilde{f},\psi_{jk}\rangle| &\leq B\, 2^{-j(\tilde{d}+1/2)} \left| \frac{\partial^{\tilde{d}} \tilde{A}\tilde{f}}{\partial t^{\tilde{d}}}(t') \right| \\
&\leq C\, 2^{-j(\tilde{d}+1/2)} \int_{\mathrm{supp}(\tilde{f})} |\tilde{f}(\tau)|\, \mathrm{dist}(\kappa(\tau), \mathrm{supp}(\hat{\psi}_{jk}))^{-\tilde{d}}\, \mathrm{d}\tau \\
&\leq C_4(k)\, 2^{-j(\tilde{d}+1/2)}\, 2^{j'/2}\, \mathrm{dist}(\mathrm{sing\,supp}\ \hat{\psi}_{j'k'}, \mathrm{supp}\ \hat{\psi}_{jk})^{-(\tilde{d}-1)}.
\end{aligned}$$

With these results, we have shown that estimate (3.14) holds, with a constant that depends on the wavenumber with the order

$$C_4(k) = \mathcal{O}(k^{\tilde{d}-1/2}). \tag{3.29}$$

3.5.3 Density of the compressed stiffness matrix

Define $E_{j,j'} := M_{j,j'} - M^{\epsilon}_{j,j'}$ as the error that is introduced by the first compression in the block matrix corresponding to the scales j and j', and $F_{j,j'} := M^{\epsilon}_{j,j'} - \hat{M}_{j,j'}$ as the error by the second compression. Then it is shown that (see [54])

$$\|E_{j,j'}\| \leq C\, a^{-2\tilde{d}-r} 2^{2Jr/2} 2^{-2d'(J-\frac{j+j'}{2})}, \tag{3.30}$$

$$\|F_{j,j'}\| \leq C\, (a')^{-\tilde{d}-r} 2^{2Jr/2} 2^{-2d'(J-\frac{j+j'}{2})}. \tag{3.31}$$

Based on these expressions, one can show that the compressed scheme is consistent with the original operator equation, and retains the order of convergence. If the errors were bounded uniformly in k, it would ensure that the compression error is independent of the wavenumber.

This means that we must choose the parameters a and a' in a suitable way. We note that expression (3.30) is established by summing the corresponding estimates of the form (3.13) for the discarded elements. This introduces a dependence on k of the order $\mathcal{O}(k^{2\tilde{d}-1/2})$, that is transferred unchanged to (3.30). It can be compensated by an asymptotic behaviour of $a = \mathcal{O}(k^p)$ if

$$k^{-2(\tilde{d}-1/2)p}\, k^{2\tilde{d}-1/2} = \mathcal{O}(1) \Leftrightarrow -2(\tilde{d}-1/2)p + 2\tilde{d} - 1/2 \leq 0$$

or

$$p \geq 1 + \frac{1}{4\tilde{d} - 2}. \tag{3.32}$$

The compression error is thus bounded uniformly in k if $a = \mathcal{O}(k^{1+\frac{1}{4\tilde{d}-2}})$. The asymptotic behaviour of the parameter a needs to be slightly larger than linear in k, but improves somewhat as the number of vanishing moments increases.

The second compression is handled similarly, leading to

$$p \geq 1 + \frac{1}{2\tilde{d} - 2}. \tag{3.33}$$

It is important to note here that we only consider the compression of the stiffness matrix. The wavenumber also influences the condition number, even after preconditioning, and will therefore have an impact on the convergence of iterative solution methods. This means that, for large values of k, the system may become increasingly ill conditioned. The uniform bound on the compression error that we have derived ensures however that, for a specific value of k, the compressed scheme retains the convergence properties of the corresponding uncompressed Galerkin scheme.

3.5.4 Wavenumber dependence of the wavelet compression

The number of nonzero elements in the compressed matrices M_J^ϵ and $\hat{M}_J$, depends on the parameters a and a' in (3.15) and (3.16). Their values determine the sparsity structure of the submatrices $\hat{M}_{jj'}$ in the stiffness matrix.

The thresholds indicate a minimal distance between the supports of wavelets corresponding to a matrix element. They are chosen such that the error introduced by discarding elements matches the discretisation error. The allowable error varies for each combination of scales j and j', and in general the rougher scales require higher accuracy, while the elements corresponding to finer scales can be less accurate. Combining the estimates for the matrix elements, and the thresholds used for discarding some of them, reveals that the compressed stiffness matrix keeps only $\mathcal{O}(N)$ elements.

As the thresholds increase, the condition on the distance between wavelets becomes stronger, and as a result the matrix will be less sparse. We see from (3.15) that the required minimal distance between the support of the wavelets corresponding to a matrix element, is directly proportional to a. While the shape of the boundary Γ and the parameterisation $\kappa(t)$ will have an influence here, we can say in first order approximation that the number of elements kept is also linear in a.

Lemma 3.5.4. *The number of nonzero elements in* $\hat{M}_J$*, as defined by* (3.17)*, is* $\mathcal{O}(a) + \mathcal{O}(a')$ *as a function of the wavenumber* k*.*

We have investigated the necessary asymptotic behaviour of the parameters a and a' for large values of k. The results (3.32) and (3.33) lead to another lemma.

Lemma 3.5.5. *In order to achieve compression that maintains the convergence properties of the uncompressed Galerkin scheme, the parameters* a *and* a' *in* (3.15) *and* (3.16) *have to be chosen such that*

$$a = \mathcal{O}(k^{1+\frac{1}{4\tilde{d}-2}}) \quad \textit{and} \quad a' = \mathcal{O}(k^{1+\frac{1}{2\tilde{d}-2}}).$$

The combination of these lemmas leads to the following statement.

Theorem 3.5.6. *The number of nonzero elements in the wavelet compressed stiffness matrix with optimal choice of* a *and* a' *in the threshold constants* (3.15) *and* (3.16) *increases asymptotically linear in* N*, with a proportionality constant of the order* $\mathcal{O}(k^{1+1/(2\tilde{d}-2)})$*.*

Note that the dependence on k of the proportionality constant means that, with increasing wavenumbers, the stiffness matrix fills up to become a dense matrix. In the high frequency regime, where N increases proportional to k, the actual number of nonzero elements in the compressed stiffness matrix grows asymptotically as $\mathcal{O}(N^2)$. The matrix looses any significant sparsity. This result will be illustrated numerically in §3.8.

3.6 Wavelet-packet based methods

3.6.1 Motivation

In the preceding section, it was shown that the wavelet method approximately requires $O(N^2)$ operations in the high frequency regime where N is chosen proportional to k. Though wavelets are suitable to sparsely represent a smooth function, or a smooth integral operator, all significant sparsity is lost in the presence of strong oscillations. In this section, we further examine the effect of oscillations on the wavelet method, and we propose a solution by using wavelet packets.

Wavelet basis functions are inherently multiscale. Aside from a subdivision of scale and position however, wavelets also induce an uncontrollable subdivision of the frequency spectrum of a function. One step of the wavelet transform divides the frequency spectrum approximately into a low frequency and high frequency part. Only the lower frequency part is further subdivided in the next step. This may not be optimal for functions with a certain fixed, inherent frequency. It appears more appropriate to *zoom in* on this specific frequency, in order to represent the function with fewer coefficients.

Wavelet packets are a generalisation of wavelets, that allow a subdivision of scale, position and of frequency spectrum. They retain the vanishing moment properties of wavelets, but add flexibility in the choice of basis functions. The increase in frequency resolution comes at a cost of spatial resolution. Nevertheless, we will see that wavelet packets can be adapted to an oscillatory problem, and enable much improved sparsity of the discretisation compared to wavelets. They lead to a sparse matrix with a number of elements that scales approximately as $O(N^{1.4})$.

3.6.2 Wavelet packets

3.6.2.1 Definition

Wavelet packets were introduced and developed by Coifman, Meyers, Quake and Wickerhauser [45, 48]. They can be defined recursively in the following way. Set $w_0(t) = \phi(t)$, and define

$$w_{2n}(t) = \sqrt{2} \sum_k h_k w_n(2t - k) \tag{3.34}$$

$$w_{2n+1}(t) = \sqrt{2} \sum_k g_k w_n(2t - k). \tag{3.35}$$

Wavelet packets can be defined on all scales and positions by

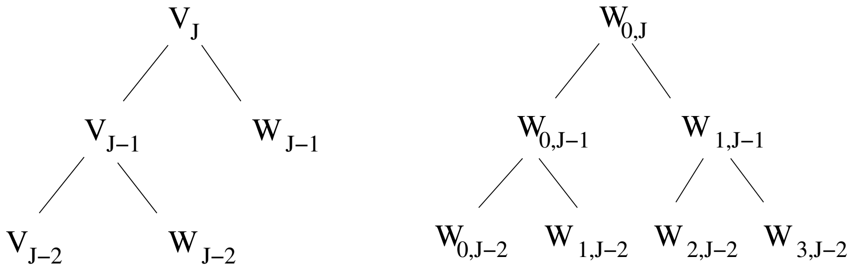
$$w_{njk}(t) = 2^{j/2} w_n(2^j t - k). \tag{3.36}$$

We represent the function spaces involved by $W_{nj} = \operatorname{span}\{w_{njk}\}$. The space of all scaling functions on scale J is $V_J = W_{0J}$. A basis of V_J can be identified by a subset Λ of the set of indices $\Xi := \{(n, j) \in \mathbb{Z}^2\}$, such that the corresponding wavelet packets w_{njk} form a basis of W_{0J}. It is convenient here to use the multi-index notation $\lambda = (n, j)$. We can expand any function $f \in V_J$ in the basis denoted by Λ as

$$f = \sum_{\lambda \in \Lambda} \sum_k v_{\lambda,k} w_{\lambda,k}. \tag{3.37}$$

A fast wavelet packet transformation can be devised, similar to the fast wavelet transform. The full wavelet packet decomposition at level $j = 0$ is the transform corresponding to the basis functions w_{n0k}, and has a computational complexity of $O(J2^J) = O(N \log N)$.

The wavelet decomposition is only a special case of (3.37). In the wavelet transform, function space $V_{j+1} = V_j \oplus W_j$ is split into two spaces, using the filters g and h. The resulting space V_j is split again, a process that is continued until the full wavelet decomposition (3.5) is obtained. In a wavelet packet transform on the other hand, one employs the same *splitting trick* also for the function space W_j containing the upper part of the frequency spectrum of V_{j+1}. This results in a highly redundant binary tree of function spaces, with V_J at the root. The trees for the wavelet and wavelet packet function spaces are shown in Figure 3.2. The frequency resolution increases downward in the tree, at a cost of decreased spatial resolution. The parameter n is an indication of the frequency content of the function space W_{nj}. The frequency properties of wavelet packets are discussed in detail in [112].



(a) Wavelet tree (b) Wavelet packet tree

Figure 3.2: Binary trees of wavelet and wavelet packet function spaces.

A two-dimensional wavelet packet transform of a matrix A can be defined by applying a one-dimensional transform to all rows and columns of the matrix successively. The resulting matrix is called the rectangular transform of A. Alternatively, a two-dimensional wavelet packet transform can be obtained by considering a quadtree of function spaces of the form $W_\lambda \times W_\mu$, for $\lambda \times \mu \in \Xi \times \Xi$. A subtree with any selection of function spaces that covers $V_J \times V_J$ leads to a two-dimensional basis. The tensor product basis functions are given by

$$w_{\mu,l,\lambda,k}(s,t) = w_{\mu,l}(s) w_{\lambda,k}(t). \tag{3.38}$$

This approach is called the square transform. The structure of the square transform W of a matrix is shown in Figure 3.3. In the following, we will denote a square subblock by $W_{\mu,\lambda}$.

3.6.2.2 Best basis algorithm of Coifman and Wickerhauser

It can be proven that, for a vector x of N elements, there exist more than 2^N possible wavelet packet bases. Coifman and Wickerhauser presented a method to find a best basis for a given criterion, such as maximal sparsity or minimal entropy [48]. The algorithm finds a global minimum for a cost function $P(\{x_\lambda\})$, among all possible wavelet packet representations $\{x_\lambda\}$ of x. It is applicable for cost functions that satisfy $P(\emptyset) = 0$, and $P(\{t_i\}) = \sum_i p(|t_i|)$ for some function p; such cost functions are said to be *additive.* For example, the choice

$$p(t) = \begin{cases} 1 & \text{if} \quad |t| > \tau \\ 0 & \text{otherwise.} \end{cases}$$

leads to the wavelet packet transform that has the smallest number of elements larger than τ.

The algorithm uses a bottom-up approach. First, a full wavelet-packet decomposition of x is computed up to scale 0. The lowest cost representation of the part of x that lies in $W_{0,1} = W_{0,0} \oplus W_{1,0}$ has an associated cost

$$Q(0,1) := \min\{P(\{x_{0,1,k}\}), P(\{x_{0,0,k}\}) + P(\{x_{1,0,k}\})\}. \tag{3.39}$$

A similar expression can be given for $Q(n,1)$, $n = 1 \dots 2^{J-1}$. The lowest cost representation of the part of x that lies in $W_{n,j}$, $j > 1$, is given by

$$Q(n,j) := \min\{P(\{x_{n,j,k}\}), Q(2n, j-1) + Q(2n+1, j-1)\}. \tag{3.40}$$

By induction, the global minimum of P is given by $Q(0,J)$. The corresponding best basis is found by remembering the arguments that minimise the expressions (3.39) and (3.40).

An extension of this algorithm to the two-dimensional case is possible only for the square transform. The cost of a subblock $W_{\lambda,\mu}$ is evaluated by $P(W_{\lambda,\mu})$. An additive cost function cannot be found for the rectangular transform, since in that case the subblocks corresponding to λ and μ, for $\lambda, \mu \in \Lambda$, overlap for each combination of λ and μ. Hence, in that case the efficient best basis algorithm can only be approximated.

3.6.2.3 The nonstandard matrix-vector product

A one-dimensional rectangular wavelet packet transform W of a regular matrix M can be written as $W = TMT^T$. The matrix-vector product $y = Mx$ can then be computed efficiently by $y = T^{-1}W(T^T)^{-1}x$. A matrix-vector product with a matrix obtained after a two-dimensional square wavelet packet transformation is more involved, but can still be defined [202]. The algorithm for the matrix-vector product is a three-step procedure. Assume that W is a two-dimensional wavelet packet transform of M. First, the representation coefficients x_λ, for $\lambda \in \Xi$, are computed for all scales $j = 0, \dots, J$. Next, for each block $W_{\mu,\lambda}$ in W that corresponds to

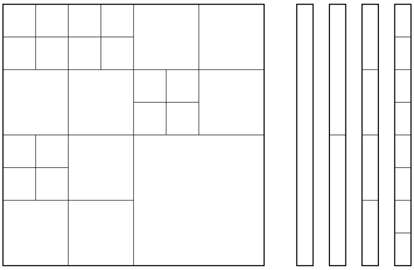


Figure 3.3: Illustration of the nonstandard matrix-vector product. The vector is represented on all scales. The matrix-vector product is performed by multiplying each subblock of M on scale j with the matching block of the vector on the same scale.

a function space $W_\mu \times W_\lambda$, compute

$$y_\mu := W_{\mu,\lambda} x_\lambda.$$

Finally, the result y is given by

$$y = \sum_\mu \sum_k y_{\mu,k} \psi_{\mu,k}.$$

The latter is easily computed by adding the inverse transformation of each y_μ. The algorithm is clarified in Figure 3.3. The rightmost part of the figure depicts the representations of the vector x on each scale j. The leftmost part of the figure shows the structure of the transformed matrix with subblocks $W_{\mu,\lambda}$. The matrix-vector product requires one to multiply each subblock on scale j of the dense matrix with the matching block on scale j of x. We note already that the structure of the matrix-vector product strongly resembles that of the $\mathcal{H}^2$-matrices that will be described in §5.4.2. The three steps described above relate directly to the three steps for a matrix-vector product with $\mathcal{H}^2$-matrices: Forward Transformation, Multiplication, and Backwards Transformation.

3.6.3 Application in boundary element methods

The use of wavelet packets for the fast solution of integral equations has been considered previously in [93, 74, 75]. Deng and Ling, and Golik independently studied wavelet packet based matrix compression. Both reported a number of significant elements after compression that scales as $O(N^{4/3})$ for the combined field integral equation (2.56) using collocation. Both approaches were based on a one-dimensional transform and, hence, only an approximation to the best basis algorithm was applied. In [93] a top-down approximation to the best basis algorithm is performed on the right hand side of the linear system (2.59). The resulting basis is used for compressing the matrix using the rectangular transform. In [74] a top-down approximation to the best basis for the rectangular transform is performed on the matrix in (2.59) itself. In [75] a one-dimensional wavelet packet basis is constructed that zooms in on the frequency given by k [75].

Here, we will consider the use of a two-dimensional wavelet packet basis using the square transform. This will increase the freedom in the choice of basis greatly. Moreover, the best basis algorithm can be applied exactly and the sparsity results can be much improved.

3.6.3.1 Choice of basis functions

It was shown in §3.4 that an optimal implementation of the wavelet method should employ biorthogonal wavelets. Wavelet packets based on biorthogonal wavelet filters are not guaranteed to be stable however [42], so we are restricted to orthogonal wavelet packets. We employ the popular Daubechies wavelets since they are compactly supported [66]. Other orthogonal wavelets with finite filters would lead to similar results.

The compression is influenced by the number of vanishing moments $\tilde{d}$. A larger value of $\tilde{d}$ leads initially to better compression, but also requires a larger number of filter coefficients h_k and g_k. This increases the computation time of the wavelet and wavelet packet transformations. A suitable tradeoff for the purposes of this study proved to be the choice $\tilde{d} = 7$.

3.6.3.2 Collocation approach for the discretisation

We consider the discretisation of the combined field integral equation (2.56),

$$(i\eta S + \frac{I}{2} + D^*)q = i\eta f + g, \tag{3.41}$$

with η proportional to k. The discretisation by collocation and by Galerkin yield very comparable results concerning the compression of the matrix. We proceed here with a collocation approach. A set of pulse basis functions is used, each with height 1 on one element of Γ and zero elsewhere. We apply a one point integration formula for the integral, as given in [108]. Define Δ_i, r_i and n_i as the width, centre position, and outgoing normal of the i-th pulse. The discretisation matrix M is then given by

$$M_{i,j} = \begin{cases} i\eta\Delta_j \frac{i}{4}\left(1 + \frac{2i}{\pi}\log\left(\frac{e^{\gamma}k\Delta_j}{4e}\right)\right) + \frac{1}{2} & \text{if} \quad i = j \\ \Delta_j\Big(i\eta\frac{i}{4}H_0^{(1)}(k|r_i - r_j|) & \\ \qquad - \left(n_i \cdot \frac{r_i - r_j}{|r_i - r_j|}\right) k\frac{i}{4}H_1^{(1)}(k|r_i - r_j|)\Big) & \text{otherwise.} \end{cases} \tag{3.42}$$

The approximation to the solution of (3.41) is found by solving

$$Mx = b, \tag{3.43}$$

where b corresponds to the pointwise evaluation of the right hand side of (3.41) on Γ. The matrix M will be transformed and compressed in order to obtain a faster matrix-vector product. The compression error can be measured in different ways. Assume x is the exact solution of (3.43), and y is the solution of the compressed problem. Then the relative error of the solution and the relative residual error are given by

$$e_S = \frac{\|x-y\|}{\|x\|} \quad \text{and} \quad e_R = \frac{\|b-My\|}{\|b\|}, \tag{3.44}$$

respectively. The residual error is a weaker error measure, but it is meaningful in practice. In computational electromagnetics applications, an approximate solution y with residual error e_R represents a current that induces the same electromagnetic field as the exact solution x to a certain precision e_R.

3.6.3.3 Matrix compression: scaling of the threshold

A compression of the transformed discretisation matrix is obtained by discarding small elements. The most straightforward implementation uses a fixed threshold value τ. In that case, the compressed matrix W^ϵ is given by

$$W^\epsilon_{i,j} = \begin{cases} W_{i,j} & \text{if} \quad |W_{i,j}| > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3.45}$$

A more advanced compression strategy utilises a scale dependent threshold, as discussed for the wavelet method in §3.4.2. Such a strategy can improve sparsity, but for the arguments in this section the simpler strategy is sufficient. A natural question that arises is how to choose the parameter τ as a function of N, in order to guarantee a fixed error. To answer this question, we will construct an estimate for the matrix compression error as a function of τ and N. We will consider the matrix obtained with the collocation scheme, and we measure the matrix compression error by

$$e = \frac{\|W - W^\epsilon\|_2}{\|W\|_2}. \tag{3.46}$$

First we determine the asymptotic behaviour of $\|M\|_2$, with M the collocation matrix whose elements are given in (3.42). The Hankel functions for large arguments behave as $|H_\nu^{(1)}(z)| \sim \frac{1}{\sqrt{z}}$, independent of the order ν [4]. The pulse width Δ_j is obviously $O(1/N) = O(h)$, and $\eta = O(k)$. The contribution of the single layer potential to the off-diagonal elements in (3.42) is therefore $O(hk/\sqrt{k})$. The contribution of D^* has the same order, so the linear combination has the same order too. The diagonal element contributions from S and D^* are each $O(1)$.

The maximum absolute column sum norm $\|M\|_1$ and the maximum absolute row sum norm $\|M\|_\infty$ are of order $O(1 + (N-1)(hk/\sqrt{k})) = O(Nhk/\sqrt{k}) = O(\sqrt{k})$. From this and from the inequality $\|M\|_2^2 \leq \|M\|_1 \|M\|_\infty$ we can deduce that

$$\|M\|_2 = O(\sqrt{k}). \tag{3.47}$$

The orthogonal wavelet transform and orthogonal wavelet packet transform can be represented by a transformation matrix T with $\|T\|_2 = 1$. Therefore, $\|W\|_2 = \|M\|_2$.

The error $\|W - W^\epsilon\|_1$ can be bounded by the worst case value $N\tau$. The infinity norm is similar, leading to $\|W - W^\epsilon\|_2 = O(N\tau)$. Combined with (3.47), this yields

$$e = \frac{\|W - W^\epsilon\|_2}{\|W\|_2} = O(N\tau/\sqrt{k}).$$

The error (3.46) will be bounded only if

$$\tau = O(N^{-1/2}). \tag{3.48}$$

The analysis for the discretisation by Galerkin is similar, and leads to the same behaviour (3.48) for the threshold.

Remark 3.6.1. *A threshold that is often proposed in the literature is* $\tau = \frac{\|M\|_1}{N}$ *[108]. This threshold has indeed the appropriate asymptotic behaviour.*

3.6.4 Computational complexity

3.6.4.1 Complexity of the matrix-vector product

The numerical results that are presented in §3.8.2 will demonstrate the improved sparsity of the discretisation matrix when using wavelet packets as basis functions. A rigorous proof of the reduced complexity would be rather involved, due to the adaptive nature of the best basis algorithm. However, an estimate is derived below, based on heuristic arguments. In addition, insights are gained into the behaviour of the method.

Denote by $\hat{w}_n(\xi)$ the Fourier transform of the wavelet packet function $w_n(t)$. Since $w_n(t)$ is compactly supported, the same can not hold for $\hat{w}_n(\xi)$. However, the main energy of $\hat{w}_n(\xi)$ is located inside a certain frequency band that depends on n. Its size can be estimated from the variance

$$\sigma_n := \inf_{\xi_0 \in \mathbb{R}} \int_0^\infty |\xi - \xi_0|^2 |\hat{w}_n(\xi)|^2 \,\mathrm{d}\xi.$$

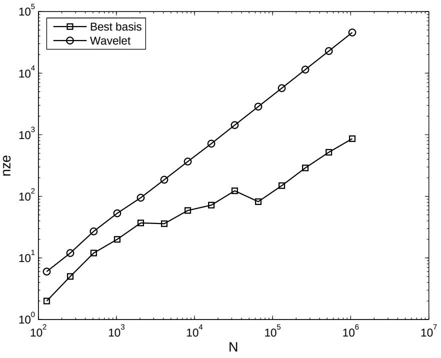


Figure 3.4: The number of nonzero elements after compression of $\cos(2\pi fx)$ with $N = 10f$ using wavelets, and wavelet packets in the best basis.

Ideally, each wavelet packet function $w_n(x)$ has a frequency spectrum inside a band of fixed size, that does not overlap with the spectrum of other basis functions. This is only approximately the case. It is shown that $\sigma_n \sim n^{2\delta}$ with $\delta > 0$ a small constant [46]. The size of the band is approximately proportional to the standard deviation

$$s_n := \sqrt{\sigma_n} = n^{\delta}. \tag{3.49}$$

Consider a function that is sampled at $N = 2^L$ equispaced points. The entire frequency spectrum is given by $0 \leq f < 2^L$. In the ideal case, $w_{n,j,k}(x)$ has a frequency spectrum of the form $f \in [2^j(\xi_0 - 1/2), 2^j(\xi_0 + 1/2)]$, i.e., a band of fixed width 2^j, with ξ_0 the average frequency of $\hat{w}_n(x)$. The basis functions $w_{n,j,k}(x)$, $n = 0, \ldots, 2^{L-j} - 1$, then cover the entire spectrum independently from each other. Now consider the function $\cos(2\pi fx)$, with a frequency f that is proportional to N. A value of ξ_0, and a corresponding value of n, can be found for any fixed scale j such that $f \in [2^j(\xi_0 - 1/2), 2^j(\xi_0 + 1/2)]$. Hence, the cosine can be represented accurately on scale j by only 2^j basis functions $w_{n,j,k}(x)$, $k = 0, \ldots, 2^j - 1$, independently of N. Both ξ_0 and n scale linearly with N.

Now assume a bandwidth of n^{δ}. Then there are $O(n^{\delta})$ intervals of the form $[2^j(\xi_0 - O(n^{\delta})), 2^j(\xi_0 + O(n^{\delta}))]$ that contain f. Hence, the total number of coefficients required to represent the cosine accurately on scale j has order $O(2^j n^{\delta}) = O(2^j N^{\delta})$. For $j = 0$, the estimate is $O(N^{\delta})$. The value of δ can be determined easily for a given wavelet family from a numerical experiment. We have computed the number of coefficients larger than a fixed threshold for the function $\cos(2\pi fx)$, $x \in [0, 1]$, with $N = 10f$. The result is shown for a wavelet approximation and a best basis approximation in Figure 3.4. The number of coefficients in the wavelet approximation scales linearly in N. For the best basis approximation, we have $\delta \approx 0.7$.

The two-dimensional discretisation matrix can be regarded as an extension of the one-dimensional cosine model, as it represents the discretisation of an oscillatory function with a frequency that is approximately fixed. The influence of the singularity may be discarded, because it can be represented locally on each scale with only few coefficients. An estimate for the number of coefficients in the two-dimensional transformed discretisation matrix is then $O(N^{2\delta}) = O(N^{1.4})$.

3.6.4.2 Total computational cost

Three phases of the solution method contribute to the computational complexity: the setup of the discretisation matrix, the wavelet packet transform and thresholding, and the iterative solution of the resulting system. The construction of the full discretisation matrix requires $O(N^2)$ operations. The computational complexity of the best basis algorithm for a matrix is $O(N^2 \log N)$, slightly larger than the setup cost. The complexity of the solution phase with an iterative method depends on the condition number of the matrix, and on the complexity of the matrix-vector product. The condition number is not significantly influenced by the compression, so the gain of our method lies solely in a faster matrix-vector product. The higher cost for the transformation will be compensated if the same system is solved for several different boundary conditions, since the setup has to happen only once. This is a common case, e.g., in electromagnetics, where incoming waves from different angles lead to different boundary conditions.

The high transformation cost can be avoided by approximating the best-basis algorithm with lower complexity methods. To that end, we considered a two-dimensional top-down approach: in each step of the wavelet packet transformation, the sparsity of a subblock in the matrix corresponding to a function space is compared to the sparsity in the representation of its four children spaces. If the sparsity is not improved, the subblock is not further transformed. The sparsity of the resulting basis is only a local minimum in the space of all possible bases, compared to the global minimum obtained by the best basis algorithm. However, the costly full wavelet packet decomposition is not required with this approach.

The high setup cost can also be avoided if the size of the elements in a wavelet packet basis can be estimated a priori. A suitable basis can then be selected that leads to a large number of small elements that do not need to be computed. Such estimates are available in the wavelet method; they are given by (3.13) and (3.14). Similar estimates for wavelet packets are not yet available. They should also incorporate the frequency resolution of the wavelet packets and the inherent frequency of the kernel function.

3.7 Wavelet quadrature

3.7.1 Integration techniques

All the wavelet based solution methods discussed so far require the computation of a large number of one-dimensional and two-dimensional integrals involving wavelet functions or scaling functions in the integrand. The integrals may be singular, and the integration domain may be large. In the wavelet method, the integrals corresponding to the wavelets with the largest support require the most accuracy, while elements corresponding to wavelets on finer scales may have larger errors [56]. A fully discrete implementation for the wavelet method was proposed in [196], requiring $O(N(\log N)^4)$ operations. An efficient method for singular and nearly singular integrals was presented in [176], enabling an $O(N)$ implementation of the wavelet method [175, 104, 86].

The wavelet coefficients of a function f involve integrals of the form

$$c_{j,k} := \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)\,\mathrm{d}x, \tag{3.50}$$

or

$$d_{j,k} := \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)\,\mathrm{d}x. \tag{3.51}$$

Since coefficients $d_{j,k}$ can be obtained from $c_{j+1,k}$ using (3.7), we focus on the former integrals involving a scaling function $\phi(x)$. We assume that the scaling function has limited support, i.e., $\mathrm{supp}(\phi(x)) = [s_1, s_2]$.

General quadrature rules for (3.50) depend on the smoothness of the integrand. Discontinuities of the integrand or of any of its derivatives may disturb the convergence of these methods. They will fail for scaling functions with only small regularity. Also, it can be computationally expensive to evaluate ϕ and f, so we would like to minimise the number of function evaluations. In some cases there is even no explicit formula for ϕ available.

The methods described in the given references assume that the wavelet function is at least piecewise smooth, for example piecewise polynomial. This allows the use of high order quadrature schemes. The singularity is removed using a regularisation [79]. Not all wavelets are smooth however. A different approach is taken in [14, 58, 88, 142, 184]. High order quadrature rules can be constructed for wavelets with arbitrarily low regularity, based solely on the defining refinement equation (3.3).

Here, we will extend this approach to singular and piecewise smooth integrands. Our goal is to develop quadrature rules that exhibit convergence characteristics that depend only on the smoothness of f, and to reuse function evaluations whenever possible. To compute the rules themselves, we will only need the coefficients h_k of the refinement equation (3.3). Our approach is based on the method originally discussed in [19] and extended in [184]. We recall the quadrature method of [184] for smooth functions in §3.7.2. In §3.7.3, the method is extended to cover piecewise smooth functions. In §3.7.4, we cover singular functions.

Throughout this section, the integration error is determined by comparison with the results of the integration package Cubpack [50].

3.7.2 An integration rule for smooth functions

3.7.2.1 The quadrature rule and its construction

The modus operandi is based on a technique described by Sweldens and Piessens in [184, 183]. These authors have constructed a quadrature rule $Q[\cdot]$ that only requires the evaluation of f in a number of quadrature points,

$$\int_{-\infty}^{\infty} f(x)\phi(x)\,\mathrm{d}x \simeq Q[f(x)] = \sum_{i=1}^{r} w_i f(x_i). \tag{3.52}$$

The convergence rate of this integration rule depends on the number of abscissae r and on the smoothness properties of f, but not on the smoothness of the scaling function ϕ.

The integrals of type (3.50) can be approximated for all values of j and k using rule (3.52), by

$$c_{j,k} \simeq 2^{-j/2} Q[f(2^{-j}(x+k))]. \tag{3.53}$$

The abscissae x_i are chosen on a regular grid that enables reusing the values $f(x_i)$ for neighbouring scaling functions. Points of the form $x_i = (i-1)2^s$, e.g., with s a negative integer, are good candidates, since any integer shift transforms the set of points $\{x_i\}$ onto itself. A real shift τ on the entire grid preserves that property.

These observations lead one to consider the points $x_i = (i-1)2^s + \tau$. In [184] it is shown that many function evaluations can be shared: if $s = 0$, each additional coefficient requires only one extra evaluation of f. All the other values that are needed for the computation of $c_{j,k}$ in (3.50), are also needed for $c_{j,k-1}$. The parameter τ is an additional degree of freedom, and can be used to increase the order of the quadrature rule.

The rule (3.52) is derived by imposing that the quadrature is exact for all polynomials of degree lower than r,

$$Q[P_l(x)] = M^l, \qquad l = 0, 1, \ldots, r-1, \tag{3.54}$$

with

$$M^l := \int_{-\infty}^{\infty} P_l(x)\phi(x)\,\mathrm{d}x. \tag{3.55}$$

The polynomials $P_l(x), l = 0, 1, \ldots, r-1$ in (3.54) form a basis for the set of polynomials of degree lower than r. The unknowns are the quadrature weights, and the matrix representing the system is found by simply evaluating $P_l(x)$ in the quadrature abscissae. Since the resulting system of equations is ill conditioned for the monomial basis $P_l(x) = x^l$, Sweldens and Piessens considered using the Chebyshev polynomials of the first kind, $T_l(x)$. These polynomials form an orthogonal basis on $[-1, 1]$ for the weight function $w(x) = (1-x^2)^{-1/2}$. When properly scaled Chebyshev polynomials are used in (3.54), the system is well conditioned. Scaling the interval $[s_1, s_2]$ to $[-1, 1]$, we have the basis polynomials

$$P_l(x) = T_l\left(2\frac{x - s_1}{s_2 - s_1} - 1\right). \tag{3.56}$$

By making use of the refinement equation (3.3) and properties of Chebyshev polynomials, an explicit formula for the moments M^l can be derived [184]. The system (3.54) can then be solved to find the quadrature weights.

3.7.2.2 Some comments on the accuracy of the quadrature rule

With a regular grid of r abscissae, we typically expect for the corresponding quadrature rule a degree of accuracy, denoted by q, of at most $q = r-1$. The integration error is then of the order $\mathcal{O}(h^{q+1})$, with h being proportional to the interval length $s_2 - s_1$, i.e., the support of ϕ. In some cases, we can achieve an order $q = r$ with a proper choice of τ. The number r depends on the parameter s that determines the regular grid. In [184], it is shown that the application of the rule to a scaling function $\phi_{j,k}(x)$ on scale j leads to a relative error on the approximation of $c_{j,k}$ of $\mathcal{O}(2^{-j(q+1)})$. As one might have expected, a finer scale, i.e., a reduction of the integration interval, leads to a more accurate result.

The technique described here leads to an interpolatory integration rule, with the scaling function as a weight function. In general, the weights for such rules can have alternating signs, which negatively impacts the stability of the computations. Even if the weight function is strictly positive, there will be at least one negative weight for each rule with a sufficiently high degree of accuracy. For weight functions that switch signs, which is not uncommon, each rule has negative and positive weights. For this reason, we will quantify the stability properties of the weights in the examples, presented further on by using the sum of absolute values as a measure [69].

If we abandon the principle of using a regular grid for the abscissae, better rules can be made by constructing Gaussian quadrature rules with ϕ as weight function. Since for Gaussian rules the weight function has to be positive, for some scaling functions $g(x) := \phi(x) + c$ is used instead. Here, the constant c is chosen such that g is a suitable nonnegative weight function. Such rules are constructed in [14, 142]. In that setting, one loses, however, the ability to reuse function evaluations.

3.7.3 Improving accuracy by composite quadrature

For a larger number of abscissae, say $r \sim 30$, the high order methods from [184] become unstable due to large quadrature weights with alternating signs. As with composite quadrature rules, the accuracy can be improved by splitting the integration interval, and by applying a lower order quadrature rule on each subinterval. Hence, we aim for a new rule on the subinterval $[a, b] \subset [s_1, s_2]$

$$\int_a^b f(x)\phi(x)\,\mathrm{d}x \simeq Q_{a,b}[f(x)]. \tag{3.57}$$

The support $[s_1, s_2]$ of the scaling function ϕ is divided into a sequence of intervals $[a_i, b_i]$. Typically, the integration subinterval $[a_i, b_i]$ has its endpoints on points of discontinuity or singularity of f.

3.7.3.1 Computation of the moments

In order to find the quadrature weights, we need to compute the moments of the scaling function on the interval $[a, b]$,

$$M^l_{a,b} := \int_a^b P_l(x)\phi(x)\,\mathrm{d}x. \tag{3.58}$$

Using refinement equation (3.3), we have, for the zeroth order moment,

$$\begin{aligned} M^0_{a,b} := \int_a^b \phi(x)\,\mathrm{d}x &= \frac{\sqrt{2}}{2}\sum_k h_k \int_{2a-k}^{2b-k} \phi(x)\,\mathrm{d}x \\ &= \frac{\sqrt{2}}{2}\sum_k h_k M^0_{2a-k,2b-k}. \end{aligned} \tag{3.59}$$

This formula expresses $M^0_{a,b}$ as a linear combination of zeroth order moments on the intervals $[2a-k, 2b-k]$. Applying (3.59) recursively for the moments in the right hand side, leads to a set of linear equations for the unknowns $M^0_{a_i,b_i}$, for different intervals $[a_i, b_i]$. Since the support of ϕ is finite, only those moments where the interval intersects the support $[s_1, s_2]$ are nonzero.

We define $S(a,b)$ as the set of all intervals generated starting from $[a,b]$, by recursively adding $[2a_i-k, 2b_i-k]\cap[s_1,s_2]$, $k=s_1,\dots,s_2$, for each interval $[a_i,b_i]$ in the set. These intervals correspond to the unknown moments in (3.59). We will first generalise equation (3.59) to moments of higher order, and then we will discuss the size of the set $S(a,b)$.

3.7.3.2 An algorithm based on using Chebyshev polynomials

For the computation of $M^l_{a,b}$, we may use the Chebyshev polynomials $T_l(x)$ scaled to the interval $[s_1,s_2]$. In this way we avoid evaluating the polynomials outside the interval $[-1,1]$. One scaling can be converted to another easily by using the following relation,

$$T_n\left(\frac{x+\lambda_1}{L_1}\right) = 2^{-n}\sum_{i=0}^{n} w_i^{(n)} T_i\left(\frac{x+\lambda_2}{L_2}\right), \tag{3.60}$$

with real parameters $\lambda_1, L_1, \lambda_2$ and L_2, and $L_1, L_2 \neq 0$. Note that the parameter $w_i^{(n)}$ depends on $\lambda_1, L_1, \lambda_2, L_2$. A stable recursive scheme to compute the coefficients $w_i^{(n)}$ in (3.60) is given in [121]. The scheme is based directly on the three-term recurrence relation for Chebyshev polynomials. A similar scheme, only slightly different, was used and derived in [184].

The moments can then be found in the following way. If $l=0$, we solve the system of equations of type (3.59). For larger l, we use (3.60) to introduce lower order Chebyshev polynomials,

$$\begin{aligned}
M^{l+1}_{a,b} &= \int_a^b T_{l+1}\left(\frac{x+\lambda_1}{L_1}\right)\phi(x)\,\mathrm{d}x \\
&= \frac{\sqrt{2}}{2}\sum_k h_k \int_{2a-k}^{2b-k} T_{l+1}\left(\frac{x+k+2\lambda_1}{2L_1}\right)\phi(x)\,\mathrm{d}x \\
&= \frac{\sqrt{2}}{2}\sum_k h_k \sum_{i=0}^{l+1}\int_{2a-k}^{2b-k} 2^{-(l+1)} w_i^{(l+1)}(k) T_i\left(\frac{x+\lambda_1}{L_1}\right)\phi(x)\,\mathrm{d}x \\
&= \frac{\sqrt{2}}{2}\sum_k h_k \sum_{i=0}^{l+1} 2^{-(l+1)} w_i^{(l+1)}(k) M^i_{2a-k,2b-k}.
\end{aligned}$$

The coefficients $w_i^{(l+1)}(k)$ are written this way in order to make the dependence on the parameter k explicit in the equation, since we applied (3.60) with a k-dependent parameter λ_2. The coefficients need to be computed once for every value of k. The parameters λ_1 and L_1 are defined here such that $T_l\left(\frac{x+\lambda_1}{L_1}\right) = T_l\left(2\frac{x-s_1}{s_2-s_1}-1\right)$.

A separation of the known and unknown components yields the equation

$$M_{a,b}^{l+1} - \frac{\sqrt{2}}{2}\sum_k h_k 2^{-(l+1)} w_{l+1}^{(l+1)}(k) M_{2a-k,2b-k}^{l+1} \qquad (3.61)$$
$$= \frac{\sqrt{2}}{2} 2^{-(l+1)} \sum_k h_k \sum_{i=0}^{l} w_i^{(l+1)}(k) M_{2a-k,2b-k}^{i}.$$

The right hand side of (3.61) is fully known and can easily be evaluated at step $l+1$.

In order to find the quadrature weights for the integration on the interval $[a,b]$, we need to solve a system of equations similar to (3.54),

$$Q\left[T_l\left(2\frac{x-s_1}{s_2-s_1}-1\right)\right] = M_{a,b}^l. \qquad (3.62)$$

The matrix entries are evaluations of Chebyshev polynomials, scaled to $[s_1,s_2]$, in the interval $[a,b]$. In order to obtain a good condition number, we will use the same matrix as in (3.54). The Chebyshev polynomials are then scaled to the interval $[a,b]$. The new right hand side can be found from the moments $M_{a,b}^l$, by combining (3.58) and (3.60),

$$Q\left[T_l\left(2\frac{x-a}{b-a}-1\right)\right] = \tilde{M}_{a,b}^l := \sum_{i=0}^{n} 2^{-n} w_i^{(n)} M_{a,b}^i. \qquad (3.63)$$

3.7.3.3 Computational complexity of the construction

The performance of the algorithm described above, depends on the cardinality of the set $S(a,b)$ defined in §3.7.3.1. The following lemma shows that the set is finite only when a and b are rational numbers.

Lemma 3.7.1. *$\#S(a,b) < \infty \iff a,b \in \mathbb{Q}$.*

Proof. As can be seen from the recursion, each interval in $S(a,b)$ can be written as $[2^n a - z, 2^n b - z] \cap [s_1,s_2]$, $n \in \mathbb{N}$, $z \in Z$. For n large enough, one endpoint of the interval will always be s_1 or s_2.

Assume a is irrational, i.e., $a \in \mathbb{R}\setminus\mathbb{Q}$. Set $a_0 := a$, and define $a_i := 2a_{i-1} - k_i$ for a sequence $\{k_i | k_i \in \mathbb{Z}\}$ such that $a_i \in [s_1,s_2]$. The cardinality of $S(a,b)$ can only be finite if the sequence $\{a_i\}$ is self-repeating. Then a has to solve $2^n a - k = 2^m a - l$, $m,n \in \mathbb{N}$, $k,l \in \mathbb{Z}$. This means $a = \frac{k-l}{2^n-2^m}$. Clearly, a cannot be irrational.

Now assume a is rational. It can be written as $a = \frac{c}{d}$, $c,d \in \mathbb{Z}$. Then $2^n a - z = \frac{2^n c - dz}{d}$ is again a rational number with the same denominator. The cardinality of the set $\{a_i\}$ is bounded by the total number of such rational numbers in $[s_1, s_2]$, $d(s_2 - s_1)$. A similar reasoning for b proves boundedness of $\#S(a, b)$, with a cardinality that is bounded by the sum of $\#\{a_i\}$ and a similarly defined set $\#\{b_i\}$. $\square$

The lemma shows that the system to be solved for the moments of the scaling function will be the smallest for integers or rational numbers a and b with a small denominator. For irrational values, it is infinitely large; obviously the algorithm can then not be applied. However, each number on a computer is represented by a rational number $a \simeq A2^{-N}$, $A \in \mathbb{Z}$. The interval sequence will self-repeat after N recursion steps, since then $2^N a - z \in \mathbb{Z}$ and $2^N b - z \in \mathbb{Z}$. An upper bound for the cardinality of $S(a, b)$ is then given by $2N(s_2 - s_1)$, i.e., the number of iterations times the number of integer shifts of $2^i a$ and $2^i b$ that lie in $[s_1, s_2]$ in each iteration i. Typically, however, the size of the set is way smaller than this upper bound: this will be illustrated with some examples further.

When the construction of a quadrature rule is required in a time-critical part of a program, it may be desirable to reduce the size of the system to be solved. This can be done by computing the moments using rounded values $\overline{a} \simeq a$ and $\overline{b} \simeq b$ that guarantee a lower cardinality $\#S(\overline{a}, \overline{b})$. The constructed rule can be seen as a rule for the integration on the interval $[\overline{a}, \overline{b}]$. If $\epsilon := \max(|a - \overline{a}|, |b - \overline{b}|)$ is the roundoff error, the integration error can be estimated by

$$\int_a^{\overline{a}} f(x)\phi(x)\,\mathrm{d}x + \int_{\overline{b}}^{b} f(x)\phi(x)\,\mathrm{d}x \leq 2\epsilon \max_{x \in [a,\overline{a}] \cup [b,\overline{b}]} (f(x)\phi(x)) = 2\epsilon M.$$

A good estimate for the constant M is just $\max(f(a)\phi(a), f(b)\phi(b))$. Experiments indicate that this error bound is sharp, giving good control of the round-off error.

3.7.3.4 Convergence of the quadrature rule

Define the integration error $E_{a,b}[\cdot]$ as

$$E_{a,b}[f(x)] := \int_a^b f(x)\phi(x)\,\mathrm{d}x - Q_{a,b}[f(x)]. \tag{3.64}$$

If $f(x) \in C^{q+1}[a, b]$ then the first $q + 1$ terms of the Taylor expansion of f around a point in $[a, b]$ are integrated exactly, and the error will depend on the $(q + 1)$-th derivative of f.

To specify this further, let $P_q(x)$ be the polynomial of degree q that interpolates the function f in $x_1, \ldots, x_r$, with $r = q + 1$. Then [31],

$$\forall x \in [s_1, s_2] : \exists \xi(x) \in [s_1, s_2] : f(x) = P_q(x) + e_q(x),$$

Table 3.2: Absolute error for the integration of the functions f_i, specified in Example 3.7.2, with CDF or Daubechies (DB) scaling functions.

$[s_1, s_2]$	CDF24 $[-1,1]$		CDF24 $[-1,1]$		DB2 $[-2,2]$	DB3 $[-3,3]$
s	f_1	f_1 split	f_2	f_2 split	f_2 split	f_2 split
0	$5.6E-2$	$1.5E-2$	$5.6E-1$	$1.5E-2$	$7.1E-2$	$1.4E-2$
-1	$4.5E-4$	$1.4E-4$	$9.9E-2$	$3.0E-4$	$2.1E-4$	$5.4E-6$
-2	$8.1E-8$	$4.6E-9$	$1.5E-2$	$4.4E-8$	$4.3E-11$	$9.6E-13$
-3	$6.7E-16$	$3.3E-16$	$1.5E-1$	$1.6E-15$	$(8.6E-6)$	$(3.0E+9)$
$\sum \lvert w_i \rvert$	4.3	2.9	4.3	2.9	78	5266

with

$$e_q(x) = \frac{\Pi(x)}{(q+1)!} f^{(q+1)}(\xi(x)), \quad \text{and} \quad \Pi(x) = \Pi_{i=1}^{r}(x - x_i).$$

The error is now given by

$$E_{a,b}[f(x)] = E_{a,b}[e_q(x)] = \frac{1}{(q+1)!} \int_a^b \phi(x)\Pi(x) f^{(q+1)}(\xi(x))\,\mathrm{d}x. \quad (3.65)$$

Estimates based on this expression are in general rather pessimistic. Moreover, the function $\xi(x)$ is not known. However, it can be seen from (3.65) that the asymptotic behaviour is essentially the same as in the rule for smooth functions. The relative error of the quadrature method with the degree of accuracy q remains $\mathcal{O}(h^{q+1})$, or $\mathcal{O}(h^r)$, $r = q + 1$. Yet, this order is to be evaluated for a smaller value of h, since $|b - a| < |s_2 - s_1|$.

For a smooth function f, this may not be the best solution. A more accurate result can be obtained by computing coefficients on a finer scale, and using the refinement equation to obtain values for the rougher scale. This would lead to an error of the order $\mathcal{O}(h^{2r})$. If the function is only piecewise smooth however, we can split the interval $[s_1, s_2]$ into pieces that correspond to the smooth parts of f. The convergence is then not adversely affected by discontinuities of f, or of any of its derivatives.

Example 3.7.2. *We consider the functions* $f_1(x) := \cos(2x) + \sin(3x)$ *and* $f_2(x) := \cos(|2x|) + \sin(|3x|)$*. We compare the integration rule of [184], discussed in §3.7.2, with the integration rule discussed in §3.7.3. The parameter* s *determines the number of abscissae* r *used in the interval* $[s_1, s_2]$ *for the first method:* $r = 2^{-s}(s_2 - s_1) + 1$*. For the second method, the interval is split at the origin, and* r *abscissae are used in both intervals. Hence, there is a total of* $2r$ *abscissae.*

The values in Table 3.2 represent the absolute error for the integrals $\int_a^b f_i(x)\phi(x)\,\mathrm{d}x$. The values in the last row represent the maximum sum of the absolute values of the weights that were used in the corresponding column. We consider three different scaling functions. The numbers 2 and 4 in the notation CDF24 represent the order of the primal and dual wavelets respectively of the CDF wavelet. The Daubechies wavelets of order 2 and 3 have very low regularity on each subinterval of their support.

The first two columns show that splitting the interval (and thus doubling the points) is not very useful for the case of a smooth function. The result is only slightly better, and does not compensate sufficiently for the extra effort. For function f_2 however, which is not smooth at the origin, the regular method shows no convergence. The second method converges rapidly to almost machine precision. Similar results are obtained for the scaling functions of two different Daubechies wavelets. The values corresponding to $s = -3$ in the last two columns indicate the presence of a large error, due to a number of abscissae greater than 30 (respectively 33 and 49). This is an illustration of the instability for large r. The values are given between parentheses. In order to get better accuracy results, the subintervals should be split into a larger number of smaller intervals.

The values in the last row, i.e., the maximum sum of the weights, are very moderate, even for the Daubechies scaling functions that switch sign. This indicates good stability properties of the constructed rules.

Example 3.7.3. *We now look at a different example to illustrate the size of* $S(a,b)$*. The computation of the moments on the interval* $[\pi/10, \pi/4]$ *for the CDF scaling functions, leads to a system with* 195 *unknowns with a representation in double precision. The condition number of the system to be solved in level* $l = 0$ *of the algorithm is only* 47*. It is smaller in the next levels corresponding to the higher order moments. Rounding the interval boundaries to the nearest multiple of* 2^{-16} *reduces the size of the system to* 57 *unknowns, and a maximum condition number of* 27*. The upper bound on the number of unknowns here is* $2N(s_2 - s_1) = 64$*. The error induced on the integration is* $5E-6$ *for the function* $f(x) = 1$*.*

This example illustrates the trade-off between computation time for the construction of the rule, and the round-off error for intervals with irrational endpoints. The error can always be made as small as needed however.

Having described the convergence as a function of q, it is also of interest to consider the convergence as a function of j, which is the scale of the scaling function $\phi_{j,k}(x)$ in the integrand. This is because in most applications we would like to match the integration error of the numerical quadrature to the discretisation error of the corresponding wavelet approximation on a given scale. The values we want to compute using rule (3.57) are given by

$$d_{j,k} = \int_{a_n}^{b_n} f(x)\phi_{j,k}(x)\,\mathrm{d}x. \tag{3.66}$$

Table 3.3: Relative error for the integration of f_1 on the interval $[a_j, b_j]$ for DB3 wavelets.

q	$j = 0$	$j = 1$	$j = 2$	$j = 3$
0	$1.1E-1$	$2.6E-1$	$1.8E-1$	$1.0E-2$
2	$4.0E-3$	$6.6E-4$	$9.6E-5$	$1.3E-5$
4	$2.1E-5$	$8.9E-7$	$3.2E-8$	$1.1E-9$
6	$8.1E-8$	$8.6E-10$	$8.7E-12$	$9.9E-14$

For the case of smooth functions in §3.7.2, seeing that h is proportional to 2^{-j}, one obtains the error estimate $O(2^{-j(q+1)})$ that was mentioned in §3.7.2.2. Here, contrary to that single case, we need to consider two cases:

1. the endpoints of the integration interval change with j, such that the same part of the scaling function is covered on each scale. In this case, $a_j = 2^{-j}(a + k)$.
2. the endpoints remain fixed as j increases, i.e., $a_j = a$.

The first case occurs, e.g., when we would like to increase the accuracy of the integration, by splitting the support of the scaling function in a finite number of subintervals. The second case occurs when f has a discontinuity (in a derivative) at a fixed point a or b. We can see that the convergence in the first case will be asymptotically similar to the case of smooth functions, i.e., $O(2^{-j(q+1)})$, albeit with a smaller constant since the integration interval is also smaller. In the second case, the error will still behave like $O(h^{q+1})$, but h does not scale as 2^{-j} initially. That only happens when the scaling parameter j becomes large enough, such that the support of the scaling function is contained entirely within the fixed interval $[a, b]$. The problem then reduces to the previous case.

Example 3.7.4. *To illustrate the above discussion, consider again the function $f_1(x)$. Table 3.3 shows the relative error for the case of Daubechies wavelets with $k = 0$.*

The error for fixed q decreases with the expected factor $2^{-(q+1)}$. For fixed j, we expect to see in Table 3.3 a convergence rate of $2^{-j(q+2+1)}/2^{j(q+1)} = 2^{2j}$. For increasing j, the convergence rates indeed approximately improve by a factor of 4 from left to right in each column.

3.7.4 An integration rule for singular functions

3.7.4.1 Functions with a known singularity

The method can be extended to work for functions with an integrable singularity, e.g., $s(t) = |t|^\alpha$, for $-1 < \alpha < 0$ or $s(t) = \log(|t|)$. First we will assume that the singularity of f is known analytically and can be subtracted, i.e., $f(x) = p(x) + q(x)s(x - x')$, $x' \in [s_1, s_2]$, where $p(x)$ is a non-singular function. We develop a quadrature rule $Q^s[\cdot]$ such that

$$\int_a^b f(x)\phi(x)\,\mathrm{d}x = \int_a^b p(x)\phi(x)\,\mathrm{d}x + \int_a^b q(x)s(x-x')\phi(x)\,\mathrm{d}x \simeq Q[p] + Q^s[q],$$

with $Q[p]$ being quadrature rule (3.52).

We demand for quadrature rule $Q^s[\cdot]$ an exact integration of the functions $s(x - x')P_l(x)$, with $P_l(x)$ from (3.56). The required moments are in their most general form given by

$$M_{a,b}^{l,m} := \int_a^b T_l\left(\frac{x + \lambda_1}{L_1}\right) s(x - m)\phi(x)\,\mathrm{d}x. \tag{3.67}$$

First, we discuss how to deal with the singularity, and consider the integration interval $(-\infty, \infty)$. Using the refinement relation, we have

$$\begin{aligned} M^{0,m} &:= \int_{-\infty}^{+\infty} s(x - m)\phi(x)\,\mathrm{d}x \\ &= \frac{\sqrt{2}}{2}\sum_k h_k \int_{-\infty}^{+\infty} s\left(\frac{x + k}{2} - m\right)\phi(x)\,\mathrm{d}x. \end{aligned} \tag{3.68}$$

Hence, in the right hand side integrals, the singularity has been shifted. For the algebraic singularity, the shifted singularity can be rewritten in the original notation $s(x - m)$,

$$\left|\frac{x + k - 2m}{2}\right|^\alpha = |x - (2m - k)|^\alpha\, 2^{-\alpha}$$

and, similarly, for the logarithmic singularity,

$$\log\left(\left|\frac{x + k - 2m}{2}\right|\right) = \log\left(|x - (2m - k)|\right) - \log(2).$$

For $s(t) = \log(|t|)$, relation (3.68) becomes

$$M^{0,m} = \frac{\sqrt{2}}{2}\sum_k h_k M^{0,2m-k} - \log(2)\frac{\sqrt{2}}{2}\sum_k h_k, \tag{3.69}$$

while for $s(t) = |t|^\alpha$ we find

$$M^{0,m} = \frac{\sqrt{2}}{2}2^{-\alpha}\sum_k h_k M^{0,2m-k}.$$

Recursive application of the above expressions for different values of m leads again to a set of linear equations in the unknown moments $M^{0,m^{(i)}}$. The parameter $m^{(i)} = 2m^{(i-1)} - k$ with $m^{(0)} = m$ grows in principle without bound. Yet, if it is large enough the integral is no longer singular. The corresponding moments can then be computed by using the techniques of §3.7.2. For accurate computations, it is better to also include the nearly singular moments as unknowns in the set of equations. Good results were obtained by including the moments for all intervals that satisfy $\operatorname{dist}(m, [a, b]) < 1$, i.e., when the distance of the singularity to the integration interval is of the same order as the size of the interval, which is $O(1)$ on scale $j = 0$.

Combined with the approach of splitted intervals, we find a linear equation for each moment. For example, for the logarithmic singularity, we have an equation of the following type,

$$M_{a,b}^{l+1,m} = 2^{-(l+3/2)} \sum_k h_k \sum_{i=0}^{l+1} w_i^{(l+1)}(k) \left(M_{2a-k,2b-k}^{i,2m-k} - \log(2) M_{2a-k,2b-k}^{l+1} \right). \tag{3.70}$$

The moments can be computed for each value of l successively, starting from $l = 0$ and the known partial moments $M_{a,b}^{l}$.

Again, the size of the system to be solved is finite only under certain conditions. Define $S(a, b, m)$ as the set of intervals and singularity locations corresponding to the moments found by applying (3.70) recursively, for which $\operatorname{dist}(m, [a, b]) < 1$. The cardinality of this set is bounded only if m is a rational number.

Lemma 3.7.5. *$\#S(a, b, m) < \infty \iff m \in \mathbb{Q}$.*

Proof. The intervals $[2^n a - z, 2^n b - z] \cap [s_1, s_2]$ with corresponding singularity $2^m - z$ are in $S(a, b, m)$ only if $2^m - z \in [s_1 - 1, s_2 + 1]$. Using the same line of reasoning as in Lemma 3.7.1 leads to the condition $m \in \mathbb{Q}$.

Conditions on a and b are not required if $a, b \neq m$, since for n large enough we have that $2^n a - z - (2^n m - z) = 2^n (a - m) > s_2 - s_1 + 1$. This means that for any $2^n m - z \in [s_1 - 1, s_2 + 1]$, $[2^n a - z, 2^n b - z] \cap [s_1, s_2] = [s_1, s_2]$. □

Note that, in order to compute the nonsingular moments, it is still required that a and b be rational numbers. In a time-critical code part, m can also be rounded to a near rational number. The error made is given by

$$\int_a^b f(x)\phi(x)(\log|x - m| - \log|x - \overline{m}|)\, \mathrm{d}x \leq M(m - a)\log|a - m| + (a - \overline{m})\log|a - \overline{m}| + (b - m)\log|b - m| + (\overline{m} - b)\log|b - \overline{m}|,$$

with $M = \max f(x)\phi(x)$. We have $x\log(x) - (x+\epsilon)\log(x+\epsilon) \approx \epsilon(\log(x)+1)$. Using this expression, we see that when $m = a$ or $m = b$, the error has order $O(\epsilon\log(\epsilon))$. Otherwise it has order $O(\epsilon)$.

Table 3.4: Absolute error for the quadrature approximation of the inner product of $\log(|x|)f_i(x)$ with CDF24, DB2, or DB3 scaling functions.

	CDF24	CDF24	DB2	DB3
r	$\log(\|x\|)f_1$	$\log(\|x\|)f_2$	$\log(\|x\|)f_2$	$\log(\|x\|)f_2$
3	$4.1E-2$	$1.6E-2$	$8.9E-1$	$1.3E-1$
5	$2.8E-4$	$7.2E-4$	$1.4E-1$	$6.0E-2$
9	$1.8E-9$	$1.5E-7$	$5.0E-4$	$3.2E-3$
13	$1.6E-13$	$6.3E-12$	$4.2E-7$	$1.5E-5$
17	$5.5E-15$	$8.9E-15$	$1.2E-10$	$2.1E-8$
$\sum \|w_i\|$	1.5	117	263	94

The error $E^s[f(x)] := \int_a^b \log(x-x')f(x)\phi(x)\,\mathrm{d}x - Q^s[f(x)]$ is given by

$$E^s[e_q(x)] = \frac{1}{(q+1)!}\int_a^b \log(x-x')\phi(x)\Pi(x)f^{(q+1)}(\xi(x))\,\mathrm{d}x. \tag{3.71}$$

This leads, asymptotically, to the same relative error $O(h^{q+1})$, with $h = 2^{-j}$ or smaller depending on the size of $[a,b]$, for scaling functions on scale j.

Example 3.7.6. *Table 3.4 lists the absolute error of the approximation obtained by the quadrature method for two singular functions,* $\log(|x|)f_1(x)$ *and* $\log(|x|)f_2(x)$*, with* $f_1(x)$ *and* $f_2(x)$ *as defined in Example 3.7.2. In this example, we compare different scaling functions for a fixed number of abscissae* r*. For the test function* f_2*, the interval* $[s_1, s_2]$ *is split at the origin in order to cope with the discontinuity of the derivative.*

It is clear from Table 3.4 that the results converge rapidly. Note that the Daubechies wavelets require a larger number of abscissae for the same absolute error, due to a wider support. The largest system that was needed to compute the required moments for this table had only dimension 6.

3.7.4.2 Functions with an unknown singularity

In some cases, one does not wish to subtract the singularity explicitly, as in §3.7.4.1, and use a separate rule for the smooth and the singular parts. It is still possible to compute an efficient quadrature rule in this situation, by requiring exactness of the integration result for the functions $T_l(x)$ and $s(x-x')T_l(x)$. The right hand side of the resulting set of equations is set up in a similar way as in §3.7.3.1 and §3.7.4.1, based on the moments $M^l_{a,b}$ and $M^{l,m}_{a,b}$. Obviously, the number of abscissae necessary to obtain a degree of accuracy doubles, as does the size of the system, i.e., $r = 2q+2$ if the above functions are used for $l = 0, 1, \ldots, q$.

Table 3.5: Absolute error for the approximation of the inner product with $\log(|x|)f_i(x)$ via rules with known (A) and unknown (B) singularity.

	CDF24 $\log(\lvert x\rvert)f_1$		DB2 $\log(\lvert x\rvert)f_1$		CDF24 $\log(\lvert x\rvert)f_2$	
q	A	B	A	B	A	B
1	$1.9E-0$	$2.0E-2$	$2.1E-0$	$1.1E-0$	$8.3E-1$	$3.5E-1$
3	$4.8E-2$	$8.2E-4$	$8.2E-1$	$4.0E-1$	$2.8E-2$	$2.5E-2$
7	$3.4E-6$	$3.0E-6$	$3.2E-3$	$1.6E-3$	$7.6E-6$	$9.5E-5$
11	$4.4E-11$	$9.4E-11$	$2.1E-6$	$7.5E-7$	$5.0E-10$	$5.9E-8$
15	$2.0E-15$	$5.8-14$	$3.7E-10$	$9.1E-11$	$1.4E-14$	$(9.7E-7)$
$\sum \lvert w_i\rvert$	1.7	1625	2.0	4180	67	$9.9E8$

Unfortunately, the matrix of the resulting system may become ill-conditioned again. The method is usable however in practice, if the required degree of accuracy for the application is not too high. If the unknown functions $p(x)$ and $q(x)$ are smooth, as in the example below, conditioning does not pose a significant problem. For the function $\log(|x|)f_2(x)$, with $f_2(x)$ defined in Example 3.7.2, we could only obtain good rules with an order of up to 11 after we split the interval at the origin. Rules with higher accuracy require different basis functions, to avoid the ill-conditioning, or the use of higher precision arithmetic.

Example 3.7.7. *We consider functions with known singularity, and compare the results with those obtained with the previous method. All calculations were performed in double precision; no additional measures were taken. The results are given in Table 3.5.*

For function $\log(|x|)f_1(x)$, there is no significant difference between the two methods for most values of q. The method for unknown singularity requires two times the number of abscissae (i.e., $r = q + 1$ for method A and $r = 2q + 2$ for method B). That explains why in some rows the result is actually better. The degree of accuracy is the same in both cases.

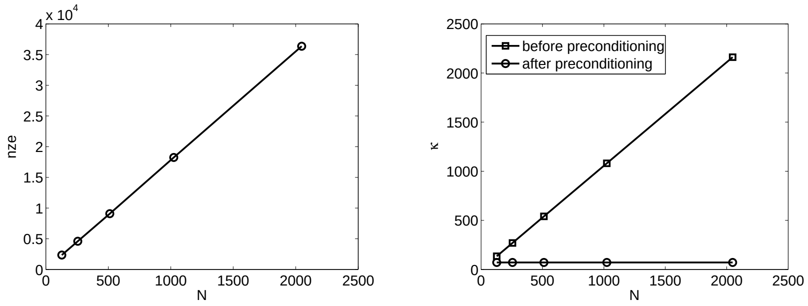
The last column illustrates the problem of ill-conditioning. The convergence of method B stops after the fourth row. Depending on the application, the error may however be already small enough at that point.

3.8 Numerical results

In this section, we illustrate the theoretical results that were obtained in this chapter with numerical examples. The fill-in of the discretisation matrix for increasing wavenumbers is illustrated in §3.8.1. The reduced complexity of the matrix-vector product using wavelet packets is discussed in §3.8.2. Finally, we illustrate the application of the quadrature rules that were constructed in §3.8.3 to the integrals arising in the boundary element method.

3.8.1 Wavenumber dependence

First, we illustrate the wavelet compression of the discretisation matrix for fixed k and increasing N. The typical structure of a sparse wavelet-transformed matrix for two-dimensional problems was shown in Figure 3.1. The number of significant elements in the discretisation matrix for a fixed value of the wavenumber k and for increasing N is shown in the left panel of Figure 3.5. The condition number before and after the preconditioning defined by (3.20) is shown in the right panel. The number of unknowns is approximately linear in N, and the condition number after preconditioning is bounded. The scattering obstacle is an ellipse with radius $R_1 = 0.3$ and $R_2 = 0.5$ along the X- and Y-axis. The wavenumber was chosen $k = 10$.



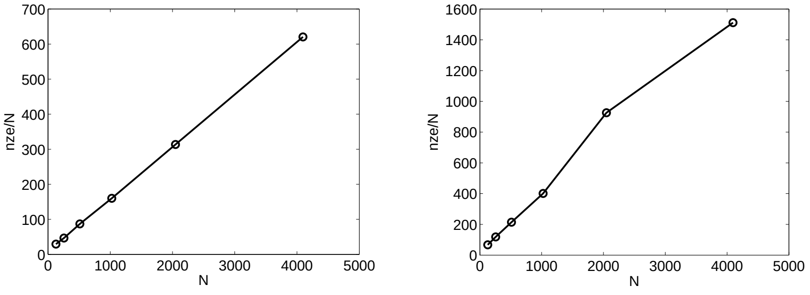
(a) Significant entries (b) Condition number

Figure 3.5: The number of significant entries (nze) and the condition numbers for scattering by an ellipse for a fixed value $k = 10$ of the wavenumber.

The situation is different when k increases proportionally to N. In that case, we have shown that the number of significant elements scales approximately as $O(kN) = O(N^2)$. In order to compare the number of significant entries for different values of k, we have chosen a simple level-independent threshold. Following the analysis in §3.6.3.3, the threshold was chosen

$$\tau = \frac{\delta}{N} \|M\|_1.$$

The parameter δ can be chosen so as to obtain a certain accuracy; here it is set to 0.1. The numerical results in Figure 3.6 show the dependence on the wavenumber for scattering by a circle and an ellipse. The circle has radius $R = 0.5$, the ellipse is the same as in the previous example. The figure shows the number of significant elements divided by N. This corresponds to the numerical values of the constant $C(k)$ in the computational complexity



(a) Circle

(b) Ellipse

Figure 3.6: The number of significant entries divided by N (nze/N) for a scattering problem on a circle and an ellipse, with 10 points per wavelength.

estimate $O(C(k)N)$ of the wavelet method. The behaviour seems to be linear in each case, $C(k) \sim k$, as predicted by the theory.

3.8.2 Wavelet-packet based methods

In this section, we will discuss the results of four numerical experiments to evaluate the proposed methods based on wavelet packets. First, we evaluate the performance of the discussed matrix compression methods for a fixed threshold that scales with N as determined in §3.6.3.3. Next, we evaluate the matrix compression for a threshold that is adaptively chosen in order to obtain a fixed error. Third, we investigate the complexity of the number of nonzero entries in the compressed matrices, and finally we discuss the condition number of the discretisation matrices.

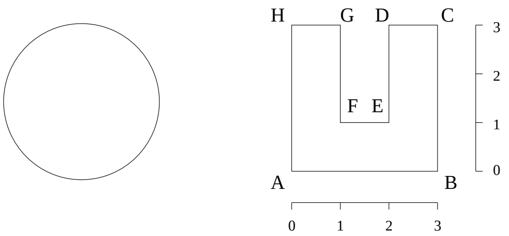


Figure 3.7: The obstacles: a circle with diameter $D = 1$, and a duct with circumference 16.

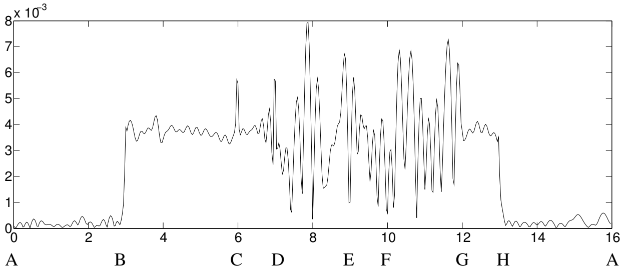


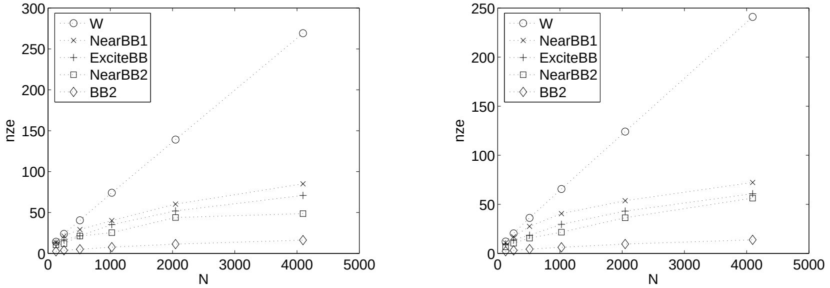
Figure 3.8: Absolute value of the solution along the boundary of the duct, counter-clockwise from the bottom-left point, with $k = 20$ and $N = 512$.

We consider two different shapes for the boundary Γ, a circle and a duct, inspired by the shapes used in [74, 199], see Figure 3.7. The circle has diameter $D = 1$, and the duct has a total circumference $D = 16$. We have obtained qualitatively similar results with other shapes. Although a simple geometry, the duct has sharp corners, and a non-convex shape with possible resonances - these are two important complications in two-dimensional wave scattering. The extension of the method to more complicated geometries, including multiple scattering configurations, is straightforward. Some numerical results involving multiple scattering are presented in [122].

The boundary condition in all examples is a plane wave e^{ikx}, incoming at an angle of 45 degrees with respect to the X-axis. The value of k is chosen proportional to N, such that there are 10 degrees of freedom for each wavelength. For completeness, we also depict a solution for one particular value of k in Figure 3.8. We compare five compression methods for each example: the regular wavelet transformation itself, the approach of [74] and [93], discussed in §3.6.1, the two-dimensional best basis algorithm and its top-down approximation that was suggested in §3.6.4.2. These methods will be referenced by (W), (NearBB1), (RhsBB), (BB2) and (NearBB2).

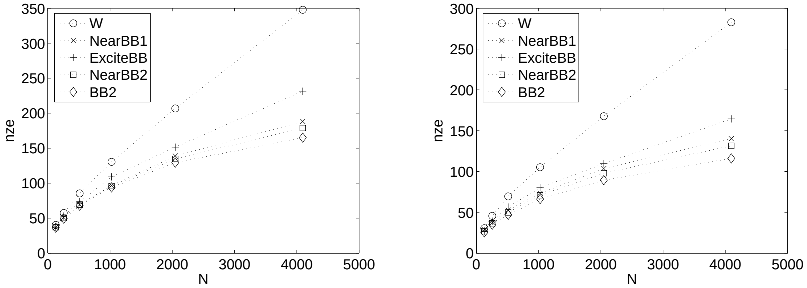
We used the two error measures (3.44), i.e., the relative solution error and the relative residual error. As a first example, we chose a threshold such that the error for the problem with size $N = 128$ was about 2.0%. The threshold was then scaled with N as suggested by the estimate (3.48). The results are given in Figure 3.9. The plotted values are the numbers of nonzero elements in the compressed discretisation matrix, divided by N. Corresponding to the previous experiment in §3.8.1, we expect the line representing the regular wavelet transformation to be linear, i.e., the number of elements grows quadratically with N. This is clearly visible in the figure.

The results show a much reduced number of significant elements of the wavelet packet transformations compared to the wavelet transformation. The residual error criterion increases the sparsity level in all cases. The



(a) nze/N for the circle, relative solution error

(b) nze/N for the circle, relative residual error

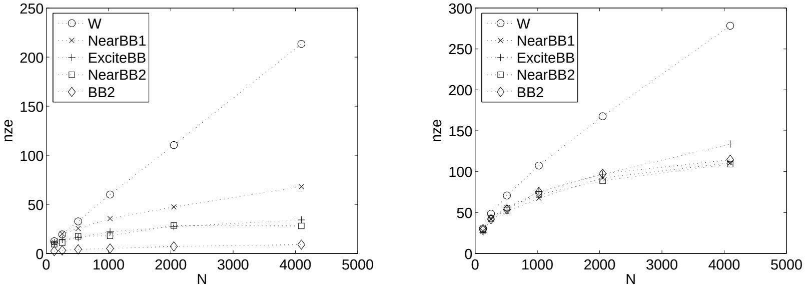


(c) nze/N for the duct, relative solution error

(d) nze/N for the duct, relative residual error

Figure 3.9: The number of nonzero elements (nze) relative to N after thresholding with a scaling threshold. The initial threshold is such that the error (e_R or e_S) is 2% for $N = 128$.

two-dimensional best basis produces the highest sparsity, and is better than the wavelet transformation by a factor of 17 at $N = 4096$ for the circle. The wavelet packet basis captures the inherent problem frequency much better than the classical wavelet basis. The factor for the duct is approximately 2.5. Due to the sharpness of the corners and the complex shape, it appears more difficult for the wavelet packet basis to capture all of the relevant frequencies. The computed errors e_R and e_S are not constant at 2% however, and tend to decrease slowly with N. All values varied between 2.5% and 1% for the current example. The errors for the wavelet-packet transformations seemed to decrease faster than for the wavelet transformation. These observations



(a) nze/N for the circle, $e_R = 2\%$ (b) nze/N for the duct, $e_R = 2\%$

Figure 3.10: The number of nonzero elements (nze) relative to N after thresholding with an adaptive threshold, such that the relative residual error is fixed at $2 \pm 0.05\%$.

indicate that the scaling threshold (3.48) is somewhat too restrictive, and that the wavelet-packet transformations not only produce more sparsity, but also a more accurate representation. The threshold can therefore be larger than for the wavelet transformation.

To quantify the maximal possible sparsity in this setting, we choose the threshold in the next example such that the error is kept constant at $2.0 \pm 0.05\%$. Of course, in practice this procedure is not feasible, but for analysis purposes it does provide some additional insight. A binary search algorithm, starting with the scaling threshold as initial value, found a suitable threshold value typically in 4 to 6 iterations. The results are given in Figure 3.10 for the residual error criterion. The threshold values scaled approximately as $O(N^{-1/3})$ for (BB2). The sparsity is much improved: for $N = 4096$, (BB2) for the circle requires $36,596$ elements to satisfy the error criterion, compared to $56,842$ in the previous example. The difference for the duct is in this case negligible: $468,775$ elements for the adaptive threshold, compared to $475,246$ for the scaling threshold. The corresponding dense matrix has 16 million elements.

One should note that for the duct, (NearBB2) and (BB2) are not significantly better than the one-dimensional (NearBB1). While (BB2) guarantees the best compression for a two-dimensional basis for a fixed threshold, the criterion in which we match the threshold to the final compression error leads to different thresholds for the different methods. The (NearBB2) method is the best one in this case by a small margin. It was verified for other obstacles that the largest gain with (BB2) is obtained for smooth obstacles. For non-smooth obstacles with corners, such as the duct, the top-

	scaling threshold					fixed error $e_R = 2\%$				
N_2	256	512	1024	2048	4096	256	512	1024	2048	4096
W	1.74	1.77	1.87	1.91	1.95	1.64	1.76	1.88	1.88	1.95
NearBB1	1.67	1.46	1.46	1.59	1.50	1.86	1.36	1.48	1.41	1.52
RhsBB	1.41	1.37	1.70	1.56	1.45	1.37	1.16	1.45	1.31	1.33
NearBB2	1.19	1.84	1.26	1.79	1.15	1.22	1.66	1.06	1.60	0.98
BB2	1.53	1.44	1.59	1.58	1.51	1.43	1.41	1.28	1.50	1.33

Table 3.6: Value of β in the estimate $S = O(N^\beta)$ for the circle, using the residual error criterion.

	scaling threshold					fixed error $e_R = 2\%$				
N_2	256	512	1024	2048	4096	256	512	1024	2048	4096
W	1.58	1.61	1.60	1.67	1.75	1.67	1.54	1.60	1.64	1.73
NearBB1	1.37	1.49	1.48	1.50	1.44	1.63	1.29	1.40	1.45	1.26
RhsBB	1.56	1.53	1.50	1.45	1.58	1.77	1.39	1.42	1.36	1.47
NearBB2	1.45	1.45	1.51	1.47	1.42	1.52	1.38	1.38	1.30	1.30
BB2	1.45	1.43	1.50	1.43	1.38	1.54	1.36	1.48	1.37	1.23

Table 3.7: Value of β in the estimate $S = O(N^\beta)$ for the duct, using the residual error criterion.

down approximations (NearBB1) and (NearBB2) perform almost equally well or sometimes even marginally better than (BB2). The other obstacles considered were an ellipse, a rounded gear wheel and an L-shaped domain.

Finally, we would like to examine the asymptotic complexity of the number of nonzero elements in the compressed matrices as a function of N. Assume we can write the number of significant elements S as $S = O(N^\beta)$, for some $\beta \in \mathbb{R}$. The value of β can then be estimated from two successive discretisations with N_1 and N_2 unknowns, and S_1 and S_2 significant elements, by $\beta \approx \log(S_1/S_2)/\log(N_1/N_2)$. Tables 3.6 and 3.7 show the results for the circle and the duct. The value of β approximates 1.5 for the scaling threshold, but is lower for the threshold that corresponds to a fixed error. The number of elements is therefore empirically $O(N^{1.5})$ for the scaling threshold, and at best $O(N)$ for the computed threshold. A complexity of $O(N^{1.4})$ appears to be a good fit. This corresponds to the estimate that was obtained by simple arguments in §3.6.4.

The condition number of the formulation (3.41) depends on the choice of the coupling parameter η. With the optimal choice $\eta = k$, the condition number is monotonously but slowly increasing. The values for the test problems are given in Table 3.8. The condition number for the duct with $N = 4096$ unknowns is 465, compared to the moderately large value of $2.5e4$ for the common choice $\eta = 1$. The latter choice leads to more irregular behaviour for the condition number. It can be seen from formula (3.42) for the matrix elements, that the contribution of $\frac{I}{2} + D^*$ to the discretisation

Table 3.8: Condition number of the discretisation matrix for different choices of the coupling parameter η in equation (3.41).

N		256	512	1024	2048	4096	8192
$\eta = 1$	circle	101.6	20.0	92.0	852	197	$3.7e3$
	duct	$1.7e2$	$1.1e3$	$6.1e2$	$1.1e3$	$2.5e4$	$1.1e4$
$\eta = k$	circle	4.3	5.4	6.7	8.5	10.7	13.4
	duct	26.8	27.2	40.4	87.5	465	448

dominates for large values of k. The ill conditioning is therefore still caused by the resonant eigenvalues. The choice $\eta = k$ reduces this behaviour.

3.8.3 Wavelet quadrature

Some examples that illustrate the convergence rates of the constructed quadrature rules were already given in §3.7. Here, we illustrate the use of the quadrature rules for the efficient construction of the discretisation matrix. Specifically, the implementation exploits the fact that function evaluations can be shared among neighbouring matrix entries.

Each element of the matrix is given by a double integral of the form

$$e_{(j',k'),(j,k)} = \int_{\Omega_{j,k}} \int_{\Omega_{j',k'}} K(t,\tau)\phi_{j,k}(t)\phi_{j',k'}(\tau)\,\mathrm{d}\tau\,\mathrm{d}t, \qquad (3.72)$$

with $\Omega_{j,k} = \operatorname{supp}(\phi_{j,k})$. The application of any type of tensor-product quadrature rule with n points per dimension requires n^2 evaluations of the kernel function, and $2n$ evaluations of the basis functions. The application of the quadrature rules that were constructed in §3.7 also requires n^2 evaluations of the kernel function, and no evaluations of the basis functions. Due to the regular grid of quadrature points, it becomes easy to share the evaluation of the kernel function among neighbouring basis functions. The singularity of the kernel function for elements that are close to the diagonal can be treated with the constructed quadrature rules by noting that

$$H_0^{(1)}(z) = J_0(z) + iY_0(z) = J_0(z) + i\left(\frac{2}{\pi}(\log(z/2)+\gamma)J_0(z) + P(z)\right),$$

where $\gamma = 0.577\ldots$ is Euler's constant, and where $P(z)$ is a smooth function.

In order to illustrate the performance gain that results from an efficient implementation of this strategy, we conducted the following numerical experiment. The full matrix is constructed for a scattering problem, using the scaling functions of a number of different wavelets as basis functions. We

Table 3.9: Number of Bessel function evaluations required for constructing a discretisation matrix with $N = 128$. The table shows the total number of function evaluations, and the number of evaluations per element (p.e.).

CDF2 ([-1,1])	s	r	not shared	p.e.	shared	p.e.
	0	3	$146,304$	8.93	$19,096$	1.16
	-1	5	$407,936$	24.9	$74,818$	4.57
	-2	9	$1,324,416$	80.8	$285,314$	18.1
DB6 ([-6,6])	s	r	not shared	p.e.	shared	p.e.
	0	13	$2,747,264$	168	$32,736$	2.00
	-1	25	$10,199,936$	623	$129,218$	7.89
	-2	49	$39,261,056$	2396	$513,426$	31.3

have counted the total number of evaluations of the Hankel and Bessel functions for two different approaches of the construction. The first approach is the classical tensor-product application of the one-dimensional rule. The second approach is the optimised implementation where function evaluations are maximally shared. The evaluation of Bessel functions is much slower than basic arithmetic operations. The number of Bessel function evaluations is therefore a good measure for the computation time.

The results are shown in Table 3.9, for different values of the scale parameter s that was defined in §3.7. The number of weights in a one-dimensional rule is $r = 2^{-s}(s_2 - s_1) + 1$, where $[s_1, s_2]$ is the support of the scaling function. The number of evaluations per matrix entry in the classical tensor-product approach is close to r^2, as could be expected. The number of evaluations per entry in the optimised approach scales like r^2, but with a much lower constant. The total number of Bessel function evaluations is one or two orders of magnitude less for this example. In fact, in the case $s = 0$, the optimised approach is almost as cheap as the pointwise evaluation of the kernel function for each matrix entry. But the accuracy is much higher - for the Daubechies scaling function at $s = 0$, each element is computed from $r^2 = 169$ values of the kernel function. Likewise, the optimised Galerkin approach is competitive with collocation methods, where the evaluation of the elements involves a single integral rather than a double integral. The accuracy of the Galerkin approach is typically much higher than in a collocation approach with the same number of basis functions.

3.9 Three-dimensional problems

The extension of the wavelet method to three-dimensional problems requires a considerable extension of the theory. Specifically, the main obstacle is the construction of a suitable wavelet basis on higher-dimensional boundaries. The norm equivalences (3.18) are not satisfied in the full range of Sobolev spaces for most proposed constructions of multidimensional wavelets. This is necessary for the preconditioning of operators with negative order. One approach that satisfies all conditions for an optimal method is suggested in literature, but has not yet been implemented [62].

Still, several constructions exist that are close to optimal, and lead to good results in practice [61, 32, 44, 105, 180]. The approach in the given references are variations of the following: starting from biorthogonal wavelets on the interval $[0, 1]$, tensor-product wavelets are constructed on $[0, 1]^d$. The boundary Γ is divided into a number of patches. Wavelets are lifted from $[0, 1]^d$ to each patch using a parameterisation. Apart from this construction, the method follows the same lines in three dimensions as in two dimensions. Numerical results for three-dimensional problems are given in [106, 193].

3.10 Conclusions

The wavelet method is the only method discussed in this thesis that achieves $O(N)$ convergence, i.e., without a logarithmic term, in the low frequency regime where the accuracy of the solution scales with the discretisation error to preserve the optimal convergence rate of the Galerkin scheme. The reduction to linear complexity is achieved by the second compression step, that discards certain singular elements of the discretisation matrix. In the physical context of the scattering problem, these elements correspond to near field interactions. We will see in the next two chapters that the fast multipole method and hierarchical matrix methods do not perform approximations of the near field.

The theory of the wavelet method also focuses on the area of preconditioning. The use of wavelets presents a fundamental solution to the typical ill-conditioning of integral equations of the first kind. The matrix-vector product reduces to a regular matrix-vector product with a sparse matrix, that can be implemented very efficiently. The construction of the sparse matrix is complicated however. It involves the evaluation of possibly singular double integrals with a very large integration domain.

We have shown that the wavelet method is not suited for high frequency problems. A different approach was suggested based on wavelet packets, and it was shown that an appropriate wavelet packet basis again leads to a sparse discretisation matrix. In this method, the oscillatory problem is represented using oscillatory basis functions.

Chapter 4

Fast multipole methods

4.1 Introduction

A different multiscale approach for scattering problems is taken by the fast multipole method (FMM). There are two major variants of the method, often referred to by *low frequency* FMM and *high frequency* FMM, that are suitable for low-frequency and high-frequency scattering problems respectively. Similar to the wavelet method, the fast multipole method is efficient in the low-frequency regime. The computational complexity is $O(N)$ for a fixed error, and $O(N \log N)$ for an error that improves with increasing N. We will see that an implementation is also possible with complexity $O(N \log^p N)$ for one matrix-vector product in the high frequency regime, where N increases proportionally to k in two dimensions. Currently, this is the best known result for general oscillatory integral equations. The purpose of this chapter is the description of the fast multipole method, in order to appreciate the similarities and differences with hierarchical matrices and wavelet based methods described in the previous and the next chapter.

The fast multipole method was pioneered by Greengard and Rokhlin in [98], originally in the context of N-body simulations. Other fast algorithms were being developed for these problems at the time, such as particle-in-cell methods [114] and tree-codes [9, 15]. The principles of the fast multipole method were far more general however, enabling an adaptive implementation [34, 37] and applications to integral equations [171, 172]. Rokhlin introduced the so-called *diagonal form of translation operators* for two- and three-dimensional problems in [172, 173], which led to an efficient implementation for oscillatory integral equations in the high frequency regime. Originally, two level fast multipole methods were suggested, yielding $O(N^{3/2})$ or $O(N^{4/3})$ complexity. A multilevel method was proposed for two-dimensional scattering problems by Lu and Chew in [152], and for three-dimensional Helmholtz and Maxwell problems in [80, 170, 63, 64]. The implementation has a computational complexity of $O(N \log^2 N)$ or $O(N \log N)$ operations for one matrix-vector product. The constants involved are large however - much of the current research in fast multipole methods consists of reducing the constants.

We start with an overview of the fast multipole method in §4.2. We describe the method in more detail in §4.3. The relevant translation operators for two-dimensional low-frequency scattering are given in §4.4. The high frequency FMM is discussed next in §4.5. We briefly discuss three-dimensional problems in §4.7 and end with some concluding remarks in §4.8.

4.2 Overview of the method

The aim of the fast multipole method is to provide a fast matrix-vector product Ax, without an explicit representation for the full matrix A. In the context of integral equations, the coefficient matrix A is the dense discretisation matrix of the boundary element method. The idea of the fast multipole method is more general however, and we will describe the main properties here for the general setting $y = Ax$. Assume that the coefficients $A_{m,n}$ of the dense matrix $A \in \mathbb{C}^{M\times N}$ are given by the evaluation of a smooth function $G(p, q)$,

$$A_{m,n} = G(p_m, q_n), \qquad 1 \le m \le M, 1 \le n \le N,$$

with points $p_m, q_n \in \mathbb{R}^d$. The function $G(p, q)$ usually corresponds to an interaction between two points p and q, for example the electrical potential energy $G(p, q) = \log(|p - q|)$, for $p, q \in \mathbb{R}^2$. The matrix-vector product $y = Ax$ corresponds to the summations

$$y_m = \sum_{n=1}^{N} A_{m,n} x_n, \qquad m = 1, \ldots, M. \tag{4.1}$$

Computing all these summations explicitly requires $O(NM)$ operations. Considerable savings can be made if the function G is separable,

$$G(x, y) = \sum_{l=1}^{L} u_l(x) v_l(y). \tag{4.2}$$

In that case, summation (4.1) can be rewritten as

$$y_m = \sum_{l=1}^{L} u_l(p_m) \sum_{n=1}^{N} v_l(q_n) x_n = \sum_{l=1}^{L} u_l(p_m) V_l. \tag{4.3}$$

The numbers V_l, $l = 1, \ldots, L$, need to be computed only once, requiring $O(LN)$ operations. The number of operations required to evaluate y_m, $m = 1, \ldots, M$, using (4.3) is $O(LM)$. Hence, the total number of operations is $O(L(N + M))$. If $L << N, M$, the computational complexity of the matrix-vector product is drastically reduced.

This is the enabling observation of the fast multipole method: the complexity of a matrix-vector product can be significantly reduced if the underlying discretised function is separable. Naturally, only few functions are separable in the sense of (4.2). Smooth functions may be approximated well by separable functions however. Consider for example the kernel function $G(x, y)$ of an integral equation, that is singular along the diagonal $x = y$. Away from the diagonal, the kernel is a smooth function, and it may be approximated locally by a separable expansion of the form (4.2). The speedup of (4.3) can be performed in the part of the matrix A where the expansion is valid. The complexity can be further reduced by considering a multilevel algorithm. In such algorithms, both the computation of the coefficients V_l and the evaluation of the summations is performed hierarchically. The key ingredients of FMM are the construction of suitable expansions, and so-called *translation operators* for translating and transforming these expansions.

4.3 Multilevel fast multipole method

The kernel function of an integral equation typically has the form $G(|x - y|)$. In a fast multipole method, this function is approximated locally by separable expansions. We describe the types of expansions in §4.3.1, and show how they can be used to expedite the matrix-vector product. This is done for a single level algorithm in §4.3.2, and for a multilevel algorithm in §4.3.3. A more detailed description can be found in [47, 99].

4.3.1 Multipole expansions and local expansions

In order to define local separable expansions, the domain of interest is subdivided into a number L of boxes C_l with centre c_l and radius r_l, $l = 1, \ldots, L$. There are two types of expansions in FMM, called *multipole expansions* and *local expansions*. The multipole expansion for a point $x \in C_l$ is an expansion with the general form

$$G(x, y) \approx \sum_{p=1}^{P} a_p(x, c_l) S_p(y - c_l), \qquad x \in C_l, y \notin C_l. \tag{4.4}$$

The coefficients a_p are called the *expansion coefficients*. The expansion approximates the interaction of $x \in C_l$ with a point outside C_l; it can be evaluated for $y \notin C_l$. The functions $S_p(y - c_l)$ are chosen to reflect the behaviour of $G(x, y)$ with y away from the box C_l. These functions are typically singular for $y = c_l$. The name *multipole* originates historically from the choice $S_0(y - c_l) = G(c_l, y)$; the first term of the expansion (4.4) is then called a monopole.

A local expansion for a point $x \notin C_l$ has the general form

$$G(x, y) \approx \sum_{p=1}^{P} b_p(x, c_l) R_p(y - c_l), \qquad x \notin C_l, y \in C_l. \tag{4.5}$$

This expansion approximates the interaction of $x \notin C_l$ with a point inside C_l; it can be evaluated for $y \in C_l$. The functions $R_p(y - c_l)$ are non-singular for $y \in C_l$. Multiple expansions of the form (4.4) or (4.5) can be added together by adding the expansion coefficients. One expansion can be transformed into another expansion that is centred around a different centre point c_m by applying a *translation operator.* There are three possible translation operators: a multipole-to-multipole operator $S|S$, a multipole-to-local operator $S|R$, and a local-to-local operator $R|R$. Usually, these operators can be represented by a matrix with dimensions $P \times P$. We will give an explicit example in §4.6.1.

4.3.2 Single level fast multipole method

Consider a matrix-vector product $Ac = d$, with elements of matrix A given by $A_{i,j} = G(x_i, x_j)$, corresponding to a set of points $\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$. The matrix-vector product is a summation $d_i = \sum_j A_{i,j} c_j$. In a FMM, the summation is carried out approximately.

Consider a covering of all points x_i by L boxes C_l with centre c_l and radius r_l, such that each point x_i is attributed to exactly one box. For one point $x_i \in C_l$, we call the interactions with the other points in C_l the *near field*, and the interactions with the points outside C_l the *far field.* We have the decomposition

$$d_i = d_{NF_i} + d_{FF_i} = \sum_{j : x_j \in C_l} A_{i,j} x_j + \sum_{j : x_j \notin C_l} A_{i,j} x_j.$$

Multipole and local expansions can be used to speed-up the evaluation of the far field interaction d_{FF_i}. Specifically, the single level fast multipole method consists of the following steps:

1. Construct a multipole expansion for each box C_l around its centre point c_l, by adding the multipole expansion coefficients of all points $x_i \in C_l$.

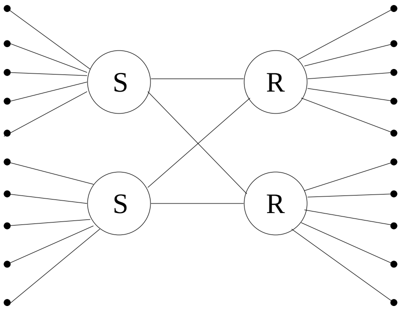


Figure 4.1: Schematic illustration of the interactions in a single level FMM: construct the multipole expansions, use multipole-to-local translation $S|R$, and evaluate the local expansions.

2. Construct a local expansion for each box C_l around its centre c_l, by transforming the multipole expansions of all boxes $C_{l'}$, $l' \neq l$, to C_l using the multipole-to-local translation operator $S|R$.

3. For each point $x_i \in C_l$, evaluate the local expansion around c_l. This is the far field d_{FF_i}. The near field d_{NF_i} is evaluated by summing the interactions with all points $x_j \in C_l$.

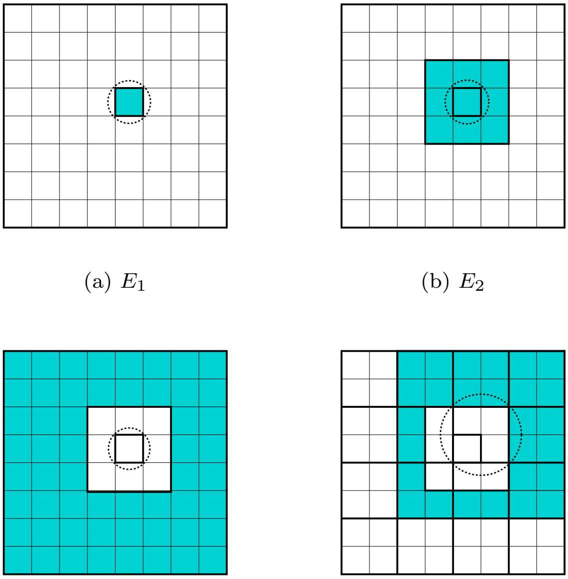
These steps are illustrated schematically in Figure 4.1.

The computational complexity of the scheme depends on a number of parameters, such as the number of terms P in the expansions, and the number of boxes L. For simplicity, assume a uniform distribution of the points over the boxes, such that each box contains $s = N/L$ points on average. The first step requires the addition of s expansions of length P for each of the L boxes, which requires $O(sPL) = O(PN)$ operations. Translation operators generally require $O(P^2)$ operations. The second step consists of the transformation of L expansions of length P for each box, and requires $O(L^2P^2)$ operations. Finally, the last step requires $O(sPL) = O(PN)$ operations for evaluating the far field for each point, and $O(Ns)$ operations for the near field. The total complexity of the single level FMM is

$$\kappa_{SL} = O(PN + P^2L^2 + PN + Ns).$$

The complexity is minimised by choosing $L = O(N^{2/3}P^{-1/3})$ and, correspondingly, $s = O(P^{1/3}N^{1/3})$. This leads to the complexity estimate

$$\kappa_{SL} = O(P^{4/3}N^{4/3}). \tag{4.6}$$



(a) E_1 (b) E_2

(c) E_3 (d) E_4

Figure 4.2: Definition of the box sets corresponding to $C_{l,j}$.

4.3.3 Multilevel fast multipole method

The complexity can be further reduced by constructing separable expansions recursively for larger and larger bounding boxes. In this approach, the boxes $C_{l,j}$ are grouped in a hierarchical manner, and multipole and local expansions are built for each box on each level j. The expansions corresponding to a large box group the interaction of many points together. Hence, we will always try to work with the expansions of the largest box possible, subject to the validity of the expansions (4.4) and (4.5). We say that two boxes $C_{l,j}$ and $C_{l',j'}$ are well separated if

$$r_{l,j} + r_{l',j'} \leq \eta \|c_{l,j} - c_{l',j'}\|, \tag{4.7}$$

with a constant $\eta < 1$. Condition (4.7) is called the *admissibility condition*. We will only approximate the kernel function by a separable expansion if the admissibility condition is satisfied.

We define four sets of boxes, associated with each box $C_{l,j}$. The sets are illustrated in Figure 4.2 for the two-dimensional case. They are:

- $E_1(l, j)$: the box $C_{l,j}$ itself;
- $E_2(l, j)$: the box $C_{l,j}$ and its direct neighbours;
- $E_3(l, j)$: boxes outside the neighbourhood of $C_{l,j}$;
- $E_4(l, j)$: boxes in the neighbourhood of the parent of $C_{l,j}$ on level $j - 1$, but not in the neighbourhood of the box $C_{l,j}$ itself.

The boxes in $E_3(l, j)$ - outside the neighbourhood of $C_{l,j}$ - satisfy the admissibility condition with $\eta = \sqrt{2}/2$. We could therefore approximate the interaction with these boxes using expansions. However, if the boxes are also outside the neighbourhood of the parent of $C_{l,j}$, it is a better choice to use the approximation on a coarser level. This cannot be done for the boxes in $E_4(l, j)$: the set $E_4(l, j)$ is therefore called the *interaction list* of $C_{l,j}$.

The multilevel fast multipole method consists of two passes, the *upward pass* and the *downward pass*. The upward pass has the following steps:

1. At the finest level $j = J$, create multipole expansions around $c_{l,J}$ for each box $C_{l,J}$.
2. At each coarser level j, with $j = J - 1, \ldots, 0$, construct a multipole expansion for each box $C_{l,j}$ by translating the multipole expansions of each child box of $C_{l,j}$, using the $S|S$ translation operator.

Multipole expansions have now been constructed for all boxes on all levels. Local expansions are constructed in the downward pass:

1. At the coarsest level $j = 0$ considered, construct a local expansion for each box $C_{l,j}$ by translating each multipole expansion of the boxes in $E_3(l, j)$ using the $S|R$ translation operator.
2. For each finer level j, with $j = 1, \ldots, J$, construct a local expansion for each box by translating the local expansion of the parent box using the $R|R$ translation operator. Add the $S|R$ translation of the expansions of each box in the interaction list $E_4(l, j)$ of $C_{l,j}$.

The total field for each point is obtained by evaluating the local expansion of the bounding box on the finest level, and by directly evaluating the near field interaction with all points in the boxes of the neighbourhood $E_2(l, j)$.

It can be shown that, using a fixed number of terms P in each expansion, the algorithm described above requires $O(N)$ steps. However, N is usually increased to improve accuracy and, hence, P should increase accordingly. One should choose $P = O(\log N)$, introducing logarithmic terms in the complexity. The total computational complexity of the scheme depends on the cost of the translation operators. The most expensive step of the algorithm is step 2 of the downward pass, i.e., the translation of each multipole expansion in the interaction list of $C_{l,j}$, for each box $C_{l,j}$. The number of boxes in $E_4(l,j)$ depends on the dimension of the problem. From Figure 4.2, there are 27 boxes in the interaction list. For three-dimensional problems, there are 189; the number grows exponentially fast with increasing dimension.

4.3.4 Application to integral equations

The discretisation matrix of an integral equation does not correspond exactly to a matrix with entries $A_{i,j} = G(x_i, x_j)$. Rather, it has elements of the form (2.61),

$$A_{i,j} = \int_\Gamma \int_\Gamma G(x,y)\phi_j(y)\phi_i(x)\,\mathrm{d}s_y\,\mathrm{d}s_x.$$

A multipole expansion for the kernel function of the form (4.4) leads to a multipole expansion with a similar form, but with different expansion coefficients. Consider the integral in x, corresponding to a basis function $\phi_i(x)$ with support in C_l. Using the multipole expansion for the kernel, the integral can be written as

$$\begin{aligned}\int_\Gamma G(x,y)\phi_i(x)\,\mathrm{d}s_x &\approx \int_\Gamma \sum_{p=1}^{P} a_p(x,c_l)S_p(y-c_l)\phi_i(x)\,\mathrm{d}s_x\\ &= \sum_{p=1}^{P} S_p(y-c_l)\int_\Gamma a_p(x,c_l)\phi_i(x)\,\mathrm{d}s_x.\end{aligned}$$

The new expansion coefficients are given by

$$a^*_{i,p} = \int_\Gamma a_p(x,c_l)\phi_i(x)\,\mathrm{d}s_x. \tag{4.8}$$

We denote the vector of expansion coefficients by $a_i \in \mathbb{C}^P$. Assume that the translation operator $S|R$ for $x \in C_l$ and $y \in C_{l'}$ is given by a matrix $T_{l,l'} \in \mathbb{R}^{P\times P}$, and define $b_i = T_{l,l'}a^*_i$. The integral for the element $A_{i,j}$ can then be written as

$$\begin{aligned}A_{i,j} &\approx \int_\Gamma \sum_{p=1}^{P} a^*_{i,p}S_p(y-c_l)\phi_j(y)\,\mathrm{d}s_y \approx \int_\Gamma \sum_{p=1}^{P} b_{i,p}R_p(y-c_{l'})\phi_j(y)\,\mathrm{d}s_y\\ &= \sum_{p=1}^{P} b_{i,p}\underbrace{\int_\Gamma R_p(y-c_{l'})\phi_j(y)\,\mathrm{d}s_y}_{c^*_{j,p}} = \sum_{p=1}^{P} b_{i,p}c^*_{j,p}.\end{aligned}$$

The approximation corresponds to a low rank approximation of a subblock of the dense discretisation matrix A. Specifically, consider the subblock $M_{l,l'}$ corresponding to the combination of N_l basis functions $\phi_i(x)$ with support in C_l, with $N_{l'}$ basis functions $\phi_j(y)$ with support in $C_{l'}$. Define the matrix $U \in \mathbb{R}^{N_l \times P}$ with elements $U_{i,p} = a^*_{i,p}$, and the matrix $V \in \mathbb{R}^{N_{l'} \times P}$ with elements $V_{j,p} = c^*_{j,p}$. We have

$$M_{l,l'} \approx U T^T_{l,l'} V^T. \tag{4.9}$$

The subblock $M_{l,l'}$ can be approximated by a low rank matrix with rank P. In the multilevel fast multipole method, the computation is more involved than in (4.9). The method more closely resembles a product of the form

$$M_{l,l'} \approx U \beta_{l,l_1} \beta_{l_1,l_2} \dots \beta_{l_{L-1},l_L} \alpha_{l_L,l'_L} \beta_{l'_L,l'_{L-1}} \dots \beta_{l'_1,l'} V^T. \tag{4.10}$$

The matrices β represent the $S|S$ and $R|R$ operators, the matrix α corresponds to the $S|R$ operator. The connection of fast multipole methods to matrix representations is explored in [182].

4.4 Expansions and translation operators

The fast multipole method was originally presented for two-dimensional problems involving a kernel function $G(x, y) = \log(|x - y|)$, which is the fundamental solution of the Laplace problem. Assume that the box C_l has centre $c_l = (c_{l1}, c_{l2}) \in \mathbb{R}^2$, and set $z_0 = c_{l1} + ic_{l2} \in \mathbb{C}$ and $z = y_1 + iy_2 \in \mathbb{C}$ for $y = (y_1, y_2) \in \mathbb{R}^2$. The kernel function can then be written as $G(x, y) = \Re(\log(z - z_0)) =: \phi(z - z_0)$. The multipole expansion for the kernel that was proposed in [98] is given by

$$\phi(z - z_0) \approx Q \log(z - z_0) + \sum_{p=1}^{P} \frac{a_p}{(z - z_0)^p}, \qquad (z - z_0) > r_l, \tag{4.11}$$

and the local expansion by

$$\phi(z) \approx \sum_{p=0}^{P} b_p (z - z_0)^p, \qquad (z - z_0) < r_{l'}. \tag{4.12}$$

Translation operators are obtained by straightforward calculations. Note that the basis functions in the multipole expansion are singular for $z = z_0$.

Similar expansions can be defined for the two-dimensional Helmholtz problem. They are based on the singular and regular functions $H^{(1)}_n(z)$ and $J_n(z)$, that are shown to be solutions to the Helmholtz equation in

Appendix A. More often however, a separable expansion for the Helmholtz kernel (2.25) is based on the following series. Assume $x \in C_l$ and $y \in C_{l'}$ and define

$$x - c_l = r_{x,l} e^{i\theta_{x,l}}, \quad y - c_{l'} = r_{y,l'} e^{i\theta_{y,l'}}, \quad \text{and} \quad c_l - c_{l'} = r_{l,l'} e^{i\theta_{l,l'}}.$$

Then we have (see, e.g., [40])

$$H_0^{(1)}(k|x-y|) = \tag{4.13}$$
$$\sum_{m=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} J_{l+m}(kr_{x,l}) e^{i(l+m)\theta_{x,l}}\, H_m(kr_{l,l'}) e^{im\theta_{l,l'}} J_l(kr_{y,l'}) e^{il\theta_{y,l'}}.$$

Truncation of this series leads to a separable representation of the kernel function. Similar to (4.9), it leads to a low rank approximation with the general form $M_{l,l'} \approx U \Sigma V^T$. The elements of U and V are defined in terms of Bessel functions $J_n(z)$ and complex exponentials, the elements of Σ in terms of Hankel functions $H_n(z)$ and complex exponentials. An interesting observation is that Σ has Toeplitz structure. As such, it can be diagonalised using the Fourier transform. This leads to a representation $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ with a diagonal matrix $\tilde{\Sigma}$. The elements of the transformed matrices $\tilde{U}$ and $\tilde{V}$ are defined in terms of plane wave basis functions. The translation of these functions can also be performed by multiplication with a diagonal matrix.

Unfortunately, there are stability issues involved in the truncation of the series (4.13). Specifically, for small arguments or for large orders, the function $H_n(z)$ can be exponentially large; it behaves as [4]

$$H_n^{(1)}(z) \sim -\sqrt{\frac{2}{\pi n}} \left(\frac{2n}{ez}\right)^n, \qquad n \to \infty. \tag{4.14}$$

The accuracy of the computations can be lost due to rounding errors. The computations involving Σ can be stabilised using a renormalisation of the Hankel functions [205]. In that case however, the Toeplitz structure of Σ is lost, and the diagonal form $\tilde{\Sigma}$ can no longer be constructed.

4.5 High frequency fast multipole method

In the high frequency regime, it is not sufficient to consider a fixed number P of terms in the expansions (4.4) and (4.5). Specifically, the number of terms should scale as $P = O(kD)$, where D is the diameter of the box in which the expansion is valid. This linear dependence on the wavenumber is a well-known result [30, 186]. The physical meaning of this condition in the context of electromagnetic scattering problems is explained in [38].

An efficient fast multipole method can still be constructed however with $O(N \log N)$ complexity, where N increases proportionally to k in two dimensions, by varying the number of terms in the expansion depending on the level of the bounding box [152, 8, 13]. In this approach, the number of terms increases linearly with the size of the box. This means that for an implementation with low computational complexity, it is required that the translation operators are diagonal. Indeed, at a coarse scale j, the number of terms in the expansions is proportional to N. Translations that require more than $O(P)$ or $O(P \log P)$ operations when $P = O(N)$ should therefore be avoided. Suitable separable representations have been proposed based on the truncation of the series (4.13).

The different lengths of the expansions at different levels require changes in the translation operators. Specifically, the translation of an expansion to a higher level with more expansion coefficients requires an interpolation procedure. The reverse translation requires a prolongation. The approximation of a subblock $M_{l,l'}$, as given by (4.10), is in this setting more accurately given by

$$M_{l,l'} \approx U I_{l,l_1} \beta_{l,l_1} I_1 \beta_{l_1,l_2} I_1 \ldots \beta_{l_{L-1},l_L} \alpha_{l_L,l'_L} \beta_{l'_L,l'_{L-1}} \ldots I_1^T \beta_{l'_1,l'} V^T. \quad (4.15)$$

The matrices I_1 and I_1^T perform one level interpolation and prolongation respectively. The matrices α and β are diagonal matrices.

There are still issues associated with this approach, related to the instability of the series (4.13) for small arguments z or large orders n of $H_n^{(1)}(z)$. This means that the high frequency fast multipole method does not converge for small values of the wavenumber. Moreover, arbitrary accuracy cannot be obtained, because increasing the number of terms may introduce roundoff errors again. Several methods have been proposed to remedy this instability [117, 13]. We will revisit the method of [13] in the next chapter on hierarchical matrices.

4.6 Numerical results

This chapter contains no elements that are not already described in literature. Still, we will illustrate the theory with some experiments, including a full implementation of the fast multipole method with linear complexity for a one-dimensional large summation problem. We investigate the numerical instability of the high frequency fast multipole method for relatively low wavenumbers, as it is an important characteristic of FMM that prevents computations with high accuracy. More elaborate numerical results will be given in the next chapter on hierarchical methods, as both methods yield qualitatively similar results.

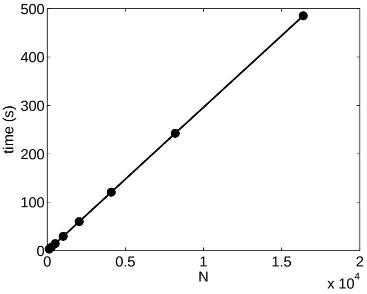


Figure 4.3: Time (in seconds) for one matrix-vector product involving interactions of the form $G(x,y) = 1/(y-x)$, with fixed-length expansions.

4.6.1 Low frequency fast multipole method

We have implemented the fast multipole method for a one-dimensional model problem with kernel function $G(x,y) = \frac{1}{y-x}$. A multipole expansion of the form (4.4) is easily obtained,

$$\begin{aligned}\frac{1}{y-x} &= \frac{1}{y-c_l-(x-c_l)} = \frac{1}{(y-c_l)\left(1-\frac{x-c_l}{y-c_l}\right)} \\ &= \frac{1}{y-c_l}\sum_{p=0}^{\infty}\frac{(x-c_l)^m}{(y-c_l)^m}, \qquad |x-c_l| < |y-c_l|.\end{aligned}$$

Similarly, a local expansion can be found for the case $|y-c_l| < |x-c_l|$,

$$\begin{aligned}\frac{1}{y-x} &= \frac{1}{y-c_l-(x-c_l)} = \frac{-1}{(x-c_l)\left(1-\frac{y-c_l}{x-c_l}\right)} \\ &= \frac{-1}{x-c_l}\sum_{p=0}^{\infty}\frac{(y-c_l)^m}{(x-c_l)^m}, \qquad |y-c_l| < |x-c_l|.\end{aligned}$$

We have the singular functions $S_p(y-c_l) = (y-c_l)^{-p-1}$ and the regular functions $R_p(y-c_l) = (y-c_l)^p$. The $R|R$ translation operator from c_l to $c_{l'}$ is derived as follows,

$$R_p(y-c_l) = (y-c_{l'}+(c_{l'}-c_l))^p = \sum_{i=0}^{p}\binom{p}{i}(c_{l'}-c_l)^{p-i}R_i(y-c_{l'}).$$

For the multipole-to-local operator $S|R$, we have $|y - c_l| < |c_l - c_{l'}|$, and

$$\begin{aligned}
S_p(y - c_l) &= (y - c_{l'} + (c_{l'} - c_l))^{-p-1} \\
&= (c_{l'} - c_l)^{-p-1}(1 + \frac{y - c_{l'}}{c_{l'} - c_l})^{-p-1} \\
&= \sum_{m=0}^{\infty} \frac{(-1)^m (m+p)!}{m!p!}(c_{l'} - c_l)^{-p-m-1}(y - c_{l'})^p \\
&\approx \sum_{m=0}^{P} \frac{(-1)^m (m+p)!}{m!p!}(c_{l'} - c_l)^{-p-m-1} R_p(y - c_{l'}).
\end{aligned}$$

Figure 4.3 shows the time required for evaluating the matrix vector product $y = Mc$, with $M_{i,j} = G(x_i, x_j)$, $i \neq j$, and $M_{ii} = 0$. A uniform distribution of points x_i was used on the interval $[0, 1]$. We used a fixed length P of the expansions, independently of N. The algorithm clearly requires a number of operations that is linear in N.

The total complexity is actually $O(P^2 N)$, with P the length of the expansions. The term P^2 arises from the matrix-vector product due to the matrix representation of the operators $S|S$, $S|R$ and $R|R$. If the error of the method should decrease with N, then the length P of the expansions should grow at least logarithmically with N, and the complexity becomes $O(N \log^2 N)$.

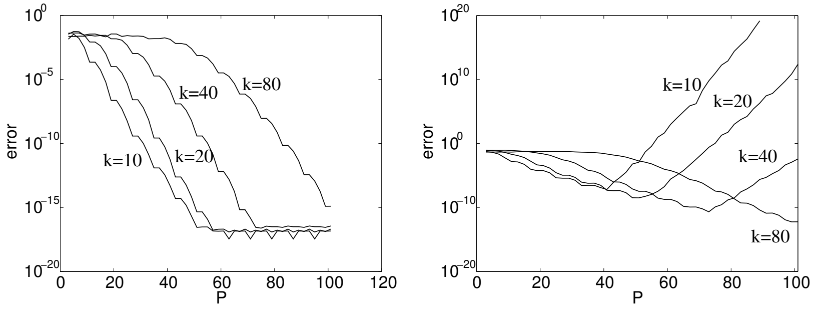
4.6.2 High frequency fast multipole method

For numerical results involving the high frequency fast multipole method for two-dimensional integral equations, we refer the reader to [8, 13, 152, 205]. Scattering problems are solved in these references for circle, ellipse and square-shaped scattering objects. The implementations have $O(N \log^2 N)$ complexity, which is supported by the numerical results. The constant involved is reported to be large.

Here, we restrict the numerical results to an illustration of the convergence properties of the separable expansions that are obtained by truncating (4.13). They correspond to an approximation of the form

$$\frac{i}{4} H_0^{(1)}(k|x - y|) \approx u^T \Sigma v, \tag{4.16}$$

with column vectors $u, v \in \mathbb{C}^P$, and a Toeplitz or diagonal matrix $\Sigma \in \mathbb{C}^{P \times P}$. The Toeplitz structure is obtained by the direct truncation of (4.13).



(a) Toeplitz matrix Σ

(b) Diagonal matrix $\tilde{\Sigma}$

Figure 4.4: Convergence behaviour of the separable kernel approximations for the high frequency FMM.

In that case, for odd $P = 2M + 1$, the elements are given by

$$u_j = J_{j-M-1}(kr_{x,l})e^{i(j-M-1)\theta_{x,l}}, \tag{4.17}$$

$$v_j = J_{j-M-1}(kr_{y,l'})e^{-i(j-M-1)\theta_{y,l'}}, \tag{4.18}$$

$$\Sigma_{j,n} = \frac{i}{4}H^{(1)}_{n-j}(kr_{l,l'})e^{i(n-j)\theta_{l,l'}}, \tag{4.19}$$

for $1 \leq j, n \leq P$. The diagonal matrix $\tilde{\Sigma}$ is obtained as the Fourier transform of Σ. This results in the approximation $\tilde{u}^T\tilde{\Sigma}\tilde{v}$ with elements given by

$$\tilde{u}_j = \frac{1}{\sqrt{P}}e^{-ikr_{x,l}cos(2\pi(j-1)/P-\theta_{x,l})},$$

$$\tilde{v}_j = \frac{1}{\sqrt{P}}e^{ikr_{y,l'}cos(2\pi(j-1)/P-\theta_{y,l'})},$$

$$\tilde{\Sigma}_{j,j} = \frac{i}{4}\sum_{m=-M}^{M}\frac{1}{i^m}H^{(1)}_{m}(kr_{l,l'})e^{im(\theta_{l,l'}-2\pi(j-1)/P)},$$

for $1 \leq j \leq P$. The accuracy of the expansions is illustrated in Figure 4.4 for an example with two well separated clusters C_l and $C_{l'}$. The figure in the left panel shows that the series starts converging only after a certain minimal number of terms. The required minimal number of terms increases proportionally to k. The figure in the right panel illustrates the numerical instability of the diagonal translation operator when more terms in the expansion are used. The instability is worse for smaller values of the wavenumber.

4.7 Three-dimensional problems

Fast multipole methods for the three-dimensional Helmholtz equation have similar properties as in the two-dimensional case. Separable expansions for the three-dimensional Helmholtz kernel e^{ikr}/r are usually defined in terms of spherical harmonics [99]. Diagonal translation operators were presented by Rokhlin in [173], based on a representation in terms of plane waves. Similar to the 2D-case, there are stability problems at low frequencies. A number of approaches have been developed for the efficient solution of low frequency scattering problems [97, 65, 135]. We refer the reader to [65] for an overview of the issues and remedies, and to [64, 99] for a full description of the method and its implementation.

A useful simplification of the three-dimensional FMM exists in the case of scattering obstacles that are nearly planar, called the *steepest descent FMM*[134, 156]. Such scattering problems arise in the development of planar electronic circuits and antennas. The resulting method requires $O(N)$ operations for a fixed accuracy, and $O(N \log N)$ operations for an accuracy that increases with N. In the steepest descent FMM, the three-dimensional problem is reduced to a small number of two-dimensional problems, that are solved using the two-dimensional high frequency FMM.

4.8 Conclusions

The fast multipole method enables a fast matrix-vector product for discretisation matrices that arise in integral equations. Both the memory requirements and the computational complexity of a matrix-vector product scale as $O(N \log^p N)$, where $p \geq 0$ is a small number. The result is obtained by constructing low rank approximations to the kernel function, and exploiting the resulting structure of the operations. Moreover, the fast multipole method is the only viable fast solution method currently known in the high frequency regime. The computational complexity is $O(N \log^p N)$, with $p \geq 1$. In this regime, low rank approximations are constructed only for small domains. On coarser scales, the rank is increased proportionally to the size of the domain. The reduced complexity is possible by using very efficient translation operators.

The construction and the analysis of fast multipole methods keep a close connection to the underlying physical problem. As a consequence, separable expansions have to be devised for each new kernel function. In the high frequency regime, this is likely unavoidable. In the low-frequency regime however, several other approaches are possible that allow a more general implementation. In the next chapter, we treat a number of approaches based on the algebraic concept of $\mathcal{H}$-matrices.

Chapter 5

Hierarchical matrix methods

5.1 Introduction

The third multiscale method we consider in this thesis is based on *hierarchical matrices*, also known as $\mathcal{H}$-matrices. Hierarchical matrices are block-structured matrices that consist of submatrices of low rank. They are well suited to approximate the discretisation matrix of integral equations and, more generally, to approximate the inverse of the sparse discretisation matrix of elliptic partial differential equations. A fast matrix-vector product is obtained by using low rank approximations to parts of the discretisation matrix. This approach bears a close resemblance to fast multipole methods. In this chapter, we explore hierarchical matrices and their application to integral equations, and we highlight some similarities and differences with fast multipole methods.

Hierarchical matrices originate in the mosaic-skeleton approach and in panel clustering methods [187, 103]. They were proposed by Hackbusch in [101, 102], with applications in multidimensional finite element methods and boundary element methods. The use of hierarchical matrices leads to a storage requirement of $O(N \log N)$ complexity, and a similar complexity for the matrix-vector product. The arithmetic of hierarchical matrices is described in [96]. Computations with even lower complexity can be performed by $\mathcal{H}^2$-matrices [23]: $\mathcal{H}^2$-matrices require $O(N)$ memory, and $O(N)$ operations for a matrix-vector product. An algebraic technique to further reduce memory requirements is proposed in [95]. Efficient algorithms for the black-box construction of hierarchical matrices for general integral equations have been described in [16, 26].

We present an overview of hierarchical matrix based methods in §5.2. We define $\mathcal{H}$-matrices and $\mathcal{H}$-matrix arithmetic in §5.3. The application to integral equations is discussed with more detail in §5.4. An adapted method for the high frequency regime based on $\mathcal{H}^2$-matrices is closely related to the high frequency fast multipole method; it is described in §5.5. We present some numerical results for two-dimensional scattering problems in §5.5. The numerical experiments in this chapter are used to discuss and illustrate the differences between the three multiscale methods that are considered in this thesis: wavelet based methods, fast multipole methods and hierarchical matrix methods. Finally, we discuss the issues for three-dimensional problems.

5.2 Overview of the method

Computations involving matrices can often be performed by fast methods with low computational complexity when the matrices have structure. Common examples are diagonal matrices, semiseparable matrices, matrices with Toeplitz or Hankel structure, and matrices that have low rank. In the latter case, a full matrix $M \in \mathbb{C}^{N\times N}$ of rank k can be written as $M = AB^T$, with $A, B \in \mathbb{C}^{N\times k}$. The matrix-vector product $z = My$, for example, can be computed efficiently as $z = A(B^T y)$. This requires $O(kN)$ operations, as opposed to $O(N^2)$ operations for the classical matrix-vector product.

A *hierarchical matrix* or $\mathcal{H}$-matrix is a block-structured matrix that consists of submatrices of rank k. The matrix-vector product can be expedited by performing small matrix-vector products of the form $A(B^T y)$ in each subblock of the matrix that has rank k. A natural representation of $\mathcal{H}$-matrices is given by storing these matrices A and B, that each have only k columns, rather than storing all elements of the dense subblock. The fast matrix-vector product is only one property of $\mathcal{H}$-matrices. Fast approximative methods for other matrix operations can be devised, such as addition, matrix multiplication and inversion. An example of the structure of a typical hierarchical matrix is shown in Figure 5.1.

In the context of integral equations, hierarchical matrices can be obtained from any locally separable approximation of the kernel function. A matrix-vector product can be performed with fixed rank k approximations in $O(N \log N)$ operations for $\mathcal{H}$-matrices, and in $O(N)$ operations for so-called $\mathcal{H}^2$-matrices. This is related to the fact that, away from the diagonal, the kernel function is a smooth function. In that sense, the method is similar to the fast multipole method; we will see that the fast multipole method actually corresponds to the use of $\mathcal{H}^2$-matrices. Hierarchical matrices differ from the fast multipole method in their algebraic nature: an $\mathcal{H}$-matrix is mainly an algebraic concept. This approach has inspired the development of black box algorithms for integral equations, rather than algorithms that require

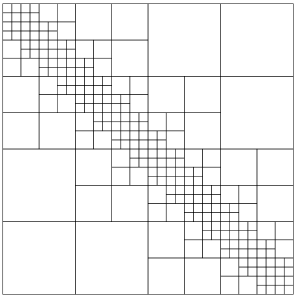


Figure 5.1: Illustration of a typical $\mathcal{H}$-matrix in integral equations. Each block is of low rank.

a separate and analytically derived expansion for each new kernel function. The $\mathcal{H}$-matrix representation suggests adaptive and purely algebraic operations, such as *adaptive recompression* to reduce memory requirements. An efficient black-box implementation for general integral equations is possible by *adaptive cross approximation* [16, 26].

The $\mathcal{H}$-matrix format can also be used in the solution of certain matrix equations, such as the Sylvester and Riccati equations, or in the solution of boundary value problems corresponding to elliptic partial differential equations. It can be shown that the inverse of a sparse finite element discretisation matrix can be well approximated by a hierarchical matrix. This knowledge can be used, even if the fundamental solution of the differential equation is not known.

5.3 $\mathcal{H}$-matrices

5.3.1 Definition of hierarchical matrices

The location of a subblock of a block-structured matrix is fully determined by the range of rows and columns that are covered. We denote a range of indices by a *cluster* τ. The number of elements in that range is denoted as $n_\tau = \#\tau$. First, we shall define the central concept of a *cluster tree*. A cluster tree T_I is a tree whose nodes are clusters, corresponding to a subset of the index set $I = \{1, \ldots, N\}$, with the following properties:

- the root of the tree contains all indices: $\tau_r = I$;
- if a cluster τ has sons, then they form a partition of the father: $\tau = \cup\{\tau' : \tau' \in \text{sons}(\tau)\}$, and $\tau_1, \tau_2 \in \text{sons}(\tau) : \tau_1 \neq \tau_2 \Rightarrow \tau_1 \cap \tau_2 = \phi$;

- each cluster τ that is not a leaf has two sons;
- a constant C_{leaf} exists independent of N such that, for each leaf τ, $0 < n_\tau \leq C_{leaf}$.

The number of nodes in the tree is bounded by $2N - 1$. The number of clusters on each level j is bounded by 2^j, where level $j = 0$ corresponds to the root of the tree.

A cluster tree represents a hierarchical subdivision of a set of indices. In order to identify a subblock of a matrix, we require two sets of indices. This leads to the definition of a *block cluster tree*. The root of a block cluster tree $T_{I \times J}$ is the couple (τ_r, σ_r), where τ_r and σ_r are the roots of cluster trees T_I and T_J respectively. The nodes have the form $\tau \times \sigma \in T_I \times T_J$. The sons of a node $\tau \times \sigma$ are given by

$$\begin{array}{ll} \{\tau \times \sigma' : \sigma' \in \mathrm{sons}(\sigma)\} & \text{if} \quad \mathrm{sons}(\tau) = \emptyset, \mathrm{sons}(\sigma) \neq \emptyset, \\ \{\tau' \times \sigma : \tau' \in \mathrm{sons}(\tau)\} & \text{if} \quad \mathrm{sons}(\tau) \neq \emptyset, \mathrm{sons}(\sigma) = \emptyset, \\ \{\tau' \times \sigma' : \tau' \in \mathrm{sons}(\tau), \sigma' \in \mathrm{sons}(\sigma)\} & \text{otherwise.} \end{array}$$

The definition of a block cluster tree $T_{I \times J}$ is such that the leaves form a partition of the index set $I \times J$.

We denote the subblock of a matrix M corresponding to the block cluster $\tau \times \sigma$ by $M|_{\tau \times \sigma}$. For increased generality, we assume a rank k approximation only in a subset of all leaves, the set of *admissible leaves* $\mathcal{L}^+$. The remaining set $\mathcal{L}^-$ of inadmissible leaves is represented by a regular dense matrix. A hierarchical matrix is defined as follows.

Definition 5.3.1. *Let $T_{I \times I}$ be a block cluster tree for the index set I. We define the set of $\mathcal{H}$-matrices of rank k as*

$$\mathcal{H}(T_{I \times I}, k) := \{M \in \mathbb{C}^{N \times N} | \, \mathrm{rank}(M|_{\tau \times \sigma}) \leq k, \quad \forall \tau \times \sigma \in \mathcal{L}^+\}.$$

A more general definition allows a variable rank in each admissible leaf.

5.3.2 $\mathcal{H}$-matrix arithmetic

The multiplication of a matrix $M \in \mathbb{C}^{N \times N}$ of rank k by a matrix U does not increase the rank. If $M = AB^T$, then $UM = (UA)B^T$ and $MU = A(U^T B)^T$. The sum of two matrices with rank k has, in general, rank $2k$. This means that the set of matrices of rank k is not closed under addition. The sum of two such matrices can efficiently be approximated however by a matrix that has rank k by truncating the *singular value decomposition* of the sum [96]. It can be shown that this approximation is optimal, in the sense that the approximation error is minimal in the Frobenius norm and in the spectral norm. The approximation can be computed in $O(k^2n + k^3)$ operations, if the matrices involved are of the form AB^T with A and B having dimensions $n \times k$.

It is a consequence that the set of $\mathcal{H}$-matrices is not closed under addition. The sum of two $\mathcal{H}$-matrices can be defined by performing the blockwise sum of the low-rank matrices, and then truncating each low rank matrix to rank k. The set of $\mathcal{H}$-matrices is not closed under multiplication either, since the blockwise multiplication of two block-structured matrices requires the addition of blockwise products. Similarly to addition, multiplication can be performed by truncating the submatrices of the result to rank k. In general, the block structure of the resulting matrix may be different from the structure of the matrices that are multiplied. Then, a new block cluster tree is required [96].

5.4 The $\mathcal{H}$-matrix method

The discretisation matrix of an integral operator equation is well suited for approximation by hierarchical matrices. We summarise the main elements of the approach in this section. A detailed description is given in [101, 102, 96] for $\mathcal{H}$-matrices, and in [23, 24, 13] for $\mathcal{H}^2$-matrices.

5.4.1 Construction of the $\mathcal{H}$-matrix

Consider the set of basis functions ϕ_i, with $i \in I = \{1, \dots, N\}$, used for solving a boundary integral equation. With any cluster τ of a cluster tree T_I, one can associate a bounding box C_τ such that

$$\operatorname{supp} \phi_i(x) \subset C_\tau, \qquad \forall i \in \tau.$$

The centre of the bounding box C_τ is denoted by c_τ, and we define the radius of C_τ as the radius of the smallest ball with centre c_τ that contains the box. In the $\mathcal{H}$-matrix method, the cluster tree T_I is constructed such that the size of the bounding boxes C_τ is as small as possible. This construction is based on purely geometric considerations; basis functions are clustered if they lie close together on $\Gamma \subset \mathbb{R}^d$. Usually, only basis functions with small support are considered.

The *admissibility condition* is defined similarly to the definition (4.7) in fast multipole methods. Two clusters are called admissible if

$$r_\tau + r_\sigma \leq \eta \| c_\tau - c_\sigma \|, \tag{5.1}$$

for a sufficiently small constant $\eta < 1$. The corresponding element $\tau \times \sigma$ of the tensor product $T_I \times T_I$ is also called admissible. The aim of the construction of the block cluster tree $T_{I\times I}$ is to maximise the size of the admissible leaf nodes. Each leaf node $\tau\times\sigma \in T_{I\times I}$ with large clusters τ and σ corresponds to a large subblock $M_{\tau\times\sigma}$ of the discretisation matrix. The larger the block, the larger the savings that result from an approximation of that block with fixed rank k.

Given a cluster tree T_I, the block cluster tree $T_{I\times I}$ is constructed using the following algorithm. Starting with $\tau\times\sigma$ equal to the root $\tau_r \times \sigma_r$, one proceeds:

- if $\tau\times\sigma$ is admissible, add it to the admissible leaves $\mathcal{L}^+$;
- if both τ and σ are leaves of T_I, add $\tau\times\sigma$ to the set of inadmissible leaves $\mathcal{L}^-$;
- otherwise, consider the combinations of the sons of τ and the sons of σ in the next iteration (proceed with τ if τ has no sons, or with σ if σ has no sons). These become sons of the node $\tau\times\sigma$ in $T_{I\times I}$.

The leaves of the block cluster tree that are not admissible usually correspond to small dense matrices along the diagonal of the matrix. Their appearance is a consequence of the admissibility condition (5.1), which is required because of the singularity of the kernel function.

A low rank approximation is then constructed for the submatrix $M_{\tau\times\sigma}$ corresponding to the admissible leaves. The representation AB^T is obtained from a separable expansion of the kernel function $G(x,y)$,

$$G(x,y) \approx \sum_{l=1}^{k} f_l(x) g_l(y). \tag{5.2}$$

The double integral corresponding to an element of the discretisation matrix can be written as a product of univariate integrals,

$$\int_\Gamma \int_\Gamma G(x,y)\phi_i(x)\phi_j(y)\,\mathrm{d}s_y\,\mathrm{d}s_x \approx \sum_{l=1}^{k} \underbrace{\left(\int_\Gamma f_l(x)\phi_i(x)\,\mathrm{d}s_x\right)}_{A_{i,l}} \underbrace{\left(\int_\Gamma g_l(y)\phi_j(y)\,\mathrm{d}s_y\right)}_{B_{j,l}} = \sum_{l=1}^{k} A_{i,l}B_{j,l}.$$

This leads to $M_{\tau\times\sigma} = AB^T$, with $A \in \mathbb{C}^{n_\tau\times k}$ and $B \in \mathbb{C}^{n_\sigma\times k}$. It can be shown that the resulting block matrix has $O(N\log N)$ memory complexity. A matrix-vector product with the resulting matrix has $O(N\log N)$ computational complexity.

5.4.2 $\mathcal{H}^2$-matrices

A different kind of separable approximation to the kernel function leads to the so-called $\mathcal{H}^2$-matrices. Consider an approximation of the form

$$G(x,y) \approx \sum_{m=1}^{k}\sum_{n=1}^{k} s_{m,n} f_m(x,c_\tau) g_n(y,c_\sigma). \tag{5.3}$$

This corresponds to an approximation of the form $M_{\tau\times\sigma} = U_\tau S_{\tau\times\sigma} V_\sigma^T$, where $S_{\tau\times\sigma,m,n} = s_{m,n}$, and

$$\begin{aligned} U_{\tau,i,m} &= \int_\Gamma f_m(x,c_\tau)\phi_i(x)\,\mathrm{d}s_x, \\ V_{\sigma,i,n} &= \int_\Gamma g_n(y,c_\sigma)\phi_i(y)\,\mathrm{d}s_y. \end{aligned}$$

The matrix $U_\tau \in \mathbb{C}^{n_\tau\times k}$ is called the *row cluster basis* for cluster τ, and $V_\sigma \in \mathbb{C}^{n_\sigma\times k}$ the *column cluster basis*. The data representing the kernel function is now stored in the small dense matrix $S \in \mathbb{C}^{k\times k}$. Compared to the regular $\mathcal{H}$-matrix method, memory is saved because the matrices U_τ and V_σ have to be stored only once.

In addition, the complexity of the matrix-vector product can be reduced if the cluster bases are nested. A cluster basis is called *nested*, if for each non-leaf cluster τ and each son cluster $\tau' \in \mathrm{sons}(\tau)$, there exist *transfer matrices* $T^U_{\tau',\tau}, T^V_{\tau',\tau} \in \mathbb{C}^{k\times k}$ such that

$$U_\tau = \begin{pmatrix} U_{\tau'}T^U_{\tau',\tau} \\ U_{\tau''}T^U_{\tau'',\tau} \end{pmatrix} \qquad \text{and} \qquad V_\sigma = \begin{pmatrix} V_{\sigma'}T^V_{\sigma',\sigma} \\ V_{\sigma''}T^V_{\sigma'',\sigma} \end{pmatrix},$$

where τ' and τ'' are the sons of τ, and similarly for σ. Nested cluster bases allow a fast matrix-vector product $z = My$ by the following algorithm [13]. We denote by y_σ the part of the vector y corresponding to the indices in the cluster σ.

1. Forward transformation (upward pass)
    - for all leaves $\sigma \in T_I$, compute $\hat{y}_\sigma = V_\sigma^T y_\sigma$
    - for each parent σ, compute $\hat{y}_\sigma = (T^V_{\sigma',\sigma})^T \hat{y}_{\sigma'} + (T^V_{\sigma'',\sigma})^T \hat{y}_{\sigma''}$
2. Multiplication (far field interaction)
    - for all $\tau \in T_I$, compute $\hat{z}_\tau = \sum_{\tau\times\sigma\in\mathcal{L}^+} S_{\tau\times\sigma}\hat{y}_\sigma$
3. Backward transformation (downward pass)
    - initialise the output vector z to zero

- for all parents τ, perform $\hat{z}_{\tau'} := \hat{z}_{\tau'} + T^U_{\tau',\tau}\hat{z}_\tau$ and $\hat{z}_{\tau''} := \hat{z}_{\tau''} + T^U_{\tau'',\tau}\hat{z}_\tau$
- for every leaf $\tau \in T_I$, perform $z_\tau := z_\tau + U_\tau \hat{z}_\tau$

4. Non-admissible blocks (near field interaction)
    - for each $\tau \times \sigma \in \mathcal{L}^-$, perform $z_\tau := z_\tau + M_{\tau\times\sigma} y_\sigma$

The matrix-vector product has $O(N)$ computational complexity for fixed rank k approximations. The method largely corresponds to the multilevel fast multipole method. The corresponding FMM phases are written between parentheses. The transfer matrices $T^U_{\tau',\tau}$ and $T^V_{\sigma',\sigma}$ correspond to the $R|R$ and $S|S$ translation operators respectively, the multiplication by $S_{\tau\times\sigma}$ corresponds to the $S|R$ translation operator. A difference is that the $\mathcal{H}^2$-matrix method allows the combination of two clusters with different sizes, whereas the $S|R$ operator in the fast multipole method is only used for boxes on the same level. The algebraic approach of $\mathcal{H}^2$-matrices have also inspired a number of black-box implementations for general kernel functions.

5.4.3 Black-box separable approximations

In the fast multipole method, separable approximations for the kernel function are usually constructed starting from a series expansion that is specific to one particular kernel function. However, approximations of the form (5.2) or (5.3) can be constructed in various other ways. The most straightforward one is perhaps a truncated Taylor series of the kernel, which, for 2D, reads

$$G(x,y) \approx \sum_{i,j} \frac{\partial^{i+j} G}{\partial y_1^i \partial y_2^j}(x, y^0) \frac{(y_1 - y_1^0)^i (y_2 - y_2^0)^j}{i! j!}.$$

This approach may require high order derivatives of the kernel, which is not always practical. An alternative is to use polynomial interpolation, which requires only evaluations of the kernel function. For one-dimensional applications, the approximation has the form

$$G(x,y) \approx \sum_{\nu=1}^{k} G(x, y^\sigma_\nu) \mathcal{L}^\sigma_\nu(y), \tag{5.4}$$

where the *Lagrange polynomials* $\mathcal{L}^\sigma_\nu(y)$ satisfy $\mathcal{L}^\sigma_\nu(y_{\nu'}) = \delta_{\nu,\nu'}$. The values y^σ_ν, $\nu = 1, \ldots, k$, are the interpolation points in the cluster σ. Higher dimensional functions can be approximated using tensor-product Lagrange polynomials. Approximations of the form (5.3) for $\mathcal{H}^2$-matrices can be obtained using interpolation in both variables,

$$G(x,y) \approx \sum_\mu \sum_\nu G(x^\tau_\mu, y^\sigma_\nu) \mathcal{L}^\tau_\mu(x) \mathcal{L}^\sigma_\nu(y).$$

The row and column cluster bases U_τ and V_σ are constructed using polynomial approximations of degree $k-1$. The existence of suitable transfer matrices is an immediate result, because the approximating functions on each level span the space of polynomials of degree $k-1$. In fact, the elements of the transfer matrices $T^U_{\tau',\tau}$ are simply given by

$$t_{\mu',\mu} = \mathcal{L}^\tau_\mu(x^{\tau'}_{\mu'}).$$

An automatic and adaptive way to construct low-rank approximations for general kernel functions is *adaptive cross approximation*. There, k rows and k columns of the regular dense matrix are computed, and a rank k approximation is constructed based on this information. For details, the reader is referred to [188] for the idea of cross approximation, and to [16, 26] for the application in hierarchical matrix methods. Memory requirements can be further reduced by *adaptive recompression*. This technique consists of a posteriori optimisation of the rank by computing the singular value decomposition of subblocks of the form AB^T, or by coarsening several subblocks into one larger block [95]. The algorithm can be performed during the construction of the $\mathcal{H}$-matrix.

For boundary integral equations, an expansion for the kernel function can also be constructed in the parameter space, because the kernel function is always evaluated on the boundary Γ. In that case, an expansion is sought for $\tilde{G}(t,\tau) := G(\kappa(t),\kappa(\tau))$, where $\kappa(t)$ is the parameterisation of Γ. This approach has advantages and disadvantages. Roughly speaking, since *less information* is required, the approximation in the parameter space may be better for a fixed number of terms than an approximation in the full space around a part of Γ. On the other hand, an approximation in the full space is more robust with respect to non-smooth boundaries. Approximations in the parameter space are usually not considered in the fast multipole method. They are an option in hierarchical matrix methods. It was shown in Chapter 3 that wavelet based methods always operate in the parameter domain.

5.5 High frequency $\mathcal{H}$-matrix methods

A recent application of $\mathcal{H}^2$-matrices to oscillatory integral equations in the high frequency regime was proposed in [13]. The method is based on expansion (4.13), and is therefore very similar to fast multipole methods in the high frequency regime. One difference is that the method allows the interaction of clusters on different levels in the hierarchy, because the standard data structures of $\mathcal{H}^2$-matrices can represent such interactions. Another difference is the approach that is used to solve the stability issues at low frequencies that were illustrated in the numerical results in §4.6.2.

Specifically, subblocks of the matrix where the expansion is unstable are replaced by the regular $\mathcal{H}$-matrix representation AB^T using adaptive cross approximation. The rank k for this subblock may have to be large, but the approximation remains stable for all values of k. It is shown in [160] that there exists a constant $C(\epsilon)$ such that expansion (4.13) in the cluster $\tau \times \sigma$ is stable if $k \min(r_\tau, r_\sigma) \geq C(\epsilon)$. Hence, asymptotically, the special treatment of some admissible leaves where this condition is not satisfied has no influence on the order of the method, because for large enough k the condition will eventually be satisfied. Then, the expansion and the efficient diagonal translation operators (or transfer matrices) can be applied.

5.6 Numerical results

We present some numerical results of hierarchical matrix methods. First, we illustrate the convergence behaviour of the boundary element method using a Galerkin discretisation in §5.6.1. This example is not specific to the $\mathcal{H}$-matrix method, but the results will be referred to in §5.6.2. There, we solve large scattering problems using $\mathcal{H}$-matrices. The implementation was based on the software library HLib [25]. Finally, we compare a number of different choices for the separable approximations and discuss the differences with other multiscale methods in §5.6.3.

5.6.1 Convergence of the boundary element method

We illustrate the results of Theorem 2.7.2 with a numerical example. Recall from the theorem that the approximation u_h to the exact solution u satisfies

$$\|u - u_h\|_{H^t(\Gamma)} \leq c h^{s-t} \|u\|_{H^s(\Gamma)}, \tag{5.5}$$

for $-d + r \leq t < \gamma$, $t \leq s$ and $r/2 \leq s \leq d$. Specifically, this means that $\|u - u_h\|_{L_2(\Gamma)} \leq c h^d \|u\|_{H^d(\Gamma)}$ if the solution is sufficiently smooth, where d is the approximation order of the basis functions. The optimal convergence rate $s - t = 2d - r$ corresponds to the choice $t = -d + r$ and $s = d$. Although the Sobolev norms involved are hard to compute, this optimal rate can be observed and is important in practice: the error of the evaluation of the single-layer potential outside Γ, i.e., the error of the actual solution to the Helmholtz equation, converges with this optimal rate. Assume that u is the exact solution of the integral equation of the first kind $Su = f$ on Γ. Recall the evaluation of the single-layer potential S in a point y outside Γ,

$$(Su)(y) = \int_\Gamma G(x, y) u(x) \, \mathrm{d}s_x, \qquad y \notin \Gamma.$$

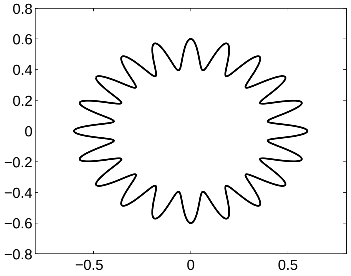


Figure 5.2: A two-dimensional scattering obstacle.

The error in the point y is given by

$$\begin{aligned}|(Su)(y) - (Su_h)(y)| &= |(G(\cdot, y), u - u_h)_{L_2(\Gamma)}| \\ &\le \|G(\cdot, y)\|_{H^{-t}(\Gamma)} \|u - u_h\|_{H^t(\Gamma)} \\ &\le c h^{s-t} \|u\|_{H^s(\Gamma)} \|G(\cdot, y)\|_{H^{-t}(\Gamma)}.\end{aligned}$$

The function $G(x, y)$ is always smooth for $x \in \Gamma$, because $y \notin \Gamma$. Hence, the factor $\|G(\cdot, y)\|_{H^{-t}(\Gamma)}$ can be bounded by a constant that depends only on the diameter of Γ. If the exact solution u is sufficiently smooth, we can choose $t = -d + r$ and $s = d$, leading to the optimal convergence rate.

We considered the integral equation $Su = f$ using the single-layer potential on the domain that is shown in Figure 5.2. The solution of the Helmholtz equation was evaluated in the origin, $y = (Su)(0)$, for a low value of the wavenumber $k = 2$. Table 5.1 shows the convergence results for this example. The convergence factor is shown in parentheses. For $d = 2$, corresponding to the use of piecewise linear hat functions, the L_2-error in the solution of the integral equation decreases at a quadratic rate. This corresponds to (5.5) with $t = 0$ and $s = 2$. The error of the solution to the Helmholtz equation in a point outside Γ decreases much faster. The optimal convergence factor is $2^5 = 32$ in this case. For $d = 3$, corresponding to the use of quadratic B-splines, the optimal rate is $2^7 = 128$. This rate is not observed in the bottom row only because maximal accuracy has already been achieved.

5.6.2 $\mathcal{H}$-matrix method

We revisit the example of scattering by the object shown in Figure 5.2, using $\mathcal{H}$-matrices. For simplicity, we consider $\mathcal{H}$-matrices with a fixed rank k, regardless of the size of the clusters. There are several ways to choose the

Table 5.1: Illustration of the convergence of the boundary element method. The table shows the L_2 error of the computed solution of the integral equation, and the error $|\tilde{y} - y|$ of the solution to the Helmholtz equation in the origin. The convergence factor is shown in parentheses.

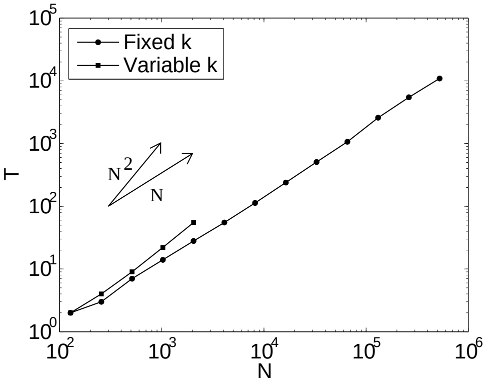
	$d = 2$		$d = 3$	
N	$\|u - u_h\|$	$\lvert\tilde{y} - y\rvert$	$\|u - u_h\|$	$\lvert\tilde{y} - y\rvert$
64	$4.4E - 0$	$3.0E - 3$	$4.0E - 0$	$3.8E - 3$
128	$1.0E - 0$ (4.4)	$3.4E - 5$ (87)	$5.4E - 1$ (7)	$1.4E - 6$ (2667)
256	$1.9E - 1$ (5.4)	$3.9E - 7$ (88)	$4.3E - 2$ (12)	$1.3E - 8$ (106)
512	$4.3E - 2$ (4.4)	$1.0E - 8$ (39)	$1.0E - 3$ (42)	$9.0E - 11$ (148)
1024	$9.9E - 3$ (4.4)	$3.0E - 10$ (34)	$1.5E - 4$ (7)	$6.1E - 12$ (15)

rank, and to measure the error. Often, a fixed tolerance error ϵ is chosen a priori, and the separable approximations are constructed such that the error of each approximation is smaller than ϵ. For larger values of N, ϵ is given a smaller value, because the discretisation error is smaller as well. Here, we will measure the error by evaluating the solution in a point outside Γ.

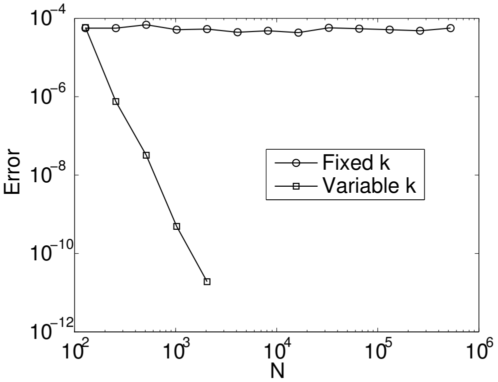
We compare two ways of choosing the rank: a fixed rank k for all values of N, and a rank that increases with N. The fixed rank was chosen $k = 18$, the latter rank was chosen such that the high accuracy of the results in Table 5.1 are preserved. The results are shown in Figure 5.3. The computation time is shown in the top panel, the accuracy of a point evaluation in the interior of Γ is shown in the bottom panel. For a fixed rank k, the error remains bounded as N increases and the computation time is approximately linear. Even for $N = 524,288$, corresponding to a dense matrix with 274 billion elements, the error is fixed at approximately $5E - 5$. This experiment illustrates the robustness of the method. In order to preserve the optimal convergence rate, the rank has to increase with N. In that case, the total time appears to increase only slightly more than linear in N. No computations were done for $N > 2048$ in this example, because the maximal obtainable precision was reached. We have plotted one function that behaves as N^2 in Figure 5.3(a) to illustrate the different slope of a linear and a quadratic function.

5.6.3 The choice of separable approximations

We investigate the influence of the approximation on the error of the solution method. Approximations may be constructed in $\mathbb{R}^d$, or in the lower-dimensional parameter domain. They can be based on analytical manipulations of a specific kernel function, or on general black-box approaches. The separable expansions of the fast multipole method are usually (but



(a) Time



(b) Error

Figure 5.3: The total time and the error of a point evaluation outside Γ for scattering by the object in Figure 5.2. The figure shows the results for a fixed rank approximation, and a variable rank approximation.

not necessarily) analytically constructed approximations in space. We have seen that the wavelet method constructs general approximations in the parameter domain. The flexible algebraic concept of $\mathcal{H}$-matrices leaves the choice.

First, if the boundary Γ is smooth, then approximations in the parameter domain are likely to be more accurate. Consider for example a polynomial interpolation procedure for two-dimensional problems: an interpolating polynomial of degree $p-1$ in the parameter domain results in a rank p approximation of the form (5.4). A two-dimensional interpolating polynomial of degree $p-1$ in y_1 and y_2, where $y=(y_1,y_2)$, results in a rank p^2 approximation. For smooth boundaries, both approximations have approximately the same error along the boundary, but the latter approach requires a much larger rank. If the boundary is not so smooth however, an approximation in space may be more accurate because its accuracy is not influenced by the shape of the boundary. As an example, consider a curve parameterised by

$$\kappa(t):[0,1]\rightarrow\Gamma:\left\{\begin{array}{l} x_1=(0.5+0.1\cos(L2\pi t))\cos(2\pi t),\\ x_2=(0.5+0.1\cos(L2\pi t))\sin(2\pi t).\end{array}\right.$$

This is the parameterisation of the object shown in Figure 5.2, for $L=20$. The parameter L determines the number of oscillations along the boundary. We implemented the $\mathcal{H}$-matrix method using polynomial interpolation in the parameter domain $[0,1]$ and in the space $\mathbb{R}^2$, for $L=100$. At $N=2048$, a rank 25 approximation for the former yields a relative error of 35% for a point evaluation outside Γ. A rank 25 approximation for the interpolation of the kernel function in two dimensions (of degree 4 in each dimension), yields a relative error of $1.2E-7$. The interpolation in the two-dimensional space is robust, and very accurate. Polynomial interpolation in the parameter domain leads to a major loss of accuracy for this example.

Finally, we illustrate the difference between a tailored separable expansion for a specific kernel function, and a general purpose black-box approximation. We compare the rank required for a fixed error $\epsilon=10^{-6}$ for increasing values of the wavenumber of the Helmholtz equation. As an example of a tailored expansion, we consider the approximation given by (4.17)-(4.19). The approximation is compared with the rank obtained by polynomial interpolation in $\mathbb{R}^2$, and with the optimal rank computed using the singular value decomposition. The results are shown in Figure 5.4. The rank increases linearly with the wavenumber for all three methods, as expected. Clearly, polynomial interpolation is not a good method for approximating the oscillatory function. In general, customised expansions lead to a smaller approximation error, although the difference is often much smaller than in the example shown. This improved accuracy comes at a cost of decreased flexibility.

The rank that is obtained using (4.17)-(4.19) is still approximately ten times larger than the optimal rank obtained with SVD for this example. Note however that the analytical expansion leads to a representation of the form $M|_{\tau\times\sigma}=U_\tau\Sigma_{\tau\times\sigma}V_\sigma^T$, where the row and column cluster bases are independent of each other. The expansion can therefore be used for the construction of $\mathcal{H}^2$-matrices. On the other hand, the singular value

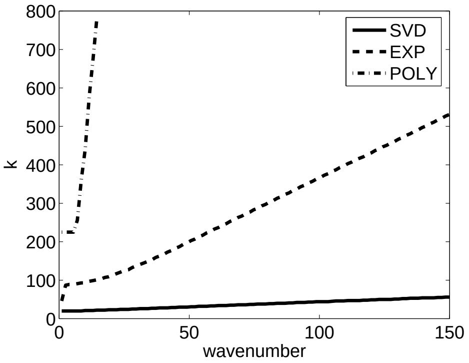


Figure 5.4: The rank required for a fixed error $\epsilon = 10^{-6}$ for a single admissible cluster as a function of the wavenumber of the Helmholtz problem. The figure compares the singular value decomposition (SVD), the fast multipole method expansion (EXP) and polynomial interpolation (POLY).

decomposition of $M|_{\tau\times\sigma}$ that was computed in this example is specific for a single block cluster $\tau \times \sigma$.

5.7 Three-dimensional problems

Hierarchical matrices have been proposed for multivariate problems from their introduction [102]. In that case, the creation of the block cluster tree $T_{I\times I}$ is more involved. In particular, $T_{I\times I}$ does not have to be a binary tree, and several different approaches for the construction of the tree have been proposed [96]. The type of clustering may have a large impact on the efficiency of the overall scheme.

Black-box separable approximations for higher dimensional kernel functions can be constructed, for example, using tensor product Lagrange interpolating polynomials. In fact, this approach is already used for two-dimensional problems, because the kernel is often approximated in the two-dimensional space, rather than in the one-dimensional parameter domain. Numerical results for three-dimensional problems are reported in [26, 95].

5.8 Conclusions

$\mathcal{H}$-matrices represent an algebraic approach to the approximation of matrices that arise in integral equations. A matrix-vector product can be obtained in $O(N \log N)$ operations for $\mathcal{H}$-matrices, and in $O(N)$ operations for $\mathcal{H}^2$-matrices, for a fixed error. The numerical results have shown that the low rank approximation is stable: very large matrices can be approximated without loss of accuracy. Several algorithms have been proposed for the black-box construction of $\mathcal{H}$-matrices for general kernel functions, and for the adaptive optimisation of the memory requirements based on purely algebraic operations. These algorithms illustrate the difference in approach compared to fast multipole methods. Fundamentally, both methods are quite similar. Fast multipole methods and $\mathcal{H}$-matrices will yield similar results for a similar problem.

In the high frequency regime, the proposed application of $\mathcal{H}^2$-matrices in literature is highly related to the high frequency fast multipole method. The numerical instability of the approach at low frequencies can be remedied by combining $\mathcal{H}^2$-matrices with $\mathcal{H}$-matrices: the subblocks corresponding to a block cluster where the separable expansion is unstable are represented by the low-rank approximation of a regular $\mathcal{H}$-matrix. The asymptotic complexity of this approach remains $O(N \log^p N)$, with $p \geq 1$, where N increases linearly with k.

Chapter 6

The evaluation of oscillatory integrals

6.1 Introduction

The discretisation of oscillatory integral equations invariably involves the evaluation of a large number of oscillatory integrals. Efficient solution methods for integral equations therefore require efficient methods for the evaluation of such integrals. Vice versa, we will see in the next chapter that the study of the properties of oscillatory integrals may actually lead to new solution methods for oscillatory integral equations. The problems of evaluating oscillatory integrals and of solving oscillatory integral equations are highly correlated. For that reason, this chapter is devoted to a study of oscillatory integrals.

Oscillatory integrals were not always explicitly present in the methods discussed so far. For example, we have seen in Chapter 2 that the elements of the discretisation matrix in a standard boundary element method are given by the double integrals (2.61). The integration domain depends on the size of the support of the basis functions. Using a fixed number of basis functions per wavelength, where the basis functions have local support, the integral may not be oscillatory at all. Still, even in that case, one may regard an entire row of the discretisation matrix as the discretisation of an oscillatory integral using many basis functions. In the wavelet method, some basis functions may span a large part of the boundary. In that case, the double integral (2.61) itself is highly oscillatory for large wavenumbers. The presence of oscillatory integrals is even more pronounced in hybrid methods, as we will see in Chapter 7.

We will model oscillatory integrals by integrals of the form

$$I[f] := \int_a^b f(x)e^{i\omega g(x)}\,\mathrm{d}x. \tag{6.1}$$

The parameter ω in (6.1) determines the frequency of the oscillations of the integrand. We assume that the functions f and g are non-oscillatory. We call f the *amplitude function*, and g the *oscillator* of (6.1). Traditional integration techniques fail for integrals of the form $I[f]$ at large frequencies. For example, classical quadrature rules of Newton-Cotes type or Gaussian type are based on polynomial interpolation. It is well known that polynomials are not suited for the approximation of oscillatory functions, and the integration error of these quadrature rules increases rapidly with increasing ω. For that reason, (6.1) is usually evaluated using composite quadrature [69]. The number of subintervals of the integration interval $[a, b]$ is chosen proportional to ω, thereby eliminating oscillation. This means that a fixed number of quadrature points is used per wavelength. It is an immediate consequence that the number of operations in this approach scales linearly in ω. For higher-dimensional integrals, the dependence is more than linear.

In the past few years, several new methods have been proposed that require only a fixed number of operations for increasing ω, and that deliver an accuracy that increases with ω [145, 129, 128, 161, 124]. These efficient methods rely on the observation that the value of $I[f]$ asymptotically depends only on the behaviour of f and g near the endpoints a and b, and near the so-called *stationary points* of g. These are all the points ξ where $g'(\xi) = 0$. Their importance lies in the fact that, locally, the integrand is not oscillatory near a stationary point. Away from the endpoints and all stationary points, the oscillations of the integrand increasingly cancel out. The foundations of the new methods can be traced back to Louis Filon [84], and to the development of the method of stationary phase and the method of steepest descent in the last two centuries [151, 181, 70]. The recent advances consist of the construction of general computational tools that allow the evaluation of $I[f]$ with arbitrarily high *asymptotic order*. It is said that an approximation $Q[f]$ to $I[f]$ has asymptotic order s, with $s \geq 0$, if

$$I[f] - Q[f] = O(\omega^{-s-1}), \qquad \omega \to \infty. \tag{6.2}$$

This means that the absolute error of the approximation decreases rapidly with increasing frequency ω. In the absence of stationary points, the relative error of a method with asymptotic order s scales as $O(\omega^{-s})$.

The first method with high asymptotic order is obtained by truncating the asymptotic expansion of $I[f]$ for large ω. This method is the natural precursor to a number of different methods with different properties. We present an overview in §6.2, and refer the reader to [132] for a more detailed discussion. We present the *numerical steepest descent method* (NSD) in §6.3. Finally, we consider multivariate integrals in §6.4. We restrict the discussion in this thesis to methods with high asymptotic order for integrals of the general form (6.1). Various other approaches for oscillatory integrals have been developed, such as exponential fitting methods and generalised quadrature rules; for these, the reader is referred to [82, 81, 137, 191].

Throughout this chapter, the integration error in the numerical experiments was determined by comparison with the results of Cubpack [50].

6.2 Methods with high asymptotic order

6.2.1 The asymptotic method

The properties of the integral $I[f]$ are revealed by the asymptotic expansion for large values of the frequency parameter ω. It was mentioned in the introduction of the chapter that the value of $I[f]$ depends on the behaviour of the amplitude f and of the oscillator g near the endpoints a and b, and near the stationary points. Here, we make this statement more precise. It is said that a stationary point ξ has *order* r if

$$\left\{ \begin{array}{ll} g^{(j)}(\xi) = 0, & j = 1, \ldots, r, \\ g^{(r+1)}(\xi) \neq 0. & \end{array} \right. \tag{6.3}$$

Assume that g has one stationary point $\xi \in (a,b)$ of order r in the interior of the integration domain $[a,b]$. It is shown in [178] that $I[f]$ has an asymptotic expansion of the form

$$I[f] \sim \sum_{j=0}^{\infty} \frac{c_j[f]}{\omega^{(j+1)/(r+1)}}, \qquad \omega \to \infty, \tag{6.4}$$

for every smooth function f that is compactly supported near ξ in (a,b). Unfortunately, the linear functionals $c_j[f]$ in (6.4) are not given in explicit form. The first coefficient $c_0[f]$ can be obtained using the method of stationary phase [178], the remaining coefficients can in some cases be determined numerically using the method of steepest descent (see Appendix B). Still, the expansion is useful: it can be shown that the first few coefficients c_j are determined by the first few derivatives of f and g at ξ. The analysis can be extended to functions f that are not compactly supported in (a,b) by considering bump functions and remainders [126, 127]. In that way, it can be shown that the asymptotic expansion of $I[f]$ is determined by the derivatives of f and g at the points a, b and ξ.

A different approach was taken by Iserles and Nørsett in [129, 128], leading to an asymptotic expansion with an explicit representation for the coefficients. The approach consists of factoring out *generalised moments* of the functional $I[f]$, given by

$$\mu_j(\omega;\xi) = I[(x-\xi)^j] = \int_a^b (x-\xi)^j e^{i\omega g(x)}\,\mathrm{d}x, \qquad j \geq 0. \tag{6.5}$$

Assume the function g has a stationary point $\xi \in (a,b)$ of order r. Then $I[f]$ has an asymptotic expansion for every smooth function f given by (see [129])

$$\begin{aligned} I[f] \sim & \sum_{j=0}^{r-1} \frac{1}{j!}\mu_j(\omega;\xi) \sum_{m=0}^{\infty} \frac{1}{(-i\omega)^m}\rho_m^{(j)}[f](\xi) \\ & - \sum_{m=1}^{\infty} \frac{1}{(-i\omega)^m} \left(\frac{e^{i\omega g(b)}}{g'(b)}\{\rho_{m-1}[f](b) - \rho_{m-1}[f](\xi)\} \right. \\ & \left. - \frac{e^{i\omega g(a)}}{g'(a)}\{\rho_{m-1}[f](a) - \rho_{m-1}[f](\xi)\} \right), \end{aligned} \tag{6.6}$$

where

$$\begin{aligned} \rho_0[f](x) &= f(x), \\ \rho_{m+1}[f](x) &= \frac{\mathrm{d}}{\mathrm{dx}} \frac{\rho_m[f](x) - \sum_{j=0}^{r-1} \frac{1}{j!}\rho_m[f]^{(j)}(\xi)(x-\xi)^j}{g'(x)}, \quad m \geq 0. \end{aligned}$$

The *asymptotic method* $Q_A[f]$ is defined by truncating expansion (6.6),

$$\begin{aligned} Q_A[f] = & \sum_{j=0}^{r-1} \frac{1}{j!}\mu_j(\omega;\xi) \sum_{m=0}^{s-j-1} \frac{1}{(-i\omega)^m}\rho_m^{(j)}[f](\xi) \\ & - \sum_{m=1}^{s} \frac{1}{(-i\omega)^m} \left(\frac{e^{i\omega g(b)}}{g'(b)}\{\rho_{m-1}[f](b) - \rho_{m-1}[f](\xi)\} \right. \\ & \left. - \frac{e^{i\omega g(a)}}{g'(a)}\{\rho_{m-1}[f](a) - \rho_{m-1}[f](\xi)\} \right). \end{aligned} \tag{6.7}$$

The size of the moments $\mu_j(\omega;\xi)$ as a function of ω is $\mu_j(\omega;\xi) = O(\omega^{-j/(r+1)})$. Specifically, we have $\mu_0(\omega;\xi) = O(\omega^{-1/(r+1)})$ by van der Corput's lemma [178]. It follows that the asymptotic method has an asymptotic error of size

$$I[f] - Q_A[f] = O(\omega^{-s-1/(r+1)}), \qquad \omega \to \infty. \tag{6.8}$$

Because we also have $I[f] = O(\omega^{-1/(r+1)})$, the method has a relative error of order $O(\omega^{-s})$.

A disadvantage of the asymptotic method is that the error is essentially uncontrollable. Asymptotic expansions may diverge after a number of terms. This means that the method may break down for low values of ω.

6.2.2 Filon-type methods

The defining characteristics of *Filon-type methods* are the construction of an approximation to the amplitude function f, and the exact integration of that approximation. Originally, Louis Filon proposed a quadratic approximation to f in 1928 [84]. This idea was generalised several times since then [153, 85]. A fundamental generalisation was realised in [129, 128], leading to a method with arbitrarily high asymptotic order.

We can describe Filon-type methods formally as follows. Assume

$$f(x) \approx \sum_{i=0}^{n} a_i[f]\phi_i(x),$$

i.e., the amplitude function can be approximated by a linear combination of given basis functions ϕ_i with coefficients $a_i[f]$ that depend linearly on f. Then we have

$$I[f] \approx \sum_{i=0}^{n} w_i a_i[f], \quad \text{with} \quad w_i := I[\phi_i].$$

Hence, the result is a quadrature rule with a classical form. The weights w_i are given by oscillatory integrals themselves; they need to be computed in a different way, or be available analytically.

Iserles and Nørsett identified suitable approximations for f that yield a method with arbitrarily high asymptotic order in [129, 128]. The approximation is such that the first coefficients of the asymptotic expansion (6.6) of the error vanish. Specifically, the exact interpolation of f and its derivatives is required at the endpoints and stationary points. This can be accomplished using Hermite interpolation. Choose points c_l and integers θ_l, $l = 1, \ldots, \nu$, and construct the polynomial $\tilde{f}$ that satisfies

$$\tilde{f}^{(j)}(c_l) = f^{(j)}(c_l), \qquad j = 0, \ldots, \theta_l, \quad l = 1, \ldots, \nu. \tag{6.9}$$

This interpolating polynomial can be written as a linear combination of function values and derivatives of f at the nodes c_l,

$$\tilde{f}(x) = \sum_{l=1}^{\nu} \sum_{j=0}^{\theta_l} f^{(j)}(c_l)\psi_{l,j}(x),$$

with

$$\psi_{l,j}^{(j)}(c_l) = 1, \quad \text{and} \quad \psi_{l,j}^{(m)}(c_n) = 0, \quad (m,n) \neq (i,j).$$

The Filon-type method is defined by

$$Q_F[f] = I[\tilde{f}] = \sum_{l=1}^{\nu} \sum_{j=0}^{\theta_l} w_{l,j} f^{(j)}(c_l), \quad \text{with} \quad w_{l,j} = I[\psi_{l,j}]. \tag{6.10}$$

The integration error depends on the number of derivatives θ_l that is interpolated in the nodes. Let $c_1 = a$, $c_\nu = b$, and assume all stationary points correspond to a node c_l. If $\theta_1, \theta_\nu \geq (s-1)$ and if $\theta_l \geq (s-1)(r+1)$ for each stationary point c_l of order r, then the error has asymptotic size

$$I[f] - Q_F[f] = O(\omega^{-s-1/(r+1)}). \tag{6.11}$$

The method has the same asymptotic accuracy as the asymptotic method. An important difference is that the Filon-type method yields good results for low values of ω as well because, by construction, it is exact for all polynomials of degree $\sum_{l=1}^{\nu} \theta_l - 1$. The accuracy can be increased arbitrarily by adding interpolation points. The evaluation of derivatives of f can be avoided by choosing the interpolation points as a function of ω [128]. A disadvantage is that the method requires knowledge of the weights, or *moments* $w_{l,j} = I[\psi_{l,j}]$. The moments are given by oscillatory integrals themselves.

6.2.3 Levin-type methods

An entirely different approach for the evaluation of $I[f]$ was pioneered by David Levin in [144, 145, 146]. Contrary to Filon-type methods, the method does not require moments for the approximation of $I[f]$. It was extended to arbitrarily high asymptotic order by Olver [161].

The *Levin-type method* of [161] can be formulated as follows. Assume we have a function $F(x)$, such that

$$\frac{d}{dx}\left[F(x)e^{i\omega g(x)}\right] = f(x)e^{i\omega g(x)}. \tag{6.12}$$

It then follows that

$$I[f] = \int_a^b f(x)e^{i\omega g(x)}\,\mathrm{d}x = \left[F(x)e^{i\omega g(x)}\right]_a^b. \tag{6.13}$$

As it turns out, in the absence of stationary points, $F(x)$ is a smooth function. From (6.12), we note that it satisfies the ordinary differential equation $L[F] = f$, with $L[F] = F' + i\omega g' F$. An approximation $v(x) = \sum_{l=1}^{\nu} a_l \psi_l(x)$ to $F(x)$ can be constructed by solving the system of collocation equations

$$\begin{aligned}
L[v](x_l) &= f(x_l), \\
\frac{dL[v]}{dx}(x_l) &= f'(x_l), \\
&\vdots \\
\frac{d^{\theta_l} L[v]}{dx^{\theta_l}}(x_l) &= f^{(\theta_l)}(x_l),
\end{aligned}$$

for $l = 1, \ldots, \nu$. The Levin-type method is defined by

$$Q_L[f] = \left[v(x) e^{i\omega g(x)} \right]_a^b. \tag{6.14}$$

If the collocation points x_l include the endpoints a and b, $c_1 = a$ and $c_\nu = b$, and if in addition $\theta_1, \theta_\nu \geq s - 1$, then the error of the method has asymptotic size

$$I[f] - Q_L[f] = O(\omega^{-s-1}). \tag{6.15}$$

The method has the same high asymptotic order as the previous methods. It does not require the knowledge of moments of I, and it works for low values of the frequency parameter ω. The accuracy can be arbitrarily increased by adding collocation points. In some cases, a choice of basis functions is available such that adding internal collocation points increases the asymptotic order [161]. A disadvantage is that the approach only works in the absence of stationary points.

6.3 Numerical steepest descent method

6.3.1 Overview of the method

The method described in this section achieves a similar high asymptotic order of accuracy as the previously discussed methods. We will see that it solves some of the problems of the other methods, and introduces some peculiarities of its own, thus adding to the spectrum of available approaches.

The method depends on two simple observations. First, the oscillatory function $e^{i\omega g(x)}$ decays exponentially fast for a complex $g(x)$ along a path with a growing imaginary part. Second, the oscillatory function $e^{i\omega g(x)}$ does *not* oscillate for complex $g(x)$ along a path with fixed real part. These observations are exploited numerically in combination with a corollary to Cauchy's Theorem: the value of a line integral of an analytic function along a path between two points in the complex plane does not depend on the exact path taken [111]. The same observations provide the foundation for the *method of steepest descent* (see appendix B). In that method, an asymptotic expansion of the form (6.4) is developed for $I[f]$. The method dates back to Cauchy and Riemann in the nineteenth century, before it was made popular by Debye in 1909 [70]. Methods in the complex plane have been considered for oscillatory integrals several times since, in specific applications or for Laplace transforms (see, e.g., [35, 68, 185]).

We will present a rather general implementation of the steepest descent method, that is also valid for small values of ω. We prove convergence estimates of the numerical scheme as a function of the frequency, and we extend the method to functions f and g that are not analytic. The implementation can be entirely numerical; hence we shall refer to the method as the *numerical steepest descent method*. We start the discussion in §6.3.2 with some practical and motivating examples that illustrate most of the theory described later. In §6.3.3 we describe and analyse the idealised setting that gives the best possible convergence. It is shown that a suitable n-point quadrature rule in that setting leads to an asymptotic order of $2n$. This setting comes with the most restrictions, but still covers many important applications. The first requirement is that the functions f and g in (6.1) be analytic in an (infinitely) large region of the complex plane containing the integration interval $[a, b]$. Further, it is assumed that there are no stationary points in $[a, b]$, and that the equation $g(x) = c$ should be "easily solvable". This rather vague description will be made more precise further on. We then proceed by relaxing the requirements one by one, until a more generally applicable method is obtained. This increase in generality will, at times, come with a loss in convergence rate. In §6.3.4 we will allow stationary points. We relax the "easy-solvability" requirement in §6.3.5. We drop the requirements that f and g should be analytic in §6.3.6 and §6.3.7.

6.3.2 Some motivating examples

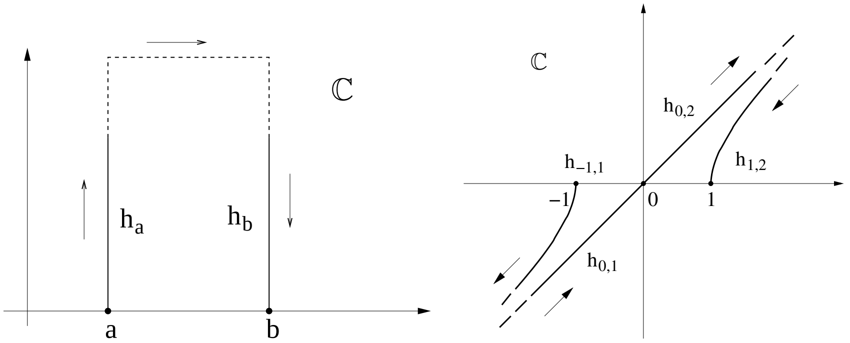
Consider the following integral, which frequently appears in Fourier analysis,

$$I_1[f] := \int_a^b f(x)e^{i\omega x}\,\mathrm{d}x. \tag{6.16}$$

This integral has the form of (6.1) with $g(x) = x$. An important observation is that the function $e^{i\omega x}$ decays rapidly for complex values of x with a positive imaginary part, since $e^{i\omega x} = e^{-\omega\Im x}e^{i\omega\Re x}$. The speed of the decay actually grows as the frequency parameter ω increases. Additionally, the function does not oscillate if the real part of the argument x remains fixed.

Based on these observations integral (6.16) can be reformulated in such a way that the difficulty - the highly oscillatory nature - is removed. To that end, the integration on interval $[a, b]$ is replaced with a path in the complex plane as illustrated in the left panel of Figure 6.1. The first, vertical part of the path is of the form $z = h_a(p) := a + ip$ for $p \in [0, P]$. The second part is horizontal and connects the points $h_a(P) := a + iP$ to the point $h_b(P) := b + iP$. Finally, the third part connects $h_b(P)$ to b with the vertical path $z = h_b(p)$ for $p \in [0, P]$. Now assume that f is analytic, and that f itself does not grow exponentially large in the complex plane. Letting P go to infinity, and using paths parameterised by $h_a(p)$ and $h_b(p)$, for $p \in [0, \infty)$, we can write (6.16) as

$$I_1[f] = e^{i\omega a}\int_0^\infty f(a + ip)e^{-\omega p}\,\mathrm{d}p - e^{i\omega b}\int_0^\infty f(b + ip)e^{-\omega p}\,\mathrm{d}p. \tag{6.17}$$



(a) $g(x) = x$ (b) $g(x) = x^2$

Figure 6.1: Illustration of the integration paths for $g(x) = x$ and $g(x) = x^2$.

The integral along the path that connects the endpoints of $h_a(P)$ and $h_b(P)$ vanishes for $P = \infty$ and can therefore be discarded. Both integrals in the right-hand side of (6.17) are well behaved. They can be evaluated efficiently by standard numerical integration techniques, e.g., by Gauss–Laguerre integration [69]. It can be expected from (6.17) that the accuracy of any numerical integration scheme will increase with increasing ω, thanks to the faster decay of the integrand. This expectation will be confirmed both theoretically and numerically in the subsequent sections. One also sees that, asymptotically, the behaviour of f around $x = a$ and $x = b$ completely determines the value of (6.16).

Next, we consider the function $g(x) = x^2$ and the corresponding integral

$$I_2[f] := \int_{-1}^{1} f(x) e^{i\omega x^2} \,\mathrm{d}x.$$

Again, we can remove the integration difficulty by a careful selection of an integration path in the complex plane. The path is drawn in the right panel of Figure 6.1. The following notation is used for the parameterisation: $h_{xj}(p) = (-1)^j \sqrt{x^2 + ip}$. Integrating along any such path for $p \in [0, \infty)$ leads to an integrand with the desired decay properties, since $e^{i\omega h_{xj}(p)^2} = e^{i\omega x^2} e^{-\omega p}$. One can see that, for general g, a similar result is obtained if the path satisfies $g(h_x(p)) = g(x) + ip$. This path can be found by using the inverse of g, if it exists, i.e., $h_x(p) = g^{-1}(g(x) + ip)$. Returning to the example function $g(x) = x^2$ however, we note that the inverse of $y = g(x)$ is multivalued: we have $x = -\sqrt{y}$ corresponding to the restriction $g_1 :=$

$g|_{[-1,0]}$, and $x = \sqrt{y}$ corresponding to $g_2 := g|_{[0,1]}$. The paths leaving -1 and arriving at 1 are uniquely determined by the requirement that $h_{x,j}(0) = x$. Hence,

$$h_{-1,1}(p) = -\sqrt{1+ip} \quad \text{and} \quad h_{1,2}(p) = \sqrt{1+ip}.$$

Contrary to the first example, the integral along the path that connects the limiting endpoints of $h_{-1,1}(p)$ and $h_{1,2}(p)$ cannot be discarded. Since $h_{-1,1}(p)$ and $h_{1,2}(p)$ have opposite signs, any connecting path should cross the real axis. Additionally we require the connecting path to be such that the integrand along the path is non-oscillatory. The solution is to pass explicitly through the point $x = 0$, via two new paths

$$h_{0,1}(p) = -\sqrt{ip} \quad \text{and} \quad h_{0,2}(p) = \sqrt{ip}.$$

The point $x = 0$ is such that the paths corresponding to the two inverses coincide at $x = 0$. We can now write $I_2[f]$ as

$$e^{i\omega} \int_0^\infty f(h_{-1,1}(p)) e^{-\omega p} h'_{-1,1}(p) \, \mathrm{d}p - \int_0^\infty f(h_{0,1}(p)) e^{-\omega p} h'_{0,1}(p) \, \mathrm{d}p \\ + \int_0^\infty f(h_{0,2}(p)) e^{-\omega p} h'_{0,2}(p) \, \mathrm{d}p - e^{i\omega} \int_0^\infty f(h_{1,2}(p)) e^{-\omega p} h'_{1,2}(p) \, \mathrm{d}p.$$

These four integrals are well behaved, although the derivatives $h'_{0,1}(p)$ and $h'_{0,2}(p)$ introduce a weak singularity of the form $1/\sqrt{p}$, for $p \to 0$. The integrands do not oscillate, and their decay is exponentially fast.

Note that $\xi = 0$ is a stationary point because $g'(\xi) = 0$. More general stationary points, where also higher order derivatives of g vanish, are treated in a similar way. Consider, e.g., $g(x) = x^3$ and its inverse $g^{-1}(y) = \sqrt[3]{y}$. The cubic root has three branches in the complex plane, and the optimal path $h_x(p) = g^{-1}(g(x) + ip)$ at the point x is found by taking the branch corresponding to the inverse of g that is valid at x, i.e., for which $h_x(0) = x$. At $\xi = 0$, we have that $g'(\xi) = g''(\xi) = 0$ and the three branches coincide. For this example, integral (6.1) can again be decomposed into 4 contributions, each of which corresponds to a non-oscillating integral. The integration path is drawn in Figure 6.2.

6.3.3 The ideal case: analytic integrand and no stationary points

6.3.3.1 An approximate decomposition of the oscillatory integral

The ideal setting for our approach has three conditions: both f and g are analytic functions, there are no stationary points in the integration interval

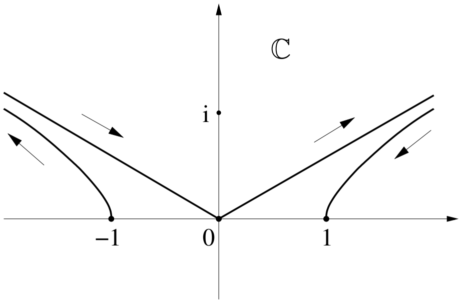


Figure 6.2: Illustration of the integration path for $g(x) = x^3$.

$[a, b]$ (i.e., $g'(x) \neq 0$), and the equation $g(x) = z$ is easily solvable, preferably by analytical means. None of these conditions is crucial in order to obtain a convergent quadrature method, as we will relax all conditions later on. But, the ideal case leads to the highest convergence rate among all cases described, and is most suited to demonstrate our approach: the problem of evaluating (6.1) can be transformed into the problem of integrating two integrals on $[0, \infty)$ with a smooth integrand that does not oscillate, and that decays exponentially fast. This will be proved in this section in Theorem 6.3.3. First, we give a basic lemma for the approximation of an integral with an integrand that becomes small in a region S of the complex plane.

Lemma 6.3.1. *Assume u is analytic in a simply connected complex region $D \subset \mathbb{C}$ with $[a, b] \subset D$, and there exists a bounded and connected region $S \subset D$ such that $|u(z)| \leq \epsilon$, $\forall z \in S$. If the shortest distance between any two points p and q of S along a curve that lies in S can be bounded from above by a constant $M > 0$, then there exists a function $F(x)$, $x \in [a, b]$, such that the integral of u can be approximated by*

$$\int_a^x u(z)\,\mathrm{d}z \approx F(a) - F(x) \tag{6.18}$$

with an error e that satisfies $|e| \leq M\epsilon$. The function F is of the form

$$F(x) = \int_{\Gamma_x} u(z)\,\mathrm{d}z \tag{6.19}$$

with Γ_x any path in D that starts at x and ends in S.

Proof. Let Γ_x be a curve in D from x to an arbitrary point in S, denoted by $q(x)$, and Γ_a a curve in D from a to $q(a) \in S$. Choose κ as the shortest path in S that connects $q(a)$ and $q(x)$. Since u is analytic in D, the integration path between a and x may be chosen as the concatenation of Γ_a, κ and $-\Gamma_x$. The integral can be written as

$$\int_a^x u(z)\,\mathrm{d}z = F(a) + \int_\kappa u(z)\,\mathrm{d}z - F(x), \quad \text{with} \quad \left|\int_\kappa u(z)\,\mathrm{d}z\right| \leq M\epsilon.$$

This proves the result. □

Note that the function F is not completely determined by the conditions of this Lemma. In particular, the endpoint $q(x)$ of Γ_x may be an arbitrary function of x.

If g is analytic, then the oscillating function $e^{i\omega g(x)}$ in the integrand of (6.1) is also analytic as a function of x. This function is small in absolute value if

$$|e^{i\omega g(x)}| \leq \epsilon \iff e^{-\omega \Im g(x)} \leq \epsilon \iff \Im g(x) \geq \frac{-\log(\epsilon)}{\omega}.$$

Hence, if the inverse of g exists, we can find a suitable region S that is required for Lemma 6.3.1 with points given by $g^{-1}(c+id)$, for $d \geq d_0 := \frac{-\log(\epsilon)}{\omega}$. Note that, in general, the inverse of an analytic function may be multi-valued. Each single-valued branch of the inverse has branch points that are located at the points ξ where $g'(\xi) = 0$, and it is discontinuous across branch cuts that extend from one branch point to another, or from a branch point to infinity. By explicitly excluding the presence of branch points locally, a single-valued branch of the inverse can be found that is analytic in a neighbourhood of $[a,b]$. We can then characterise the error of the decomposition given in Lemma 6.3.1 for the particular case of integral (6.1) as a function of ω.

Theorem 6.3.2. *Assume f and g are analytic in a bounded and open complex neighbourhood D of $[a,b]$, and assume $g'(z) \neq 0$, $z \in D$. Then there exists an approximation of the form (6.18) for (6.1), with an error that has order $O(e^{-\omega d_0})$ as a function of ω, for a real constant $d_0 > 0$.*

Proof. Define $S := \{z : \Im g(z) \geq d_0\} \cap D$ with $d_0 > 0$. A positive constant d_0 can always be found such that S is non-empty because g is analytic. In order to prove this, consider a point $x \in [a,b]$. Since g is analytic at x, the equation $g(z) = g(x) + id_0$ always has a solution z for sufficiently small $d_0 > 0$ [111]. Additionally, d_0 can be chosen small enough such that $z \in D$, because D contains an open neighbourhood of x. The necessary geometrical conditions on S required by Lemma 6.3.1 follow from the continuity properties of g. We have

$$\forall x \in S : |f(x)e^{i\omega g(x)}| \leq |f(x)|e^{-\omega d_0}.$$

Since S is finite (because D is bounded), there exists a constant $C > 0$ such that $|f(x)| \leq C$, $x \in S$. The result is established by Lemma 6.3.1 with $u(x) = f(x)e^{i\omega g(x)}$ and $\epsilon = Ce^{-\omega d_0}$. □

Theorem 6.3.2 shows that the error in the approximation $I \approx F(a) - F(b)$ for (6.1) decays exponentially fast as the frequency parameter ω increases. It only requires that f and g are analytic in a finite neighbourhood of $[a, b]$. The function F is given by an integral along a curve that originates in x, and leads to a point z such that $g(z)$ has a positive imaginary part. The result follows from the observation that the integrand has exponential decay along such a path.

6.3.3.2 An exact decomposition of the oscillatory integral

Next, we will take the second observation into account: $e^{i\omega g(x)}$ does not oscillate along a path where $g(x)$ has a fixed real part. This will lead to a particularly useful choice for the path Γ_x in the definition (6.19) of F.

Let $h_x(p)$ be a parameterisation for Γ_x, $p \in [0, P]$, then we find a suitable path as the solution to

$$g(h_x(p)) = g(x) + ip, \quad x \in [a, b].$$

If the inverse of g exists, we have the unique solution $h_x(p) = g^{-1}(g(x) + ip)$. The path $h_x(p)$ is also called the *steepest descent path* (see appendix B). This can be understood as follows. Define $k(x, y) := ig(z) = u(x, y) + iv(x, y)$, with $z = x + iy$. Then we have $e^{i\omega g(z)} = e^{\omega k(x,y)}$. It can be shown that the path is such that $v(x, y) = v(x_0, y_0)$ is constant, and that the descent of $u(x, y)$ is maximal. In particular, the direction of steepest descent coincides with $-\nabla u$ at each point $z = x + iy$.

Using this path in the definition of F, the decomposition becomes

$$\int_a^x f(z)e^{i\omega g(z)}\, \mathrm{d}z \approx F(a) - F(x) = e^{i\omega g(a)} \int_0^P f(h_a(p))e^{-\omega p} h_a'(p)\, \mathrm{d}p \\ - e^{i\omega g(x)} \int_0^P f(h_x(p))e^{-\omega p} h_x'(p)\, \mathrm{d}p.$$

The integrands in the right-hand side do not oscillate, and they decay exponentially fast as the integration variable p or the frequency parameter ω increases.

In the following Theorem, we will consider the limit case $P \to \infty$ in which the error of the approximation vanishes. This will require stronger analyticity conditions for both f and g. Additionally, the function f can no longer be assumed to be bounded. The result of the theorem will hold if f does not grow faster than polynomially in the complex plane along the suggested integration path.

Theorem 6.3.3. *Assume that the functions f and g are analytic in a simply connected and sufficiently (infinitely) large complex region D containing the interval $[a, b]$, and that the inverse of g exists on D. If the following conditions hold in D:*

$$\exists m \in \mathbb{N} : |f(z)| = O(|z|^m), \quad \textit{and} \tag{6.20}$$

$$\exists \omega_0 \in \mathbb{R} : |g^{-1}(z)| = O(e^{\omega_0 |z|}), \quad |z| \to \infty, \tag{6.21}$$

then there exists a function $F(x)$, for $x \in [a, b]$, such that

$$\int_a^x f(z) e^{i\omega g(z)} \,\mathrm{d}z = F(a) - F(x), \qquad \forall \omega > (m+1)\omega_0, \tag{6.22}$$

where $F(x)$ is of the following form,

$$F(x) := \int_{\Gamma_x} f(z) e^{i\omega g(z)} \,\mathrm{d}z, \tag{6.23}$$

with Γ_x a path that starts at x. A parameterisation $h_x(p)$, $p \in [0, \infty)$, for Γ_x exists such that the integrand of (6.23) is $O(e^{-\omega p})$.

Proof. In this proof, we will use $u(z)$ to denote the integrand of (6.1). Using the fact that $|u(z)| = |f(z) e^{i\omega g(z)}| = |f(z)| e^{-\omega \Im g(z)}$, and conditions (6.20) and (6.21), we can state

$$c + id \in D \Rightarrow |u(g^{-1}(c + id))| = O(e^{(m\omega_0 - \omega)d}), \quad d \to \infty. \tag{6.24}$$

If $\omega > m\omega_0$, then (6.24) characterises the exponential decay of the integrand in the complex plane. We will now choose an integration path from the point a to the region where the integrand becomes small, and from that region back to the point $x \in [a, b]$. We will show that the contribution along the line that connects both paths can be discarded. This will establish the existence of Γ_a and Γ_x in (6.23), and the independence of Γ_a and Γ_x.

Assume an integration path for I that consists of three connected parts, parameterised as $h_a(p)$ and $h_x(p)$ with $p \in [0, P]$, and $\kappa(p)$ with $p \in [a, x]$. The parameterisations can be chosen differentiable and satisfy $h_a(0) = a$, $h_x(0) = x$, $h_a(P) = \kappa(a)$ and $h_x(P) = \kappa(x)$. We have

$$\begin{aligned} \int_a^x u(z) \,\mathrm{d}z = &\int_0^P u(h_a(p)) h_a'(p) \,\mathrm{d}p \\ &+ \int_a^x u(\kappa(p)) \kappa'(p) \,\mathrm{d}p - \int_0^P u(h_x(p)) h_x'(p) \,\mathrm{d}p. \end{aligned} \tag{6.25}$$

Since the inverse function g^{-1} exists, we can choose the points $h_a(P)$ and $h_x(P)$ as follows: $h_a(P) = g^{-1}(g(a) + iP)$ and $h_x(P) = g^{-1}(g(x) + iP)$. Hence, by (6.24),

$$|u(h_a(P))| = O(e^{(m\omega_0 - \omega)P}) \quad \text{and} \quad |u(h_x(P))| = O(e^{(m\omega_0 - \omega)P}).$$

We will now show that, as $P \to \infty$, the second integral vanishes. Equation (6.25) is then of the form (6.22), with Γ_a and Γ_x parameterised by $h_a(p)$ and $h_x(p)$ respectively, $p \in [0, \infty)$.

The contribution of the integral along $\kappa(p)$ is bounded by

$$\left| \int_a^x u(\kappa(p))\kappa'(p)\,\mathrm{d}p \right| \leq \max_{p\in[a,x]} |u(\kappa(p))| \max_{p\in[a,x]} |\kappa'(p)|\,|x-a|. \tag{6.26}$$

By selecting the path $\kappa(p) = g^{-1}(g(p) + iP)$, we have from (6.24): $|u(\kappa(p))| = O(e^{(m\omega_0-\omega)P})$, $p \in [a, x]$. We can write the second factor in the bound (6.26) as

$$\kappa'(p) = \frac{\partial g^{-1}}{\partial y}(g(p) + iP)\frac{\mathrm{d}g}{\mathrm{d}p}(p).$$

The derivative of $g(p)$ with respect to p is bounded on $[a, b]$ because g is analytic. The factor $\frac{\partial g^{-1}}{\partial y}(g(p) + iP)$ is bounded by $O(e^{\omega_0 P})$. Combining the asymptotic behaviour of the factors in (6.26), the second term in (6.25) vanishes for $P \to \infty$ and for all $x \in [a, b]$ if $\omega > (m + 1)\omega_0$. This proves the result. □

Remark 6.3.4. *Note that f and g should be analytic in a simply connected region D that contains the paths $h_a(p)$, $h_b(p)$ and $\kappa(p)$ in order to apply Cauchy's Theorem. The unique existence of the inverse of g is a necessary condition: if $g'(z) = 0$ with $z \in D$, then the point z is a branch point of the inverse function. The path $\kappa(p)$ may cross the branch cut that originates at z, and Cauchy's Theorem cannot be applied.*

Remark 6.3.5. *Conditions (6.20) and (6.21) are sufficient but not necessary. For example, the limit case also applies when $f(x) = e^x$ and $g(x) = x$. If however $f(x) = e^{-x^2}$ and $g(x) = x$, the integrand always diverges at infinity along the steepest descent path, regardless of the value of ω. In that case, the path should be truncated at a finite distance from the real axis. The accuracy of the decomposition is then described by Theorem 6.3.2, i.e., the error decays exponentially fast.*

Remark 6.3.6. *The decomposition $I[f] = F(a) - F(b)$ is similar to the result of the form $I[f] = F_L(b)e^{i\omega g(b)} - F_L(a)e^{i\omega g(a)}$ in the Levin-type method of [161] discussed in §6.2.3. Although the approach is very different, both methods yield nearly the same function: we have $F(x) = -F_L(x)e^{i\omega g(x)}$.*

6.3.3.3 Evaluation of $F(x)$ by Gauss–Laguerre quadrature

Next, we consider the evaluation of $F(x)$ as defined by (6.23). The parameterisation of the path $h_x(p)$ solves the equation

$$g(h_x(p)) = g(x) + ip. \tag{6.27}$$

The integrand of (6.1) along this path is non-oscillatory and exponentially decaying,

$$f(h_x(p))e^{i\omega g(h_x(p))} = f(h_x(p))e^{i\omega g(x)}e^{-\omega p}.$$

In the simplest, yet important case $g(x) := x$ the suggested path is given by $h_x(p) = x + ip$.

An efficient approach for infinite integrals with exponentially decaying integrand is Gauss–Laguerre quadrature [69]. Laguerre polynomials are orthogonal w.r.t. e^{-x} on $[0, \infty]$. A Gauss–Laguerre rule with n points is exact for polynomials up to degree $2n - 1$. The integral $F(x)$ with the suggested path can be written as

$$\begin{aligned} F(x) &= \int_0^\infty f(h_x(p))e^{i\omega(g(x)+ip)}h_x'(p)\,\mathrm{d}p \\ &= e^{i\omega g(x)}\int_0^\infty f(h_x(p))h_x'(p)e^{-\omega p}\,\mathrm{d}p \\ &= \frac{e^{i\omega g(x)}}{\omega}\int_0^\infty f(h_x(q/\omega))h_x'(q/\omega)e^{-q}\,\mathrm{d}q \end{aligned}$$

with $q = \omega p$ in the last expression. Applying a Gauss–Laguerre quadrature rule with n points x_i and weights w_i yields a quadrature rule

$$F(x) \approx Q_S[f;x] := \frac{e^{i\omega g(x)}}{\omega}\sum_{i=1}^{n} w_i f(h_x(x_i/\omega))h_x'(x_i/\omega). \tag{6.28}$$

The rule requires the evaluation of f in a complex neighbourhood of x.

Theorem 6.3.7. *Assume functions f and g satisfy the conditions of Theorem 6.3.3. Let I be approximated by the quadrature formula*

$$I \approx Q_{NSD}[f] := Q_S[f;a] - Q_S[f;b], \tag{6.29}$$

where Q_S is evaluated by an n-point Gauss–Laguerre quadrature rule as specified in (6.28)*. Then the quadrature error behaves asymptotically as $O(\omega^{-2n-1})$.*

Proof. A formula for the error of the n-point Gauss–Laguerre quadrature rule applied to the integral $\int_0^\infty f(x)e^{-x}\mathrm{d}x$ is given by [69]

$$E = \frac{(n!)^2}{(2n)!}f^{(2n)}(\zeta), \quad \zeta \in [0, \infty).$$

Table 6.1: Absolute error of the approximation of $I[f]$ by $Q_S[f;a]-Q_S[f;b]$ with n quadrature points for the functions $f(x)=1/(1+x)$ and $g(x)=x$ on $[0,1]$. The last row shows the value of $\log_2(e_{40}/e_{80})$: this value should approximate $2n+1$.

$\omega \setminus n$	1	2	3	4	5
10	$1.0E-3$	$3.1E-5$	$1.9E-6$	$1.7E-7$	$2.1E-8$
20	$1.2E-4$	$1.1E-6$	$2.3E-8$	$7.5E-10$	$3.2E-11$
40	$1.7E-5$	$3.9E-8$	$2.1E-10$	$2.0E-12$	$2.8E-14$
80	$2.0E-6$	$1.2E-9$	$1.7E-12$	$4.2E-15$	$1.6E-17$
rate	3.1	5.0	6.9	8.9	10.8

Using this formula, one can derive an expression for the error $E := F(a) - Q_S[f;a]$:

$$\begin{aligned} E &= \frac{e^{i\omega g(a)}}{\omega}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q/\omega))h_a'(q/\omega))}{\mathrm{d}q^{2n}}\right|_{q=\zeta} \\ &= \frac{e^{i\omega g(a)}}{\omega^{2n+1}}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q))h_a'(q))}{\mathrm{d}q^{2n}}\right|_{q=\zeta/\omega} \end{aligned} \tag{6.30}$$

with $\zeta \in \mathbb{C}$. The error behaves asymptotically as $O(\omega^{-2n-1})$. The absolute error for the approximation to (6.1) is composed of 2 contributions of the form (6.30), and, hence, has the same high order of convergence. □

Remark 6.3.8. *The number of function evaluations required to evaluate $Q_{NSD}[f]$ is $2n$, and the asymptotic order of the method is also $2n$. The Filon type method using derivatives of f up to order $n-1$ at the points a and b requires $2n$ function values and derivatives in* (6.10)*, and has asymptotic order n. For the same amount of* data*, the asymptotic order of the numerical steepest descent method is twice as large as the order of the Filon-type method. A similar observation holds when comparing with the Levin-type method discussed in §6.2.3.*

Example 6.3.9. *We end this section with a numerical example to illustrate the sharpness of our convergence result. The absolute error for different values of ω and of n is given in Table 6.1 for the functions $g(x)=x$ and $f(x)=1/(1+x)$ on $[0,1]$. The parameterisation for Γ_x is given by $h_x(p) = x+ip$. The behaviour as a function of ω follows the theory until machine precision is reached. The relative error scales only slightly worse, since $I[f]=O(\omega^{-1})$.*

One should note that decomposition (6.22) is exact for all positive values of the parameter $\omega > (m+1)\omega_0 > 0$. The conditions from Theorem 6.3.3

yield the minimal frequency parameter $(m+1)\omega_0$. The method itself is therefore not asymptotic, only the convergence estimate is. Table 6.1 shows an absolute error of $2.1E-8$ (relative error $1.4E-7$) for $\omega = 10$ with a number of quadrature points as small as $n = 5$. The corresponding integral is not highly oscillatory at all. In order to achieve the same absolute error with standard Gaussian quadrature on $[0,1]$, we had to choose a rule with 10 points. Considering the fact that we evaluate both $Q_S[f;a]$ and $Q_S[f;b]$ with $n = 5$ points, the amount of work is the same. Thus, even at relatively low frequencies, our approach is competitive with conventional quadrature on the real axis. For higher frequencies, obviously, the new approach may be many orders of magnitude faster.

6.3.4 The case of stationary points

6.3.4.1 A new decomposition for the oscillatory integral

At a stationary point ξ, the derivative of g vanishes and the integrand $f(x)e^{i\omega g(x)}$ does not oscillate, at least locally. The contribution of the integrand and its derivatives at ξ can therefore not be neglected. The Theorems of §6.3.3 do not apply, because the inverse of g does not exist uniquely due to the branch point at ξ.

In order to illustrate the problem, consider the following situation. Assume that the equation $g'(x) = 0$ has one solution ξ and $\xi \in [a,b]$. Now define the restrictions

$$g_1 := g|_{[a,\xi]} \quad \text{and} \quad g_2 := g|_{[\xi,b]} \tag{6.31}$$

of g to the intervals $[a,\xi]$ and $[\xi,b]$ respectively. Then, the unique inverse of g on $[a,b]$ does not exist, but a single-valued branch g_1^{-1} can be found that satisfies $g_1^{-1}(g_1(x)) = x$, $x \in [a,\xi]$. This branch is analytic everywhere except at the point ξ, and along a branch cut that can be chosen arbitrarily but that always originates at ξ. Similarly, a single-valued branch g_2^{-1} exists that satisfies $g_2^{-1}(g_2(x)) = x$, $x \in [\xi,b]$. Both branches satisfy $g(g_i^{-1}(z)) = z$, $i = 1,2$, in their domain of analyticity. The integrand is small in the region S_1 with points of the form $g_1^{-1}(c+id)$, $d \geq d_0$, or in the region S_2 with points of the form $g_2^{-1}(c+id)$, $d \geq d_0$. It is easy to see that S_1 and S_2 are not connected: applying g on both sides of the equality $g_1^{-1}(y) = g_2^{-1}(z)$ leads to $y = z$, which is only possible if $z = \xi \notin S_1, S_2$. The path (6.27) that solves $g(h_x(p)) = g(x) + ip$, as suggested in §6.3.3, leads to a path in S_1 for a, and to a path in S_2 for b.

The solution is therefore to split the integration interval into the two subintervals $[a,\xi]$ and $[\xi,b]$. This procedure can be repeated for any number of stationary points. The analogues of Theorems 6.3.2 and 6.3.3 can be stated as follows.

Theorem 6.3.10. *Assume that the functions f and g are analytic in a bounded and open complex neighbourhood D of $[a,b]$. If the equation $g'(x) = 0$ has only one solution ξ in D and $\xi \in (a,b)$, then there exist functions $F_j(x)$, $j = 1,2$, such that*

$$\int_s^t f(z)e^{i\omega g(z)}\,\mathrm{d}z = F_1(s) - F_1(\xi) + F_2(\xi) - F_2(t) + O(e^{-\omega d_0}), \quad d_0 > 0,$$

for $s \in [a,\xi]$ and $t \in [\xi,b]$, where $F_j(x)$ is of the form

$$F_j(x) := \int_{\Gamma_{x,j}} f(z)e^{i\omega g(z)}\,\mathrm{d}z \tag{6.32}$$

with $\Gamma_{x,j}$ a path that starts at x.

Proof. Define $g_2(x)$ as in (6.31). A decomposition for $\int_\xi^t f(x)e^{i\omega g_2(x)}\,\mathrm{d}x$ can be found using the proof of Theorem 6.3.2 with two modifications. First, the equation $g(z) = g(x) + id_0$ now has at least two solutions locally around $x = \xi$. We choose the solution that corresponds to the single-valued branch g_2^{-1} of the inverse of g that satisfies $g_2^{-1}(g(x)) = x$, $x \in [\xi,b]$. The branch cut can always be chosen such that it does not prevent from applying Cauchy's Theorem. Secondly, the set S in the proof is now defined as $S := \{z : \Im g(z) \geq d_0 \,\text{and}\, g_2^{-1}(g(z)) = z\} \cap D$, i.e., the set is restricted to one connected part of D where the integrand is small, as opposed to the set of all points where the integrand is small. The latter set would not be connected in this case. With these modifications, the proof shows the existence of F_2 such that

$$\int_\xi^t f(z)e^{i\omega g_2(z)}\,\mathrm{d}z = F_2(\xi) - F_2(t) + O(e^{-\omega d_0}).$$

The same reasoning can be applied in order to find a decomposition on the interval $[a,\xi]$. This leads to the result. □

The next Theorem is the limit case of Theorem 6.3.10 where the error vanishes. The notation g_1^{-1} denotes a branch of the multi-valued inverse of g that satisfies $g_1^{-1}(g_1(x)) = x$, $x \in [a,\xi]$. The notation g_2^{-1} is similar.

Theorem 6.3.11. *Assume that the functions f and g are analytic in a simply connected and sufficiently (infinitely) large complex region D containing the interval $[a,b]$. Assume further that the equation $g'(x) = 0$ has only one solution ξ in D and $\xi \in (a,b)$. Define g_1 and g_2 as in (6.31). If the following conditions hold:*

$$\exists m \in \mathbb{N} : |f(z)| = O(|z|^m),$$
$$\exists \omega_0 \in \mathbb{R} : |g_1^{-1}(z)| = O(e^{\omega_0|z|}) \text{ and } |g_2^{-1}(z)| = O(e^{\omega_0|z|}), \quad |z| \to \infty,$$

then there exist functions $F_j(x)$, $j = 1, 2$, *of the form* (6.32) *such that*

$$\int_s^t f(z)e^{i\omega g(z)}\,\mathrm{d}z = F_1(s)-F_1(\xi)+F_2(\xi)-F_2(t), \quad \forall\omega > (m+1)\omega_0, \tag{6.33}$$

for $s \in [a, \xi]$ *and* $t \in [\xi, b]$. *A parameterisation* $h_{\xi,j}(p)$, $p \in [0, \infty)$, *for* $\Gamma_{x,j}$ *exists such that the integrand of* (6.32) *is* $O(e^{-\omega p})$.

Theorems 6.3.10 and 6.3.11 are easily extended to the case where $\xi = a$ (or $\xi = b$), by discarding the two terms $F_1(a) - F_1(\xi)$ (or $F_2(\xi) - F_2(b)$).

Example 6.3.12. *We consider the function* $g(x) = (x - 1/2)^2$, *with a stationary point at* $\xi = 1/2$. *The inverse of* g, *i.e.,* $g^{-1}(y) = 1/2 \pm \sqrt{y}$, *is a two-valued function. One branch is valid on the interval* $[0, \xi]$, *the other on* $[\xi, 1]$. *The paths suggested by* (6.27) *on* $[0, \xi]$ *that originate at the endpoints* 0 *and* ξ *respectively are parameterised by*

$$h_{0,1}(p) = 1/2 - \sqrt{1/4 + ip} \quad \textit{and} \quad h_{\xi,1}(p) = 1/2 - \sqrt{ip}$$

The paths on $[\xi, 1]$ *for the points* $1/2$ *and* 1 *are parameterised by*

$$h_{\xi,2}(p) = 1/2 + \sqrt{ip} \quad \textit{and} \quad h_{1,2}(p) = 1/2 + \sqrt{1/4 + ip}$$

These paths correspond to the two inverse functions. We have found the decomposition $I = F_1(a) - F_1(\xi) + F_2(\xi) - F_2(b)$.

Note that the paths $h_{\xi,1}$ and $h_{\xi,2}$ that originate in the point ξ introduce a numerical problem. Their derivatives, that appear in the integrand of the line integral, behave like $1/\sqrt{p}$, $p \to 0$ at ξ. This weak singularity is integrable, but prevents convergence of the Gauss–Laguerre quadrature rules. We will require a new method to evaluate $F_j(\xi)$.

6.3.4.2 The evaluation of $F_j(x)$ by generalised Gauss–Laguerre quadrature

The previous example showed a numerical problem for the evaluation of $F_j(x)$ by numerical quadrature: the integrand of $F_j(\xi)$ along the path suggested by (6.27) becomes weakly singular at the stationary point ξ. A similar singularity occurs if higher order derivatives of $g(\xi)$ also vanish. Assume that $g^{(k)}(\xi) = 0$, $k = 1, \ldots, r$. The Taylor expansion of g is then

$$g(x) = g(\xi) + 0 + \ldots + 0 + g^{(r+l)}(\xi)\frac{(x-\xi)^{r+1}}{(r+1)!} + O((x-\xi)^{l+2}).$$

The path $h_{\xi,j}(p)$ solves the equation $g(h_{\xi,j}(p)) = g(\xi) + ip$. Its behaviour at $p = 0$ is

$$h_{\xi,j}(p) \sim \xi + \sqrt[r+1]{\frac{(r+1)!\,p}{g^{(r+1)}(\xi)}i}. \tag{6.34}$$

The derivative has a singularity of the form $p^{\frac{1}{r+1}-1}$, $p \to 0$.

Fortunately, these types of singularities can be handled efficiently by generalised Gauss–Laguerre quadrature. Generalised Laguerre polynomials are orthogonal with respect to the weight function $x^\alpha e^{-x}$, $\alpha > -1$ [69]. Function $F_j(\xi)$ with optimal path $h_{\xi,j}(p)$ is given by

$$\begin{aligned} F_j(\xi) &= \int_0^\infty f(h_{\xi,j}(p)) e^{i\omega(g(\xi)+ip)} h'_{\xi,j}(p)\,\mathrm{d}p \\ &= \frac{e^{i\omega g(\xi)}}{\omega} \int_0^\infty f(h_{\xi,j}(q/\omega)) h'_{\xi,j}(q/\omega) e^{-q}\,\mathrm{d}q. \end{aligned} \tag{6.35}$$

Generalised Gauss–Laguerre quadrature will be used with n points x_i and weights w_i that depend on the value of $\alpha = 1/(r+1) - 1 = -r/(r+1)$. The function $F_j(x)$ is then approximated by

$$Q^\alpha_{S_j}[f;\xi] := \frac{e^{i\omega g(\xi)}}{\omega} \sum_{i=1}^n w_i\, f(h_{\xi,j}(x_i/\omega))\, h'_{\xi,j}(x_i/\omega)\, x_i^{-\alpha}. \tag{6.36}$$

This expression is similar to (6.28) but includes the factor $x_i^{-\alpha}$ to regularise the singularity.

Theorem 6.3.13. *Assume functions f and g satisfy the conditions of Theorem 6.3.11. Assume that $g^{(k)}(\xi) = 0$, $k = 1, \ldots, r$ and $g^{(r+1)}(\xi) \neq 0$. Let the function $F_j(\xi)$ be approximated by the quadrature formula*

$$F_j(\xi) \approx Q^\alpha_{S_j}[f;\xi]$$

with $\alpha = -r/(r+1)$. Then the quadrature error behaves asymptotically as $O(\omega^{-2n-1/(r+1)})$.

Proof. The error formula for an n-point generalised Gauss–Laguerre quadrature rule is

$$\frac{n!\Gamma(n+\alpha+1)}{(2n)!} f^{(2n)}(\zeta), \quad 0 < \zeta < \infty. \tag{6.37}$$

We can repeat the arguments of the proof of Theorem 6.3.7. An expression for the error $e := F_j(\xi) - Q^\alpha_{S_j}[f;\xi]$ can be derived by using (6.37). This leads to

$$\begin{aligned} e &= \frac{e^{i\omega g(\xi)}}{\omega} \frac{n!\Gamma(n+\alpha+1)}{(2n)!} \left.\frac{\mathrm{d}^{2n}(f(h_{\xi,j}(q/\omega)) h'_{\xi,j}(q/\omega) q^{-\alpha})}{\mathrm{d}q^{2n}}\right|_{q=\zeta} \\ &= \frac{e^{i\omega g(\xi)}}{\omega^{2n+1}} \frac{n!\Gamma(n+\alpha+1)}{(2n)!} \left.\frac{\mathrm{d}^{2n}(f(h_{\xi,j}(q)) h'_{\xi,j}(q) (\omega q)^{-\alpha})}{\mathrm{d}q^{2n}}\right|_{q=\zeta/\omega} \end{aligned}$$

with $\zeta \in \mathbb{C}$. Hence, the error is asymptotically of the order $O(\omega^{-2n-1-\alpha})$. $\square$

Remark 6.3.14. *Generalised Gauss–Laguerre quadrature converges rapidly only if the function* $v(x)$ *in an integrand of the form* $v(x)x^\alpha e^{-x}$ *has polynomial behaviour. Depending on* f*, the function* $f(h_{\xi,j}(p))$ *may not resemble a polynomial very well, due to the root in* (6.34) *for small* p*. An alternative to generalised Gauss–Laguerre quadrature with* $\alpha = -1/2$ *is to remove the singularity by the transformation* $u = \sqrt{p}$ *or* $p = u^2$*. The same transformation also removes the square root behaviour of* $h_{\xi,j}(p)$*. The integrand after the transformation decays like* e^{-u^2}*. In that case, variants of the classical Hermite polynomials that are orthogonal w.r.t.* e^{-u^2} *on the half-range interval* $[0, \infty)$ *can be used, with corresponding Gaussian quadrature rules as constructed by Gautschi [87]. A similar convergence analysis yields the order* $O(\omega^{-n-1/(r+1)})$ *in this case.*

We can now characterise the approximation of (6.1) in the presence of several stationary points.

Theorem 6.3.15. *Assume that* f *and* g *are analytic in a sufficiently large region* $D \subset \mathbb{C}$*, and that the equation* $g'(x) = 0$ *has* l *solutions* $\xi_i \in (a, b)$*. Define* $r_i := (\min_{k>1} g^{(k)}(\xi_i) \neq 0) - 1$ *and* $r := \max_i r_i$*. If the conditions of Theorem 6.3.11 are satisfied on each subinterval* $[\xi_i, \xi_{i+1}]$*, and on* $[a, \xi_1]$ *and* $[\xi_r, b]$*, then* (6.1) *can be approximated by*

$$\begin{aligned} I[f] \approx Q_{NSD}[f] := Q_{S_0}[f;a] - Q_{S_0}^{\alpha_1}[f;\xi_1] + \sum_{i=1}^{l-1} \big(Q_{S_i}^{\alpha_i}[f;\xi_i] \\ -Q_{S_i}^{\alpha_{i+1}}[f;\xi_{i+1}]\big) + Q_{S_l}^{\alpha_l}[f;\xi_l] - Q_{S_l}[f;b] \end{aligned} \tag{6.38}$$

with $\alpha_i = -r_i/(r_i+1)$*, with a quadrature error of the order* $O(\omega^{-2n-1/(r+1)})$*.*

Proof. This follows from a repeated application of the decomposition given by Theorem 6.3.11, and from the approximation of each term $F_i(x)$ by $Q_{S_i}^{\alpha_i}[f;x]$ as in Theorem 6.3.13. □

Theorem 6.3.15 can easily be extended to the case where $g'(a) = 0$ or $g'(b) = 0$. If, e.g., $g'(a) = 0$, we can set $\xi_1 = a$ and use the general decomposition (6.38) with the first two terms left out.

Example 6.3.16. *We return to the Example 6.3.12 of this section in order to illustrate the convergence results. The approximation of* (6.1) *for the function* $g(x) = (x - 1/2)^2$ *is given by*

$$\begin{aligned} I[f] \approx Q_{NSD}[f] =& Q_{S_1}[f;0] - Q_{S_1}^{-1/2}[f;1/2] \\ &+ Q_{S_2}^{-1/2}[f;1/2] - Q_{S_2}[f;1]. \end{aligned}$$

Theorem 6.3.15 predicts an error of the order $O(\omega^{-2n-1/2})$*. The sharpness of this estimate can be verified by the results in Table 6.2.*

Table 6.2: Absolute error of the approximation of $I[f]$ by $Q_{NSD}[f]$ using (generalised) Gauss-Laguerre quadrature with $f(x) = 1/(1+x)$ and $g(x) = (x - 1/2)^2$ on $[0, 1]$. The last row shows the value of $\log_2(e_{80}/e_{160})$: this value should approximate $2n + 1/2$.

$\omega \setminus n$	1	2	3	4	5
10	$4.7E-3$	$7.1E-4$	$1.7E-4$	$4.9E-5$	$1.7E-5$
20	$7.8E-4$	$5.6E-5$	$7.2E-6$	$1.3E-6$	$2.7E-7$
40	$1.2E-4$	$2.8E-6$	$1.5E-7$	$1.2E-8$	$1.3E-9$
80	$1.6E-5$	$1.0E-7$	$1.7E-9$	$5.0E-11$	$2.1E-12$
160	$2.3E-6$	$3.4E-9$	$1.6E-11$	$1.3E-13$	$1.6E-15$
rate	2.8	4.9	6.8	8.6	10.4

Remark 6.3.17. *There exists a useful technique that can be used for the evaluation of $F(x)$ when x is* near *a stationary point ξ. In that case, the integral $F(x)$ is nearly singular, and it may be expensive to evaluate the integral numerically. The proposed technique is as follows: instead of $F(x)$, one evaluates $F(\xi)$, which can be done cheaply. Following Remark 6.3.6, we have $F(x) = -F_L(x)e^{i\omega g(x)}$. The derivatives of $F_L(x)$ of all order can be computed based on expression* (6.12)*, requiring only the value $F_L(x)$ itself. Hence, the Taylor series of $F(x)$ can be constructed around ξ, requiring only the value of $F(\xi)$. The desired value $F(x)$ is obtained from the Taylor approximation.*

6.3.4.3 The case of complex stationary points

So far, we have required the stationary point $\xi \in [a, b]$ to be real. But even for functions g that are real-valued on the real axis, the equation $g'(x) = 0$ may have complex solutions. The value of $g'(x)$ on $[a, b]$ can become very small, if a complex stationary point ξ lies close to the real axis. We may therefore expect that such a point contributes to the value of the integral (6.1). Here, we will not pursue the extension of the theory to the case of complex stationary points in any detail. Instead, we will restrict ourselves to a number of remarks that address some of the relevant issues.

A first observation is that Theorem 6.3.10 can still be applied, if the region D is chosen small enough such that it does not contain ξ. This means that the contribution of ξ to the value of I, if any, decays exponentially fast as ω increases. Still, for small values of ω, the error may become prohibitively large if ξ lies close to the real axis.

In order to resolve this problem, one must first know which stationary points can contribute to the error of the approximations of §6.3.4. In general, the question can be answered by inspecting the integration paths. A stationary point contributes if it lies in the interior of the domain bounded by the integration interval on the real axis and the complex integration path (including the limiting connecting part at infinity). In order to obtain an exact decomposition, the integration path should be changed to pass through ξ explicitly. Specifically, the decomposition should include two additional terms for ξ of the form (6.35).

As a final remark, we note that the integral of the form (6.35) has a factor $e^{i\omega g(\xi)}$ with $g(\xi) = c + id$ complex. If $d > 0$ then the contribution decays exponentially as $e^{-\omega d}$. We know from Theorem 6.3.10 that the error introduced by discarding complex stationary points should decay exponentially. Hence, complex stationary points for which $d \leq 0$ cannot contribute to the value of I.

6.3.5 The case where the oscillator is not easily invertible

Theorems 6.3.3 and 6.3.11 continue to hold for paths different from the one implicitly defined by (6.27). The value of $F(a)$ does not depend on the path taken, and does not even depend on the limiting endpoint of the path, as long as the imaginary part of $g(x)$ grows infinitely large. We have merely suggested (6.27) as it yields a non-oscillatory integrand with exponential decay, suitable for Gauss-Laguerre quadrature. Other integration techniques may be applied for other paths with different numerical properties. We will not explore these possibilities in depth here.

We restrict the discussion to an approach that is useful when the inverse function of g is known to exist, but when the suggested path is not easily obtained by analytical means. As ω increases, we see from (6.28) that $Q_S[f;a]$ requires function values in a complex region around a of diminishing size. Therefore, it is reasonable to assume that approximating the path defined by (6.27) locally around $x = a$ is acceptable. Use of the first order Taylor approximation $g(x) \approx g(a) + g'(a)(x-a)$ to replace the left hand side of (6.27) leads to the path

$$h_a(p) = a + \frac{ip}{g'(a)}. \tag{6.39}$$

The second order Taylor approximation leads to the path

$$h_a(p) = a - \frac{g'(a) - \sqrt{g'(a)^2 + 2ipg^{(2)}(a)}}{g^{(2)}(a)}.$$

In the case of stationary points the path can be approximated by using (6.34),

$$h_{\xi,i}(p) = \xi + \sqrt[r+1]{\frac{(r+1)!\,p}{g^{(r+1)}(\xi)}}\,i.$$

The general expression for the integral along the approximate path is given by

$$F(a) = \int_0^\infty f(h_a(p))e^{i\omega g(h_a(p))}h_a'(p)\,\mathrm{d}p.$$

Computing $F(a)$ by Gauss-Laguerre quadrature yields a numerical approximation with an error given by

$$\begin{aligned} E &= \omega^{-1}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q/\omega))e^{i\omega g(h_a(q/\omega))}h_a'(q/\omega)e^{q})}{\mathrm{d}q^{2n}}\right|_{q=\zeta} \\ &= \omega^{-2n-1}\frac{(n!)^2}{(2n)!}\left.\frac{\mathrm{d}^{2n}(f(h_a(q))h_a'(q)e^{i\omega g(h_a(q))}e^{\omega q})}{\mathrm{d}q^{2n}}\right|_{q=\zeta/\omega}. \end{aligned}$$

The order of convergence is not necessarily $O(\omega^{-2n-1})$ in this case because the derivative still depends on ω. However, the function $e^{i\omega g(h_a(q))}$ is a good approximation to $e^{i\omega g(a)}e^{-\omega q}$ and we can expect the quadrature to converge. This will be illustrated further on.

The results can be improved to preserve the original convergence rate of $O(\omega^{-2n-1})$ at the cost of a little extra work to determine the optimal path. The optimal path depends only on $g(x)$ and on the interval $[a,b]$, and can therefore be reused for different functions f. The extra computations have to be done once for each combination of $g(x)$ and $[a,b]$.

The Taylor approximation of the path can be used to generate suitable starting values for a Newton-Raphson iteration, applied to find the root x of the equation

$$g(x) - g(a) - ip = 0. \tag{6.40}$$

For the set of n (fixed) values for p that are required by the quadrature rule, the iteration yields the points $x = h_a(p)$ on the path. The values of $h_a'(p)$, i.e., $\frac{\mathrm{d}x}{\mathrm{d}p}$, are found by taking the derivative of (6.40) with respect to p,

$$g'(x)\frac{\mathrm{d}x}{\mathrm{d}p} = i. \tag{6.41}$$

With the Newton-Raphson method, the points on the optimal path and the derivatives at these points can be found to high precision. Since the Taylor approximation is already a good approximation for large ω, the required number of iterations is small.

Table 6.3: Absolute error of approximation of $F(a) - F(b)$ by Gauss-Laguerre quadrature with $f(x) = 1/(1+x)$ and $g(x) = (x^2+x+1)^{1/3}$ on $[0,1]$ and second order Taylor approximation of the optimal path. The last row shows the value of $\log_2(e_{160}/e_{320})$.

$\omega \setminus n$	1	2	3	4	5
20	$1.4E-2$	$2.7E-3$	$7.4E-4$	$2.4E-4$	$8.9E-5$
40	$2.5E-3$	$2.6E-4$	$4.6E-5$	$1.0E-5$	$2.5E-6$
80	$3.8E-4$	$1.8E-5$	$1.7E-6$	$2.0E-7$	$2.9E-8$
160	$5.2E-5$	$1.1E-6$	$4.0E-8$	$2.1E-9$	$1.5E-10$
320	$6.7E-6$	$6.8E-8$	$7.7E-10$	$1.6E-11$	$4.4E-13$
rate	3.0	4.0	5.7	7.0	8.4

Table 6.4: The same example as in Table 6.3, but using Newton-Raphson iterations to compute the optimal path. The number of iterations per quadrature point varied between 1 and 4. The last row shows the value of $\log_2(e_{320}/e_{640})$: this value should approximate $2n+1$.

$\omega \setminus n$	1	2	3	4	5
20	$1.1E-2$	$2.4E-3$	$7.4E-4$	$2.5E-4$	$7.5E-5$
40	$2.1E-3$	$2.4E-4$	$4.4E-5$	$1.0E-5$	$2.4E-6$
80	$3.3E-4$	$1.5E-5$	$1.2E-6$	$1.5E-7$	$2.3E-8$
160	$4.5E-5$	$6.1E-7$	$1.8E-8$	$8.7E-10$	$6.2E-11$
320	$5.9E-6$	$2.1E-8$	$1.8E-10$	$2.7E-12$	$6.2E-14$
640	$7.2E-7$	$6.7E-10$	$1.5E-12$	$6.3E-15$	$4.3E-17$
rate	3.0	5.0	6.9	8.8	10.5

Example 6.3.18. *We consider the second order Taylor approximation of the path for $f(x) = 1/(1+x)$ and $g(x) = (x^2+x+1)^{1/3}$. The absolute error is shown in Table 6.3. Use of the Newton-Raphson iteration for the same example yields an error of order $O(\omega^{-2n-1})$. This is shown in Table 6.4. The number of iterations per quadrature point varied between* 1 *and* 4*.*

6.3.6 Filon-type methods for a non-analytic function f

If $f(x)$ is not analytic in a complex region surrounding $[a,b]$, then the method presented thus far will not work. If $f(x)$ is piecewise analytic (e.g., piecewise polynomial), the integration can be split into the integrals corresponding to the analytic parts of f. More generally however, we need to resort to another approach. For a suitable analytic function $\tilde{f}$ that approximates f, we can expect the integral

$$I[\tilde{f}] = \int_a^b \tilde{f}(x) e^{i\omega g(x)} \,\mathrm{d}x$$

to approximate the value of I. This leads to *Filon-type methods* that were discussed already in §6.2.2. Since polynomials are analytic, we can use the Hermite interpolating polynomials satisfying (6.9) as basis functions for the analytic approximation $\tilde{f}$ of f. The weights $w_{l,j} = I[\psi_{l,j}]$ of the Filon-type method (6.10) can be evaluated using the numerical steepest descent method. (Note that the method also enables the computation of the moments in the *asymptotic method* (6.7).)

The asymptotic order of the Filon-type method using Hermite interpolation was given by (6.11). It is lower than the asymptotic order of the numerical steepest descent method using the same amount of data (see Remark 6.3.8). Unfortunately, the asymptotic order of the Filon-type method can not be improved using steepest descent for approximations of f. We can improve on the Filon-type method however in a different way. Thanks to the decomposition of (6.1) as $I[f] = F(a) - F(b)$, it is possible to use different approximations around a and b, and, hence, to approximate $F(a)$ and $F(b)$ independently. Since $F(a)$ only depends on the behaviour of f around a, the approximating Hermite polynomial can have much lower degree. In the theorem below, we show that we can obtain the same asymptotic order s of the Filon-type method with two independently constructed polynomials of degree $s-1$ instead of with one polynomial of degree $2s-1$. For notational convenience, we define the integral $S[f;x]$ by

$$S[f;x] := F(x), \tag{6.42}$$

as the line integral along the optimal path $h_x(p)$ for the oscillator g.

Theorem 6.3.19. *Assume that f is a smooth function, and g is analytic. Let $f_a(x)$ and $f_b(x)$ be the Hermite interpolating polynomials of degree $s-1$ that satisfy*

$$f_a^{(k)}(a) = f^{(k)}(a) \quad \textit{and} \quad f_b^{(k)}(b) = f^{(k)}(b), \quad k = 0, \ldots, s-1.$$

Then the approximation $I[f] \approx S[f_a;a] - S[f_b;b]$ has asymptotic order s.

Proof. First we consider the approximation with the Hermite interpolating polynomial $\tilde{f}(x)$ of degree $2s-1$ that satisfies $\tilde{f}^{(k)}(a) = f^{(k)}(a)$ and $\tilde{f}^{(k)}(b) = f^{(k)}(b)$, $k = 0, \ldots, s-1$. Since $\tilde{f}$ is analytic, it can be used to approximate (6.1) as $I[f] \approx I[\tilde{f}] = S[\tilde{f};a] - S[\tilde{f};b]$. This approximation has an asymptotic error of $O(\omega^{-s-1})$ by [129, Theorem 2.3].

Now consider the approximation of $S[\tilde{f};a]$ by $S[f_a;a]$. Since $\tilde{f}(x)$ is a polynomial, we can write $S[\tilde{f};a]$ as

$$S[\tilde{f};a] = \sum_{k=0}^{2s-1} \tilde{f}^{(k)}(a) \frac{S[(x-a)^k;a]}{k!} := \sum_{k=0}^{2s-1} \tilde{f}^{(k)}(a) \frac{\mu_k(a)}{k!}.$$

The moments $\mu_k(a) = S[(x-a)^k;a]$ are given explicitly by

$$\begin{aligned} \mu_k(a) &= \int_0^\infty (h_a(p)-a)^k e^{i\omega g(h_a(p))} h_a'(p)\,\mathrm{d}p \\ &= \int_0^\infty \frac{e^{i\omega g(a)}}{\omega} (h_a(q/\omega)-a)^k e^{-q} h_a'(q/\omega)\,\mathrm{d}q. \end{aligned} \tag{6.43}$$

Although q goes to infinity, the behaviour for small q/ω dominates due to the factor e^{-q} (this follows from Watson's Lemma [3, 200]). Since $(h_a(q/\omega) - a) \sim \omega^{-1}$, we see that $\mu_k(a) \sim \omega^{-k-1}$. For $S[f_a;a]$, we have

$$S[f_a;a] = \sum_{k=0}^{s-1} f_a^{(k)}(a) \frac{\mu_k(a)}{k!}. \tag{6.44}$$

The first discarded moment, $\mu_s(a)$, has asymptotic size $O(\omega^{-s-1})$. The approximation of $S[\tilde{f};b]$ by $S[f_b;b]$ has an error of the same order. This concludes the proof. □

There are two ways to proceed: either $f_a(x)$ can be evaluated explicitly in the quadrature evaluation of $S[f_a;a]$, or the moments (6.43) can be precomputed with the previous techniques and used in the summation (6.44). The latter leads to a *localised Filon-type* quadrature rule for $I[f]$, using function values of f at a and b,

$$I[f] \approx Q_{LF}[f] := \sum_{j=0}^{s-1} w_{1,j} f^{(j)}(a) + \sum_{j=0}^{s-1} w_{2,j} f^{(j)}(b), \tag{6.45}$$

with the weights given by

$$w_{1,j} = S\left[\frac{(x-a)^j}{j!};a\right], \quad \text{and} \quad w_{2,j} = -S\left[\frac{(x-b)^j}{j!};b\right].$$

The quadrature rule has asymptotic order s, like the regular Filon-type method. For a fixed frequency, the localised Filon-type method is exact for polynomials up to degree $s-1$, while the regular Filon-type method is exact for polynomials up to degree $2s-1$. Hence, the simplified construction comes at a cost; the order of accuracy as a function of ω is the same, but one can expect the coefficient to be much larger.

We can generalise the result to include stationary points. The same reasoning applies, but we need to interpolate more derivatives in order to achieve a similar convergence rate. The number of derivatives depends on the order r of the stationary point. We use the notation $S_j[f;\xi]$ to denote the line integral corresponding to the path $h_{\xi,j}(p)$.

Theorem 6.3.20. *Assume that* g *is analytic and that* $g^{(k)}(\xi) = 0$, $k = 1, \dots, r$, *and* $g^{(r+1)}(\xi) \neq 0$. *Let* f *be sufficiently smooth, and let* $f_\xi(x)$ *be the Hermite interpolating polynomial of degree* $s(r+1)-1$ *that satisfies*

$$f_\xi^{(k)}(\xi) = f^{(k)}(\xi), \quad j = 0, \dots, s(r+1)-1.$$

Then the sequence $S_j[f_\xi;\xi]$ *converges for increasing values of* s *to a limit with an error of order* $O(\omega^{-s-1/(r+1)})$.

Proof. The proof follows essentially the same lines as the proof of Theorem 6.3.19. Define the moments $\mu_{j,k}(\xi) = S_j[(x-\xi)^k;\xi]$, given by

$$\mu_{j,k}(\xi) = \int_0^\infty \frac{e^{i\omega g(\xi)}}{\omega} (h_{\xi,j}(q/\omega) - \xi)^k e^{-q} h'_{\xi,j}(q/\omega)\, \mathrm{d}q. \tag{6.46}$$

The derivative of the parameterisation $h_{\xi,j}$ in the integrand has an integrable singularity of the form $(q/\omega)^{-r/(r+1)}$ at the stationary point ξ, and leads to a factor $\omega^{r/(r+1)}$. By (6.34) we have $(h_{\xi,j}(q/\omega) - \xi) \sim \omega^{-1/(r+1)}$. This makes $\mu_{k,j}(\xi) \sim \omega^{r/(r+1)-k/(r+1)-1} = \omega^{(-k-1)/(r+1)}$. The first discarded moment $\mu_{k,j}(\xi)$ in the sum $S_j[f_\xi;\xi]$ of the form (6.44) has the index $k = s(r+1)$, which leads to the result. □

Assume there is one stationary point $\xi \in (a,b)$ of order r. Then we can extend the definition of quadrature rule (6.45) to

$$I[f] \approx Q_{LF}[f] = \sum_{j=0}^{s-1} w_{1,j} f^{(j)}(a) + \sum_{j=0}^{s(r+1)-1} w_{2,j} f^{(j)}(\xi) + \sum_{j=0}^{s-1} w_{3,j} f^{(j)}(b),$$

with the weights given by

$$w_{1,j} = S_1\left[\frac{(x-a)^j}{j!};a\right], \tag{6.47}$$

$$w_{2,j} = -S_1\left[\frac{(x-\xi)^j}{j!};\xi\right] + S_2\left[\frac{(x-\xi)^j}{j!};\xi\right], \tag{6.48}$$

$$w_{3,j} = -S_2\left[\frac{(x-b)^j}{j!};b\right]. \tag{6.49}$$

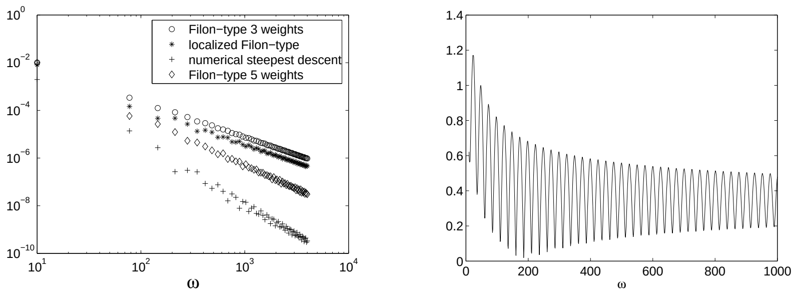
The rule has an absolute error of order $O(\omega^{-s-1/(r+1)})$, and a relative error of order $O(\omega^{-s})$.

Table 6.5: Absolute error of the approximation of $I[f]$ for $f(x) = 1/(1+x)$ and $g(x) = (x-1/3)^2$ on $[0,1]$. We approximate f by interpolating m derivatives. The last row shows the value of $\log_2(e_{1280}/e_{2560})$: this value should approximate $(m+2)/2$ for odd m, and $(m+3)/2$ for even m.

$\omega \setminus m$	0	1	2	3	4
160	$1.0E-4$	$1.8E-4$	$9.5E-7$	$9.7E-7$	$8.6E-9$
320	$6.5E-5$	$6.5E-5$	$1.7E-7$	$1.7E-7$	$7.6E-10$
640	$2.8E-5$	$2.3E-5$	$3.1E-8$	$3.0E-8$	$6.7E-11$
1280	$8.1E-6$	$8.2E-6$	$5.4E-9$	$5.4E-9$	$5.9E-12$
2560	$3.2E-6$	$2.9E-6$	$9.5E-10$	$9.5E-10$	$5.2E-13$
rate	1.4	1.5	2.5	2.5	3.5

Example 6.3.21. *We consider the functions* $f(x) = 1/(1+x)$ *and* $g(x) = (x-1/3)^2$ *on* $[0,1]$. *Since* f *is analytic, we could use the previous techniques. However, here we will only use the values of the first few derivatives of* f *at* 0 *and* 1 *and at the stationary point* $\xi = 1/3$. *The results are shown in Table 6.5 for varying degrees of interpolation. The convergence rate is limited to the convergence rate at the stationary point. According to Theorem 6.3.20, in order to obtain an error of order* $O(\omega^{-s-1/(r+1)})$, *we need to interpolate up to the derivative of order* $m = s(r+1)-1$. *Hence, solving the latter expression for* s, *we expect a convergence rate of* $(m+2)/(r+1)$. *The rate is actually higher in the columns with even* m, *due to the cancellation of the moments at* ξ *with odd index. For a more general function* g *there is no exact cancellation, but the difference of the moments at* ξ, *i.e.,* $\mu_{1,k}(\xi) - \mu_{2,k}(\xi)$, *can have lower order than predicted by Theorem 6.3.20. This cancellation does not occur if the stationary point* ξ *is the endpoint of the integration interval.*

Example 6.3.22. *We make a numerical comparison between the regular Filon-type method, the localised Filon-type method and the numerical steepest descent method for* $f(x) = 1/(1+x^2)$ *and* $g(x) = (x-1/2)^2$ *on* $[-1,1]$. *Filon-type methods for this integral suffer from Runge's phenomenon: the interpolation error for the function* f *is large [169]. We choose* $s = 1$, *i.e., we use only function values of* f *in* $\{-1, 1/2, 1\}$ *and no derivatives. The order of the Filon-type methods is then* $O(\omega^{-3/2})$. *We choose* $n = 1$ *in Theorem 6.3.15. The order of the numerical steepest descent method is then* $O(\omega^{-5/2})$, *using* 4 *evaluations of* f *in the complex plane. We also interpolate two additional derivatives at* $1/2$ *for the Filon-type method: this yields a quadrature rule with* 5 *weights, and order* $O(\omega^{-2})$. *The results are illustrated in Figure 6.3.*



(a) Absolute error for four methods.

(b) The absolute error for numerical steepest descent, scaled by ω^3.

Figure 6.3: A numerical comparison between the regular and localised Filon-type methods, and the numerical steepest descent method (Example 6.3.22).

6.3.7 Generalisation to a non-analytic function g

If $g(x)$ is piecewise analytic, the integration interval can be split into subintervals where the function is analytic. Otherwise we can try to approximate $g(x)$ by an analytic function $\tilde{g}(x)$ on $[a, b]$. We should take care not to introduce new stationary points, and make sure that we accurately approximate all stationary points of $g(x)$. This can be accomplished using *comonotone polynomial approximations* [166, 167]. Alternatively, we can approximate $g(x)$ locally around the special points, possibly by using different functions for each point. This will turn out to be easier and will yield the same convergence rate.

When $g(x)$ is smooth, it can be approximated arbitrarily well by an analytic function $\tilde{g}(x)$ on $[a, b]$, using for example comonotone polynomial approximations with sufficiently high degree. Hence, there exist analytic $\tilde{g}$ such that the integral

$$I[f, \tilde{g}] := \int_a^b f(x) e^{i\omega\tilde{g}(x)}\, \mathrm{d}x = S[f, \tilde{g}; a] - S[f, \tilde{g}; b], \tag{6.50}$$

is arbitrarily close to the value of $I[f, g]$. The notation $S[f, g; x]$ is used to denote the line integral at x along the optimal path corresponding to g. Owing to decomposition (6.50), it becomes possible to do Hermite interpolation in a and b separately by different polynomials.

Theorem 6.3.23. *Assume that f and $\tilde{g}$ are analytic. Let $g_a(x)$ be the Hermite interpolating polynomial of degree s that satisfies*

$$g_a^{(k)}(a) = \tilde{g}^{(k)}(a), \quad k = 0, \ldots, s.$$

Then we have $S[f, \tilde{g}; a] - S[f, g_a; a] = O(\omega^{-s-1})$, $\omega \to \infty$.

Proof. In order to determine the asymptotic size of the error e, it can be written as

$$\begin{aligned} e &\sim \int_0^\infty f(h_a(p))(e^{i\omega\tilde{g}(h_a(p))} - e^{i\omega g_a(h_a(p))})h_a'(p)\,\mathrm{d}p \qquad (6.51) \\ &= \int_0^\infty f(h_a(p))e^{i\omega g_a(h_a(p))}(e^{i\omega(\tilde{g}(h_a(p))-g_a(h_a(p)))} - 1)h_a'(p)\,\mathrm{d}p \\ &= \frac{e^{i\omega g_a(a)}}{\omega}\int_0^\infty f(h_a(\frac{q}{\omega}))e^{-q}(e^{i\omega(\tilde{g}(h_a(\frac{q}{\omega}))-g_a(h_a(\frac{q}{\omega})))} - 1)h_a'(\frac{q}{\omega})\,\mathrm{d}q, \end{aligned}$$

where $h_a(p)$ is a parameterisation that agrees with the optimal paths for the oscillators $\tilde{g}$ and g_a at $p = 0$ up to the first few derivatives. This is possible because $g_a^{(k)}(a) = \tilde{g}^{(k)}(a)$, $k = 0, \ldots, s$. Using a Taylor approximation around a, we have

$$\tilde{g}(x) - g_a(x) = (\tilde{g}^{(s+1)}(a) - g_a^{(s+1)}(a))\frac{(x-a)^{s+1}}{(s+1)!} + O((x-a)^{s+2}).$$

Because $h_a(q/\omega) - a \sim \omega^{-1}$, we have

$$e^{i\omega(\tilde{g}(h_a(q/\omega))-g_a(h_a(q/\omega)))} - 1 \sim i\omega(\tilde{g}(h_a(q/\omega)) - g_a(h_a(q/\omega))) \sim \omega^{-s}.$$

The error e is therefore of order $O(\omega^{-s-1})$. □

The value of $I[f, \tilde{g}]$, defined by (6.50), is completely determined by the derivatives of $\tilde{g}$ at a and b. If $I[f, \tilde{g}] - I[f, g]$ is small, it follows from Theorem 6.3.23 that $\tilde{g}$ should satisfy $\tilde{g}^{(j)}(a) = g^{(j)}(a)$ and $\tilde{g}^{(j)}(b) = g^{(j)}(b)$, $j = 0, \ldots, s$, for some maximal order s that depends on the smoothness of g. Hence, $\tilde{g}$ need not be explicitly constructed.

At a stationary point ξ, more derivatives are needed. The convergence rate depends on the order r of the stationary point.

Theorem 6.3.24. *Assume that f and $\tilde{g}$ are analytic and that $\tilde{g}^{(k)}(\xi) = 0$, $k = 1, \ldots, r$, and $\tilde{g}^{(r+1)}(\xi) \neq 0$. Let $g_\xi(x)$ be the Hermite interpolating polynomial of degree $(s+1)(r+1) - 1$ that satisfies*

$$g_\xi^{(k)}(\xi) = \tilde{g}^{(k)}(\xi), \quad k = 0, \ldots, (s+1)(r+1) - 1.$$

Then the approximation of $S_j[f, \tilde{g}; \xi]$ by $S_j[f, g_\xi; \xi]$ has an error of order $O(\omega^{-s-1/(r+1)})$.

Table 6.6: Absolute error of the approximation of $S[f,\tilde{g};a]$ by $S[f,g_a;a]$ for $f(x) = 1/(1+x)$ and $g(x) = (x-1/2)^2(x-2)e^{x^2}$ at $a = 0$. We approximate g by interpolating m derivatives. The last row shows the value of $\log_2(e_{400}/e_{800})$: this value should approximate $m+1$.

$\omega \setminus m$	1	2	3	4
100	$6.1E-5$	$7.6E-7$	$1.3E-8$	$1.9E-10$
200	$1.5E-5$	$9.5E-8$	$8.4E-10$	$6.1E-12$
400	$3.8E-6$	$1.2E-8$	$5.3E-11$	$1.9E-13$
800	$9.6E-7$	$1.5E-9$	$3.3E-12$	$6.0E-15$
rate	2.0	3.0	4.0	5.0

Proof. The proof follows the same lines as the proof of Theorem 6.3.23. The difference is that, similar to the situation in the proof of Theorem 6.3.20, we have $h_{\xi,j}(q/\omega) - \xi \sim \omega^{-1/(r+1)}$ and $h'_{\xi,j}(q/\omega) \sim \omega^{r/(r+1)}$. This leads to

$$e^{i\omega(\tilde{g}(h_{\xi,j}(q/\omega)) - g_\xi(h_{\xi,j}(q/\omega)))} - 1 \sim \omega^{-s}.$$

The error estimate for this case is analogous to (6.51) in the proof of Theorem 6.3.23. Adding all contributions, it is of order $O(\omega^{-1-s+r/(r+1)}) = O(\omega^{-s-1/(r+1)})$. ☐

Example 6.3.25. *We illustrate the convergence with two examples. The function $g(x) = (x-1/2)^2(x-2)e^{x^2}$ is approximated by a polynomial of degree m in the end points $a = 0$ and $b = 1$, and in the stationary point $\xi = 1/2$. The resulting errors are displayed in Tables 6.6 and 6.7. Table 6.6 shows the error in approximating only $\tilde{F}(a)$. Table 6.7 shows the error of the approximation of I. The latter error is dominated by the error made at the stationary points but follows the theory. As in the last example for a non-analytic function f, the convergence rate is actually higher for even m, because the difference of the terms at ξ in the decomposition of I can have lower order than predicted by Theorem 6.3.24. Note that it is not possible to approximate $g(x)$ by a fixed constant since in that case also $e^{i\omega g_a(x)} = e^{i\omega c}$ reduces to a constant. At a stationary point with r vanishing derivatives, the minimal number of derivatives to interpolate is $r+1$.*

6.4 Multivariate numerical steepest descent

6.4.1 Overview

Multivariate oscillatory integrals exhibit a number of features that are not seen in the univariate case. Still, the main observations that were made

Table 6.7: Absolute error of the approximation of $I[f,g]$ by $I[f,\tilde{g}]$ for $f(x) = 1/(1+x)$ and $g(x) = (x-1/2)^2(x-2)e^{x^2}$ on $[0,1]$. We approximate g by interpolating m derivatives. The last row shows the value of $\log_2(e_{400}/e_{800})$: this value should approximate $m/2$ for odd m, and $(m+1)/2$ for even m.

$\omega \setminus m$	2	3	4
100	$1.6E-4$	$2.7E-4$	$1.8E-7$
200	$5.5E-5$	$9.8E-6$	$3.2E-8$
400	$2.0E-5$	$3.5E-6$	$5.6E-9$
800	$6.9E-6$	$1.2E-6$	$9.9E-10$
rate	1.5	1.5	2.5

regarding strong oscillations for one-dimensional integrals remain the same, and methods can be devised with high asymptotic order. In this section, we consider integrals on a bounded and connected domain $S \in \mathbb{R}^n$, with the general form

$$I_n[f] := \int_S f(\mathbf{x}) e^{i\omega g(\mathbf{x})} \,\mathrm{d}\mathbf{x}, \tag{6.52}$$

where both f and g are smooth n-dimensional functions. Integral (6.52) is a straightforward generalisation of (6.1).

As in the univariate case, the value of $I_n[f]$ is determined by the behaviour of f and g near a number of critial points. The equivalent of stationary points are those points ξ where all derivatives of the oscillator vanish: $\nabla g(\xi) = 0$. The point is said to be *degenerate* if the Hessian of the oscillator is singular. The endpoints in the univariate case correspond to the corner points of the integration domain S, and in general to all points on the boundary where the surface of S is not smooth. A new set of points are the so-called *resonance points*. These are points on the boundary where the oscillator is orthogonal to the boundary: $\nabla g(\xi) \perp \partial S$. Their importance lies in the fact that, locally, the integrand of (6.52) does not oscillate *along the boundary*. The boundary ∂S may also have an entire curve of resonance points. In fact, even more degenerate cases can be found where each point on the surface is a resonance point. The integrand is then only oscillatory in the interior of S.

The methods that were discussed in this chapter can all be extended to a multivariate setting with varying levels of generality. The asymptotic method and, correspondingly, Filon-type methods were constructed for n-dimensional polytopes in [131]. The presence of critical points and resonance points was explicitly avoided in this approach by the so-called *nonresonance condition*. The approach was subsequently extended to include critical and resonance points for a number of relevant two-dimensional problems in [130]. The multivariate extension of Levin-type methods is described in [162]. In that approach, the oscillator and the integration domain may be arbitrary, subject to the nonresonance condition. Here, we discuss the extension of the numerical steepest descent method to multivariate integrals. The results are obtained by repeated one-dimensional integration. In the process, resonance points are identified as stationary points in lower-dimensional integrals.

6.4.2 Extension to two-dimensional integrals

As motivating examples that illustrate the general case, we extend the results of the one-dimensional steepest descent approach to a number of two-dimensional oscillatory integrals. The problems that arise are introduced one by one, in a series of examples that become exceedingly more general. First, we consider the integration on a rectangular domain which will be handled by repeated one-dimensional integration. Next, we generalise to smooth integration boundaries, which introduces possible resonance points. Finally, we study an example with a critical point in the interior of S. Such points appear as stationary points in each integration variable.

In this section, we will assume that all considered functions f and g are such that the error in the decompositions vanishes. This assumption is made in this section purely for the sake of clarity and brevity. The theory will be described without this assumption in §6.4.3.

6.4.2.1 Rectangular domains in two dimensions

The simplest extension of the one-dimensional method to multivariate integrals is the use of repeated one-dimensional integration on a rectangular domain. In order to illustrate the basic approach, we restrict the discussion to a strictly monotonically increasing function g. Consider therefore the double integral

$$I_2 = \int_a^b \int_c^d f(x,y) e^{i\omega(x+y)} \,\mathrm{d}y \,\mathrm{d}x, \tag{6.53}$$

with f analytic in both variables x and y. For a fixed value of x, the inner integration in y can be written as a finite sum of contributions by applying Theorem 6.3.2. We have

$$\int_c^d f(x,y) e^{i\omega(x+y)} \,\mathrm{d}y = G(x,c) - G(x,d).$$

An expression for $G(x,y)$ is given by

$$G(x,y) = e^{i\omega(x+y)} \int_0^\infty f(x, v_y(x,q)) \frac{\partial v_y}{\partial q}(x,q) e^{-\omega q} \,\mathrm{d}q, \tag{6.54}$$

where $v_y(x,q)$ is found as the solution to $g(x,v_y(x,q)) = g(x,y) + iq$. The particular oscillator $g(x,y) = x + y$ in this example leads to the path $v_y(x,q) := y + iq$. An important observation is that the function $G(x,y)$ is analytic as a function of x, because all factors in expression (6.54) are analytic in x. In addition, $G(x,y)$ is an oscillatory function of x with the oscillator $g_1(x) := x$. Hence, the integration of $G(x,y)$ in x can also be written as a sum of contributions. The optimal path is given by $u_x(p) := x+ip$. We arrive at

$$I_2 = \int_a^b (G(x,c) - G(x,d)) \,\mathrm{d}x = [F(a,c) - F(b,c)] - [F(a,d) - F(b,d)]\,,$$

where the function $F(x,y)$ is given by

$$\begin{aligned} F(x,y) &= e^{i\omega(x+y)} \int_0^\infty \int_0^\infty f(u_x(p), v_y(u_x(p),q)) \frac{\partial u_x}{\partial p}(p) \\ &\qquad\qquad \frac{\partial v_y}{\partial q}(u_x(p),q) e^{-\omega(p+q)} \,\mathrm{d}q \,\mathrm{d}p \\ &= e^{i\omega(x+y)} \int_0^\infty \int_0^\infty f(x+ip, y+iq) i^2 e^{-\omega(p+q)} \,\mathrm{d}q \,\mathrm{d}p. \end{aligned} \tag{6.55}$$

The value of I_2 is found by summing contributions from each of the corner points of the rectangular domain. These contributions are given by a double integral with a non-oscillating integrand that decays exponentially fast as a function of both integration variables. They can be evaluated efficiently using, e.g., tensor-product Gauss-Laguerre quadrature.

6.4.2.2 Smooth boundaries in two dimensions

The double integral (6.53) is generalised by considering integration boundaries for y that depend on x. The simplest of those extensions is a simplex. We therefore consider the evaluation of the following integral first,

$$I_2 := \int_a^b \int_a^x f(x,y) e^{i\omega(x+y)} \,\mathrm{d}y \,\mathrm{d}x. \tag{6.56}$$

Applying Theorem 6.3.2 for the inner integration in y leads to

$$\int_a^x f(x,y) e^{i\omega(x+y)} \,\mathrm{d}y = G(x,a) - G(x,x),$$

with $G(x,y)$ again given by (6.54). The term $G(x,x)$ did not appear before; it is given by

$$G(x,x) = e^{i\omega 2x} \int_0^\infty f(x, x+iq) i e^{-\omega q} \,\mathrm{d}q. \tag{6.57}$$

This means that the oscillators in x of $G(x,a)$ and of $G(x,x)$ are different: they are respectively given by $g_1(x) := x$ and $g_2(x) := 2x$. A decomposition can be written for the integration in x, applying Theorem 6.3.2 for both terms separately. This leads to

$$I_2 = \int_a^b (G(x,a) - G(x,x))\,\mathrm{d}x = [F_1(a,a)-F_1(b,a)]-[F_2(a,a)-F_2(b,b)],$$

with $F_1(x,y) = F(x,y)$ corresponding to the integral of $G(x,a)$, and with F_2 given by

$$F_2(x,x) = e^{i\omega 2x}\int_0^\infty\int_0^\infty f(x+\frac{p}{2}i, x+\frac{p}{2}i+iq)\frac{i^2}{2}e^{-\omega(p+q)}\,\mathrm{d}q\,\mathrm{d}p.$$

This expression is obtained by following the paths $v_y(x,q) = y + iq$ and $u_x(p) = x + \frac{p}{2}i$. Although function F_2 is a function of only one variable x, the notation $F_2(x,x)$ is used for later notational convenience. Note that all contributions in the total decomposition are given by the evaluation of a function F_1 or F_2 at a corner point of the simplex.

A new difficulty arises when the boundaries of the integration in y are more general. Assume analytic functions $c(x)$ and $d(x)$ are given and define the double integral

$$I_2 := \int_a^b\int_{c(x)}^{d(x)} f(x,y)e^{i\omega(x+y)}\,\mathrm{d}y\,\mathrm{d}x. \tag{6.58}$$

Decomposing the inner integration in y now leads to

$$\begin{aligned}\int_{c(x)}^{d(x)} f(x,y)e^{i\omega(x+y)}\,\mathrm{d}y &= G(x,c(x)) - G(x,d(x))\\ &= e^{i\omega(x+c(x))}\int_0^\infty f(x,c(x)+iq)ie^{-\omega q}\,\mathrm{d}q\\ &\quad - e^{i\omega(x+d(x))}\int_0^\infty f(x,d(x)+iq)ie^{-\omega q}\,\mathrm{d}q.\end{aligned}$$

The oscillator of $G(x,c(x))$ is $g_1(x) := g(x,c(x)) = x + c(x)$. Although the partial derivatives of the original function $g(x,y) = x + y$ do not vanish anywhere, the function $g_1(x)$ may have stationary points:

$$\frac{\mathrm{d}}{\mathrm{d}x}g(x,c(x)) = \frac{\partial g}{\partial x} + \frac{\partial g}{\partial y}\frac{\mathrm{d}c}{\mathrm{d}x} = 0 \iff \nabla g\cdot\nabla\begin{bmatrix} x \\ c(x)\end{bmatrix} = 0,$$

Function $g_1(x)$ has a stationary point in x if the gradient of $g(x,y)$ is orthogonal to the tangent line of the boundary, which is parameterised by $[x\,c(x)]^T$. Such points are called *resonance points*.

Assume that $g_1(x)$ has l_c stationary points $\xi_{c,i} \in (a,b)$, $i = 1,\ldots,l_c$, and $g_2(x) := g(x,d(x)) = x + d(x)$ has l_d stationary points $\xi_{d,i} \in (a,b)$, $i = 1,\ldots,l_d$. Set $\xi_{c,0} := a$, $\xi_{c,l_c+1} := b$, $\xi_{d,0} := a$ and $\xi_{d,l_d+1} := b$. Then we can write I_2 as

$$\begin{aligned} I_2 = &\sum_{i=1}^{l_c+1} \left[F_{1,i}(\xi_{c,i-1}, c(\xi_{c,i-1})) - F_{1,i}(\xi_{c,i}, c(\xi_{c,i})) \right] \\ &- \sum_{i=1}^{l_d+1} \left[F_{2,i}(\xi_{d,i-1}, d(\xi_{d,i-1})) - F_{2,i}(\xi_{d,i}, d(\xi_{d,i})) \right]. \end{aligned}$$

The contributions come from the boundary points $(a,c(a))$, $(b,c(b))$, $(a,d(a))$ and $(b,d(b))$, and also from other points on the boundary, given by $(\xi_{c,i}, c(\xi_{c,i}))$ and $(\xi_{d,i}, d(\xi_{d,i}))$. The latter are all the points where the gradient of g is orthogonal to the boundary; they are resonance points.

Note that for a simplex we have $c(x) = a$ and $d(x) = x$. For the particular choice of oscillator $g(x,y) := x - y$, we have $g_2(x) := g(x,d(x)) = 0$. In other words, the function $G(x,d(x))$ is not oscillatory at all! The gradient of g is orthogonal to the boundary at all points (x,x); each point is a resonance point. For this case, the integration in x cannot be written as a sum of contributions. However, there is no need for a decomposition, as the integral $\int_a^b G(x,d(x))\mathrm{d}x$ can be evaluated by, e.g., regular Gaussian quadrature on the real line $[a,b]$.

6.4.2.3 Stationary points

A final complication that may arise in decomposing highly oscillatory two-dimensional integrals into a sum of contributions, is the presence of stationary points where $\nabla g = 0$. Consider the model integral

$$I_2 := \int_a^b \int_c^d f(x,y) e^{i\omega(x^2 - xy - y^2)}\,\mathrm{d}y\,\mathrm{d}x, \tag{6.59}$$

with $a,c < 0$ and $b,d > 0$. We have $g(x,y) = x^2 - xy - y^2$ and $\nabla g(0,0) = 0$ in the internal point $(0,0)$. In the following, we will derive a decomposition for I_2 as a sum of contributions of the form $F_{jkl}(x,y)$. Each function F_{jkl} is evaluated in a special point that is to be determined. The index j denotes the path for y: $v_{y,j}(x,q)$. The combination of index j and index k denotes the different oscillators in x that result: $g_{jk}(x)$. Finally, index l is used to denote the path for x: $u_{x,jkl}(p)$. The general form of the contribution F_{jkl}

will be shown to be

$$F_{jkl}(x,y) = e^{i\omega g_{jk}(x)} \int_0^\infty \int_0^\infty f(u_{x,jkl}(p), v_{y,j}(u_{x,jkl}(p),q)) \\ \frac{\partial u_{x,jkl}}{\partial p}(p) \frac{\partial v_{y,j}}{\partial q}(u_{x,jkl}(p),q) e^{-\omega(p+q)}\, \mathrm{d}q\, \mathrm{d}p. \tag{6.60}$$

Stationary points in y

For any $x \in [a,b]$, function $g(x,y)$ has a stationary point in y given by $y = -x/2$, since $\frac{\partial g}{\partial y}(x,-x/2) = 0$. We can write the integral (6.59) as

$$\int_c^d f(x,y) e^{i\omega g(x,y)} \mathrm{d}y = \int_c^{-x/2} f(x,y) e^{i\omega(x^2-xy-y^2)} \mathrm{d}y \\ + \int_{-x/2}^d f(x,y) e^{i\omega(x^2-xy-y^2)} \mathrm{d}y. \tag{6.61}$$

For this decomposition, we have assumed that $c \le -b/2$ and $-a/2 \le d$, as illustrated in Figure 6.4. The problem has now become similar to the problem of a smooth boundary treated earlier. Consider the first integral in the right-hand side of (6.61). By Theorem 6.3.10, there exists a decomposition

$$\int_c^{-x/2} f(x,y) e^{i\omega(x^2-xy-y^2)} \mathrm{d}y = G_1(x,c) - G_1(x,-x/2).$$

The path for y is found by solving $g(x, v_{y,1}(x,q)) = g(x,y) + iq$, leading to

$$v_{y,1}(x,q) = -x/2 - 1/2\sqrt{x^2 + 4xy + 4y^2 - 4iq}.$$

The function $G_1(x,y)$ is given in its general form by

$$G_1(x,y) = e^{i\omega g(x,y)} \int_0^\infty f(x, v_{y,1}(x,q)) \frac{\partial v_{y,1}}{\partial q}(x,q) e^{-\omega q}\, \mathrm{d}q,$$

and thus $G_1(x,c)$ has an oscillator $g_{11}(x) := x^2 - cx - c^2$, with a stationary point at $x = c/2$. The latter corresponds to the point $(c/2, c)$ on the integration boundary. The oscillator for $G_1(x,-x/2)$ is $g_{12}(x) := 5/4x^2$, with a stationary point at $x = 0$. This corresponds to the internal point $(0,0)$.

Similarly, the second integral in the right-hand side of (6.61) can be written as

$$\int_{-x/2}^d f(x,y) e^{i\omega(x^2-xy-y^2)} \mathrm{d}y = G_2(x,-x/2) - G_2(x,d).$$

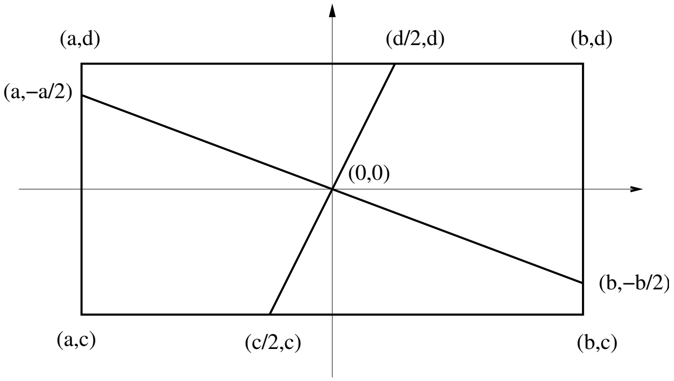


Figure 6.4: The points that contribute to the double integral I_2 for $g(x, y) = x^2 - xy - y^2$ on the rectangle $[a, b] \times [c, d]$.

The path for y differs from the path used to obtain the function G_1. We denote the path by $v_{y,2}(x, q)$, and note that it is given by

$$v_{y,2}(x, q) = -x/2 + 1/2\sqrt{x^2 + 4xy + 4y^2 - 4iq}.$$

We can define oscillators $g_{21}(x) := 5/4x^2$ and $g_{22}(x) := x^2 - dx - d^2$, corresponding to the functions $G_2(x, -x/2)$ and $G_2(x, d)$ respectively. They have a stationary point at $x = 0$ and $x = d/2$.

Stationary points in x

We have already shown that I_2 can be written as a sum of four integrals of the form

$$\int_a^b G_j(x, s_{jk}(x))\,\mathrm{d}x, \quad j = 1, 2, \quad k = 1, 2,$$

where each $G_j(x, s_{jk}(x))$ has an oscillator in x of the form $g_{jk}(x) := g(x, s_{jk}(x))$, with one stationary point $x = \xi_{jk}$. Applying Theorem 6.3.10 shows the existence of two functions $F_{jk1}(x, y)$ and $F_{jk2}(x, y)$ such that

$$\begin{aligned}\int_a^b G_j(x, s_{jk}(x))\,\mathrm{d}x =& F_{jk1}(a, s_{jk}(a)) - F_{jk1}(\xi_{jk}, s_{jk}(\xi_{jk})) \\ &+ F_{jk2}(\xi_{jk}, s_{jk}(\xi_{jk})) - F_{jk2}(b, s_{jk}(b)).\end{aligned}$$

The paths for x are found by solving $g_{jk}(u_{jkl}(p)) = g_{jk}(x_{jkl}) + ip$, $j = 1, 2$, $k = 1, 2$, $l = 1, 2$. Analytic expressions are easily derived: for the oscillator

$g_{21}(x) = g(x, -x/2) = 5/4x^2$, evaluated at $x_{211} = \xi_{21} = 0$, we find

$$g_{21}(u_{211}(p)) = g_{21}(x_{211}) + ip \Rightarrow \frac{5}{4}u_{211}^2(p) = ip \Rightarrow u_{211}(p) = \sqrt{\frac{4}{5}ip}.$$

We have arrived at a decomposition for I_2 with 16 functions of the form (6.60). Substituting the functions $s_{11}(x) = c$, $s_{12}(x) = s_{21}(x) = -x/2$ and $s_{22}(x) = d$ into the general form, the total decomposition is given by

$$\begin{aligned} I_2 = {} & F_{111}(a, c) - F_{111}(c/2, c) + F_{112}(c/2, c) - F_{112}(b, c) \\ & - F_{121}(a, -a/2) + F_{121}(0, 0) - F_{122}(0, 0) + F_{122}(b, -b/2) \\ & + F_{211}(a, -a/2) - F_{211}(0, 0) + F_{212}(0, 0) - F_{212}(b, -b/2) \\ & - F_{221}(a, d) + F_{221}(d/2, d) - F_{222}(d/2, d) + F_{222}(b, d). \end{aligned}$$

There is one evaluation in each corner point, there are two evaluations in the points where ∇g is orthogonal to the boundary, and there are four evaluations in the central stationary point $(0, 0)$ where ∇g vanishes in all integration variables. All relevant points are shown in Figure 6.4. The lines connecting $(a, -a/2)$ with $(b, -b/2)$, and $(c/2, c)$ with $(d/2, d)$ are given by $y = -x/2$ and $x = y/2$ respectively: they correspond to curves along which the partial derivative of $g(x, y)$ with respect to x or y vanishes. They intersect in the stationary point.

6.4.3 A decomposition of multivariate highly oscillatory integrals

In the previous section, we have illustrated the issues that arise in identifying the individual contributions to oscillatory integrals in two dimensions. These examples will motivate and clarify the results for the general n-dimensional case. First, we prove a decomposition for a one-dimensional integral of an n-dimensional function in section §6.4.3.1. Next, a decomposition of multivariate integrals is obtained by repeated one-dimensional integration in §6.4.3.2.

6.4.3.1 A decomposition for one variable

The decomposition of a one-dimensional integral is given in Theorem 6.3.2 for the case without stationary points, and in Theorem 6.3.10 in the presence of a stationary point. Here, we will refine Theorem 6.3.2 and obtain an expression for the error of the decomposition.

Lemma 6.4.1. *Assume that the functions f and g are analytic in an open complex neighbourhood D of $[a, b]$. If $g'(x) \neq 0$, $x \in (a, b)$, then there exists* *a function* $F(x)$*,* $x \in [a,b]$*, and a constant* $d_0 > 0$ *such that*

$$\int_a^x f(z)e^{i\omega g(z)}\,\mathrm{d}z = F(a) - F(x) + E(x), \tag{6.62}$$

with $F(x)$ *and* $E(x)$ *of the form*

$$F(x) = e^{i\omega g(x)} \int_0^{d_0} f(h_x(p))e^{-\omega p}\frac{dh_x}{dp}(p)\,\mathrm{d}p, \tag{6.63}$$

$$E(x) = e^{-\omega d_0} \int_a^x f(\kappa(z))e^{i\omega g(z)}\frac{d\kappa}{dz}(z)\,\mathrm{d}z. \tag{6.64}$$

Proof. We will prove the existence of decomposition (6.62) by the explicit construction of a new integration path for the integral. The construction of the path is illustrated in Figure 6.5. The first part of the new path is parameterised by $z = h_a(p)$, $p \in [0, d_0]$, such that $e^{i\omega g(h_a(p))} = e^{i\omega g(a)}e^{-\omega p}$. This means that the parameterisation $h_a(p)$ should satisfy $g(h_a(p)) = g(a) + ip$. The second part is parameterised by $z = \kappa(y)$, $y \in [a,x]$, such that $g(\kappa(y)) = g(y) + d_0 i$. Finally, the last part is parameterised by $z = h_x(p)$, $p \in [0, d_0]$, such that $g(h_x(p)) = g(x) + ip$.

Assuming that these paths exist and lie in D, we have by Cauchy's theorem

$$\begin{aligned}\int_a^x f(z)e^{i\omega g(z)}\,\mathrm{d}z &= \int_0^{d_0} f(h_a(p))e^{i\omega g(h_a(p))}\frac{dh_a}{dp}(p)\,\mathrm{d}p \\ &+ \int_a^x f(\kappa(y))e^{i\omega g(\kappa(y))}\frac{d\kappa}{dy}(y)\,\mathrm{d}y \\ &- \int_0^{d_0} f(h_x(p))e^{i\omega g(h_x(p))}\frac{dh_x}{dp}(p)\,\mathrm{d}p.\end{aligned}$$

This decomposition has the form of (6.62).

It remains to show that such a path exists. Here, we will prove this is the case for an integration over $[a,b]$ with $g'(a) = g'(b) = 0$. The easier case of an interval $[a,x]$ with no stationary points, or with a single stationary point at a, is proven along the same lines.

Since g' is analytic in D, any compact singly connected subset of D will contain at most a finite number of isolated zeros of g'. Since $g'(x) \neq 0$ for $x \in (a,b)$, one can always construct such a subset D_0 with $[a,b] \subset \mathrm{int}D_0$, containing no zeros of g' except a and b. Consider $g(D_0)$, with boundary $\partial g(D_0)$, and set d_0 to be the minimum vertical distance defined as

$$d_0 = \min\{\Im z \text{ for } z \in \partial g(D_0) \cap i\,\mathbb{C}^+ \text{ satisfying } \Re z \in g([a,b])\}\,. \tag{6.65}$$

Since g is analytic and non-constant, we have that the image of $[a,b]$ is strictly in the interior of the compact region $g(D_0)$. Hence, the minimum in (6.65) is well-defined and $d_0 > 0$.

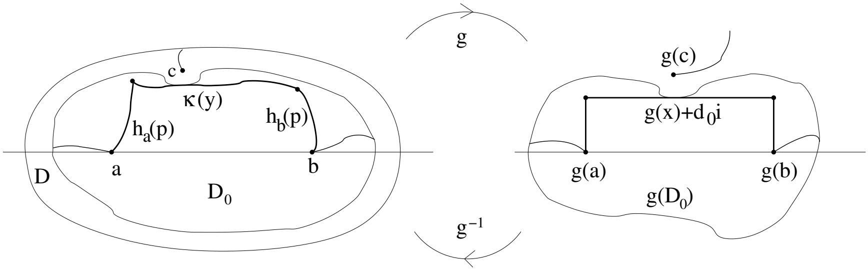


Figure 6.5: Illustration for Lemma 4.1 for the case where $g'(a) = g'(b) = g'(c) = 0$. The figure shows the connection between the domain D_0 and its image $g(D_0)$.

The inverse of g is typically a multivalued function on $g(D_0)$ with branch points at each point $g(z)$ where $g'(z) = 0$ [111]. In the present case those branch points are $g(a)$ and $g(b)$. Function g can be made uniquely invertible by selecting non-intersecting branch cuts connecting those points to $\partial g(D_0)$. These cuts can always be constructed in such a way that they do not intersect the rectangle. Define g^{-1} as the branch that satisfies $g^{-1}(g(x)) = x$ for $x \in [a, b]$. Then, the inverse of the rectangular path lies entirely within D_0, and hence, within the region of analyticity of f and g. ☐

A general decomposition in the presence of multiple stationary points can be obtained by repeatedly applying Lemma 6.4.1 on suitable subintervals. Note that the value of d_0 in the definition of F and E is determined by the size of D_0, or, more precisely, by the presence of stationary points $z \in \mathbb{C}$ that lie close to the interval $[a, b]$, and by the region of analyticity D of f and g. In most cases, d_0 may be quite large or even infinite.

The following theorem gives a decomposition for a one-dimensional integral with an n-dimensional integrand. A function in n variables is called analytic if it is analytic in each variable. We denote such a function f here by $f(\mathbf{x}, y)$, with $\mathbf{x} \in \mathbb{C}^{n-1}$ and $y \in \mathbb{C}$. A similar notation is used for g.

Theorem 6.4.2. *Assume f and g are n-dimensional functions that are analytic for $\mathbf{x}$ in an open complex neighourhood of a closed domain $B \subset \mathbb{R}^{n-1}$, and y in an open complex neighbourhood $D(\mathbf{x})$ of $[a(\mathbf{x}), b(\mathbf{x})]$. If $\frac{\partial g}{\partial y}(\mathbf{x}, y) \neq 0$, for $\mathbf{x} \in B$, $a(\mathbf{x}) < y < b(\mathbf{x})$, and if*

$$\frac{\partial g}{\partial y}(\mathbf{x}, a(\mathbf{x})) \neq 0 \quad \text{or} \quad \frac{\partial g}{\partial y}(\mathbf{x}, a(\mathbf{x})) \equiv 0, \quad \forall \mathbf{x} \in B, \ \text{and} \tag{6.66}$$

$$\frac{\partial g}{\partial y}(\mathbf{x}, b(\mathbf{x})) \neq 0 \quad \text{or} \quad \frac{\partial g}{\partial y}(\mathbf{x}, b(\mathbf{x})) \equiv 0, \quad \forall \mathbf{x} \in B, \tag{6.67}$$

then there exist functions F *and* E*, such that*

$$\int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x},y)e^{i\omega g(\mathbf{x},y)}\,\mathrm{d}y = F(\mathbf{x},a(\mathbf{x}))-F(\mathbf{x},b(\mathbf{x}))+E(\mathbf{x}), \quad \forall x \in B, \tag{6.68}$$

and with F *and* E *of the form*

$$F(\mathbf{x},y) = e^{i\omega g(\mathbf{x},y)} \int_0^{d_0} f(\mathbf{x},h(\mathbf{x},p))e^{-\omega p}\frac{\partial h}{\partial p}(\mathbf{x},p)\,\mathrm{d}p, \tag{6.69}$$

$$E(\mathbf{x}) = e^{-\omega d_0} \int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x},\kappa(\mathbf{x},y))e^{i\omega g(\mathbf{x},y)}\frac{\partial \kappa}{\partial y}(\mathbf{x},y)\,\mathrm{d}y, \tag{6.70}$$

with $d_0 > 0$*. The functions* F *and* E *are analytic in* $\mathbf{x}$ *in an open neighbourhood of* B *if* $a(\mathbf{x})$ *and* $b(\mathbf{x})$ *are analytic.*

Proof. For a fixed value of $\mathbf{x} \in B$, we can apply Lemma 6.4.1. This yields two functions of y, $F_1(y;\mathbf{x})$ and $E_1(y;\mathbf{x})$, such that

$$\int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x},y)e^{i\omega g(\mathbf{x},y)}\,\mathrm{d}y = F_1(a(\mathbf{x});\mathbf{x}) - F_1(b(\mathbf{x});\mathbf{x}) + E_1(b(\mathbf{x});\mathbf{x}).$$

These functions can be identified with (6.69) and (6.70) by $F(\mathbf{x},y) := F_1(y;\mathbf{x})$ and $E(\mathbf{x}) := E_1(b(\mathbf{x});\mathbf{x})$. However, as the constant $d_0(\mathbf{x})$ still depends on $\mathbf{x}$, it remains to be proven that it can be chosen independently of $\mathbf{x}$.

Recall that the region D_0 in the proof of Lemma 6.4.1 was chosen such that it contains no zeros of g', except possibly a and b. The size of the region D_0 and, hence, of the constant d_0, is restricted only by the analyticity of f and g, and by the presence of isolated stationary points other than a and b. In the current multivariate application of the lemma, this means that $D_0(\mathbf{x})$ is chosen such that it contains no zeros of $\frac{\partial g}{\partial y}$, except possibly $(\mathbf{x},a(\mathbf{x}))$ and $(\mathbf{x},b(\mathbf{x}))$. Now consider a (complex) curve $c(\mathbf{x})$ of stationary points, i.e., $\frac{\partial g}{\partial y}(\mathbf{x},c(\mathbf{x})) \equiv 0$, $\mathbf{x} \in B$. The value $d_0(\mathbf{x})$ could become arbitrarily small if $c(\mathbf{x})$ lies arbitrarily close to $[a(\mathbf{x},b(\mathbf{x})]$. However, conditions (6.66) and (6.67), together with the closedness of B, guarantee that $c(\mathbf{x})$ either coincides with $a(\mathbf{x})$ or $b(\mathbf{x})$, or it is an isolated stationary point for all $\mathbf{x} \in B$ including $\mathbf{x} \in \partial B$. Therefore, $d_0(\mathbf{x})$ can be bounded from below by a constant $d_0 > 0$.

Finally, we note that all factors in the expressions for F and E are analytic in $\mathbf{x}$, and the integral of an analytic function is again analytic if the integration boundaries are given by a constant, or by an analytic function [111]. Hence, $F(\mathbf{x},a(\mathbf{x}))$, $F(\mathbf{x},b(\mathbf{x}))$ and $E(\mathbf{x})$ are analytic in $\mathbf{x} \in B$ if the boundary functions $a(\mathbf{x})$ and $b(\mathbf{x})$ are analytic. By a similar reasoning as in the previous paragraph, d_0 can be chosen small enough, but still positive, such that F and E are analytic at least in an open complex neighbourhood of B. ☐

Remark 6.4.3. *Condition* (6.66) *requires that the boundary function* $a(\mathbf{x})$ *does not cross a* curve $c(\mathbf{x})$ *of stationary points in* y*: either* $a(\mathbf{x})$ *and* $c(\mathbf{x})$ *are disjunct, or they coincide. If* $a(\mathbf{x}_1) = c(\mathbf{x}_1)$ *at a single point* $\mathbf{x}_1 \in B$*, then the constant* d_0 *may become arbitrarily small. The function* $F(\mathbf{x}, y)$ *can still be shown to exist, but it may not be possible to evaluate the function using the path of steepest descent due to the presence of stationary points in the complex plane. Aside from the numerical singularity at such points, crossing a stationary point means that the line integral that connects the endpoints of the paths for* a *and* b *can no longer be discarded. Still, the function* $F(\mathbf{x}, y)$ *can be evaluated using any other path that yields exponential decay, as long as the total decomposition is justified by Cauchy's Theorem, and the integration path does not cross any stationary points. An example of this special case will be given in §6.4.5.*

We can now describe the total decomposition in the presence of real stationary points. For n-dimensional functions, the equation $\frac{\partial f}{\partial y}(\mathbf{x}, y) = 0$ has $(n-1)$-dimensional solutions $y = s_i(\mathbf{x})$, $i = 1, \ldots, l$. As in the one-dimensional case, the integration region will be subdivided, using these solutions as new boundaries.

Theorem 6.4.4. *Assume* f *and* g *are* n*-dimensional functions that are analytic for* $\mathbf{x}$ *in an open complex neighbourhoof of a closed domain* $B \subset \mathbb{R}^{n-1}$*, and* y *in an open complex neighbourhood* $D(\mathbf{x})$ *of* $[a(\mathbf{x}), b(\mathbf{x})]$*. Assume further that* $\frac{\partial g}{\partial y}(\mathbf{x}, s_i(\mathbf{x})) = 0$*,* $i = 1, \ldots, l$*, and* $\frac{\partial g}{\partial y}(\mathbf{x}, y) \neq 0$ *otherwise. If* $s_0(\mathbf{x}) := a(\mathbf{x}) \leq s_1(\mathbf{x}) \leq \ldots \leq s_l(\mathbf{x}) \leq s_{l+1}(\mathbf{x}) := b(\mathbf{x})$*, and* $a(\mathbf{x})$ *and* $b(\mathbf{x})$ *satisfy* (6.66)-(6.67)*, then there exist functions* F_i *and* E_i *of the form* (6.69) *and* (6.70) *such that*

$$\int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x}, y) e^{i\omega g(\mathbf{x}, y)} \, \mathrm{d}y = \sum_{j=1}^{l+1} [F_j(\mathbf{x}, s_{j-1}(\mathbf{x})) - F_j(\mathbf{x}, s_j(\mathbf{x}))] + \sum_{j=1}^{l+1} E_j(\mathbf{x}), \quad \forall x \in B. \tag{6.71}$$

Proof. We can write the integral as

$$\int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x}, y) e^{i\omega g(\mathbf{x}, y)} \, \mathrm{d}y = \int_{a(\mathbf{x})}^{s_1(\mathbf{x})} \cdot \, \mathrm{d}y + \int_{s_1(\mathbf{x})}^{s_2(\mathbf{x})} \cdot \, \mathrm{d}y + \ldots + \int_{s_l(\mathbf{x})}^{b(\mathbf{x})} \cdot \, \mathrm{d}y$$

The result follows from the repeated application of Theorem 6.4.2. ☐

6.4.3.2 Repeated one-dimensional integration

The results of the previous subsection can be used in a recursive setting in order to obtain a decomposition for an n-dimensional integral,

$$I_n := \int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_2(x_1)} \int_{a_3(x_1,x_2)}^{b_3(x_1,x_2)} \cdots \int_{a_n(x_1,\ldots,x_n)}^{b_n(x_1,\ldots,x_n)} f(\mathbf{x}) e^{i\omega g(\mathbf{x})} \, \mathrm{d}\mathbf{x}. \tag{6.72}$$

The decomposition of the inner integral in x_n can be obtained by Theorem 6.4.4. Assume that the equation $\frac{\partial g}{\partial x_n}(\mathbf{x}, x_n) = 0$ has l solutions $x_n = s_i(\mathbf{x})$, $i = 1, \ldots, l$. Then, the decomposition of the inner integral in x_n has the form of (6.71). The functions $F_i(\mathbf{x}, s_j(\mathbf{x}))$ are analytic in $\mathbf{x}$, and have an oscillator of the form $g(\mathbf{x}, s_j(\mathbf{x}))$. Define $s_0(\mathbf{x}) = a_n(\mathbf{x})$ and $s_{l+1}(\mathbf{x}) = b_n(\mathbf{x})$; then every function F_j, for $j = 1, \ldots, l+1$, leads to two oscillators,

$$g_{j,1}(\mathbf{x}) = g(\mathbf{x}, s_{j-1}(\mathbf{x})) \quad \text{and} \quad g_{j,2}(\mathbf{x}) = g(\mathbf{x}, s_j(\mathbf{x})). \tag{6.73}$$

Obviously $g_{j,1}(\mathbf{x}) = g_{j-1,2}(\mathbf{x})$. These oscillators are $(n-1)$-dimensional analytic functions. The first index i denotes the subinterval of $[a_n(\mathbf{x}), b_n(\mathbf{x})]$, the second index denotes an endpoint of that interval.

In the following we will denote an oscillator compactly by g_λ, where λ is a multi-index. An integral corresponding to g_λ can be decomposed again using Theorem 6.4.4. If $g_\lambda(\mathbf{x}, x_{n-1})$ has l_λ stationary points $s_{\lambda,i}(\mathbf{x})$, $i = 1, \ldots, l_\lambda$, this yields $l_\lambda + 1$ functions $F_{\lambda,i}$, $i = 1, \ldots, l_\lambda + 1$. Denote by $s_{\lambda,0}(\mathbf{x}) := a_{n-1}(\mathbf{x})$ and $s_{\lambda,l_\lambda+1}(\mathbf{x}) := b_{n-1}(\mathbf{x})$. Each contribution has the form $F_{\lambda,i}(\mathbf{x}, s_{\lambda,i-1}(\mathbf{x}))$ or $F_{\lambda,i}(\mathbf{x}, s_{\lambda,i}(\mathbf{x}))$. The oscillators can be defined recursively by

$$g_{\lambda,i,1}(\mathbf{x}) := g_\lambda(\mathbf{x}, s_{\lambda,i-1}(\mathbf{x})) \quad \text{and} \quad g_{\lambda,i,2}(\mathbf{x}) := g_\lambda(\mathbf{x}, s_{\lambda,i}(\mathbf{x})). \tag{6.74}$$

These oscillators are $(n-2)$-dimensional analytic functions. The definitions can be extended recursively, applying Theorem 6.4.4 for each integration variable until integral I_n is fully written as a sum of integrals that are no longer oscillatory. Extending our notation, each recursive step adds two layers of indices to λ: the decomposition of an integral with oscillator g_λ yields the functions $F_{\lambda,i}$, $i = 1, \ldots, l_\lambda + 1$, and the evaluation of $F_{\lambda,i}$ in the endpoints leads to the new oscillators $g_{\lambda,i,1}$ and $g_{\lambda,i,2}$. After the final recursive step, we have functions $F_{\lambda'}$ with $\mathrm{size}(\lambda') = 2n-1$, evaluated in points x_λ with $\mathrm{size}(\lambda) = 2n$ of the form

$$x_\lambda = g_\lambda(a) = (a, f_1(a), f_2(a, f_1(a)), f_3(a, f_1(a), f_2(a, f_1(a))), \ldots), \tag{6.75}$$

with $a = a_1$ or $a = b_1$. Examples will be given in §6.4.5. The functions f_i can either be one of the boundary functions a_j or b_j of I_n, or a curve of stationary points for one integration variable. In the following theorem, we use $F_{\lambda'}$ to denote the function that is evaluated at x_λ (i.e., λ' is λ with the last index omitted).

Theorem 6.4.5. *Assume f and g are n-dimensional functions that are analytic in a complex neighbourhood of the integration region of I_n, given by* (6.72)*, with all boundary functions a_i and b_i analytic, $i = 2, \dots, n$. Define the functions g_λ recursively by* (6.73) *and* (6.74)*. If the following condition holds,*

$$\forall \lambda, \exists y : \frac{\partial g_\lambda}{\partial y}(\mathbf{x}, y) \neq 0, \tag{6.76}$$

then there exist functions $F_{\lambda'}$ and points $\mathbf{x}_\lambda$ such that

$$I_n = \sum_{\text{size}(\lambda)=2n} s_\lambda F_{\lambda'}(\mathbf{x}_\lambda) + O(e^{-\omega d_0}), \tag{6.77}$$

with $s_\lambda = \pm 1$ and with a constant $d_0 > 0$.

Proof. The construction of the functions F_λ and the points $\mathbf{x}_\lambda$ follows from the recursive description given earlier in this section, based on applying Theorem 6.4.4 repeatedly for all integration variables. Condition (6.76) guarantees that each oscillator encountered for an integration variable y is not independent of y. It remains to show in this proof that the error of the full decomposition decays exponentially fast as $O(e^{-\omega d_0})$ with a constant $d_0 > 0$.

Consider the decomposition of the integration in x_n of an n-dimensional oscillatory integrand, as given by Theorem 6.4.4. The error expression E_j has the form of (6.70),

$$E_j(\mathbf{x}) = e^{-\omega d_{0,j}} \int_{a(\mathbf{x})}^{b(\mathbf{x})} f(\mathbf{x}, \kappa_j(\mathbf{x}, x_n)) e^{i\omega g(\mathbf{x}, x_n)} \frac{\partial \kappa_j}{\partial x_n}(\mathbf{x}, x_n) \, \mathrm{d}x_n. \tag{6.78}$$

Function f is analytic on a (finite) complex neighbourhood of the integration domain, and can therefore be bounded uniformly on that domain by a constant $M > 0$. Additionally, we have $|e^{i\omega g(\mathbf{x}, x_n)}| \leq 1$ since $g(\mathbf{x}, x_n)$ is real. Finally, in order to bound the third factor $\frac{\partial \kappa_j}{\partial x_n}(\mathbf{x}, x_n)$, recall that $\kappa_j(\mathbf{x}, x_n) := g_n^{-1}(g(\mathbf{x}, x_n) + d_{0,j} i)$, where $g_n^{-1}(y)$ represents the inverse of g with respect to x_n. We have

$$\frac{\partial \kappa_j}{\partial x_n}(\mathbf{x}, x_n) = \frac{\partial g_n^{-1}}{\partial y}(g(\mathbf{x}, x_n) + d_{0,j} i) \frac{\partial g}{\partial x_n}(\mathbf{x}, x_n).$$

The derivative of g is bounded, because g is analytic on the (finite) integration domain. The derivative of g_n^{-1} can only be unbounded if

$g'(\mathbf{x}, \kappa(\mathbf{x}, x_n)) = 0$. This situation occurs when there is a stationary point along the path for the error integral. By construction, this is never the case. Hence, the third factor of (6.78) can also be bounded by a constant $N > 0$. Combining these observations, we have

$$\left| \int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_2(x_1)} \int_{a_3(x_1,x_2)}^{b_3(x_1,x_2)} \cdots \int_{a_{n-1}(\mathbf{x})}^{b_{n-1}(\mathbf{x})} E_j(\mathbf{x})\, \mathrm{d}x_{n-1} \ldots \mathrm{d}x_1 \right| \leq DMNe^{-\omega d_{0,j}},$$

with D the size of the integration domain.

The decomposition for the integration in x_n yields $l+1$ functions F_i, when there are l stationary points $s_i(\mathbf{x})$ in x_n. From expression (6.69) for F_i, we see that each contribution to I_n is of the form

$$\int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_2(x_1)} \int_{a_3(x_1,x_2)}^{b_3(x_1,x_2)} \cdots \int_{a_{n-1}(\mathbf{x})}^{b_{n-1}(\mathbf{x})} \tilde{f}(\mathbf{x}) e^{i\omega\tilde{g}(\mathbf{x})}\, \mathrm{d}\mathbf{x}.$$

Each contribution has the form of I_{n-1}. The line of arguments can therefore be repeated in order to bound the error for the decomposition in x_{n-1}, and recursively for $x_{n-2}, \ldots, x_1$. The constant d_0 in (6.77) is obtained as the smallest of the $d_{0,j}$ constants. □

Remark 6.4.6. *Condition* (6.76) *explicitly excludes those case where* $\frac{\partial g_\lambda}{\partial y}(\mathbf{x}, y) \equiv 0$. *In that case,* $g_\lambda(\mathbf{x}, y) = f(\mathbf{x})$ *is independent of* y*, and hence, it is not an oscillator for the variable* y*. The corresponding integral cannot be decomposed. However, since the integral is not oscillatory, this case does not pose a problem: it can be evaluated using standard integration techniques. If* $\frac{\partial g_\lambda}{\partial z}(\mathbf{x}, z, y) \neq 0$*, the recursive procedure can be continued for the oscillatory integral in the variable* z*.*

Remark 6.4.7. *Throughout this section we have assumed that equations of the form* $\frac{\partial g}{\partial x_n}(\mathbf{x}, x_n) = 0$ *have* l *solutions, where* l *is a constant independent of* $\mathbf{x}$*. If* l *depends on the value of* $\mathbf{x}$*, then the integration region can always be split into a number of regions where* l *is a constant. This may introduce integrals for which conditions* (6.66) *and* (6.67) *in Theorem 6.4.2 cannot hold. Still, the decomposition can be computed following Remark 6.4.3. A numerical example of this special case will be given in §6.4.5.*

6.4.3.3 Integration on closed volumes

The procedure to locate the special points is simplified when the integration region is a closed and smooth n-dimensional volume without corner points. In order to see this, note that there are many equivalent ways of writing an integral over a closed and smooth volume in the general form of (6.72). In particular, the integration boundary functions a_i and b_i are not unique: they correspond to a certain parameterisation of the volume, of which there are infinitely many. However, a different choice of integration boundary functions leads to a different set of critical points x_λ, as identified by the recursive procedure described in §6.4.3.2. Although the resulting decomposition will be correct, we can expect that some of these points are merely an artefact of our arbitrary choice of boundary functions. Indeed, one can verify that such points appear twice in the decomposition, and that $\mathbf{x}_\lambda = \mathbf{x}_\mu$, $F_{\lambda'}(\mathbf{x}_\lambda) = F_{\mu'}(\mathbf{x}_\mu)$ and $s_\lambda = -s_\mu$. Hence, the artificial contributions cancel out. They need not be computed. The relevant points are the critical points in the interior, and the resonance points on the boundary.

6.4.4 The construction of cubature rules using derivatives

The decomposition of an n-dimensional integral as described in §6.4.3 can be written as

$$I_n[f] := \int_S f(\mathbf{x}) e^{i\omega g(\mathbf{x})}\, \mathrm{d}\mathbf{x} = \sum_{\mathrm{size}(\lambda)=2n} s_\lambda F_{\lambda'}[f](\mathbf{x}_\lambda) + O(e^{-\omega d_0}), \quad (6.79)$$

with

$$\begin{aligned} F_{\lambda'}[f](\mathbf{x}_\lambda) :=& e^{i\omega g(\mathbf{x}_\lambda)} \int_0^{d_0} \cdots \int_0^{d_0} f(h_1(p_1), h_2(h_1(p_1), p_2), \ldots) \\ & e^{-\omega(\sum p_i)} \frac{\partial h_1}{\partial p_1}(p_1) \frac{\partial h_2}{\partial p_2}(h_1(p_1), p_2) \ldots \\ & \frac{\partial h_n}{\partial p_n}(\ldots, h_{n-1}(\ldots, p_{n-1}), p_n)\, \mathrm{d}p_1 \ldots \mathrm{d}p_n, \end{aligned} \quad (6.80)$$

The functions h_i represent the optimal paths with respect to the oscillators that are implied by the multi-index λ. This is a generalisation of the two-dimensional form given by (6.60). If the function f is easily evaluated for complex arguments, tensor-product Gauss-Laguerre rules can be used to obtain an accurate approximation to each of the $F_{\lambda'}[f](\mathbf{x}_\lambda)$ values. This is a straightforward extension of Theorems 6.3.7 and 6.3.13. Alternatively, the function value $F_{\lambda'}[f](\mathbf{x}_\lambda)$ can be approximated by approximating f locally around the point $\mathbf{x}_\lambda$. That is the approach taken in this section. The result is a cubature rule that requires only function values and derivatives of f at $\mathbf{x}_\lambda$. The use of tensor-product Gauss-Laguerre quadrature rule to evaluate the weights of the cubature rule will be illustrated in §6.4.5.

6.4.4.1 A localised Filon-type method

The multivariate extension of the Filon-type method, discussed in §6.2.2, is straightforward: if f can be approximated by a linear combination of basis functions, $f(\mathbf{x}) = \sum_{i=1}^{N} a_i \phi_i(\mathbf{x})$, then $I_n[f]$ can be approximated by

$$I_n[f] \approx Q_F[f] := \sum_{i=1}^{N} w_i a_i, \quad \text{with} \quad w_i := I_n[\phi_i]. \tag{6.81}$$

Similarly to the univariate case, a polynomial basis is suggested in [131, 130], such that the value of f and a number of its derivatives are interpolated in the critical points $\mathbf{x}_\lambda$. Depending on the number of critical points and the number of derivatives interpolated, the degree of the basis functions may have to be high. Owing to our decomposition of the integral into a sum of independent contributions however, the contributions can be approximated separately, i.e., there is no need for a global approximation of f. This will lead to a cubature rule with the same order of accuracy, but using a much lower degree of polynomials.

We will now construct an approximation for $F_{\lambda'}[f](\mathbf{x}_\lambda)$ as given in (6.80). Define the multi-index $i = i_1 i_2 \ldots i_n$ with $|i| = i_1 + i_2 + \ldots + i_n$, and denote $(x_1 - y_1)^{i_1} \ldots (x_n - y_n)^{i_n}$ by $(\mathbf{x} - \mathbf{y})^i$. Then we can write the Taylor series of f in the following way,

$$F_{\lambda'}[f](\mathbf{x}_\lambda) = \sum_{|i| \leq \infty} \frac{F_\lambda[(\mathbf{x} - \mathbf{x}_\lambda)^i](\mathbf{x}_\lambda)}{i_1! i_2! \ldots i_n!} f^{(i)}(\mathbf{x}_\lambda), \tag{6.82}$$

where $f^{(i)}(\mathbf{x})$ is used to denote

$$f^{(i)}(\mathbf{x}) = \frac{\partial^{i_1} \partial^{i_2} \ldots \partial^{i_n} f}{\partial x_1^{i_1} \partial x_2^{i_2} \ldots \partial x_n^{i_n}}(x_1, x_2, \ldots, x_n).$$

The truncated Taylor series can be used in order to obtain a convergent cubature rule for (6.81). Assume that the total order of the derivative at point $\mathbf{x}_\lambda$ is limited by d_λ, then we propose the cubature rule

$$Q_{LF}[f] := \sum_{\text{size}(\lambda)=2n} \sum_{|i| \leq d_\lambda} w_{\lambda,i} f^{(i)}(\mathbf{x}_\lambda), \tag{6.83}$$

with the weights given by

$$w_{\lambda,i} := s_\lambda \frac{F_{\lambda'}[(\mathbf{x} - \mathbf{x}_\lambda)^i](\mathbf{x}_\lambda)}{i_1! i_2! \ldots i_n!}. \tag{6.84}$$

Remark 6.4.8. *The method of constructing the cubature rule is as follows. The oscillator $g(\mathbf{x})$ and the integration domain S determine the location of the critical points $\mathbf{x}_\lambda$ and the optimal paths, by the recursive procedure described in §6.4.3. Hence, the abscissae $\mathbf{x}_\lambda$ depend only on the oscillator and on the domain. The value of the weights is found by evaluating* (6.84) *along these paths. The weights depend in general on ω. Finally, an approximation to $I_n[f]$ is obtained by evaluating the function f and its derivatives in the abscissae and evaluating* (6.83).

6.4.4.2 Convergence properties

In order to obtain the asymptotic order as a function of ω of the cubature rule (6.83), we will first examine the error in the truncation of (6.82). The size of the truncation error will determine the integration error.

Lemma 6.4.9. *Consider the point* $\mathbf{x}_\lambda$ *with* $\mathrm{size}(\lambda) = 2n$*, and the oscillator* g_λ *obtained by repeated one-dimensional integration. If the oscillator for integration variable* x_j *has a stationary point of order* r_j*,* $j = 1, \ldots, n$*, then we have, with* i *the multi-index* $i_1 \ldots i_n$*,*

$$\left|F_{\lambda'}[(\mathbf{x}-\mathbf{x}_\lambda)^i](\mathbf{x}_\lambda)\right| = O(\omega^{-\alpha_{\lambda,i}}), \quad \textit{with} \quad \alpha_{\lambda,i} = \sum_{j=1}^{n} \frac{i_j+1}{r_j+1}.$$

The technical proof is omitted. It is based on a repeated application of the reasoning in the proof of 6.3.20 for each integration variable. Integration variable x_j contributes a factor $\omega^{-(i_j+1)/(r_j+1)}$; the sum of all contributions yields the result.

Theorem 6.4.10. *The approximation of* $I_n[f]$ *by the cubature rule* (6.83) *has an error of the order*

$$I_n[f] - Q_{LF}[f] = O(\omega^{-\alpha}), \quad \textit{with} \quad \alpha = \min_{\mathrm{size}(\lambda)=2n} \min_{|i|=d_\lambda+1} \alpha_{\lambda,i}. \tag{6.85}$$

Proof. From Lemma 6.4.9, we see that the error in the truncation of (6.82) is asymptotically equivalent to the asymptotic order of the first discarded term. The latter is given by $\alpha_{\lambda,i}$ with $|i| = d_\lambda + 1$. Hence, the order of the truncation error is found as the minimum for all λ and i of $\alpha_{\lambda,i}$, with $|i| = d_\lambda + 1$. The order of the error $I_n[f] - Q_{LF}[f]$ is the same. The exponentially decaying error $e^{-\omega d_0}$ in (6.79) may be discarded because, asymptotically, it vanishes faster than any power of ω^{-1}. ☐

Remark 6.4.11. *The convergence rate may actually be faster than the rate predicted by Theorem 6.4.10. This is due to the cancellation of moments at stationary points. In particular, it may be that* $\mathbf{x}_\lambda = \mathbf{x}_\mu$*, and that* $F_{\lambda'}[(\mathbf{x}-\mathbf{x}_\lambda)^i](\mathbf{x}_\lambda) - F_{\mu'}[(\mathbf{x}-\mathbf{x}_\mu)^i](\mathbf{x}_\mu) = o(\omega^{-\alpha_{\lambda,i}})$ *and* $o(\omega^{-\alpha_{\mu,i}})$*, i.e., the difference of the moments at the special point* $\mathbf{x}_\lambda$ *can have lower order than the moments themselves.*

6.4.5 Numerical results

In this section, we illustrate the convergence of the constructed cubature rules for some arbitrary functions f. The integration domains considered are the right half of a circle in §6.4.5.1, the unit ball in §6.4.5.2 and a rectangular domain in §6.4.5.3 and §6.4.5.4. We consider the Fourier oscillator and more general oscillators that lead to an internal stationary point.

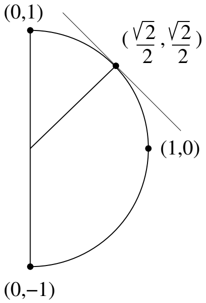


Figure 6.6: The integration domain for example 1. The gradient of the oscillator $\nabla g = [1 \ \ 1]^T$ is orthogonal to the tangent line at the point $(\sqrt{2}/2, \sqrt{2}/2)$.

6.4.5.1 Half of a circle

We consider an integral over half a circle, written as

$$I_2 := \int_0^1 \int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} \left(\cos(x_1 x_2) + \frac{1}{2 + x_1 + x_2} \right) e^{i\omega(x_1+x_2)} \, \mathrm{d}x_2 \, \mathrm{d}x_1.$$

The integration domain is shown in Figure 6.6. The set of critical points consists of the points $(0, -1)$ and $(0, 1)$, because they are boundary points of the piecewise smooth integration domain, and the point $(\sqrt{2}/2, \sqrt{2}/2)$, because the gradient of the Fourier oscillator is orthogonal to the circle at that point. This also follows from the analysis following §6.4.3: we have

$$\begin{cases} g_{11}(x) &= x - \sqrt{1-x^2} \\ g_{12}(x) &= x + \sqrt{1-x^2}, \end{cases}$$

with stationary points at $-\sqrt{2}/2$, and $+\sqrt{2}/2$ respectively. Since $x = -\sqrt{2}/2$ lies outside the integration domain, the special points are

$$\begin{cases} \mathbf{x}_{1111} &= (0, -1) \\ \mathbf{x}_{1112} &= (1, 0), \end{cases}$$

and

$$\begin{cases} \mathbf{x}_{1211} &= (0, 1) \\ \mathbf{x}_{1212} &= (\sqrt{2}/2, \sqrt{2}/2) \\ \mathbf{x}_{1221} &= (\sqrt{2}/2, \sqrt{2}/2) \\ \mathbf{x}_{1222} &= (1, 0). \end{cases}$$

Table 6.8: Absolute error of the approximation of I_2 by a cubature rule using derivatives of maximal order d. The last row shows the value of $\log_2(e_{400}/e_{800})$. The theoretically predicted asymptotic lower bound is shown between parentheses. The rules in columns $1-3$ require 3, 9 and 18 function evaluations respectively.

$\omega \setminus d$	0	1	2
50	$6.1E-5$	$1.4E-5$	$1.3E-6$
100	$1.0E-5$	$2.6E-6$	$1.1E-7$
200	$2.4E-6$	$4.5E-7$	$9.4E-9$
400	$3.9E-7$	$7.9E-8$	$8.3E-10$
800	$6.2E-8$	$1.4E-8$	$7.3E-11$
rate	2.7 (2.0)	2.5 (2.5)	3.5 (3.0)

This corresponds to the total decomposition

$$\begin{aligned}[][F_{111}(\mathbf{x}_{1111}) - F_{111}(\mathbf{x}_{1112})] - [F_{121}(\mathbf{x}_{1211}) - F_{121}(\mathbf{x}_{1212})\\ +F_{122}(\mathbf{x}_{1221}) - F_{122}(\mathbf{x}_{1222})].\end{aligned}$$

The two contributions from the point $(1,0)$ cancel out, $F_{111}(\mathbf{x}_{1112}) = F_{122}(\mathbf{x}_{1222})$; they are an artefact from the chosen parameterisation of which the point appears to be a boundary point.

The moments $F_{121}[(\mathbf{x} - \mathbf{x}_{1212})^i](\mathbf{x}_{1212})$ and $F_{122}[(\mathbf{x} - \mathbf{x}_{1221})^i](\mathbf{x}_{1221})$ have a stationary point of order $r_1 = 1$ in the variable x_1, due to the stationary point $\sqrt{2}/2$ of g_{12}; the other moments are regular. Using a fixed number of derivatives d at each point, the moments at $(\sqrt{2}/2, \sqrt{2}/2)$ will asymptotically be the largest. From Lemma 6.4.9, the moment $F_{121}[(\mathbf{x} - \mathbf{x}_{1212})^i](\mathbf{x}_{1212})$ with $i = (0, d)$ scales as $O(\omega^{-1-(d+1)/2})$. Hence, the first discarded moment with minimal order has order $\omega^{-1-(d+2)/2}$. By Theorem 6.4.10, this is the leading order of the integration error. This is illustrated in Table 6.8. The columns with d even have a higher convergence rate than predicted due to the (partial) cancellation of moments.

The total number of weights in the cubature formula for the rightmost column ($d = 2$) is 18: there are 3 critical points, and the evaluation of 6 partial derivatives with total order less than or equal to 2 is required in each point. The value of I_2 itself scales as the zero-th order moments, $\omega^{-3/2}$. Hence, the convergence rate of the relative error is 1.5 smaller than the rate shown for the absolute error.

The weights were evaluated using tensor-product rules. Following Remark 6.3.14, half-range Gauss-Hermite rules were used for evaluating one-dimensional integrals with a singularity due to a stationary point [87]. Hence, we expect a convergence rate of the relative error of $O(\omega^{-n})$, where n is the number of quadrature rules used in each dimension. The absolute

Table 6.9: Absolute error of the approximation of the zero-th order moment $F_{121}[(\mathbf{x}-\mathbf{x}_{1212})^0](\mathbf{x}_{1212})$ by a tensor-product of Gauss-Laguerre and half-range Gauss-Hermite rules with n points in each dimension. The last row shows the value of $\log_2(e_{100}/e_{50})$. The theoretically predicted asymptotic lower bound is shown between parentheses.

$\omega \setminus n$	1	2	3
25	$2.3E-05$	$4.6E-08$	$2.3E-10$
50	$4.1E-06$	$4.1E-09$	$1.0E-11$
100	$7.2E-07$	$3.6E-10$	$4.5E-13$
rate	2.5 (2.5)	3.5 (3.5)	4.5 (4.5)

error scales as $O(\omega^{-n-3/2})$. This is confirmed by the results in Table 6.9. Note that this approach is possible for more general f, and that the results require much less operations than the construction of the appropriate cubature rule. If applicable, and if high accuracy and efficiency is required, this approach is preferable over the use of a cubature rule. For $\omega = 100$ and $n = 3$, 9 function evaluations were required by the 2D tensor-product rule for an absolute error of $4.5E-13$ and a relative error of $5.8E-10$.

6.4.5.2 The unit ball

We consider an integral over the unit ball, written as

$$I_3 := \int_{-1}^{1}\int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}}\int_{-\sqrt{1-x_1^2-x_2^2}}^{\sqrt{1-x_1^2-x_2^2}} e^{x_1+x_2^2x_3}(3x_3+\cos(x_2))e^{i\omega(x_1+x_2+x_3)}\,\mathrm{d}\mathbf{x}.$$

There are no corners in this example, since the integration domain is smooth everywhere. The critical points are those where the gradient of the Fourier oscillator is orthogonal to the boundary: $(-\sqrt{3}/3, -\sqrt{3}/3, -\sqrt{3}/3)$ and $(\sqrt{3}/3, \sqrt{3}/3, \sqrt{3}/3)$.

The decomposition for the integration variable x_3 yields the oscillators

$$g_{11}(x_1,x_2) = x_1 + x_2 - \sqrt{1-x_1^2-x_2^2}, \quad \text{and}$$
$$g_{12}(x_1,x_2) = x_1 + x_2 + \sqrt{1-x_1^2-x_2^2}.$$

These oscillators have a curve of stationary points in x_2, given by $x_2 = \pm\sqrt{2-2x_1^2}$. The relevant oscillators after the second decomposition are

$$g_{1112}(x_1) = g_{1121}(x_1) = x_1 - \sqrt{2-2x_1^2}, \quad \text{and}$$
$$g_{1212}(x_1) = g_{1221}(x_1) = x_1 + \sqrt{2-2x_1^2},$$

Table 6.10: Absolute error of the approximation of I_3 on the unit ball by a cubature rule using derivatives of maximal order d. The last row shows the value of $\log_2(e_{400}/e_{800})$. The theoretically predicted asymptotic lower bound is shown between parentheses. The rules in columns $1-3$ require 2, 8 and 20 function evaluations respectively.

$\omega \setminus d$	0	1	2
100	$2.6E-5$	$2.4E-6$	$1.2E-7$
200	$3.2E-6$	$3.6E-7$	$8.0E-9$
400	$3.9E-7$	$5.2E-8$	$5.5E-10$
800	$5.0E-8$	$3.8E-9$	$2.7E-11$
1600	$6.3E-9$	$5.2E-10$	$1.7E-12$
rate	3.0 (2.5)	2.9 (3.0)	4.0 (3.5)

with the stationary points $x_1 = \pm\sqrt{3}/3$.

At the point $(\sqrt{3}/3, \sqrt{3}/3, \sqrt{3}/3)$, there is a stationary point of order 1 in the variables x_1 and x_2. The size of the moments hence scales as $\omega^{-(i_1+1)/2-(i_2+1)/2-(i_3-1)}$. With the restriction $|i| = i_1 + i_2 + i_3 = d + 1$ from Theorem 6.4.10, the leading order of the error is given by $\omega^{-(d+3)/2-1}$. The rate is higher in the columns with d even. The size of I_2 scales as the zero-th order moments, $\omega^{-1/2-1/2-1} = \omega^{-2}$. The convergence rate of the relative error is therefore 2 less than that shown for the absolute error.

Note that for $\omega = 1600$ and $d = 0$, only two function evaluations are required for an absolute error of $6.3E-9$ and a relative error of $1.4E-3$. The computation of the two weights in this case took less than a second of computation time. For comparison, a general purpose integration package was used on the same computer for the case $\omega = 10$, using polar coordinates: it took $100,000$ function evaluations to obtain an absolute error of $1E-7$. Assuming that the number of function evaluations scales at a cubic rate with respect to the frequency, due to the presence of oscillations in three dimensions, a comparable error for the case $\omega = 1600$ would require roughly 400 billion function evaluations.

6.4.5.3 A rectangular domain with critical points

We consider a two-dimensional integral with the more general oscillator that was used in §6.4.2.3,

$$I_2 := \int_0^{0.5} \int_0^1 \frac{1}{1+x+y} e^{i\omega(x^2-xy-y^2)} \, \mathrm{d}y \, \mathrm{d}x.$$

The stationary points in x are given by $x = y/2$, and the stationary points in y are given by $y = -x/2$. There is a critical point $(0, 0)$ where $\nabla g = 0$.

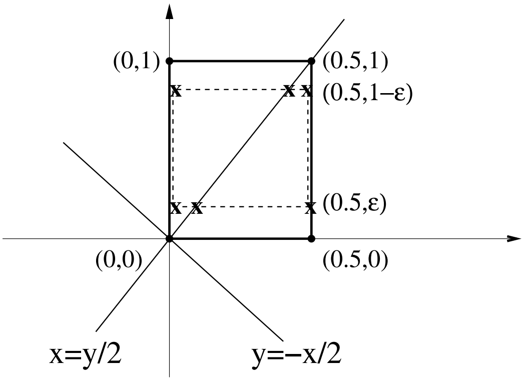


Figure 6.7: The points that contribute to the double integral I_2 for $g(x,y) = x^2 - xy - y^2$ on the rectangle, indicated by the •-s; the points that contribute to the integral over $[0, 0.5] \times [\epsilon, 1-\epsilon]$ are indicated by x-symbols.

The integration domain is illustrated in Figure 6.7. The contributions for this example integral come from the four corner points of the integration region: $(0,0)$, $(0.5,0)$, $(0.5,1)$ and $(0,1)$. The example was constructed however such that in the origin $(0,0)$ the conditions of Theorem 6.4.2 are not satisfied.

In order to find the decomposition, consider first the integral I_ϵ with the same integrand on $[0, 0.5] \times [\epsilon, 1-\epsilon]$. The decomposition of I_ϵ consists of 8 contributions associated with the six points indicated by x-symbols in Figure 6.7,

$$\begin{aligned} I_\epsilon \sim & F_{111}(0,\epsilon) - F_{111}(\epsilon/2,\epsilon) + F_{112}(\epsilon/2,\epsilon) - F_{112}(0.5,\epsilon) \\ & - F_{121}(0,1-\epsilon) + F_{121}(0.5-\epsilon/2, 1-\epsilon) \\ & - F_{122}(0.5-\epsilon/2, 1-\epsilon) + F_{122}(0.5, 1-\epsilon). \end{aligned}$$

The first index a in F_{abc} is the same for all contributions since it denotes the inverse of g with respect to y, which is unique on the integration domain. The second index b denotes $y = \epsilon$ when $b = 1$, and $y = 1-\epsilon$ when $b = 2$. The third index denotes the two inverses of g with respect to x, corresponding to the regions on the left and on the right of the line $x = y/2$ respectively. In the limiting case $\epsilon \to 0$, we find that

$$I_2 \sim F_{112}(0,0) - F_{112}(0.5,0) - F_{121}(0,1) + F_{121}(0.5,1).$$

The example was constructed such that the decomposition of the inner integral in y does not satisfy the conditions of Theorem 6.4.2. Indeed,

Table 6.11: Absolute error of the approximation of I_2 by a cubature rule using derivatives of maximal order d. The last row shows the value of $\log_2(e_{400}/e_{800})$. The theoretically predicted asymptotic lower bound is shown between parentheses. The rules in columns 1, 2 and 3 require 4, 12 and 24 function evaluations respectively.

$\omega \setminus d$	0	1	2
50	$1.4E-03$	$1.6E-04$	$1.5E-05$
100	$5.2E-04$	$4.0E-05$	$2.7E-06$
200	$1.9E-04$	$1.0E-05$	$4.8E-07$
400	$6.7E-05$	$2.6E-06$	$8.6E-08$
800	$2.4E-05$	$6.5E-07$	$1.5E-08$
rate	1.48 (1.5)	1.99 (2.0)	2.49 (2.5)

the boundary function $y = 0$ coincides with the curve of stationary points $y = -x/2$ in exactly one point $x = 0$. Following remark 6.4.3, we cannot evaluate the contribution $F_{112}(0, 0)$ using the path of steepest descent for y due to the presence of complex stationary points. However, an alternative integration path can be constructed. There are two disjunct regions where the integrand becomes arbitrarily small, corresponding to the two inverses of g with respect to y. These regions can be characterised by $g_1^{-1}(x, y + ip)$ for $p > 0, y \leq -x/2$ and $g_2^{-1}(x, y+ip)$ for $p > 0, y \geq -x/2$. The integration path for y at the point $(0.5, 0)$ leads to the latter region. The integration path for y at the point $(0, 0)$ should therefore lead to the same region: the line integral that connects the paths can then be discarded. An equivalent condition is that the imaginary part of $g(x, y)$ should be positive along the total integration path for y, including the discarded connecting part. For this particular example, we arbitrarily chose a linear path for y from $y_0 = 0$ to the point $y_1 = 1 - 1i$.

The results are shown in Table 6.11. Since $(0, 0)$ is a stationary point for both integration variables, the absolute error is the largest for the contribution of the origin. The size of the first discarded term scales as $O(\omega^{(-d-3)/2})$ by Theorem 6.4.10 and hence this is also the size of the absolute error. The convergence rate of the relative error is 1 less than the rate shown for the absolute error.

6.4.5.4 A degenerate critical point

In the final example, we consider a degenerate critical point. Consider the integral

$$I_{deg} := \int_{-1}^{1} \int_{-1}^{1} \frac{1}{3+x+y} e^{i\omega(x^3+y^3)} \, \mathrm{d}y \, \mathrm{d}x,$$

Table 6.12: Absolute error of the approximation of I_{deg} by a cubature rule using derivatives of maximal order d. The last row shows the value of $\log_2(e_{400}/e_{800})$. The theoretically predicted asymptotic lower bound is shown between parentheses. The rules in columns 1, 2 and 3 require 9, 27 and 54 function evaluations respectively.

$\omega \setminus d$	0	1	2
50	$5.3E-03$	$2.5E-04$	$2.9E-05$
100	$2.7E-03$	$9.9E-05$	$9.0E-06$
200	$1.3E-03$	$3.9E-05$	$2.8E-06$
400	$6.7E-04$	$1.6E-05$	$8.9E-07$
800	$3.4E-04$	$6.1E-06$	$2.9E-07$
rate	0.98 (1.0)	1.34 (1.33)	1.63 (1.66)

that has a stationary point of order $r = 2$ in both integration variables x and y at the origin $(0, 0)$. There are additional contributions from the corner points $(-1, -1)$, $(-1, 1)$, $(1, -1)$ and $(1, 1)$, and from the boundary points $(-1, 0)$, $(1, 0)$, $(0, -1)$ and $(0, 1)$. Cubature rules can be constructed using function evaluations at these 9 critical points.

The leading order of the size of I_{deg} as a function of ω is determined by the contribution of the degenerate critical point $(0, 0)$. Since $r = 2$ for x and for y, we expect that the value of $|I_{deg}|$ behaves as $O(\omega^{-1/3}\omega^{-1/3}) = O(\omega^{-2/3})$. The leading order of the error of the cubature rule behaves as $O(\omega^{-(2+d)/3})$. The theoretically predicted convergence rate is confirmed by the results in Table 6.12.

6.5 Conclusions

Traditionally, the evaluation of oscillatory integrals is considered a hard problem, because the number of operations usually increases linearly with the frequency. For multivariate integrals, the total number of operations increases more than linearly. The analysis of oscillatory integrals have led to a number of efficient methods that require only a fixed number of operations, and that yield increasingly accurate approximations for the integral. All methods exploit the same observations: the value of an oscillatory integral is determined largely by the behaviour of the integrand near the stationary points, and near the endpoints of the integration interval.

The same observations hold for the oscillatory integrals that arise in scattering problems. Moreover, they also hold for the integral equation itself. In the next chapter, we investigate the possibilities of exploiting these observations in the solution method.

Chapter 7

A hybrid asymptotic boundary element method

7.1 Introduction

The limitations of a traditional boundary element method for the solution of integral equations in the high frequency regime have become clear in the previous chapters. The method requires a large number of unknowns in order to represent the solution. The discretisation matrix is dense, and even with an efficient implementation of the high frequency fast multipole method or hierarchical matrix method, each matrix-vector product requires $O(N \log N)$ operations, with N proportional to k. It is at present still an open problem how the condition number of the system behaves for increasing wavenumber k. This dependence has a direct impact on the number of iterations required in iterative solution methods, and hence, on the total solution time. Likewise, it is unknown how the accuracy of the solution depends on the wavenumber. Similar to finite element methods for the Helmholtz equation, the boundary element method at high frequencies exhibits the so-called *pollution error*. One may regard the pollution error as an error in the numerical wavenumber of the discretisation; for a detailed discussion, in the context of finite element methods, see [12] and the references therein. The pollution error may require that N grows faster than linearly with k. A fundamental cause of these problems is that the typical piecewise polynomial basis functions in boundary element methods are not well adapted for oscillatory problems. For example, the solution space contains many functions that are completely impossible as wave solutions.

There are various other approaches for problems involving short wavelengths. At increasing frequencies, asymptotic methods become increasingly effective. The solution is expanded in an asymptotic expansion in terms of the small parameter $1/k$. Rather than solving for an oscillatory function $u(x)$, one can solve for the phase $\phi(x)$ and amplitude $A(x)$ such that $u(x) = A(x)e^{\phi(x)} + O(1/k^\alpha)$. This leads to the *eikonal equation* for the phase and *transport equation* for the amplitude. Both are nonlinear equations which can be solved numerically [163]. Linearised approximations yield raytracing methods, Geometrical Optics [27] and wavefront methods [195]. The *Physical Optics* approximation consists of using the Geometrical Optics approximation as the density function in the single layer potential. The *Geometrical Theory of Diffraction* was developed to extend asymptotic methods to model diffraction for a number of canonical problems [136, 22]. A disadvantage of these asymptotic methods is that the error is essentially uncontrollable. The higher order coefficients of the expansion are hard to obtain. Moreover, asymptotic expansions are not very flexible, especially for more complex geometries.

A new direction of research is the combination of finite element methods and asymptotic methods. This can be achieved by considering basis functions that incorporate the asymptotic form of the solution at large frequencies. This approach combines the fine error control of finite element methods with the accuracy of asymptotic methods for large frequencies. The asymptotic behaviour of the solution to the problem of scattering by smooth convex obstacles was analysed in [155]. Motivated by these results, a hybrid scheme was considered by Abboud et al. in [1]. The authors report an overall solution method that requires $O(k^{1/3})$ number of basis functions as a function of the wavenumber, a huge improvement over the linear dependence on k. The basis functions are piecewise polynomials, multiplied by plane waves in a number of directions. Similar hybrid methods with even better results are proposed in [2, 29, 141, 140, 92, 78]. A number of operations that is independent of the wavenumber, for a fixed error, is achieved by Bruno, Geuzaine, Monro and Reitich in [29] for the scattering by smooth convex obstacles, and by Langdon and Chandler-Wilde in [141] for scattering on a half-plane. Elements of the Geometric Theory of Diffraction are used in [92] to model diffraction. In this chapter, we combine a similar approach with the insights of Chapter 6 on the behaviour of oscillatory integrals. As a result, we obtain a small, and highly sparse discretisation matrix. In addition, the accuracy of a large part of the solution actually *increases* with increasing frequency.

7.2 Overview of the method

The hybrid asymptotic boundary element method is inspired by the progress in evaluation methods for oscillatory integrals. Recall the oscillatory integral

(6.1) from Chapter 6,

$$I[f] := \int_a^b f(x) e^{ikg(x)} \, \mathrm{d}x, \tag{7.1}$$

with f and g smooth functions. We will denote the frequency parameter by k in this chapter, for notational correspondence to the wavenumber of the Helmholtz equation. Several methods were discussed in Chapter 6 that yield increasingly accurate approximations for $I[f]$ for increasing k, using a fixed number of operations. It turns out that the value of the integral is determined by the behaviour of f and g near a set of contributing points - the endpoints and the stationary points.

We can apply these methods to the oscillatory integrals that arise in the discretisation of an oscillatory integral equation. Even more interesting however, is that we may apply the methods to the integral equation itself. Assume we know the phase of the kernel function of an integral equation, and, in addition, we know the phase of the *solution* of that equation. We can model a general integral equation satisfying these conditions as

$$(Aq)(x) = \int_0^1 G(x, y; k) e^{ikg_1(x,y)} q(y; k) e^{ikg_2(y)} \, \mathrm{d}y = f(x), \tag{7.2}$$

where both the kernel function $G(x, y; k)$ and the solution $q_s(y; k)$ are smooth functions that may still depend on k. For a fixed value of x, the integral in (7.2) closely resembles the model integral (7.1). The oscillator $g(y; x)$ for the integration variable y is found as the sum of the phase of the kernel function and the phase of the solution, $g(y; x) = g_1(x, y) + g_2(y)$. The value of the integral $(Aq)(x)$ is determined by the behaviour of the integrand near the endpoints and the stationary points of g. Now consider an approximation of the form $q_c(y) = \sum_j c_j \phi_j(y)$ of the solution $q(y)$, in a collocation approach with collocation points x_i. Each value $(Aq)(x_i)$ is determined by the behaviour of $G(x, y; k)$ and $q_c(y)$ near the endpoints of $[0, 1]$ and the stationary points of $g(y; x_i)$. The behaviour of the kernel function is known. If the basis functions $\phi_j(y)$ have compact support, then the behaviour of $q_c(y)$ is determined by only a small number of coefficients c_j, corresponding to the basis functions that are nonzero in the contributing points. For each particular fixed value x_i, it does not matter what the value of the other coefficients is in the approximation of q. An immediate consequence of this insight is that it should be possible to derive a *sparse discretisation matrix* for equation (7.2) for large values of k.

There are still some issues associated with this approach. First, the phase of the solution is not known for general problems. For scattering problems however, it is known for the case of a smooth and convex scattering obstacle. In that case, the phase of the solution is asymptotically determined by the phase of the boundary condition. Second, there is an important difference between (7.2) and (7.1): the smooth amplitude function in (7.2) depends on k. We will see that this dependence influences the convergence of Filon-type methods for the evaluation of integral (7.2).

We commence in §7.3 by generalising the results of Filon-type methods in the previous chapter to model integrals that resemble (7.2) for the scattering problems of interest. We formulate a suitable high frequency boundary integral equation in §7.4, and examine the asymptotic behaviour of the solution. The efficient quadrature rules are used to obtain a sparse discretisation matrix for the integral equation in §7.5. We examine the convergence of the quadrature rules as a function of k. Finally, we illustrate the new method with numerical results in §7.6.

7.3 Specialised quadrature rules

Filon-type methods for the evaluation of oscillatory integrals lead to a quadrature rule involving derivatives. The quadrature rules that were derived in Chapter 6 only apply to the model integral (7.1). Intuitively however, one sees that the ideas underlying the method can be readily generalised to any oscillatory integral. The value of an oscillatory integral is determined by the behaviour of the integrand near the endpoints of the integration interval, and near the points where the integrand locally does not oscillate. In order to construct similar quadrature rules for more general integrals, one requires knowledge of the phase of the integral. In this section, we will construct such rules for a family of integrals that will arise in the scattering problem discussed later. In particular, the integrand involves an oscillatory Hankel function.

7.3.1 A generalised model form

Consider the oscillatory integral

$$I_H[f] = \int_a^b f(x) H_\nu^{(1)}(k g_1(x)) e^{ikg_2(x)} \, \mathrm{d}x, \tag{7.3}$$

where f, g_1 and g_2 are smooth functions, and $H_\nu^{(1)}(z)$ is the Hankel function of the first kind of order ν. The Hankel function of order zero $H_0^{(1)}(z)$ has a logarithmic singularity at $z = 0$. Hankel functions of higher order have algebraic singularities of the form $1/z^\nu$, $z \to 0$ [4].

For large arguments, the Hankel functions behave like an oscillatory complex exponential with a decaying amplitude,

$$H_\nu^{(1)}(z) \sim \sqrt{\frac{2}{\pi z}} e^{i(z - \frac{1}{2}\nu\pi - 1/4\pi)}, \qquad -\pi < \arg z < 2\pi, \quad |z| \to \infty. \tag{7.4}$$

Hence, the oscillator of the integrand of (7.3) is approximately given by

$$g(x) = g_1(x) + g_2(x), \tag{7.5}$$

up to the addition of a constant. The Hankel function decays exponentially fast for complex arguments with a positive imaginary part, as can be seen from the asymptotic behaviour (7.4). This means that the approach of §6.3 using the path of steepest descent is applicable. Hence, we may conjecture that a quadrature rule exists of the form

$$I_H[f] \approx Q_H[f] := \sum_{i=0}^{l} \sum_{j=0}^{d_i} w_{i,j}^H f^{(j)}(x_i). \tag{7.6}$$

In the remainder of the section, we will prove this conjecture, determine the quadrature abscissae x_i and show how the weights can be computed efficiently.

7.3.2 Construction of the quadrature rule

We start by stating some assumptions on the functions f and g. These assumptions are a.o. needed to guarantee the integrability and analyticity of the integrand in (7.3). First, we assume that f is analytic in an open complex neighbourhood D of $[a, b]$, so that $[a, b] \subset \operatorname{int} D$. Likewise, we assume that g_1 and g_2 are non-singular and analytic in D, except possibly along a branch cut that extends from a or b to the boundary of the region D, i.e., a and b may be branch points but not singular points. We assume furthermore that $g(x)$, defined by (7.5), is strictly monotonic on the open interval (a, b) and hence invertible, but possibly $g'(a) = 0$ or $g'(b) = 0$. Also, we assume that $g_1(x) \neq 0$, $x \in (a, b)$. Finally, if $\nu > 0$ and $g_1(\xi) = 0$ we assume that f behaves like

$$f(x) \sim (x - \xi)^{\nu - 1 + \epsilon}, \quad x \to 0, \quad \text{with} \quad \epsilon > 0. \tag{7.7}$$

Condition (7.7) guarantees that the integrand of $I_H[f]$ is integrable. Subject only to condition (7.7) and the analyticity requirements, the integration interval of (7.3) can always be split into a number of subintervals that satisfy the conditions. The assumptions guarantee that the integrand of $I_H[f]$ is analytic on $[a, b]$ except possibly in the points a and b. In particular, this will allow us to apply Cauchy's integral theorem to select the integration path of (7.3).

Theorem 7.3.1. *Under the assumptions stated above, the integral $I_H[f]$ can be approximated by a sum of contributions*

$$I_H[f] = S^H[f; a] - S^H[f; b] + O(e^{-kd_0}), \tag{7.8}$$

with $d_0 > 0$*, and with the contributions given by the integrals*

$$S^H[f;x] = \int_0^P f(h_x(p)) H_\nu^{(1)}(kg_1(h_x(p))) e^{ikg_2(h_x(p))} h_x'(p)\,\mathrm{d}p, \tag{7.9}$$

where $h_x(p)$ *satisfies*

$$g(h_x(p)) = g(x) + ip. \tag{7.10}$$

The proof is almost identical to the proof of Lemma 6.4.1 and is omitted; it differs mainly in the special treatment of the Hankel function based on the asymptotic expression (7.4).

We note from the asymptotic behaviour (7.4) that the integrand of the line integral $S^H[f;x]$ in (7.9) is non-oscillatory and exponentially decaying in the integration variable p,

$$H_\nu^{(1)}(kg_1(h_x(p))) e^{ikg_2(h_x(p))} \sim \sqrt{\frac{2}{\pi k g_1(h_x(p))}} e^{ikg(x)} e^{i(-\frac{1}{2}\nu\pi - 1/4\pi)} e^{-kp},$$

for $k \to \infty$. The size of the constant d_0 is related to the size of the region of analyticity of f and g (see Theorem 6.3.2). In the numerical examples of the scattering problem, given in §7.6, we will be able to choose the limit case $P = \infty$. The error of decomposition (7.8) then vanishes even at low frequencies.

We proceed in a similar way as for the construction of the localised Filon-type method in §6.3.6. Since f is analytic in D, it has an absolutely convergent Taylor series. By the linearity of S^H, we may write

$$S^H[f;x_0] = \sum_{j=0}^{\infty} f^{(j)}(x_0) S^H\left[\frac{(x-x_0)^j}{j!};x_0\right].$$

Now, consider a subdivision of $[a,b]$ into subintervals $[a_i,b_i]$, $i = 1,\ldots,l$, such that on each subinterval the conditions of Theorem 7.3.1 are satisfied. Truncating the Taylor series of f at each special point a_i and b_i after a finite number of terms, we arrive at a quadrature rule $Q_H[f]$ of the form (7.6), with weights given by

$$w_{0,j}^H = S_1^H\left[\frac{(x-a)^j}{j!};a\right], \tag{7.11}$$

$$w_{i,j}^H = -S_i^H\left[\frac{(x-x_i)^j}{j!};x_i\right] + S_{i+1}^H\left[\frac{(x-x_i)^j}{j!};x_i\right], \quad i = 1,\ldots,l-1, \tag{7.12}$$

$$w_{l,j}^H = -S_l^H\left[\frac{(x-b)^j}{j!};b\right]. \tag{7.13}$$

The weights can be explicitly computed very efficiently, by using Gauss-Laguerre quadrature or similar techniques. The accuracy of these methods improves rapidly as a function of k, due to the faster decay of the integrands as k increases. For the purposes of our application, this advantageous characteristic is not even needed. It suffices already that the number of operations for a fixed accuracy is bounded with respect to k. We therefore choose to focus on the convergence properties of the quadrature rule itself, rather than on the convergence of methods to compute the weights.

7.3.3 Convergence properties of the quadrature rule

We discuss the properties of the specialised quadrature rule $Q_H[f]$ as a function of k. It is clear that the rule is exact by construction for polynomials of degree less than or equal to

$$p = \min_i d_i. \tag{7.14}$$

For more general functions, the accuracy as a function of k is determined by the asymptotic size of the weights. We will show that the size of the weights decreases both with increasing frequency and with increasing order of the corresponding derivative. The order of accuracy of the quadrature rule is therefore equal to the asymptotic size of the first weight that is discarded by truncation. In order to quantify this size, we require a few technical lemmas.

Lemma 7.3.2. *Assume x_0 is a stationary point that has order r. The parameterisation of the path* (7.10) *behaves as*

$$h_{x_0}(p) = x_0 + O(p^{1/(r+1)}), \qquad p \to 0, \tag{7.15}$$

$$h'_{x_0}(p) = O(p^{1/(r+1)-1}), \qquad p \to 0. \tag{7.16}$$

Proof. Since $g^{(j)}(x_0) = 0$, $j = 1, \ldots, r$, we can write the Taylor series of g as

$$g(x) = g(x_0) + g^{(r+1)}(x_0)\frac{(x-x_0)^{r+1}}{(r+1)!} + O((x-x_0)^{r+2}).$$

The path $h_{x_0}(p) = g^{-1}(g(x) + ip)$ solves $g(h_{x_0}(p)) = g(x) + ip$, and hence

$$h_{x_0}(p) \sim x_0 + \sqrt[r+1]{\frac{ip(r+1)!}{g^{(r+1)}(x_0)}}, \qquad p \to 0. \tag{7.17}$$

The second result follows by differentiation. Note that the complex root is multi-valued: the correct root is selected by using the analytic continuation of the inverse g_i^{-1} that satisfies $g_i^{-1}(g(x)) = x$ on $[a_i, b_i]$ in expression (7.10). $\square$

The size of the weights follows from the size of the line integrals $S^H\left[\frac{(x-x_0)^j}{j!};x\right]$. Recall that the integral may be singular if $g_1(x)=0$. We will assume for the sake of brevity that, in that case, $g_1'(x)\neq 0$. This condition is always satisfied by the applications in §7.6.

Lemma 7.3.3. *Let $S^H[f;x]$ be defined by (7.9) with $P=\infty$, and g defined by (7.5). Assume that $g'(x_0)\neq 0$, i.e., x_0 is not a stationary point. If $g_1(x_0)\neq 0$, we have*

$$\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-j-3/2}), \qquad k\to\infty.$$

If $g_1(x_0)=0$ and $g_1'(x_0)\neq 0$, the integral is singular and we have

$$\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-j-1}), \qquad j\geq\nu, \quad k\to\infty.$$

Proof. We write the integral $S^H[(x-x_0)^j;x_0]$ as

$$S^H[(x-x_0)^j;x_0] = \int_0^\infty u(p)e^{-kp}\,\mathrm{d}p = \frac{1}{k}\int_0^\infty u(q/k)e^{-q}\,\mathrm{d}q, \tag{7.18}$$

with

$$u(p) = (h_{x_0}(p)-x_0)^j h_{x_0}'(p)H_\nu^{(1)}(kg_1(h_{x_0}(p)))e^{ikg_2(h_{x_0}(p))}e^{kp}. \tag{7.19}$$

It is a consequence of Watson's Lemma that the asymptotic expansion of the integral can be obtained by integrating the asymptotic expansion of $\frac{1}{k}u(q/k)$ as $k\to\infty$ term by term in (7.18) [21, 200]. Generalising Watson's Lemma, this remains true for integrals of the form $\int_0^\infty u(p)h(kp)\,\mathrm{d}p$ where $h(z)\sim\log(z)^n z^s e^{-z}$, $n\geq 0$, $s\in\mathbb{Z}$, $z\to 0$, if the integrand is integrable [20]. This means that the singularity of the Hankel function has no influence on the asymptotic expansion.

First, consider the case $g_1(x_0)\neq 0$. Then, combining the asymptotic behaviour of the Hankel function for large arguments (7.4) with the results (7.15)-(7.16) of Lemma 7.3.2 for $r=0$, we have $u(q/k)\sim k^{-j-1/2}$. From (7.18) we can conclude $\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-j-3/2})$.

Next, consider the case $g_1(x_0)=0$. If $g_1'(x_0)\neq 0$, then we have $g_1(h_{x_0}(p))\sim p^{1/(r+1)}=p$. It follows that $H_\nu^{(1)}(kg_1(h_{x_0}(q/k)))=O(1)$, $k\to\infty$. Hence, by the generalisation of Watson's Lemma, we may conclude $\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-j-1})$, $j\geq\nu$. ☐

The corresponding lemma for stationary points is very similar. The difference is due to the different behaviour of the parameterisation as described by Lemma 7.3.2.

Lemma 7.3.4. *Let $S^H[f;x]$ be defined by* (7.9) *with $P=\infty$, and g defined by* (7.5)*. Assume that x_0 is a stationary point of order r. If $g_1(x_0) \neq 0$, then we have*

$$\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-(j+1)/(r+1)-1/2}), \qquad k \to \infty.$$

If $g_1(x_0)=0$ and $g_1'(x_0) \neq 0$ then we have

$$\left|S^H[(x-x_0)^j;x_0]\right| = O(k^{-(j+1+r/2)/(r+1)}), \qquad j \geq \nu, \quad k \to \infty.$$

Proof. Consider again the function $u(p)$, given by (7.19). Assume first that $g_1(x_0) \neq 0$. We have $(h_{x_0}(q/k)-x_0)^j \sim k^{-j/(r+1)}$ and $h_{x_0}'(q/k) \sim k^{r/(r+1)}$. Since $kg_1(h_{x_0}(q/k)) \sim kg_1(h_{x_0}(0)) = kg_1(x_0)$, we also have $H_\nu^{(1)}(kg_1(h_{x_0}(q/k))) \sim k^{-1/2}$. Combined in (7.18), and by applying Watson's Lemma, this yields the first result.

The case where $g_1(x_0)=0$ is slightly different. Since $g_1'(x_0) \neq 0$, we have $g_1(h_{x_0}(q/k)) \sim k^{-1/(r+1)}$ and, hence, $kg_1(h_{x_0}(q/k)) \sim k^{r/(r+1)}$. The Hankel function therefore yields the factor $k^{-(r/2)/(r+1)}$ instead of $k^{-1/2}$ as in the first case. ☐

The convergence of the quadrature rule (7.6) as a function of k can now be established. Note that the results of Lemma 7.3.3 agree with those of Lemma 7.3.4 if we take the order of a regular point to be $r=0$. Hence, we need not distinguish between stationary points and regular (end)points.

Lemma 7.3.5. *The error of the approximation of $S_i^H[f;x_0]$ by $Q_i^S[f;x_0]$, for $x_0 \in [a_i,b_i]$, is*

$$S_i^H[f;x_0] - Q_i^S[f;x_0] = S_i^H[f;x_0] - \sum_{j=0}^{d_i} w_{i,j}^H f^{(j)}(x_0) = O(k^{-\alpha_i}), \tag{7.20}$$

for $k \to \infty$. If $g_1(x_0) \neq 0$ then $\alpha_i := (d_i+2)/(r+1)-1/2$. If $g_1(x_0)=0$ and $g_1'(x_0) \neq 0$, then $\alpha_i := (d_i+2+r/2)/(r+1)$.

Proof. Since the weights decay as a function of k, and as a function of the order of derivative j, the error of the quadrature scheme is asymptotically determined by the size of the first discarded weight. The result follows from Lemma's 7.3.3 and 7.3.4 by setting $j=d_i+1$. ☐

The theorem that characterises the accuracy of the complete quadrature rule follows immediately.

Theorem 7.3.6. *Consider the approximation of $I_H[f]$ by $Q_H[f]$. The error has asymptotic order $\alpha-1$ with $\alpha = \min_i \alpha_i$, and where α_i is specified in Lemma 7.3.5.*

Table 7.1: Absolute error of the approximation of $I_H[f]$ by $Q_H[f]$, with $f(x) = (x-1)$, $g_1(x) = x$ and $g_2(x) = x^2 + x^3 - x$. The last row shows the value of $\log_2(e_{400}/e_{800})$: this value should approximate $d_0/2 + 5/4$.

$k \setminus d_0$	0	1	2	3
100	$1.2E-3$	$2.8E-5$	$1.3E-6$	$2.6E-8$
200	$5.1E-4$	$8.6E-6$	$2.9E-7$	$4.1E-9$
400	$2.2E-4$	$2.6E-6$	$6.4E-8$	$6.2E-10$
800	$9.3E-5$	$7.8.1E-7$	$1.4E-8$	$9.7E-11$
rate	1.23	1.73	2.20	2.68

Consider the integral $\int_0^1 \cos(x-1) H_0^{(1)}(kx) e^{ik(x^2+x^3-x)}\,\mathrm{d}x$. The total oscillator for this integral is $g(x) = x^2 + x^3$. There are two quadrature points: there is a singularity and a stationary point of order 1 at $x = 0$, and a regular endpoint at $x = 1$. The weights $w_{0,j}^H$ and $w_{1,j}^H$ are given by (7.11) and (7.13) respectively. From Lemma 7.3.4 we have $|w_{0,j}^H| = O(k^{-(j+1)/2-1/4})$ and from Lemma 7.3.3 we have $|w_{1,j}^H| = O(k^{-j-3/2})$. Using d_0 and d_1 derivatives, the error has order $\min\{O(k^{-(d_0+2)/2-1/4}), O(k^{-(d_1+1)-3/2})\}$ by Theorem 7.3.6. We choose $d_1 = \max\{0, \lceil (2d_0-5)/4 \rceil\}$ to match the errors. Table 7.1 shows the convergence of the quadrature rule $Q_H[f]$ as a function of k and d_0. The integration error was determined by comparison with the results of Cubpack [50].

7.4 High frequency scattering problems

7.4.1 High frequency integral equation formulation

We will describe the method for the exterior problem of the two-dimensional Helmholtz equation, subject to a Dirichlet boundary condition on the boundary. We find a solution in terms of the single-layer potential

$$(Sq)(x) = \int_\Gamma \frac{i}{4} H_0^{(1)}(k|x-y|) q(y)\,\mathrm{d}s_y, \tag{7.21}$$

where q is the density function defined on the boundary Γ of the scattering obstacle. The density function q is found as the solution to the integral equation (2.47) of the first kind,

$$(Sq)(x) = u^i(x), \qquad x \in \Gamma. \tag{7.22}$$

The same method applies to the combined field integral equation (2.56) that is solvable for all values of the wavenumber. An important observation is that $q = \frac{\partial u}{\partial n}$, i.e., the density function is exactly the (exterior) normal derivative of the total solution $u = u^i + u^s$. This means that the solution to equation (7.22) is directly related to a physical property of the problem. For example, in electromagnetics, the normal derivative of the electric field is proportional to the induced current on the surface of the conducting obstacle [159].

The density function q is highly oscillatory for large values of the wavenumber k. The solution of equation (7.22) therefore requires, generally, a large number of unknowns. In some cases however, one has a priori information about the phase of the solution. For example, if the obstacle is convex, and if the incoming wave is a plane wave, then the phase of the solution q is approximately the same as the phase of the incoming wave. Assume the incoming wave is $u^i(x) = u^i_s(x)e^{ikg^i(x)}$. Then we can write

$$q(x) = q_s(x)e^{ikg^i(x)}, \qquad x \in \Gamma, \tag{7.23}$$

where $q_s(x)$ is a non-oscillatory function, at least approximately. In physical terms, the oscillations of the induced current on a perfectly conducting surface tend to follow the oscillations of the incoming electromagnetic wave. This is the reason why the problem should be formulated such that the solution q corresponds to a physical variable - only in that case is the phase known in the form of (7.23). This was noted in [29]; the integral equation formulation of this section follows the same pattern as [29].

Substituting (7.23) in (7.22) yields

$$\int_\Gamma \frac{i}{4} H_0^{(1)}(k|x-y|) q_s(y) e^{ikg^i(y)} \, \mathrm{d}s_y = u^i(x) = u^i_s(x) e^{ikg^i(x)}.$$

Dividing by the oscillatory factor in the right hand side, and introducing a periodic parameterisation $\kappa : [0,1] \to \Gamma$ for Γ, we have the integral equation

$$\int_0^1 \frac{i}{4} H_0^{(1)}(k|x-\kappa(\tau)|) e^{ik(g^i(\kappa(\tau))-g^i(x))} |\nabla \kappa(\tau)| q_s(\tau) \, \mathrm{d}\tau = u^i_s(x). \tag{7.24}$$

The unknown in (7.24) is $q_s(\tau)$, defined in the parameter domain for simplicity. The unknown is non-oscillatory, and one can therefore solve (7.24) using a coarse discretisation for $q_s(\tau)$.

7.4.2 Asymptotic behaviour of the solution

In the past decades, a lot of effort has been invested in studying the asymptotic behaviour of the solution q to (7.22) as a function of the wavenumber, concentrating mainly on the scattering of a plane wave (see, e.g., [136, 155] and references therein). For smooth and convex obstacles, there are three

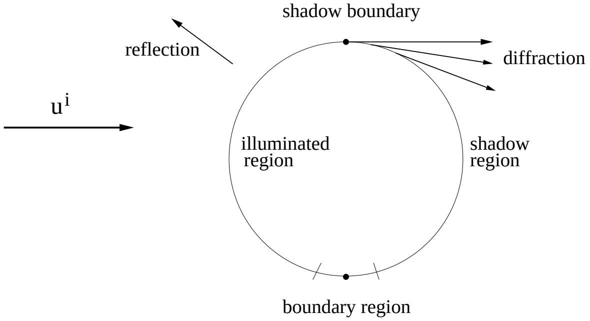


Figure 7.1: Reflection and diffraction effects in the scattering of an incoming wave u^i by a smooth and convex obstacle.

important regions with different properties, illustrated in Figure 7.1: the illuminated region, the shadow region, and the transitional shadow boundary region. In the illuminated region, the scattered wave is described asymptotically by geometrical optics: a wave is reflected such that the angle of incidence and the angle of reflection are identical. A wave tangential to the shadow boundary causes diffraction. The density function decays rapidly away from the shadow boundary into the shadow region, due to the continuous emission of diffracted waves. In the deep shadow region, the density function vanishes.

The asymptotic behaviour of q_s reflects these three regions. Assume an incoming plane wave in the direction α of the form $u^i(x) = e^{ik\alpha\cdot x}$ with $|\alpha| = 1$. It was proved in [155] that q_s has an asymptotic expansion, for $\tau \in [0,1]$, of the form

$$q_s(\tau) \sim \sum_{m,n\geq 0} k^{2/3-n-2m/3} b_{m,n}(\alpha,\tau)\Psi^{(n)}(k^{1/3}Z(\alpha,\tau)). \tag{7.25}$$

We recall the main characteristics of this expansion that are needed in our analysis. For a more complete discussion, we refer to [155]. The function $Z \in C^\infty$ is infinitely smooth, and has a simple root at the two shadow boundary points. The shadow boundary points are characterised by $\alpha \cdot \nu = 0$, with ν the exterior normal to Ω. The function Z is positive when $\alpha \cdot \nu < 0$, i.e., in the illuminated region, and negative in the shadow region. The function $\Psi(z)$ is smooth for positive arguments, with

$$\Psi(z) \sim z, \qquad z \to \infty, \tag{7.26}$$

and exponentially but oscillatory decaying for large negative arguments. This means that, as $k \to \infty$, we can derive the asymptotic properties from the leading order of (7.25),

$$|q_s(\tau)| = \begin{cases} O(k), & \text{illuminated region,} \\ O(k^{2/3}), & \text{shadow boundary,} \\ O(e^{-k^{1/3}d(\tau)}), & \text{shadow region.} \end{cases} \tag{7.27}$$

There is a k-dependent transition region near the shadow-boundary, since the order of the size of q_s changes smoothly from 1 to 2/3. The behaviour in the shadow region is also k-dependent, as the solution is oscillatory decaying. Motivated by (7.25), we introduce a transition region of size $O(k^{-1/3})$ around the shadow boundary points, and define the shadow boundary regions as

$$T_{B1}(k) = [t_{sb1} - D_1 k^{-1/3}, t_{sb1} + C_1 k^{-1/3}], \tag{7.28}$$

$$T_{B2}(k) = [t_{sb2} - C_2 k^{-1/3}, t_{sb2} + D_2 k^{-1/3}], \tag{7.29}$$

with constants $C_1, C_2, D_1, D_2 > 0$ independent of k, but small enough such that $T_{B1}(k)$ and $T_{B2}(k)$ are non overlapping, and with t_{sb1} and t_{sb2} the locations of the two shadow boundary points in the parameter domain $[0, 1]$. The illuminated region is defined as

$$T_I(k) = (t_{sb1} + C_1 k^{-1/3}, t_{sb2} - C_2 k^{-1/3}). \tag{7.30}$$

The shadow region is the remaining part of the interval $[0, 1]$.

The size of the transition region is related to the behaviour of the argument $k^{1/3} Z(\omega, \tau)$ of the function $\Psi^{(n)}$ in (7.25). Since $Z(\omega, \tau)$ has a simple zero at t_{sb1}, we have

$$Z(\omega, \tau) \approx Z'(\omega, t_{sb1})(\tau - t_{sb1}), \qquad \tau \to t_{sb1}.$$

Hence, $|k^{1/3} Z(\omega, \tau)| = O(1)$ for $\tau \in T_{B1}(k)$. We can therefore state

$$|q_s(\tau)| = \begin{cases} O(k), & \tau \in T_I(k), \\ O(k^{2/3}), & \tau \in T_{B1}(k) \cup T_{B2}(k), \\ O(e^{-k^{1/3}d(\tau)}), & \tau \in [0, 1] \setminus (T_I(k) \cup T_{B1}(k) \cup T_{B2}(k)) . \end{cases} \tag{7.31}$$

We also state the size of the first order derivative for further reference,

$$|q_s'(\tau)| = \begin{cases} O(k), & \tau \in T_I(k), \\ O(k), & \tau \in T_{B1}(k) \cup T_{B2}(k). \end{cases} \tag{7.32}$$

7.5 A high frequency boundary element method

The collocation of integral equation (7.24) in a point x_n leads to a one-dimensional and oscillatory integral in the integration variable τ. In this section, we show how an efficient quadrature rule can be used for the discretisation of that collocation integral. First, we discuss both the classical boundary element approach and a discretisation based on the quadrature rule in §7.5.1. We show that the quadrature rule can not always be applied in §7.5.2, and we arrive at a method combining both approaches in §7.5.3.

7.5.1 Collocation approach for the discretisation

Consider a collocation scheme for integral equation (7.24), with a set of N distinct collocation points $x_n = \kappa(t_n)$, $t_n \in [0,1]$, $n = 1, \dots, N$. The classical way to proceed is to look for an approximation q_c to solution q_s in the form

$$q_c(\tau) = \sum_{m=1}^{N} c_m \phi_m(\tau), \tag{7.33}$$

where the ϕ_m functions are a set of linearly independent basis functions with support $\Omega_m := \operatorname{supp}(\phi_m)$. The number of basis functions may be small, since the exact solution q_s is not oscillatory. Collocating equation (7.24), with q_s replaced by q_c, in the points t_n leads to the equations

$$\int_0^1 \frac{i}{4} H_0^{(1)}(k|\kappa(t_n) - \kappa(\tau)|) e^{ik(g^i(\kappa(\tau)) - g^i(\kappa(t_n)))} |\nabla \kappa(\tau)| q_c(\tau)\, \mathrm{d}\tau = u_s^i(x_n), \tag{7.34}$$

for $n = 1, \dots, N$. The collocation approach therefore leads to a linear system $Ac = b$ of size $N \times N$, where the elements of the discretisation matrix A are given by

$$A_{n,m} = \int_{\Omega_m} \frac{i}{4} H_0^{(1)}(k|\kappa(t_n) - \kappa(\tau)|) e^{ik(g^i(\kappa(\tau)) - g^i(\kappa(t_n)))} |\nabla \kappa(\tau)| \phi_m(\tau)\, \mathrm{d}\tau, \tag{7.35}$$

and the right hand side by $b_n = u_s^i(x_n)$. The discretisation matrix A is dense, but small compared to the classical boundary element discretisation for the original equation. Hence, this is a big improvement over the direct discretisation of (7.22). Since the elements $A_{n,m}$ are given by oscillatory integrals in (7.35), they can be computed efficiently using the numerical steepest descent technique described in §6.3. This yields an efficient total solution method, that remains efficient when k increases.

However, there are still some issues associated with this approach. Since the matrix is dense, the method requires the evaluation of N^2 integrals. Although N may be rather small, the computational cost can still be high.

Interestingly, it was observed in [91] that many of the elements are small, and can in fact be discarded, reducing the computation time. A second, and more important issue is that the results of [91] indicate that the error of the scheme increases with increasing wavenumber. Here, we examine a different discretisation of (7.34) that aims to address these issues, based on the quadrature rule developed in §7.3 and motivated by the accuracy of the rules for high wavenumbers. Owing to the small number of required quadrature points, the resulting discretisation matrix will be highly sparse.

Based on the collocation integral (7.34), we define the oscillatory integral

$$I_c[f;t_n] := \int_0^1 \frac{i}{4} H_0^{(1)}(k|\kappa(t_n)-\kappa(\tau)|) e^{ik(g^i(\kappa(\tau))-g^i(\kappa(t_n)))} |\nabla\kappa(\tau)| f(\tau)\, \mathrm{d}\tau. \tag{7.36}$$

The integral $I_c[f;t_n]$ is highly similar to the model integral $I_H[f]$ that was introduced in §7.3. In particular, one can find a quadrature rule such that

$$I_c[f;t_n] \approx Q_c[f;t_n] := \sum_{i=0}^{l_n} \sum_{j=0}^{d_{n,i}} w_{n,i,j}^c f^{(j)}(\tau_{n,i}). \tag{7.37}$$

The weights are found by evaluating line integrals $S^c[f;t_n]$ in the complex plane similar to $S^H[f]$. A difference of $Q_c[f;t_n]$ compared to $Q_H[f]$ is the additional factor $\frac{i}{4}|\nabla\kappa(\tau)|$ in the integrand. Assuming an analytic parameterisation κ, this factor can simply be included in the weight function of the rule. Hence, the construction and the convergence properties of $Q_c[f;t_n]$ are described by the corresponding analysis for $Q_H[f]$ in §7.3. The weights $w_{n,i,j}^c$ depend on k and on t_n, the constants l_n and $d_{n,i}$ and the points $\tau_{n,i}$ depend on t_n only.

The quadrature points are found from the oscillator. The oscillator of (7.36) is known explicitly; it is given by

$$\begin{aligned} g(\tau;t_n) =& \sqrt{(\kappa_1(t_n)-\kappa_1(\tau))^2 + (\kappa_2(t_n)-\kappa_2(\tau))^2} \\ &+ g^i(\kappa_1(\tau),\kappa_2(\tau)) - g^i(\kappa_1(t_n),\kappa_2(t_n)), \end{aligned} \tag{7.38}$$

with $\kappa(t) = (\kappa_1(t),\kappa_2(t))$. The quadrature points $\tau_{n,i}$ of $Q_c[f;t_n]$ are the points τ where the integrand becomes singular (and hence non-analytic), and the stationary points of the oscillator $g(\tau;t_n)$. These points are derived by a straightforward, but technical analysis of $g(\tau;t_n)$. There are no contributing endpoints, as the integrand is periodic on the closed curve Γ. The location of the quadrature points is illustrated in Figure 7.2 for the scattering of a plane wave by a circular obstacle. There is one stationary point if t_n lies in the illuminated region, and there are three stationary points if t_n lies

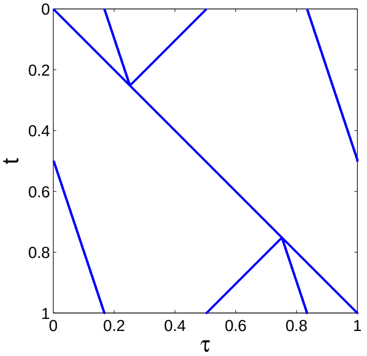


Figure 7.2: The location of the contributing points of the collocation integral for the scattering of a plane wave by a circular obstacle. Each row corresponds to a fixed value $t \in [0, 1]$, with the shadow boundary points at 0.25 and 0.75, and the illuminated region in between. The singular points are located along the diagonal. The remaining points correspond to stationary points.

in the shadow region. Two of these points coalesce into one stationary point of order $r = 2$ exactly at the shadow boundary. Figure 7.2 is illustrative for more general convex shapes and incoming waves. If the incoming wave is not a plane wave however, then the point where two stationary points coalesce may differ from the point where the incoming wave is tangential to the boundary. In the following, we consistently use the term *shadow boundary* to refer to the point where two stationary points coalesce into one, as that point determines the numerical properties of the scheme.

We now describe how the quadrature rule (7.37) can be used in the discretisation. The derivatives of q_c can be written in terms of the basis functions ϕ_m,

$$q_c^{(j)}(\tau) = \sum_{m=1}^{N} c_m \phi_m^{(j)}(\tau).$$

Hence, applying the quadrature rule to q_c yields a matrix B with entries

$$B_{n,m} = \begin{cases} \sum_{i:\tau_{n,i} \in \Omega_m} \sum_{j=0}^{d_{n,i}} w_{n,i,j}^c \phi_m^{(j)}(\tau_{n,i}), & \exists i \in [0, l_n] : \tau_{n,i} \in \Omega_m, \\ 0 & \text{otherwise} \end{cases} \tag{7.39}$$

The entry $B_{n,m}$ is nonzero only if at least one quadrature point $\tau_{n,i}$ exists that lies in the support of the basis function ϕ_m. The number of nonzero points therefore depends on the size of the supports of the basis functions. If all basis functions are local, then the structure of B will resemble the structure shown in Figure 7.2.

However, the value $Q_c[q_c; t_n]$ is not always a good approximation for the collocation integral (7.34). It turns out that the quadrature rule can not always be used, depending on the collocation point t_n. Therefore, we first examine the accuracy of $Q_c[q_c; t_n]$. We will formulate a combined approach in §7.5.3 that uses the quadrature rule only where it is sufficiently accurate.

7.5.2 Convergence of the specialised quadrature rule as a function of k

The convergence of quadrature rule $Q_c[f; t_n]$ as a function of k can be derived from the results for $Q_H[f]$ discussed in §7.3.3. These results were derived with the assumption that the function f is independent of k. The solution q_s of integral equation (7.24) depends on k; an important issue in the derivation of our hybrid scheme is what can be said about the accuracy of the quadrature rule $Q_c[q_s; t_n]$ when applied to the exact solution q_s. As it turns out, in that case, the accuracy of the rule depends on the location of the quadrature points.

First, we consider the case of a function that is independent of k. We will show that for a given collocation point t_n and a given value of i, the quadrature point $\tau_{n,i}$ yields a contribution to the value of Q_c in (7.37) with an accuracy that increases with k. Define the partial sums

$$S(d; f, n, i) = \sum_{j=0}^{d} w^c_{n,i,j} f^{(j)}(\tau_{n,i}).$$

The definition of S is such that one can write $Q_c[f; t_n] = \sum_{i=0}^{l_n} S(d_{n,i}; f, n, i)$, where $l_n + 1$ is the total number of quadrature points for a given collocation point t_n. The partial sum $S(d_{n,i}; f, n, i)$ may be regarded as the contribution of the quadrature point $\tau_{n,i}$ to the total value of the quadrature approximation. This local contribution converges to a fixed value $s(k)$ for increasing d, and the convergence becomes faster with increasing k. In order to see this, consider a point $\tau_{n,i}$ that is not a stationary point. By Lemma 7.3.3, we have $w^c_{n,i,j} = O(k^{-j-3/2})$. If f is independent of k, it follows that

$$|S(d+1; f, n, i) - S(d; f, n, i)| = O(k^{-(d+1)-3/2}). \tag{7.40}$$

The error of the partial sum using d derivatives decreases with increasing wavenumber.

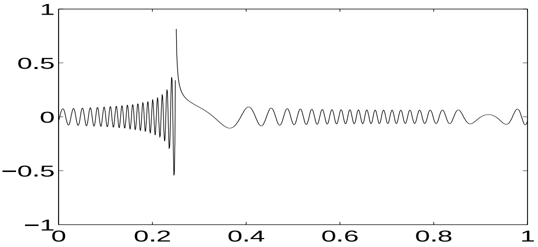


Figure 7.3: The real part of the integrand for $t_n = t_{sb1}$ for a circular obstacle and $k = 100$.

When the function f depends on k, the asymptotic behaviour of f and its derivatives at $\tau_{n,i}$ should be included in the derivation of estimate (7.40). We shall elaborate this for the case of f equal to the exact solution q_s. The necessary information for this elaboration was derived in §7.4.2. We will show that the error of the quadrature rule may actually be $O(1)$ in k , depending on the location of the quadrature points $\tau_{n,i}$.

Consider first the case $\tau_{n,i} \in T_I(k)$, i.e., the quadrature point lies in the illuminated region. If $\tau_{n,i} = t_n$, then $\tau_{n,i}$ is a singular point. From Lemma 7.3.3, we have $|w^c_{n,i,0}| = O(k^{-1})$ and $|w^c_{n,i,1}| = O(k^{-2})$. We have $|q_s(\tau_{n,i})| = O(k)$ from (7.31) and $|q'_s(\tau_{n,i})| = O(k)$ from (7.32). Combined, this yields the estimate

$$|S(1; q_s, n, i) - S(0; q_s, n, i)| = O(k^{-1}), \qquad \tau_{n,i} = t_n \in T_I(k).$$

This means that the accuracy increases with increasing k: the error when using only the first term scales as $O(k^{-1})$. Using Lemma 7.3.4, it can be found that the error of the first term for a stationary point $\tau_{n,i} \neq t_n$ is $O(k^{-1/2})$.

Next, consider the case $\tau_{n,i} \in T_{B1}(k)$. At the shadow boundary $\tau_{n,i} = t_n = t_{sb1}$, we have $|w^c_{n,i,0}| = O(k^{-2/3})$ and $|w^c_{n,i,1}| = O(k^{-1})$ from Lemma 7.3.4, and $|q_s(t_n)| = O(k^{2/3})$ and $|q'_s(t_n)| = O(k)$ from (7.31) and (7.32). The exponents cancel exactly; we have

$$|S(1; q_s, n, i) - S(0; q_s, n, i)| = O(1), \qquad \tau_{n,i} = t_{sb1}.$$

The accuracy of the contribution $S(d; f, n, i)$ does not improve with increasing k for a fixed d, although the partial sum may still converge if more derivatives are used. Similar observations hold for $\tau_{n,i} \neq t_{sb1}$ in the shadow boundary region. It turns out that the quadrature rule is not as useful in the shadow boundary region as in the illuminated region.

There is however an important remark that can be made here. Due to the square root in the definition (7.38) of the oscillator, the singular point

$\tau_{n,i} = t_n$ is a branch point of the oscillator. A consequence is that the left and right limit of the derivatives of the oscillator may differ. Indeed, the oscillator has a stationary point of order $r = 2$ at $t_n = t_{sb1}$ only in the right limit. The left limit of the derivative of the oscillator is nonzero. This is illustrated in Figure 7.3: the integrand is highly oscillatory to the left of the shadow boundary, and not oscillatory to the right. It turns out that the quadrature rule can be used on the left interval $[0, t_{sb1}]$ and in the illuminated region. It is not suited for the intermediate interval $[t_{sb1}, t_{sb1} + C_1 k^{-1/3}]$. Due to the stationary point, the integrand is not oscillatory in that interval.

7.5.3 A sparse discretisation

The basis functions for the discretisation are chosen corresponding to the behaviour of the solution in the three different regions identified in §7.4.2. Recall that the solution is smooth in the illuminated region, and oscillatory but exponentially decaying in the shadow region. First, following [29], we approximate the solution by zero in the shadow region. We choose a fixed number of basis functions in the illuminated region. Finally, we also choose a fixed number of basis functions in the transitional shadow boundary region, independently of k. Based on the asymptotic expansion (7.25), one can see that this corresponds to using a fixed number of basis functions per oscillation of the solution. It was already noted in §7.4.2 that the argument $k^{1/3} Z(\alpha, \tau)$ of the function Ψ is bounded in k, $\tau \in T_{B1}(k)$. The oscillatory behaviour of the exact solution q_s in $T_{B1}(k)$ is due to the oscillations of Ψ for negative arguments. Since the argument of Ψ is bounded, the number of possible oscillations is also bounded, independently of k.

In our implementation, we have chosen to use cubic B-splines as basis functions. The nodes of the splines are the collocation points of the collocation method. They are chosen equidistantly in the regions $T_{B1}(k)$, $T_{B2}(k)$ and $T_I(k)$. Owing to the small number of quadrature points $\tau_{n,i}$ for each collocation point t_n, the discretisation matrix is highly sparse. Since the quadrature rule, applied to the exact solution q_s, may not provide sufficient accuracy for quadrature points in the shadow boundary region, we propose the following scheme.

If $t_n \in T_I(k)$, then

- the quadrature rule is applied for the singular point in the illuminated region,
- the stationary points lie in the shadow region, and they are discarded.

For cubic splines, the singular point t_n lies in the support of only three separate basis functions. The three corresponding matrix entries are given

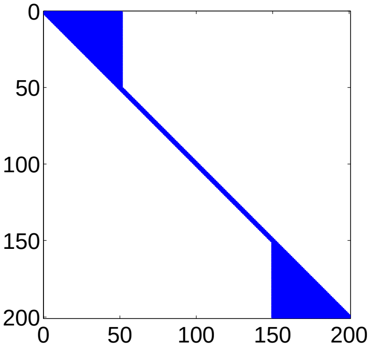


Figure 7.4: Illustration of the sparse discretisation matrix using cubic B-splines for the scattering of a plane wave by a circular obstacle. The middle part of the matrix is tridiagonal.

by (7.39). The contributions of the stationary points are discarded because the solution is approximated by zero in the shadow region.

If $t_n \in T_{B1}(k)$, then

- the quadrature rule is applied on the interval $[0, t_n]$,
- a classical dense discretisation is used on the interval $[t_n, t_{sb1} + C_1 k^{-1/3}]$,
- the quadrature rule is applied on the interval $[t_{sb1} + C_1 k^{-1/3}, 1]$.

The quadrature rule on $[0, t_n]$ reduces to the contribution of the singular point t_n. The corresponding weights have the form of (7.13) with S_l^H replaced by S_i^c,

$$w_{n,s,j}^c = -S_{i_s}^c \left[\frac{(x-t_n)^j}{j!}; t_n \right],$$

where i_s is the index such that $t_n = \tau_{n,i_s}$. The quadrature rule on $[t_{sb1} + C_1 k^{-1/3}, 1]$ consists of the contributions of the stationary points $\tau_{n,i}$ outside the shadow boundary region, and of the endpoint $t_r := t_{sb1} + C_1 k^{-1/3}$. The weights corresponding to that endpoint have the form of (7.11),

$$w_{n,r,j}^c = S_{i_r+1}^c \left[\frac{(x-t_r)^j}{j!}; t_r \right],$$

where i_r is the index such that $t_r \in [\tau_{n,i_r}, \tau_{n,i_r+1}]$. Finally, the dense discretisation in the interval $[t_n, t_r]$ leads to elements of the form

$$\sigma_{n,m} = \int_{\Omega_m \cap [t_n,t_r]} \frac{i}{4} H_0^{(1)}(k|\kappa(t_n) - \kappa(\tau)|) \qquad (7.41)$$
$$e^{ik(g^i(\kappa(\tau)) - g^i(\kappa(t_n)))} |\nabla\kappa(\tau)| \phi_m(\tau)\, \mathrm{d}\tau,$$

The only difference compared to (7.35) is that the integration domain may be cut at the boundaries of $[t_n, t_r]$. Summarising, the elements of the discretisation matrix for $t_n \in T_{B1}(k)$ can be written as

$$C_{n,m} = \begin{cases} \sigma_{n,m} & \text{if } \Omega_m \cap [t_n,t_r] \neq \varnothing \\ \quad + \sum_{j=0}^{d_{n,i_s}} w^c_{n,s,j} \phi_m^{(j)}(t_n) & \quad \text{if } t_n \in \Omega_m \\ \quad + \sum_{j=0}^{d_{n,i_r}} w^c_{n,r,j} \phi_m^{(j)}(t_r) & \quad \text{if } t_r \in \Omega_m \\ \quad + \sum_{i:t_r<\tau_{n,i}\in\Omega_m} \sum_{j=0}^{d_{n,i}} w^c_{n,i,j} \phi_m^{(j)}(\tau_{n,i}), & \exists i : t_r < \tau_{n,i} \in \Omega_m, \\ \sum_{i:t_r<\tau_{n,i}\in\Omega_m} \sum_{j=0}^{d_{n,i}} w^c_{n,i,j} \phi_m^{(j)}(\tau_{n,i}), & \exists i : t_r < \tau_{n,i} \in \Omega_m, \\ 0 & \quad \text{otherwise.} \end{cases}$$

The case $t_n \in T_{B2}(k)$ can be treated similarly. The structure of the sparse matrix C is illustrated in Figure 7.4. The two small dense parts correspond to the dense discretisation in the intervals $[t_n, t_{sb1} + C_1 k^{-1/3}]$ and $[t_{sb2} - C_2 k^{-1/3}, t_{sb2}]$. For simplicity, we have chosen the constant C_1 large enough such that, for $t_n \in T_{B1}$, all stationary points $\tau_{n,i} \in T_{B1}(k)$ also lie in the shadow boundary region. The constant C_2 was chosen similarly. One can show that the required integrals of the form (7.41) are not oscillatory. Due to the stationary point of order $r = 2$, the integrand behaves as $e^{ikc(\tau - t_{sb1})^3}$. The argument of the exponential is bounded in k, since by construction

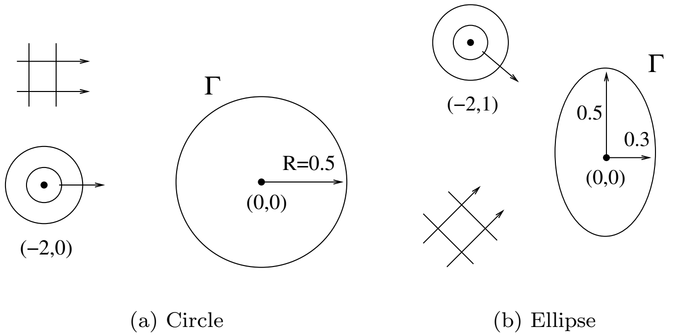
$$|\tau - t_{sb1}| \leq \max\{C_1, D_1\} k^{-1/3}.$$

Hence, there is only a bounded number of oscillations in the integrals for increasing k. The integrals can therefore be evaluated with a number of operations that is independent of k. In our implementation, these integrals were evaluated using Cubpack [50]. Since the weights of the quadrature rule can be evaluated efficiently as well, and because the number of unknowns is fixed, the matrix in Figure 7.4 can be computed with a total number of operations that is independent of k.

7.6 Numerical results

7.6.1 Convergence and total solution time

We consider the scattering by two convex obstacles, a circle and an ellipse, shown in Figure 7.5. We use two types of boundary conditions: a plane



(a) Circle (b) Ellipse

Figure 7.5: Illustration of two smooth convex scattering obstacles. The boundary conditions are plane waves, or circular waves originating from a point source.

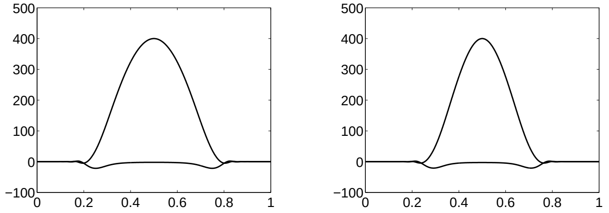
wave, modelled in the form $u^i(x) = e^{ik\alpha\cdot x}$, and a point source, modelled by $u^i(x) = H_0^{(1)}(|x - x_0|)$, with x_0 a point in the exterior Ω^+ of the obstacle. The circle and ellipse are parameterised by

$$\kappa(t) = \left\{ \begin{array}{l} R\cos(2\pi t) \\ R\sin(2\pi t) \end{array} \right. \quad \text{and} \quad \kappa(t) = \left\{ \begin{array}{l} R_1\cos(2\pi t) \\ R_2\sin(2\pi t) \end{array} \right.$$

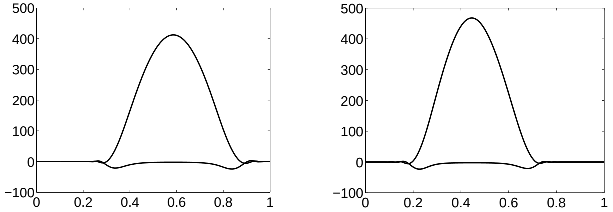
respectively. The boundary conditions for the ellipse are deliberately chosen to yield a non-symmetric problem. We use N cubic B-spline basis functions, defined on the interval $[t_{sb1} - D_1 k^{-1/3}, t_{sb2} + D_2 k^{-1/3}]$. A fixed number N_1 of these functions are defined on the shadow boundary regions T_{B1} and T_{B2}. The collocation points are the nodes of the spline function, chosen equidistantly in the intervals T_{B1}, T_I and T_{B2} respectively.

The smooth function q_c is illustrated in Figure 7.6 for the different scattering problems. The mild oscillatory behaviour of the function near the shadow boundary is illustrated in the left panel of Figure 7.7, showing only the real part of the solution. Two spikes are present near the shadow boundary, with a peak value that scales as $O(k^{2/3})$ as predicted by our estimate (7.27). The dashed line shows the effect of doubling k. The $O(k)$ behaviour in the illuminated region is clear from the imaginary part illustrated in the right panel of Figure 7.7

Table 7.2 shows the timings for an implementation of the algorithm of §7.5.3 in Matlab. In all examples considered, the time actually decreases with increasing wavenumber k. This is due to the fact that, at larger frequencies, the weights of the specialised quadrature rule $Q_c[f; t_n]$ are easier to compute. In a classical boundary element method, and using 10 unknowns per wavelength, the case $k = 100000$ corresponds to a dense matrix



(a) Circle, plane wave (b) Circle, point source



(c) Ellipse, plane wave (d) Ellipse, point source

Figure 7.6: The real and imaginary part of the smooth function q_s for different scattering problems. The real part is the lower curve in each example. The wavenumber is $k = 200$.

with $N = 500000$ unknowns.

7.6.2 Error of the solution

In applications, one is usually interested in the quantity q_s/k. For example in electromagnetics, the quantity q_s/k is proportional to the induced current on the surface of the obstacle, with a proportionality constant that is independent of k. The exact solution q_s of the scattering problem is known only for the case of scattering of a plane wave by a circle. The relative error $(q_c - q_s)/q_s$ and the absolute error $(q_c - q_s)/k$ for this case are illustrated in Figure 7.8 for a number of different values for k. We have chosen to use derivatives up to order $d_{n,i} = d = 1$ in each quadrature rule for this example. The error decreases rapidly with increasing k in the illuminated region. This is due to the higher accuracy of the quadrature rule $Q_c[f;t]$ at larger frequencies. The relative error tends to 100% in the deep shadow

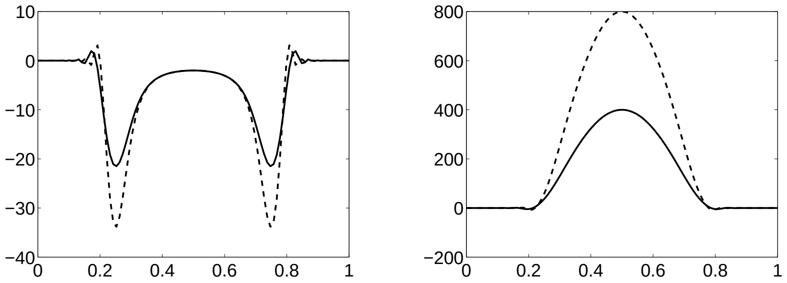


Figure 7.7: The real part (left) and imaginary part (right) of q_s for the scattering of a plane wave by a circle. The dense line corresponds to $k = 200$, the dashed line to $k = 400$.

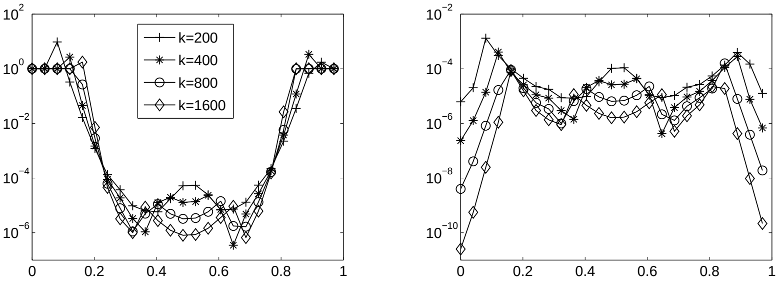
Table 7.2: Total solution time in seconds for the different scattering problems. All parameters are kept fixed, except the wavenumber k. We used $d = 2$ derivatives and $N = 150$ unknowns.

	Circle		Ellipse	
k	Plane wave	Point source	Plane wave	Point source
200	$232s$	$546s$	$322s$	$616s$
400	$226s$	$531s$	$308s$	$596s$
800	$224s$	$525s$	$303s$	$589s$
1600	$221s$	$521s$	$299s$	$583s$
10000	$217s$	$505s$	$290s$	$567s$
100000	$211s$	$495s$	$272s$	$542s$

region because we have approximated the solution by 0 in that region. One can verify from the figures that the absolute error in that region is still quite small compared to the average value of the function $q_s(\tau)$.

It is important to know how the error can be controlled, i.e., how the parameters of the method can be chosen to achieve a given accuracy. The parameters are:

- N, N_1: the total number of unknowns, and the number of unknowns in the shadow boundary region respectively;
- C_1, C_2, D_1, D_2: these parameters control the size of the shadow boundary region in definitions (7.28)-(7.30);
- $d_{n,i}$: the number of derivatives used in the quadrature rule $Q_c[f; t_n]$ given by (7.37).



(a) Relative error $(q_c - q_s)/q_s$ (b) Absolute error $(q_c - q_s)/k$

Figure 7.8: Absolute and relative error for the scattering of a plane wave by a circle for different values of k. We have used $d = 1$ derivative in the quadrature rule $Q_c[f;t]$.

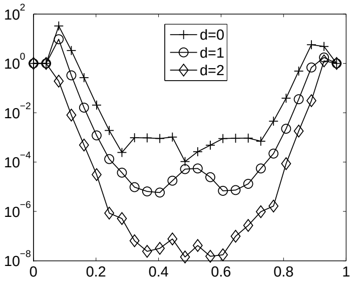


Figure 7.9: Comparison of the relative error for different values of d, the number of derivatives used in the quadrature rule $Q_c[f;t]$.

Increasing each of these parameters decreases the error, although the effect of increasing each parameter independently is bounded. For example, increasing N indefinitely does not yield arbitrary accuracy, because the accuracy of the quadrature rule is independent of N. The constants C_1, C_2, D_1 and D_2 can be chosen so large that the shadow boundary region covers the whole boundary. In that case, the method almost reduces to the regular boundary element method. The accuracy of the quadrature rule is increased by increasing $d_{n,i}$. We have considered a fixed number of derivatives $d = d_{n,i}$ in each example. Figure 7.9 shows the relative error for different values of d. The accuracy of the method greatly increases with increasing d, while the sparsity structure of the matrix remains exactly the same. For a cubic spline, $d = 2$ is the largest possible value since higher order derivatives are discontinuous.

Finally, we note that the matrices are well conditioned. The matrices that were constructed to produce the results in Table 7.2 have size 150×150. The largest condition number in these example was 167. Since the matrices are small, the corresponding system of equations can be readily solved using a direct solver.

7.7 Three-dimensional problems

The enabling properties of the scattering problems are the same in two and three dimensions. First, for smooth and convex scatterers, the phase of the solution is known a priori. Second, the value of oscillatory integrals in one or two dimensions is determined by the behaviour of the integrand in only a small part of the integration domain. Hence, it is reasonable to expect that a sparse discretisation matrix exists for three-dimensional high frequency scattering problems. An implementation may be based on the cubature rules for multivariate oscillatory integrals that were constructed in Chapter 6. Their application to three-dimensional scattering problems is the subject of future research.

7.8 Conclusions

The method that was proposed in this chapter computes the solution of a high frequency scattering problem in a number of operations that is bounded, independently of the wavenumber. The discretisation matrix is small and sparse, the accuracy of the solution increases with increasing wavenumber. All these properties are the opposite of the properties of the multiscale methods that were discussed in the previous chapters. Rather than focusing on a multiscale approximation of the discretisation matrix, the result is achieved by focusing on the specific properties of highly oscillatory integrals.

The method is currently still rather restricted however. Specifically, the phase of the solution should be known a priori. This problem was avoided in this chapter by considering only smooth convex scatterers. The approach of [29] was extended to multiple scattering configurations using an iterative approach [90]. It is reasonable to expect that the same will hold for our approach. Multiple scattering and scattering by concave objects is the subject of future research.

Chapter 8

Conclusions and suggestions for future research

In this chapter, we formulate the conclusions of this thesis, and present some suggestions for further research. First, we briefly summarise our main contributions:

- the analysis of the matrix compression in the wavelet method for increasing wavenumbers;
- a new method using wavelet packets with improved compression for oscillatory problems;
- the construction of efficient quadrature rules for integrals involving wavelets or scaling functions in the integrand;
- an efficient numerical method for the evaluation of univariate highly oscillatory integrals using the path of steepest descent;
- the extension of this method to multivariate integrals;
- a hybrid asymptotic boundary element method for highly oscillatory integral equations for scattering problems.

8.1 Multiscale methods

We have considered three different multiscale methods for the solution of integral equations: the fast multipole method, hierarchical matrix methods and wavelet based methods. One could place the first two methods under the common denominator of low-rank approximation methods. We have seen that both methods essentially achieve a speed-up of the matrix-vector product by approximating subblocks of the discretisation matrix with a matrix of low rank, that is represented with less data than the original dense matrix. These approximations are possible thanks to the smoothness of the kernel function of the integral operator away from the diagonal. Wavelet based methods exploit the smoothness of the kernel function in a different way. Rather than *agglomerating* basis functions that are defined on a fine scale, wavelet methods focus on refining a coarse discretisation.

At low frequencies, the wavelet method is the only method that achieves $O(N)$ computational complexity, where the accuracy of the solution scales like the discretisation error for increasing N. However, the absence of logarithmic terms in the asymptotic complexity is less significant in practice, and the other multiscale methods are competitive. Due to the approximation of the kernel function in the parameter space, the wavelet method is more sensitive to irregularly shaped boundaries. The approximation of the kernel function in the full domain surrounding the scattering obstacle avoids this dependence in the low-rank approximation methods. The specific multipole and local expansions that are constructed in the fast multipole method for a given kernel function are usually more accurate than black-box separable approximations, at a cost of decreased flexibility. The best solution method therefore depends very much on the application at hand.

Of the methods that we considered, we found that, at high frequencies, the high frequency fast multipole method is the only viable efficient solution method, together with the intimately related implementation of $\mathcal{H}^2$-matrices. There, subblocks of the discretisation matrix are approximated by low rank matrices, where the rank scales linearly with the size of the subblock. A fast matrix-vector product is made possible by the efficient hierarchical construction of the low rank approximations of these subblocks, using so-called diagonal translation operators. In addition, computations are maximally shared among different blocks. Additional care is required in this method for relatively low values of the wavenumbers, or for high accuracy computations, due to the numerical instability of the efficient diagonal translation operators.

We have shown that the wavelet method does not scale to high frequency problems. Using wavelet packets, the complexity of the matrix-vector product can be reduced however from $O(N^2)$ to approximately $O(N^{1.4})$. The method is not as efficient overall as the high frequency fast multipole method. The study of wavelet packets does however support the idea that oscillatory problems require oscillatory basis functions. The choice of wavelet packets actually corresponds to choosing a specific set of oscillatory basis functions. The basis functions can be matched adaptively to the problem, in order to obtain a sparse discretisation matrix.

8.2 Hybrid methods

One may ask whether solving high frequency problems using the classical piecewise polynomial basis functions is a good idea at all. Such basis functions are not adapted to the oscillatory nature of the problem. On the other hand, asymptotic methods for very high frequency problems lack the robustness and fine error control of finite element methods. It is therefore reasonable to expect that the future of solution methods in the high frequency regime lies in a combination of the two. So far in the available literature, a number of approaches have been suggested that enrich the approximation space of the smooth basis functions with oscillatory functions. The use of oscillatory wavelet packet basis functions that was considered in this thesis also resembles this idea.

For the specific case of smooth and convex obstacles, the combination of finite element methods with asymptotic methods was taken to a new direction in Chapter 7. The scattering problem was reformulated such that the difficulty - the highly oscillatory nature - was removed. Based on a thorough analysis of the properties of oscillatory integrals, and using corresponding efficient evaluation techniques elaborated in Chapter 6, the resulting scheme has a number of surprising properties. The number of operations required for the solution of the scattering problem is bounded independently in k. The discretisation matrix is small and sparse, and the accuracy of the solution may even increase with increasing frequency. Scattering problems have been solved in Chapter 7 for values of the wavenumber that are much larger than currently feasible with even the most efficient multiscale methods on standard computers.

8.3 Future research directions

8.3.1 Oscillatory integrals

The recent development of efficient evaluation techniques for highly oscillatory integrals, that are described and summarised in Chapter 6, represents a genuine progress in our understanding of the numerical treatment of oscillatory problems. These algorithms will likely enable many new applications in the future, only one of which was explored in the hybrid method in Chapter 7.

The algorithms can still be improved and extended in many ways.

- A possible direction of research is a robust implementation of the numerical steepest descent method for univariate and multivariate integrals involving complex stationary points. For example, we have seen that the path of steepest descent may not be suitable due to the possible presence of complex stationary points along the path, when a real stationary point coincides with a corner point of the integration domain in two dimensions.

- Neither conditions involving the complex plane or the knowledge of certain moments are required in Levin-type methods; an interesting innovation would therefore be the extension of Levin-type methods to integrals with stationary points. A useful starting point is the relation that was identified between Levin-type methods and the steepest descent method (Remark 6.3.6).

- The asymptotic expansion of univariate oscillatory integrals can be obtained using the steepest descent method (see Appendix B). Results regarding the asymptotic expansion of multivariate integrals involving degenerate critical points do not appear to be available [21, 178, 204]. The asymptotic analysis of the multivariate numerical steepest descent method may lead to new results in this field.

- Efficient algorithms should be constructed for a number of variations of the model integral. Consider, for example, problems where the oscillator is not known a priori, or where the oscillator is only known approximately. Alternatively, other types of oscillators can be considered, such as a cosine, or Bessel functions.

- Although originally developed in the context of integral equations, the use of our quadrature methods may turn out to be advantageous in very different applications, e.g., in the solution of oscillatory differential equations, in the J-matrix method for quantum scattering and in the Wave Based method for the simulation of sound and vibrations in mechanical structures. Current research directions involving these methods are summarised in Appendix C.

- Finally, we strongly believe that the algorithms will turn out to be useful in a rich range of applications. Yet, a wider acceptance of the methods in the engineering community will require easy accessibility of the algorithms through optimised, robust and user friendly software packages.

8.3.2 Boundary element methods

An important advantage of classical boundary element method formulations over hybrid methods, is that no additional knowledge is required of the asymptotic behaviour of the problem. The classical method is well understood and robust, and is being used in many implementations for the numerical simulation of scientific problems.

- It remains important to improve existing fast solution methods for boundary element method discretisations of integral equations, and to better understand the behaviour of the standard method at high frequencies. For example, the efficiency of an implementation also depends on the value of the condition number, which influences the number of matrix-vector products that are required in an iterative solution method. Likewise, the importance of the pollution error in boundary element methods remains to be determined.

- The separable approximation of the kernel function in the high frequency fast multipole method is essentially based on the discretisation of an oscillatory integral. It may be possible, based on the insights regarding oscillatory integrals that are developed in Chapters 6 and 7, to improve this discretisation, thereby reducing the constants in the computational complexity of the FMM for oscillatory integral equations.

- Another possible direction of research is the extension of the $O(1)$ hybrid method that was proposed in Chapter 7 to more general scattering configurations. Multiple scattering or concave scatterers may perhaps be simulated using an iterative approach, such as the one considered in [90]. For these scattering problems, the solution no longer has the form of a smooth function times a certain oscillatory function. Rather, the solution consists of a (possibly infinite) sum of functions with this form, corresponding to the superposition of reflected, diffracted, and refracted waves in the medium. The issues that need to be considered are the asymptotic behaviour of each term in the sum, and the extension of the hybrid method to the new model form of the solution.

- The application area can be further extended by considering three-dimensional problems, Maxwell's equations and non-perfectly conducting or non-smooth obstacles. Extensions of the hybrid method may be based on the existing asymptotic analysis of such scattering problems.

- It is at present very much an open question whether hybrid methods can be formulated for general problems with industrially relevant complexity, such as a planar antenna consisting of several electronic elements. The problem is complicated further by the presence of edges and sharp corners. The asymptotic form of the solution for these problems may become very complex. The development and implementation of $O(1)$ complexity methods for real-life problems with high wavenumbers remains a fascinating challenge.

Appendix A

Scattering by the circle

The scattering of a plane wave by a circular obstacle has a known, analytical solution. Despite its simplicity, the problem is well suited to illustrate the boundary element method, since it already exhibits most of the properties of more general scattering problems. Scattering by a circle is frequently used as an example in the thesis. In this appendix, we show how the analytical solution is obtained. The interested reader is referred to [107] for a more detailed analysis.

A.1 Solution by separation of variables

The two-dimensional scattering by a circle is equivalent to the scattering by an infinitely long cylinder with circular cross section in three dimensions. The cylindrical symmetry of this problem is expressed best in cylindrical coordinates. Figure A.1 shows the convention of cylindrical coordinates that will be used. The three-dimensional scalar Helmholtz equation in cylindrical coordinates, applied to the function $\psi(\rho, \phi, z)$, is given by

$$\frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho \frac{\partial \psi}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 \psi}{\partial \phi^2} + \frac{\partial^2 \psi}{\partial z^2} + k^2 \psi = 0. \tag{A.1}$$

Using the method of separation of variables, we look for a solution of the form

$$\psi(\rho, \phi, z) = R(\rho)\Phi(\phi)Z(z). \tag{A.2}$$

Substituting (A.2) into (A.1), and subsequently dividing by ψ, yields

$$\frac{1}{\rho R}\frac{d}{d\rho}\left(\rho \frac{dR}{d\rho}\right) + \frac{1}{\rho^2 \Phi}\frac{d^2 \Phi}{d\phi^2} + \frac{1}{Z}\frac{d^2 Z}{dz^2} + k^2 = 0. \tag{A.3}$$

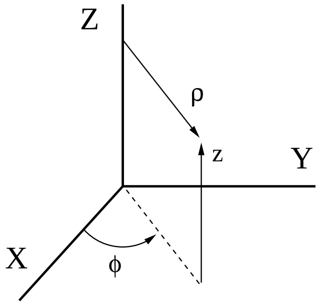


Figure A.1: Cylindrical coordinates.

The third term in (A.3) is independent of ρ and of ϕ, and therefore it must be independent of z as well, because the equation sums to zero. Hence, we can write

$$\frac{1}{Z}\frac{d^2 Z}{dz^2} = -k_z^2, \tag{A.4}$$

with k_z a constant. This is an harmonic equation, giving rise to solutions of the form $\cos(k_z z)$, $\sin(k_z z)$, $e^{\pm i k_z z}$ or a linear combination of these functions. We will denote such harmonic functions in general by $Z(z) = h(k_z z)$. Substituting (A.4) into (A.3), and multiplying by ρ^2, yields

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right) + \frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} + (k^2 - k_z^2)\rho^2 = 0. \tag{A.5}$$

Now the second term is independent of ρ. Hence, we can write

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} = -n^2,$$

and we have $\Phi(\phi) = h(n\phi)$. Define a constant k_ρ such that

$$k_\rho^2 + k_z^2 = k^2. \tag{A.6}$$

The remaining equation (A.5) can then be written as

$$\rho\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right) + [(k_\rho\rho)^2 - n^2]R = 0.$$

This is *Bessel's equation* of order n. The solution can be the *Bessel function of the first kind* $J_n(k_\rho\rho)$, the *Bessel function of the second kind* $Y_n(k_\rho\rho)$, or a linear combination of these functions. Usual combinations are the Hankel function of the first and second kind,

$$H_0^{(1)}(k_\rho \rho) = J_n(k_\rho \rho) + iY_n(k_\rho \rho) \quad \text{and} \quad H_0^{(2)}(k_\rho \rho) = J_n(k_\rho \rho) - iY_n(k_\rho \rho).$$

We denote the possible Bessel functions by $R(\rho) = B_n(k_\rho \rho)$. In summary, we can form a solution to the Helmholtz equation as

$$\psi_{k_\rho,n,k_z} = B_n(k_\rho \rho)h(n\phi)h(k_z z), \tag{A.7}$$

subject to condition (A.6). More general solutions can be obtained by summing functions of the form (A.7).

The Hankel function of the first kind $H_n^{(1)}(k_\rho \rho)$ represents an outward travelling wave.[1] Outward travelling solutions therefore correspond to the choice $B_n(k_\rho \rho) = H_n^{(1)}(k_\rho \rho)$. Such solutions satisfy the required Sommerfeld radiation condition (2.18) at infinity. The Bessel function of the first kind $J_n(k_\rho \rho)$ is the only Bessel function that is non-singular at $\rho = 0$. Fields that are finite at $\rho = 0$ therefore correspond to the choice $B_n(k_\rho \rho) = J_n^{(1)}(k_\rho \rho)$.

A.2 An analytical solution

Consider a plane wave propagating in the positive x-direction,

$$u^i = e^{ikx} = e^{ik\rho\cos(\phi)}.$$

This wave is finite at $\rho = 0$. One can express the plane wave as an infinite series of solutions of the form (A.7) by

$$e^{ikx} = \sum_{n=-\infty}^{\infty} a_n J_n(k\rho)e^{-in\phi}, \qquad \text{with} \quad a_n = (-i)^{-n}.$$

The total solution is given by $u = u^s + u^i$. The scattered wave is an outgoing wave, and therefore it can be written in the form

$$u^s = \sum_{n=-\infty}^{\infty} b_n(-i)^{-n} H_n^{(1)}(k\rho)e^{-in\phi}.$$

Hence, the total field is given by

$$u = \sum_{n=-\infty}^{\infty} (-i)^{-n}(J_n(k\rho) + b_n H_n^{(1)}(k\rho))e^{-in\phi}. \tag{A.8}$$

---

[1] This is a consequence of the time dependence of the form $e^{-i\omega t}$. With the other choice $e^{i\omega t}$, $H_n^{(2)}(k_\rho \rho)$ would represent an outward travelling wave.

Assume the Dirichlet condition $u = 0$ is imposed at the boundary $\rho = a$ of the cylinder with radius a. One obtains from (A.8) that this condition is satisfied if

$$b_n = \frac{-J_n(ka)}{H_n^{(1)}(ka)}. \tag{A.9}$$

Expression (A.8) with the coefficients b_n given by (A.9) is an analytical expression for the total solution of the exterior Dirichlet problem.

A.3 Eigenvalues

The eigenvalues and eigenfunctions of the scattering problem are also known analytically. The single-layer potential has the eigenvalues

$$\lambda_p = \frac{ia\pi}{2} J_p(ka) H_p^{(1)}(ka), \qquad p = 0, 1, \ldots$$

with corresponding eigenspaces $\{\psi_p, \psi_{-p}\}$, where

$$\psi_p = \frac{e^{ikp}}{\sqrt{2\pi}}.$$

It follows from the properties of Bessel functions that, asymptotically, the rate of decay of the eigenvalues is given by

$$\lambda_p \sim 1/p, \qquad p \to \infty.$$

The operator has resonant frequencies for those values of k where ka corresponds to a zero of the Bessel or Hankel function of some order p. This is more likely to occur if ka is large. Note that the eigenfunctions corresponding to increasing values of p are increasingly oscillatory.

Appendix B

The method of steepest descent

The numerical steepest descent method that was presented in Chapter 6 builds on the ideas of the *method of steepest descent*, see [21, 204]. This method was invented independently by Cauchy and by Riemann, and has been continuously used since the description by Debye in 1909 [70]. It is closely related to the method of stationary phase that was introduced by Lord Kelvin in 1887 [151]. The method leads to an asymptotic expansion for the oscillatory integral $I[f]$, as given by (6.1), for large values of ω. Usually, the result is obtained after a series of analytical manipulations that are specific for the integrand. Here, we describe a black-box method to find the asymptotic expansion.

B.1 Watson's Lemma

The first step in the method of steepest descent is to identify the critical points of the integrand. In the case treated in Chapter 6, the integration on $[a, b]$ for real-valued functions f and g, the critical points are the endpoints a and b, and all real stationary points in $[a, b]$. The next step is to deform the integration path onto the path of steepest descent at all critical points, subject to the validity of the deformation by Cauchy's integral theorem. Finally, the asymptotic expansion is obtained by summing the asymptotic expansions of each line integral along a path of steepest descent. The expansion for a typical line integral is found using Watson's lemma.

Lemma B.1.1 (Watson's Lemma). *Assume $f(p)$ is a locally integrable* *function on* $(0,\infty)$ *and* $f(p) = O(e^{ap})$, $p \to \infty$, *with* $a \in \mathbb{R}$. *If*

$$f(p) \sim \sum_{k=0}^{\infty} c_k p^{a_k}, \qquad p \to 0,$$

with $\Re a_k > -1$, *then we have, for* $\omega \to \infty$,

$$\int_0^\infty e^{-\omega p} f(p)\,\mathrm{d}p \sim \sum_{k=0}^{\infty} c_k \int_0^\infty e^{-\omega p} p^{a_k}\,\mathrm{d}p = \sum_{k=0}^{\infty} \frac{c_k \Gamma(a_k+1)}{\omega^{a_k+1}}.$$

Watson's Lemma essentially states that it is justified to perform the term-by-term integration of the series of $f(p)$ for $p \to 0$.

The difference between the numerical approach described in Chapter 6 and the steepest descent method, is that no asymptotic expansion is constructed in the numerical approach. Instead, the integral representation along the path of steepest descent is kept and is evaluated numerically. The use of Gaussian quadrature leads to the very high order of convergence, but only when the integrand is evaluated exactly on the path of steepest descent. A generally applicable iterative method can be used to find the necessary points on the path to high precision with few iterations.

B.2 An asymptotic expansion for $I[f]$

We will show how the steepest descent method can be used to obtain the coefficients of an asymptotic expansion of the form (6.4) for the oscillatory integral $I[f]$ in the presence of stationary points. Note that it is sufficient to consider the expansion of the generalised moments $\mu_j(\omega;\xi)$ defined by (6.5),

$$\mu_j(\omega;\xi) = I[(x-\xi)^j] = \int_a^b (x-\xi)^j e^{i\omega g(x)}\,\mathrm{d}x. \tag{B.1}$$

The full expansion can be obtained from the uniform expansion (6.6), that was constructed by Iserles and Nørsett in [129].

Consider an oscillator g with one stationary point $\xi = 0$ of order $r = 1$ in the interval $[-1,1]$. The required moment $\mu_0(\omega;\xi)$ can be written as

$$\mu_0(\omega;0) = F_1(-1) - F_1(0) + F_2(0) - F_2(1), \tag{B.2}$$

where the functions F_j have the form

$$F_j(x) = \int_0^\infty e^{-\omega p} h'_{x,j}(p)\,\mathrm{d}p, \tag{B.3}$$

and the functions $h_{x,j}(p)$ satisfy $g(h_{x,j}(p)) = g(x) + ip$ on $[-1,0]$ and $[0,1]$ respectively. Justified by Watson's Lemma, the asymptotic expansion of $F_j(x)$ is obtained by the term-by-term integration of the series representation of $h'_{x,j}(p)$, for $p \to 0$.

B.2.1 An asymptotic expansion for $F_j(x)$

Away from the stationary point $\xi = 0$, $h_{x,j}(p)$ can be developed into a regular Taylor series: $h_{x,j}(p) \sim \sum_{k=0}^{\infty} a_{j,k}p^k$, $p \to 0$. By construction, we have $a_{j,0} = x$. The derivatives of $h_{x,j}(p)$ are found by taking the derivatives of $g(h_{x,j}(p)) = g(x) + ip$ with respect to p. The first order derivative was already given by (6.41). It is straightforward to derive expressions for the higher order derivatives using the product and chain rules. Then, we have

$$F_j(x) = \int_0^{\infty} e^{-\omega p} \sum_{k=0}^{\infty} a_{j,k}p^k \,\mathrm{d}p \sim \sum_{k=0}^{\infty} b_{j,k}\, \omega^{-k-1}, \qquad k \to \infty,$$

for $x \neq \xi$, and with $b_{j,k} = \Gamma(k+1)a_{j,k}$.

The difficulty lies in the expansions of $F_j(0)$, $j = 1, 2$, because $h'_{\xi,j}(p)$ is singular at $p = 0$. The singularity is known however, and we can find a series of the form

$$h_{\xi,j}(p) = \sum_{k=0}^{\infty} c_{j,k}\, p^{k/2}, \qquad p \to 0. \tag{B.4}$$

By construction we know that $c_{j,0} = \xi = 0$. The remaining coefficients $c_{j,k}$ can be obtained via a series expansion of the equation $g(h_{\xi,j}(p)) = g(\xi)+ip$. Assume without loss of generality that $g(\xi) = 0$, such that we can write $g(x) = \sum_{k=2}^{\infty} g_k x^k$. This leads to

$$\sum_{k=2}^{\infty} g_k \left(\sum_{l=1}^{\infty} c_{j,l}\, p^{l/2} \right)^k = ip. \tag{B.5}$$

Even though (B.5) is a highly nonlinear problem, it is straightforward to obtain the values $c_{j,k}$. Equating the coefficient in p of both the left and right hand side immediately yields

$$g_2 c_{j,1}^2 = i.$$

The correct choices of the root, for $a_2 > 0$, are

$$c_{1,1} = -\sqrt{i/a_2}, \tag{B.6}$$

$$c_{2,1} = \sqrt{i/a_2}. \tag{B.7}$$

Rewriting (B.5) using the Cauchy product repeatedly leads to

$$\sum_{k=2}^{\infty} g_k \sum_{i_1=0}^{k} \sum_{i_2=0}^{k-i_1} \sum_{i_3=0}^{k-i_1-i_2} \cdots \binom{k}{i_1} \binom{k-i_1}{i_2} \binom{k-i_1-i_2}{i_3} \cdots$$
$$c_{j,1}^{i_1}\, p^{i_1/2}\, c_{j,2}^{i_2}\, p^{2i_2/2}\, c_{j,3}^{i_3}\, p^{3i_3/2}, \cdots . \tag{B.8}$$

with the added condition $\sum_l i_l = k$ to avoid spurious terms (consider for example $i_l = 0$, $l = 1, \ldots, \infty$). Equating the total coefficient of increasing powers of $p^{1/2}$ to zero leads to a recursive scheme for $c_{j,k}$. Specifically, the coefficient in $p^{N/2}$ is given by

$$\sum_{k=2}^{N} \sum_{(\sum_l i_l = k) \& (\sum_l l i_l = N)} \binom{k}{i_1} \binom{k-i_1}{i_2} \binom{k-i_1-i_2}{i_3} \cdots \prod_l c_{j,l}^{i_l}. \tag{B.9}$$

The second summation in (B.9) is a sum over all sets of indices $(i_1, i_2, \ldots)$ that satisfy the conditions

$$\sum_l i_l = k, \tag{B.10}$$

$$\sum_l l \; i_l = N. \tag{B.11}$$

These conditions state that the indices sum to k, a necessary condition that was present already in (B.8), and that the total exponent should be $N/2$. The coefficient c_L with $L > N$ cannot occur in this expression due to condition (B.11). The coefficient c_N does not occur because conditions (B.10) and (B.11) cannot both be satisfied. Finally, for $N > 2$, the expression is linear in c_{N-1}. The term corresponds to the only possible combination $k = 2$, $i_1 = 1$ and $i_{N-1} = 1$, and it is given by $2g_2c_1c_{N-1}$. Thus, each value of $N > 2$ leads to an explicit expression for $c_{j,N-1}$ in terms of a_k, $k = 1, \ldots, N$, and in terms of products of powers of $c_{j,l}$, $l = 1, \ldots, N-2$.

The total expansion of $F_j(0)$ is obtained by starting the recursion with $c_{1,1}$ given by (B.6) or $c_{2,1}$ given by (B.7) respectively. We have

$$\begin{aligned} F_j(\xi) &= \int_0^\infty e^{-\omega p} h'_{\xi,j}(p)\,\mathrm{d}p = \int_0^\infty e^{-\omega p} \sum_{k=1}^{\infty} c_{j,k} \frac{k}{2} p^{k/2-1}\,\mathrm{d}p \\ &\sim \sum_{k=1}^{\infty} \Gamma(k/2+1) c_{j,k}\,\omega^{-k/2} \sim \sum_{k=1}^{\infty} d_{j,k}\,\omega^{-k/2}, \qquad k \to \infty. \end{aligned}$$

The expansion for $F_2[1](0) - F_1[1](0)$ has terms in $\omega^{-j-1/2}$ only, with j integer, because the coefficients $c_{1,2j}$ and $c_{2,2j}$ are equal.

Finally, note that higher order stationary points can be treated similarly. We have $g(x) = \sum_{k=r+1}^{\infty} a_k x^k$, and $h_{j,k}(p) = \sum_{k=1}^{\infty} c_k p^{k/(r+1)}$. The recursion is started by selecting the two correct roots of the equation $a_1 c_1^{r+1} = i$.

B.2.2 Computational issues

Expression (B.9) is not entirely explicit. Indeed, the conditions (B.10) and (B.11) correspond to solving a knapsack-type problem. As a result, the number of operations required to obtain the next coefficient using (B.9) increases exponentially with N. However, the explicit expressions for $c_{j,k}$ have to be found only once, and they can then be programmed. Although the number of terms in the expressions also increases exponentially with N for general oscillators g, an implementation for the first fixed N terms can be very rapid.

For completeness, we list the first four explicit expressions for $c_{j,k}$:

$$\begin{aligned}
c_{j,1} &= (-1)^j \sqrt{i/g_2}, \\
c_{j,2} &= (-g_3 c_{j,1}^2)/(2g_2), \\
c_{j,3} &= -(g_4 c_{j,1}^4 + 3g_3 c_{j,1}^2 c_{j,2} + g_2 c_{j,2}^2)/(2g_2 c_{j,1}), \\
c_{j,4} &= -(4g_4 c_{j,1}^3 c_{j,2} + 3g_3 c_{j,1}^2 c_{j,3} + 2g_2 c_{j,2} c_{j,3} + 3g_3 c_{j,1} c_{j,2}^2)/(2g_2 c_{j,1}).
\end{aligned}$$

The expression for the nineteenth coefficient consists of 624 terms.

Appendix C

Applications in computational science and engineering

We point out in this section how the methods developed in this thesis can be used for other applications in computational science and engineering. First we consider three different physical models, in sections C.1, C.2 and C.3, each leading to a different type of integral to be computed. Next, in sections C.4 and C.5, we consider some efforts related to different geometries: non smooth polygonal surfaces and multiple scatterers. Some of this work is still ongoing and preliminary, although the initial results are promising. Most of this work is in collaboration with researchers from other departments of the K.U.Leuven, or from abroad. The model discussed in §C.1 was provided by W. Desmet, B. Pluymers and C. Vanmaele (Noise and Vibration Research Group, PMA). The problems treated in §C.2 and §C.5 were suggested by G. Vandenbosch (Telemic, ESAT). The ideas in §C.3 are being worked out with W. Vanroose (Dept. Computer Science). Finally, the scattering problem for domains with corners, discussed in §C.4, is being studied in collaboration with S. Chandler-Wilde and S. Langdon (University of Reading).

C.1 The Wave Based method in acoustics

The Wave Based method is a method for the numerical modelling of the propagation and scattering of acoustic waves in mechanical structures involving vibrations, such as a vehicle [76]. The method uses plane wave basis

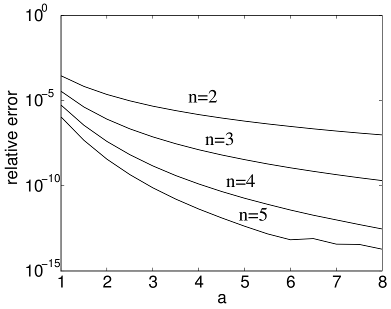


Figure C.1: Relative error for the evaluation of an integral of the form (C.2), using n quadrature points along each steepest descent path. The parameter a is a linear scaling factor for the wavenumbers k_x, k_y and k_b.

functions, possibly attenuated, that satisfy the Helmholtz equation in free space. Any combination of these oscillatory functions therefore automatically satisfies the Helmholtz equation. The solution to a specific problem is found by enforcing the correct boundary condition on the boundaries of the mechanical structure, which leads to a dense matrix. As such, the Wave Based method resembles boundary element methods. A difference is that the matrix of the Wave Based method may be rather ill conditioned, necessitating the very accurate computation of the matrix elements.

Each element of the dense matrix is given by an oscillatory integral, involving harmonic functions and Hankel functions (see, e.g., [168]). These integrals have the form

$$I = \frac{-i}{8k_b^2} \int_\Gamma (c_1 \sin(k_x x) + c_2 \cos(k_x x)) e^{-ik_y(y-Ly)} H_0^{(2)}(k_b r) ds_\Gamma, \quad \text{(C.1)}$$

with $r = \sqrt{(x-x_0)^2 + (y-y_0)^2}$ the distance to an excitation point (x_0, y_0). The efficient evaluation of this integral is considered one of the computational challenges in the overall acoustic simulation algorithm [189]. Writing the factors $\cos(k_x x)$ and $\sin(k_x x)$ in terms of complex exponentials, the integrand of (C.1) can be written as a sum of integrands that behave as

$$f(x) e^{i(\pm k_x x - k_y(y-Ly) + k_b r)}, \quad \text{(C.2)}$$

with $f(x)$ a smooth function. This is a generalization of the model form (6.1), because of the different constants k_x, k_y and k_b, which may be complex valued. The steepest descent path corresponding to the oscillator in (C.2) can still be computed however.

Table C.1: Absolute error of the approximation of I_M by Filon-type quadrature with $(n+1)/2$ weights, for the example function $g(\beta) = (\beta+2)^{-3/2}$.

$R \setminus n$	1	3	5
10	$8.1E-06$	$5.9E-07$	$7.9E-08$
20	$2.7E-07$	$5.1E-09$	$1.8E-10$
40	$8.5E-09$	$4.1E-11$	$3.6E-13$

Numerical results are shown in Figure C.1 for a representative example integral. The parameter a in this figure is a linear scaling factor for the constants k_x, k_y and k_b. The results show that a very high accuracy can be obtained with a minimal computational effort. Moreover, the accuracy for a fixed number of operations greatly increases with increasing frequency.

C.2 Maxwell's equations

The Maxwell equations are a system of partial differential equations and model electromagnetic phenomena (see §2.2.3). A linear system of integral equations corresponding to the Maxwell equations can be derived in terms of a scalar kernel function. Numerical simulation using this system of boundary integral equations is practical only if the Green's function can be evaluated efficiently [198]. For planar, multi-layered structures, this requires the evaluation of integrals of the form

$$I_M = \int_0^\infty \beta g(\beta) J_0(\beta R)\, \mathrm{d}\beta, \tag{C.3}$$

with R a constant, and with g a smooth function [72]. The smooth function g is obtained after subtracting the known branch point and pole singularities of a non-smooth function $\tilde{g}(\beta)$. The subtracted integrals can be evaluated analytically.

The Bessel function of the first kind $J_0(z)$ can be written as $J_0(z) = \Re H_0^{(1)}(z)$. Integral (C.3) with $J_0(\beta R)$ replaced by $H_0^{(1)}(\beta R)$ is similar to the model form (6.1). It then becomes clear that, for $R > 0$, the behaviour of g near $\beta = 0$ determines the value of the integral. A localised Filon-type quadrature rule can be constructed that uses the function values $g^{(j)}(0)$,

$$I_M \approx \sum_{j=0}^{n} w_j g^{(j)}(0), \qquad \text{with } w_j := \Re \int_0^\infty ip \frac{(ip)^j}{j!} J_0(ipR) i\, \mathrm{d}p.$$

It turns out that $w_0 = 0$, and $w_1 = -1/R^3$. Hence, $I_M \approx -g'(0)/R^3$ is a good approximation for large R. Table C.1 shows results for different

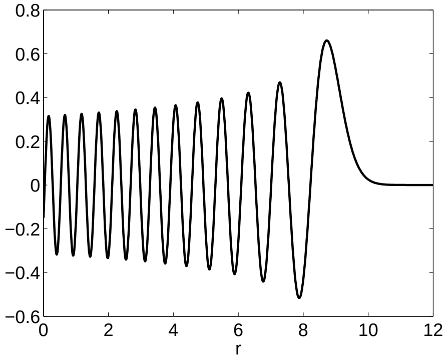


Figure C.2: Example integral for an element of the discretisation matrix in the modified J-matrix method (see [194], Fig. 1).

values of R and using higher order approximations. The weights w_j with even index j are always zero.

Asymptotic expansions have been used in other applications in electromagnetics for a long time, based on the steepest descent method (see [83]). The differences between the numerical approach described in Chapter 6 and the steepest descent method were summarised in §B.1. Use of the new techniques may lead to higher accuracy in these applications, especially for lower frequencies.

C.3 The modified J-matrix method in quantum scattering

Oscillatory phenomena arise naturally in quantum physics. The wave-like nature of small particles is reflected in their mathematical description by wave functions. These wave functions are found as the solution to Schrödingers' equation. The J-matrix method can be used to solve Schrödingers' equation for a given positive energy E and a potential $V(r)$ [110]. The solution in the free field, away from the influence of the potential $V(r)$, is qualitatively known. It needs to be matched quantitavely to the near field, the field that is in the range of the potential. In the J-matrix method, the solution is expanded into an oscillatory basis in the near field, which is then matched to the parameterised asymptotic form of the free-space solution. This results in a dense matrix.

A modification of the J-matrix method was proposed in [194]. There,

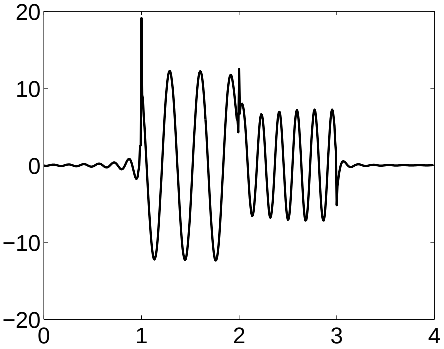


Figure C.3: Real part of the solution to the scattering problem for the case of a square scattering obstacle. The sides of the square have unit length. Two sides are in shadow, two sides are lit by the incoming plane wave.

the size of the matrix was reduced by using an approximation for the matrix elements in the so-called *far interaction* region, that lies in the range of the potential $V(r)$ between the near field and the free field. The method of stationary phase was applied to the oscillatory integrals that determine the matrix elements. As shown in Appendix B, the method of steepest descent allows the computation of additional coefficients of the asymptotic expansion. This may lead to more accurate computations, or to an even smaller dense matrix. An example integral is shown in Figure C.2. Contributions come from the boundary point $r = 0$, and from the stationary point at the so-called *turning point*, here at $r \approx 9.11$.

C.4 High frequency scattering by convex polygons

The hybrid method that was presented in Chapter 7 is specific for smooth and convex obstacles. An extension of the method to domains with corners would already greatly enhance the application possibilities. Corners in the obstacles introduce a singularity in the solution of the integral equation. Moreover, the diffraction of waves at the corner points introduces additional oscillatory behaviour of the solution. For the case of convex polygons, this oscillatory behaviour was determined as a function of the wavenumber in [36]. These authors formulate a Galerkin method that has a non-oscillatory solution. The elements of the Galerkin discretisation matrix

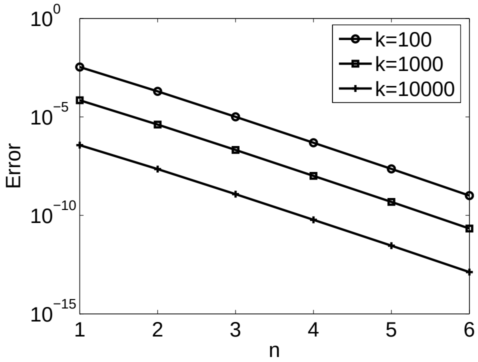


Figure C.4: Convergence of the multivariate numerical steepest descent method for integral (C.4), corresponding to one element of the Galerkin discretisation matrix. Paramter n is the number of quadrature points per dimension.

are given by integrals of the form

$$\int_{\Omega_m}\int_{\Omega_n}\Big(e^{ik(\sigma_m(s+x_m)+\sigma_n(t+x_n))}\eta H_0^{(1)}(kR)+ik\left[(a_lb_j-b_la_j)(t+x_n)\right.$$
$$\left.+b_l(c_l-c_j)-a_l(d_l-d_j)\right]H_1^{(1)}(kR)/R\Big)\,\mathrm{d}t\,\mathrm{d}s, \qquad \text{(C.4)}$$

with parameters that depend on the shape of the polygon. A collocation scheme can also be applied, leading to integrals of the form

$$\int_{\Omega_n}\Big(e^{ik(\sigma_m(s_m+x_m)+\sigma_n(t+x_n))}\eta H_0^{(1)}(kR)+ik\left[(a_lb_j-b_la_j)(t+x_n)\right.$$
$$\left.+b_l(c_l-c_j)-a_l(d_l-d_j)\right]H_1^{(1)}(kR)/R\Big)\,\mathrm{d}t,$$

with s_m the collocation point [10].

An efficient evaluation method for these integrals would lead to a solution method for scattering problems involving convex polygonal scatterers that is independent of the wavenumber. The evaluation method may be based on the results for multivariate oscillatory integrals in Chapter 6. Figure C.4 shows the result of applying the multivariate numerical steepest descent method to an integral of the form (C.4). The example integral corresponds to two basis functions defined on adjacent sides of a square. The figure shows the absolute error as a function of the number of quadrature points per dimension. High accuracy is obtained at a very low computational cost.

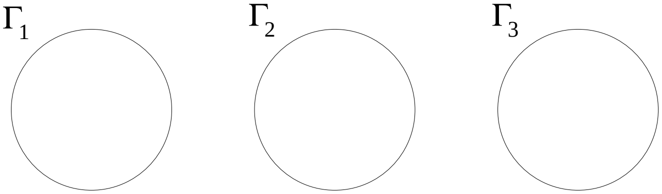


Figure C.5: Example of a multiple scattering configuration.

The solution of the scattering problem is illustrated for the case of a square in Figure C.3. The solution was computed with 500 piecewise linear basis functions. The figure illustrates the oscillatory behaviour of the solution, and the spikes at the corner points.

C.5 Multiple scattering

The results and the theory of this thesis were formulated for scattering problems involving a single scattering obstacle. The integral equation formulations can also be extended to cover multiple scattering configurations. In that case, the integral operator has the form

$$(Au)(x) = \sum_{i=1}^{L} \int_{\Gamma_i} G(x,y) u_i(y) \, \mathrm{d}s_y, \tag{C.5}$$

with $u = (u_1, u_2, \ldots, u_L)$ consisting of L density functions corresponding to L distinct obstacles. For the example configuration shown in Figure C.5, we have $L = 3$ obstacles with a circular shape.

The boundary element method can be used to solve integral equations involving operators of the form (C.5). The solution has the form

$$u_i(x) = \sum_{j=1}^{N_i} c_{ij} \phi_{ij}(x), \qquad i = 1, \ldots, L, \tag{C.6}$$

with basis functions ϕ_{ij} that are defined on Γ_i. We applied the boundary element method using wavelet packet basis functions on each boundary Γ_i, $i = 1, \ldots, L$ [122]. The results are shown in Figure C.6, in the same way as the numerical results of the wavelet packet method were presented in §3.8.2. The figure shows the number of nonzero matrix elements for increasing wavenumbers after compression of the transformed discretisation matrix, divided by the total number of basis functions $N := \sum_{i=1}^{L} N_i$. The

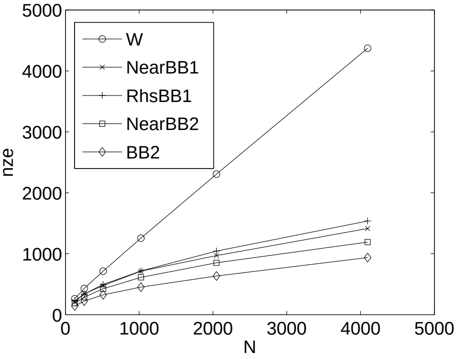


Figure C.6: Comparison of the sparsity of the discretisation matrix after a number of wavelet and wavelet packets transforms, for the scattering configuration shown in Figure C.5.

figure is similar to Figures 3.9 and 3.10. This result for multiple scattering confirms the earlier findings. The best result is obtained with the two-dimensional best basis algorithm (BB2). The other wavelet packets transforms (NearBB1,RhsBB1,NearBB2), that have a lower total computational cost, still outperform the regular wavelet method (W) for higher frequencies.

The ability to model multiple scattering increases the applicability of the boundary element method. For example, when the distance between the circular obstacles in Figure C.5 has the same order as the wavelength, this scattering problem models the typical interference pattern of waves.

Bibliography

[1] T. Abboud, J.-C. Nédélec, and B. Zhou. Méthode des équations intégrales pour les hautes fréquences. *C. R. Acad. Sci. Paris*, 318:165–170, 1994.

[2] K. R. Aberegg and A. F. Peterson. Application of the integral equation-asymptotic phase method to two-dimensional scattering. *IEEE Trans. Antennas Propagat.*, 43(5):534–537, 1995.

[3] M. J. Ablowitz and A. S. Fokas. *Complex variables: introduction and applications.* Cambridge University Press, Cambridge, 1997.

[4] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover Publications, New York, 1965.

[5] R. A. Adams. *Sobolev spaces.* Academic press, New York, 1975.

[6] M. Ainsworth, W. McLean, and T. Tran. The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling. *SIAM J. Numer. Anal.*, 36(6):1901–1932, 1999.

[7] S. Amini. On the choice of the coupling parameter in boundary integral formulations of the exterior acoustic problem. *Appl. Anal.*, 35:75–92, 1990.

[8] S. Amini and A. T. J. Profit. Multi-level fast multipole solution of the scattering problem. *Engineering Analysis with Boundary Elements*, 27(5):547–564, 2003.

[9] A. W. Appel. An efficient program for many-body simulation. *SIAM J. Sci. Stat. Comput.*, 6:446–449, 1985.

[10] S. Arden, S. Langdon, and S. N. Chandler-Wilde. A collocation method for high frequency scattering by convex polygons. *J. Comput. Appl. Math.*, 2005. To appear.

[11] K. E. Atkinson. *The numerical solution of integral equations of the second kind.* Cambridge University Press, Cambridge, 1997.

[12] I. M. Babuška and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wavenumbers? *SIAM Rev.*, 42(3):451–484, 2000.

[13] L. Banjai and W. Hackbusch. $\mathcal{H}$- and $\mathcal{H}^2$-matrices for low and high frequency Helmholtz equation. Technical Report 17, Max-Planck-Institut für Mathematik in den Naturwissenschaften, 2005.

[14] A. Barinka, T. Barsch, S. Dahlke., and M. Konik. Some remarks on quadrature formulas for refinable functions and wavelets. *ZAMM Z. Angew. Math. Mech*, 81(12):839–855, 2001.

[15] J. E. Barnes and P. Hut. A hierarchical $o(n \log n)$ force-calculation algorithm. *Nature*, 324(6270):446–449, 1986.

[16] M. Bebendorf. Approximation of boundary element matrices. *Numer. Math.*, 86(4):565–589, 2000.

[17] M. Bebendorf. Hierarchical LU decomposition based preconditioners for BEM. *Computing*, 74:225–247, 2005.

[18] G. Beylkin. Wavelets and fast numerical algorithms. In *Proceedings of Symposia in Applied Mathematics*, volume 47, 1993.

[19] G. Beylkin, R. R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.*, 44:141–183, 1991.

[20] N. Bleistein. Asymptotic expansions of integral transforms of functions with logarithmic singularities. *SIAM J. Math. Anal.*, 8(4):655–672, 1977.

[21] N. Bleistein and R. Handelsman. *Asymptotic expansions of integrals.* Holt, Rinehart and Winston, New York, 1975.

[22] B. D. Bonner, I. G. Graham, and V. P. Smyshlyaev. The computation of conical diffraction coefficients in high-frequency acoustic wave scattering. *SIAM J. Numer. Anal.*, 43(3):1202–1230, 2005.

[23] S. Börm. $\mathcal{H}^2$-matrices - Multilevel methods for the approximation of integral operators. *Comput. Vis. Sci.*, 7(3-4):173–181, 2004.

[24] S. Börm. Approximation of integral operators by $\mathcal{H}^2$-matrices with adaptive bases. *Computing*, 74(3):249–271, 2005.

[25] S. Börm and L. Grasedyck. HLib - a library for $\mathcal{H}$- and $\mathcal{H}^2$-matrices, 1999.

[26] S. Börm and L. Grasedyck. Hybrid cross approximation of integral operators. *Numer. Math.*, 101(2):221–249, 2005.

[27] M. Born and E. Wolf. *Principles of optics.* Cambridge University Press, Cambridge, 1999.

[28] H. Brakhage and P. Werner. Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung. *Arch. der Math.*, 16:325–329, 1965.

[29] O. P. Bruno, C. A. Geuzaine, J. A. Monro, and F. Reitich. Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: the convex case. *Phil. Trans. R. Soc. Lond. A*, 362(1816):629–645, 2004.

[30] O. M. Bucci and G. Franceschetti. On the degrees of freedom of scattered fields. *IEEE Trans. Ant. Prop.*, 37(7):918–926, 1989.

[31] R. L. Burden and J. D. Faires. *Numerical analysis.* PWS-Kent Publishing Company, Boston, 4 edition, 1989.

[32] C. Canuto, A. Tabacco, and K. Urban. The wavelet element method part I: Construction and analysis. *Appl. Comp. Harm. Anal.*, 6(1):1–52, 1999.

[33] B. Carpentieri, I. S. Duff, L. Giraud, and G. Sylvand. Combining fast multipole techniques and an approximate inverse preconditioner for large electromagnetism calculations. *SIAM J. Sci. Comput.*, 27(3):774–792, 2006.

[34] J. Carrier, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm for particle simulations. *SIAM J. Sci. Stat. Comput.*, 9(4):669–686, 1988.

[35] S. N. Chandler-Wilde and D. C. Hothersall. Efficient calculation of the Green function for acoustic propagation above a homogeneous impedance plane. *J. Sound Vibration*, 180:705–724, 1995.

[36] S. N. Chandler-Wilde and S. Langdon. A Galerkin boundary element method for high frequency scattering by convex polygons.

[37] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comput. Phys.*, 155:468–498, 1999.

[38] W. C. Chew. Computational electromagnetics: the physics of smooth versus oscillatory fields. *Phil. Trans. R. Soc. A*, 362(1816):579–602, 2004.

[39] W. C. Chew, J. M. Chin, C. C. Lu, E. Michielssen, and J. M. Song. Fast solution methods in electromagnetics. *IEEE Trans. Antennas and Propagation*, 45(3):533–543, 1997.

[40] W. C. Chew, J.-M. Jin, E. Michielssen, and J. Song. *Fast and efficient algorithms in computational electromagnetics.* Artech House, Boston, 2001.

[41] A. Cohen, W. Dahmen, and R. DeVore. Adaptive wavelet methods - beyond the elliptic case. *Found. of Comp. Math*, 2(3):203–245, 2002.

[42] A. Cohen and I. Daubechies. On the instability of arbitrary biorthogonal wavelet packets. *SIAM J. Math. Analysis*, 24(5):1340–1354, 1993.

[43] A. Cohen, I. Daubechies, and J.-C. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 55:485–560, 1992.

[44] A. Cohen and R. Masson. Wavelet adaptive method for second order elliptic problems. *Numer. Math.*, 86(2):193–238, 2000.

[45] R. R. Coifman, Y. Meyer, S. R. Quake, and M. V. Wickerhauser. Signal processing and compression with wavelet packets. In Y. Meyer and S. Roques, editors, *Progress in Wavelet Analysis and Applications*, pages 77–93. Editions Frontières, 1993.

[46] R. R. Coifman, Y. Meyer, and M. V. Wickerhauser. Size properties of wavelet packets. In M. B. Ruskai, G. Beylkin, R. R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Raphael, editors, *Wavelets and Their Applications*, pages 453–470. Jones and Bartlett, Boston, 1992.

[47] R. R. Coifman, V. Rokhlin, and S. Wandzura. The fast multipole method for the wave equation: A pedestrian prescription. *IEEE Antennas and Propagation Magazine*, 35:7–12, 1993.

[48] R. R. Coifman and M. V. Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 32:712–718, 1992.

[49] D. Colton and R. Kress. *Integral equation methods in scattering theory.* Wiley, New York, 1983.

[50] R. Cools and A. Haegemans. Algorithm 824: CUBPACK: A package for automatic cubature; framework description. *ACM Transactions on Mathematical Software*, 29(3):287–296, 2003.

[51] M. Costabel. Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Math. Anal.*, 19(3):613–626, 1988.

[52] W. Dahmen. Wavelet and multiscale methods for operator equations. In A. Iserles, editor, *Acta Numerica*, volume 6, pages 55–228. Cambridge Univ. Press, Cambridge, 1997.

[53] W. Dahmen, H. Harbrecht, and R. Schneider. Adaptive methods for boundary integral equations - complexity and convergence estimates. Technical Report IGPM Report #250, RWTH Aachen, 2005.

[54] W. Dahmen, H. Harbrecht, and R. Schneider. Compression techniques for boundary integral equations – asymptotically optimal complexity estimates. *SIAM J. Numer. Anal.*, 43:2251–2271, 2006.

[55] W. Dahmen and A. Kunoth. Multilevel preconditioning. *Numer. Math.*, 63:315–344, 1992.

[56] W. Dahmen, A. Kunoth, and R. Schneider. Operator equations, multiscale concepts and complexity. In J. Renegar, M. Shub, and S. Smale, editors, *Lectures in Applied Mathematics*, volume 32, pages 225–261. American Mathematical Society, 1996.

[57] W. Dahmen, A. Kunoth, and K. Urban. A wavelet Galerkin method for the Stokes problem. *Computing*, 56(3):259–302, 1996.

[58] W. Dahmen and C. A. Micchelli. Using the refinement equation for evaluating integrals of wavelets. *SIAM J. Numer. Anal.*, 30(2):507–537, 1993.

[59] W. Dahmen, S. Prößdorf, and R. Schneider. Wavelet approximation methods for periodic pseudodifferential equations on smooth manifolds. Part II. Fast solution and matrix compression. *Adv. Comput. Math.*, 1:259–335, 1993.

[60] W. Dahmen, S. Prößdorf, and R. Schneider. Multiscale methods for pseudo-differential equations on smooth closed manifolds. In C. K. Chui, L. Montefusco, and L. Puccio, editors, *Wavelets: Theory, Alorithms and Applications*, volume 5 of *Wavelet Analysis and Its Applications*, pages 385–424. Academic Press, 1994.

[61] W. Dahmen and R. Schneider. Composite wavelet bases for operator equations. *Math. Comp.*, 68(228):1533–1567, 1999.

[62] W. Dahmen and R. Schneider. Wavelets on manifolds I: construction and domain decomposition. *SIAM J. Math. Anal.*, 31(1):184–230, 1999.

[63] E. Darve. The fast multipole method I: Error analysis and asymptotic complexity. *SIAM J. Numer. Anal.*, 38(1):98–128, 2000.

[64] E. Darve. The fast multipole method: Numerical implementation. *J. Comput. Phys.*, 160(1):195–240, 2000.

[65] E. Darve and P. Havé. Efficient fast multipole method for low-frequency scattering. *J. Comput. Phys.*, 197(1):341–363, 2004.

[66] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure and Appl. Math.*, 41:909–996, 1988.

[67] I. Daubechies. *Ten lectures on wavelets.* SIAM, Philadelphia, 1992.

[68] K. T. R. Davies, M. R. Strayer, and G. D. White. Complex-plane methods for evaluating highly oscillatory integrals in nuclear physics. I. *J. Phys. G: Nucl. Phys.*, 14(7):961–972, 1988.

[69] P. J. Davis and P. Rabinowitz. *Methods of numerical integration.* Computer Science and Applied Mathematics. Academic Press, New York, 1984.

[70] P. Debye. Näherungsformeln für die Zylinderfunktionen für grosse Werte des Arguments und unbeschränkt veranderliche Werte des Index. *Math. Anal.*, 67:535–558, 1909.

[71] G. Deliège. *Flexible implementation of the finite element method applied to 3D coupled problems considering convective effects.* PhD thesis, K.U.Leuven, 2003.

[72] F. J. Demuynck. *The expansion wave concept.* PhD thesis, K.U.Leuven, 1995.

[73] F. J. Demuynck, G. A. E. Vandenbosch, and A. Van de Capelle. The expansion wave concept, part I: Efficient calculation of spatial Green's functions in a stratified dielectric medium. *IEEE Trans. Antennas Propagat.*, 46(3):397–406, 1998.

[74] H. Deng and H. Ling. Fast solution of electromagnetic integral equations using adaptive wavelet packet transform. *IEEE Trans. Antennas Propagat.*, 47(4):674–682, 1999.

[75] H. Deng and H. Ling. On a class of predefined wavelet packet bases for efficient representation of electromagnetic integral equations. *IEEE Trans. Antennas Propagat.*, 47(12):1772–1779, 1999.

[76] W. Desmet. *A wave based prediction technique for coupled vibroacoustic analysis.* PhD thesis, K.U.Leuven, 1998.

[77] W. Desmet, B. V. Hal, P. Sas, and D. Vandepitte. A computationally efficient prediction technique for the steady-state dynamic analysis of coupled vibro-acoustic systems. *Advances in Engineering Software*, 33:527 – 540, 2002.

[78] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev. A hybrid numerical-asymptotic boundary integral method for high-frequency acoustic scattering. Technical Report BICSP 1/06, University of Bath, 2006.

[79] M. Duffy. Quadrature over a pyramid or cube of integrands with a singularity at the vertex. *SIAM J. Numer. Anal.*, 19(6):1260–1262, 1982.

[80] M. E. Epton and B. Dembart. Multipole translation theory for the three-dimensional Laplace and Helmholtz equations. *SIAM J. Sci. Comput.*, 16(4):865–897, 1995.

[81] G. A. Evans and K. C. Chung. Some theoretical aspects of generalised quadrature methods. *J. Complexity*, 19:272–285, 2003.

[82] G. A. Evans and J. R. Webster. A comparison of some methods for the evaluation of highly oscillatory integrals. *J. Comput. Appl. Math.*, 112(1):55–69, 1999.

[83] L. B. Felsen and N. Marcuvitz. *Radiation and Scattering of Waves.* Prentice-Hall Inc., New Jersey, 1973.

[84] L. N. G. Filon. On a quadrature formula for trigonometric integrals. *Proc. Roy. Soc. Edinburgh*, 49:38–47, 1928.

[85] E. A. Flinn. A modification of Filon's method of numerical integration. *J. Assoc. Comp. Mach.*, 7:181–184, 1960.

[86] T. Gantumur and R. Stevenson. Computation of singular integral operators in wavelet coordinates. *Computing*, 76:77–107, 2006.

[87] W. Gautschi. *Orthogonal polynomials: computation and approximation.* Clarendon Press, Oxford, 2004.

[88] W. Gautschi, L. Gori, and F. Pitolli. Gauss quadrature for refinable weight functions. *Appl. Comput. Harmon. Anal.*, 8(3):249–257, 2000.

[89] I. M. Gel'fand and G. E. Shilov. *Generalized functions.* Academic Press, New York, 1964.

[90] C. Geuzaine, O. Bruno, and F. Reitich. On the O(1) solution of multiple-scattering problems. *IEEE Trans. Magn.*, 41(5):1488–1491, 2005.

[91] E. Giladi. Asymptotically derived boundary element method for the Helmholtz equation. In *Proceedings of the 7th International Conference on Mathematical and Numerical Aspects of Wave Propagation (WAVES2005)*, pages 420–422, 2005.

[92] E. Giladi and J. B. Keller. A hybrid numerical asymptotic method for scattering problems. *J. Comput. Phys.*, 174(1):226 – 247, 2001.

[93] W. L. Golik. Wavelet packets for fast solution of electromagnetic integral equations. *IEEE Trans. Antennas Propagat.*, 46(5):618–624, 1998.

[94] I. G. Graham and W. McLean. Anisotropic mesh refinement, the conditioning of Galerkin boundary element matrices and simple preconditioners. *SIAM J. Numer. Anal.*, 2006. To appear.

[95] L. Grasedyck. Adaptive recompression of $\mathcal{H}$-matrices for BEM. *Computing*, 74(3):205 – 223, 2005.

[96] L. Grasedyck and W. Hackbusch. Construction and Arithmetics of $\mathcal{H}$-matrices. *Computing*, 70(4):295–334, 2003.

[97] L. Greengard, J. Huang, V. Rokhlin, and S. Wandzura. Accelerating fast multipole methods for the Helmholtz equation at low frequencies. *IEEE Computational Science and Engineering*, 5(3):32–38, 1998.

[98] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.

[99] N. Gumerov and R. Duraiswami. *Fast multipole methods for the Helmholtz equation in three dimensions.* Elsevier, Amsterdam, 2004.

[100] W. Hackbusch. *Integral equations: Theory and numerical treatment.* Birkhäuser, Basel, 1995.

[101] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[102] W. Hackbusch and B. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic. Part II: applications to multidimensional problems. *Computing*, 64(1):21–47, 2000.

[103] W. Hackbusch and Z. Novak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.*, 54:463–491, 1989.

[104] H. Harbrecht, M. Konik, and R. Schneider. Fully discrete wavelet Galerkin schemes. *Engineering Analysis with Boundary Elements*, 27(5):439–454, 2003.

[105] H. Harbrecht and R. Schneider. Biorthogonal wavelet bases for the boundary element method. *Math. Nachr.*, 269:167–188, 2004.

[106] H. Harbrecht and R. Schneider. Wavelet Galerkin schemes for boundary integral equations - implementation and quadrature. *SIAM J. Sci. Comput.*, 27(4):1347–1370, 2006.

[107] R. F. Harrington. *Time-harmonic electromagnetic fields.* McGraw-Hill, New York, 1961.

[108] R. F. Harrington. *Field computation by moment methods.* Macmillan, New York, 1968.

[109] R. F. Harrington and J. R. Mautz. H-field, E-field and combined field solution for conducting bodies of revolution. *Arch. Elektron. Ubertragungstech*, 32(4):157–164, 1978.

[110] E. J. Heller and H. A. Yamani. J-matrix method: Appplication to S-wave electron-hydrogen scattering. *Phys. Rev. A*, 9(3):1209–1214, 1974.

[111] P. Henrici. *Applied and computational complex analysis Volume I.* Wiley & Sons, New York, 1974.

[112] N. Hess-Nielsen and M. V. Wickerhauser. Wavelets and time-frequency analysis. *Proceedings of the IEEE*, 84(4):523–540, 1996.

[113] E. Hille and J. D. Tamarkin. On the characteristic values of linear integral equations. *Acta Math.*, 57(1):1–76, 1931.

[114] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles.* McGraw-Hill, New York, 1981.

[115] G. C. Hsiao and W. L. Wendland. A finite element method for some integral equations of the first kind. *J. Math. Anal. Appl.*, 58:449–481, 1977.

[116] G. C. Hsiao and W. L. Wendland. *Encyclopedia of computational mechanics. Volume I: Fundamentals*, chapter Boundary element methods: foundation and error analysis, pages 339–373. John Wiley & Sons, 2004.

[117] B. Hu, W. Chew, E. Michielssen, and J. Zhao. Fast inhomogeneous plane wave algorithm for the fast analysis of two-dimensional scattering problems. *Radio Sci.*, 35(1):31–43, 2000.

[118] D. Huybrechs, J. Simoens, and S. Vandewalle. A note on wave number dependence of wavelet matrix compression for integral equations with oscillatory kernel. *J. Comput. Appl. Math.*, 172(2):233–246, 2004.

[119] D. Huybrechs and S. Vandewalle. The efficient evaluation of singular and oscillatory integrals arising in boundary element methods. In K. Chen, editor, *Advances in Boundary Integral Methods. Proceedings of the Fifth UK Conference on Boundary Integral Methods*, pages 20–30. University of Liverpool, 2005.

[120] D. Huybrechs and S. Vandewalle. Quadrature formulae for wavelet approximations of piecewise smooth or singular functions. *J. Comput. Appl. Math.*, 180(1):119–135, 2005.

[121] D. Huybrechs and S. Vandewalle. A two-dimensional wavelet packet transform for matrix compression of integral equations with highly oscillatory kernel. *J. Comput. Appl. Math.*, 2005. To appear.

[122] D. Huybrechs and S. Vandewalle. A wavelet-packet transformation for the fast solution of oscillatory integrals. In *Proceedings of the 7th International Conference on Mathematical and Numerical Aspects of Waves*, pages 399–401, 2005.

[123] D. Huybrechs and S. Vandewalle. The construction of cubature rules for multivariate highly oscillatory integrals. *Math. Comp.*, 2006. To appear.

[124] D. Huybrechs and S. Vandewalle. On the evaluation of highly oscillatory integrals by analytic continuation. *SIAM J. Numer. Anal.*, 2006. To appear.

[125] D. Huybrechs and S. Vandewalle. A sparse discretisation for integral equation formulations of high frequency scattering problems. Technical Report TW 447, K.U. Leuven, 2006.

[126] A. Iserles. On the numerical quadrature of highly-oscillating integrals I: Fourier transforms. *IMA J. Num. Anal.*, 24(3):365–391, 2004.

[127] A. Iserles. On the numerical quadrature of highly-oscillating integrals II: Irregular oscillators. *IMA J. Num. Anal.*, 25(1):25–44, 2005.

[128] A. Iserles and S. P. Nørsett. On quadrature methods for highly oscillatory integrals and their implementation. *BIT*, 44(4):755–772, 2004.

[129] A. Iserles and S. P. Nørsett. Efficient quadrature of highly oscillatory integrals using derivatives. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 461:1383–1399, 2005.

[130] A. Iserles and S. P. Nørsett. On the computation of highly oscillatory multivariate integrals with critical points. Technical Report NA08, University of Cambridge, 2005.

[131] A. Iserles and S. P. Nørsett. Quadrature methods for multivariate highly oscillatory integrals using derivatives. *Math. Comp.*, 2005. To appear.

[132] A. Iserles, S. P. Nørsett, and S. Olver. Highly oscillatory quadrature: the story so far. In *Proceedings of ENuMath 2005*, Santiago de Compostela, Berlin, 2006. Springer Verlag.

[133] S. Jaffard. Wavelet methods for fast resolution of elliptic equations. *SIAM J. Numer. Anal.*, 29(4):965–986, 1992.

[134] V. Jandhyala, E. Michielssen, S. Balasubramaniam, and W. C. Chew. A combined steepest descent-fast multipole algorithm for the fast analysis of three-dimensional scattering by rough surfaces. *IEEE Transactions on Geoscience and Remote Sensing*, 36(3):738–748, 1998.

[135] L. J. Jiang and W. C. Chew. Low-frequency fast inhomogeneous plane-wave algorithm (LF-FIPWA). *Microwave Opt. Technol. Lett.*, 40(2):117–122, 2004.

[136] J. B. Keller. Geometrical theory of diffraction. *J. Opt. Soc. Am.*, 52:116–130, 1962.

[137] K. J. Kim, R. Cools, and L. G. Ixaru. Quadrature rules using first derivatives for oscillatory integrands. *J. Comput. Appl. Math.*, 140(1-2):479–497, 2002.

[138] R. Kress. *Linear integral equations.* Springer-Verlag, Berlin, 1989.

[139] D. Lahaye. *Algebraic Multigrid for Two-Dimensional Time-Harmonic Magnetic Field Computations.* PhD thesis, K.U.Leuven, 2001.

[140] S. Langdon and S. N. Chandler-Wilde. Implementation of a boundary element method for high frequency scattering by convex polygons. In K. Chen, editor, *Proc. 5th U.K. Conf. on Boundary Integral Methods*, pages 2–11, 2005.

[141] S. Langdon and S. N. Chandler-Wilde. A wavenumber independent boundary element method for an acoustic scattering problem. *SIAM J. Numer. Anal.*, 43(6):2450–2477, 2006.

[142] D. Laurie and J. De Villiers. Orthogonal polynomials for refinable linear functionals. *Math. Comp.*, 2004. To appear.

[143] R. Leis. Zur Dirichletschen Randwertaufgabe des Aussenraumes der Schwingungsgleichung. *Math. Z.*, 90:209–211, 1965.

[144] D. Levin. Procedure for computing one- and two-dimensional integrals of functions with rapid irregular oscillations. *Math. Comp.*, 38(158):531–538, 1982.

[145] D. Levin. Fast integration of rapidly oscillatory functions. *J. Comput. Appl. Math.*, 67(1):95–101, 1996.

[146] D. Levin. Analysis of a collocation method for integrating rapidly oscillatory functions. *J. Comput. Appl. Math.*, 87(1):131–138, 1997.

[147] J. Lighthill. *Waves in fluids.* Cambridge University Press, Cambridge, 1978.

[148] P. Linz. *Analytical and numerical methods for Volterra equations.* SIAM, Philadelphia, 1985.

[149] J. L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications*, volume 1. Springer, Berlin, 1972.

[150] G. Little and J. R. Reade. Eigenvalues of analytic kernels. *SIAM J. Math. Anal.*, 15(1):133–136, 1984.

[151] Lord Kelvin (W. Thomson). On the wave produced by a single impulse in water of any depth or in a dispersive medium. *Philos. Mag.*, 23:252–255, 1887.

[152] C.-C. Lu and W. C. Chew. A multilevel algorithm for solving a boundary integral equation of wave scattering. *Microwave Opt. Technol. Lett.*, 7(10):466–470, 1994.

[153] Y. L. Luke. On the computation of oscillatory integrals. *Proc. Cambridge Phil. Soc.*, 50:269–277, 1954.

[154] W. McLean. *Strongly elliptic systems and boundary integral equations.* Cambridge University Press, Cambridge, 2000.

[155] R. B. Melrose and M. E. Taylor. Near peak scattering and the corrected Kirchhoff approximation for a convex obstacle. *Adv. in Math.*, 55(3):242–315, 1985.

[156] E. Michielssen and W. C. Chew. The fast steepest descent path algorithm for analyzing scattering from two-dimensional objects. *Radio Science*, 31(5):1215–1224, 1996.

[157] R. Miller. *Nonlinear Volterra integral equations.* Benjamin/Cummings publishing, San Francisco, 1971.

[158] J.-C. Nédélec. Integral equations with non-integrable kernels. *Integral Equations Operator Theory*, 4:563–572, 1982.

[159] J.-C. Nédélec. *Acoustic and Electromagnetic Equations*, volume 144 of *Applied Mathematical Sciences.* Springer, Berlin, 2001.

[160] S. Ohnuki and W. C. Chew. Truncation error analysis of multipole expansions. *SIAM J. Sci. Comput.*, 25(4):1293–1306, 2003.

[161] S. Olver. Moment-free numerical integration of highly oscillatory functions. *IMA J. Num. Anal.*, 26(2):213–227, 2006.

[162] S. Olver. On the quadrature of multivariate highly oscillatory integrals over non-polytope domains. *Numerische Mathematik*, 2006. To appear.

[163] S. Osher and C.-W. Shu. High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations. *SIAM J. Numer. Anal.*, 28(4):907–922, 1991.

[164] P. Oswald. *Multilevel finite element approximations.* B.G. Teubner, Stuttgart, 1994.

[165] O. Panich. On the question of the solvability of the exterior boundary-value problems for the wave equation and maxwells equations. *Usp. Mat. Nauk*, 20A:221–226, 1965.

[166] E. Passow and L. Raymon. Monotone and comonotone approximation. *Proc. Amer. Math. Soc.*, 42:340–349, 1974.

[167] E. Passow, L. Raymon, and J. A. Roulier. Comonotone polynomial approximation. *J. Approx. Th.*, 11:221–224, 1974.

[168] B. Pluymers, W. Desmet, D. Vandepitte, and P. Sas. On the use of a wave based prediction technique for steady-state structural-acoustic radiation analysis. *Journal of Computer Modeling in Engineering & Sciences*, 7(2):173–184, 2005.

[169] M. J. D. Powell. *Approximation theory and methods.* Cambridge University Press, Cambridge, 1981.

[170] J. Rahola. Diagonal forms of the translation operators in the fast multipole algorithm for scattering problems. *BIT*, 36(2):333–358, 1996.

[171] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *J. Comput. Phys.*, 60:187–207, 1985.

[172] V. Rokhlin. Rapid solution of integral equations of scattering theory in two dimensions. *J. Comput. Phys.*, 86(2):414–439, 1990.

[173] V. Rokhlin. Diagonal forms of translation operators for the Helmholtz equation in three dimensions. *Appl. Comput. Harmon. Anal.*, 1(1):82–93, 1993.

[174] G. Schmidlin, C. Lage, and C. Schwab. Rapid solution of first kind boundary integral equations in R^3. *Engineering Analysis with Boundary Elements*, 27(5):469–490, 2003.

[175] R. Schneider. *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung großer vollbesetzter Gleichungssysteme.* B. G. Teubner, Stuttgart, 1998.

[176] C. Schwab. Variable order composite quadrature of singular and nearly singular integrals. *Computing*, 53:173–194, 1994.

[177] L. Schwartz. *Théorie des Distributions.* Hermann, Paris, 1966.

[178] E. M. Stein. *Harmonic analysis: Real-variable methods, orthogonality and oscillatory integrals.* Princeton University Press, Princeton, 1993.

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

[180] R. Stevenson. Composite wavelet bases with extended stability and cancellation properties. Technical Report 1345, Utrecht University, 2006.

[181] G. G. Stokes. On the numerical calculation of a class of definite integrals and infinite series. *Camb. Philos. Trans.*, 9:166–187, 1856.

[182] X. Sun and N. Pitsianis. A matrix version of the fast multipole method. *SIAM Rev.*, 43(2):189–200, 2001.

[183] W. Sweldens and R. Piessens. Asymptotic error expansion of wavelet approximations of smooth functions II. *Numer. Math.*, 68(3):377–401, 1994.

[184] W. Sweldens and R. Piessens. Quadrature formulae and asymptotic error expansions for wavelet approximations of smooth functions. *SIAM J. Numer. Anal.*, 31(4):1240–1264, 1994.

[185] A. Talbot. The accurate numerical inversion of Laplace transforms. *J. Inst. Math. Appl.*, 23(1):97–120, 1979.

[186] G. Toraldo Di Francia. Degrees of freedom of an image. *J. Opt. Soc. Am.*, 59:799–804, 1969.

[187] E. E. Tyrtyshnikov. Mosaic-skeleton approximations. *Calcolo*, 33:47–57, 1996.

[188] E. E. Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton approach. *Computing*, 64:367–380, 2000.

[189] B. Van Hal. *Automation and performance optimization of the wave based method for interior structural-acoustic problems.* PhD thesis, K.U.Leuven, 2004.

[190] B. Van Hal, W. Desmet, and D. Vandepitte. A coupled finite element - wave based approach for the steady state dynamic analysis of acoustic systems. *Journal of Computational Acoustics*, 11(2):255 – 283, 2003.

[191] G. Vanden Berghe and L. G. Ixaru. *Exponential fitting.* Kluwer Academic Publishers, Dordrecht, 2004.

[192] G. A. Vandenbosch and F. J. Demuynck. The expansion wave concept, part II: A new way to model mutual coupling in microstrip arrays. *IEEE Trans. Antennas Propagat.*, 46(3):407–413, 1998.

[193] B. Vandereycken. 3D-simulatie van electromagnetische golven met wavelets. Master's thesis, K.U. Leuven, 2005.

[194] W. Vanroose, J. Broeckhove, and F. Arickx. Modified J-matrix method for scattering. *Phys. Rev. Lett.*, 88:010404, 2002.

[195] V. Vinje, E. Iversen, and H. Gjøstdal. Traveltime and amplitude estimation using wavefront construction. *Geophysics*, 58:1157–1166, 1993.

[196] T. von Petersdorff and C. Schwab. Fully discrete multiscale Galerkin BEM. In W. Dahmen, A. Kurdila, and P. Oswald, editors, *Multiscale wavelet methods for PDEs*. Academic Press, 1997.

[197] T. von Petersdorff, C. Schwab, and R. Schneider. Multiwavelets for second kind integral equations. *SIAM J. Numer. Anal.*, 34(6):2212–2227, 1997.

[198] M. Vrancken. *Full wave integral equation based electromagnetic modelling of 3D metal structures in planar stratified media*. PhD thesis, K.U.Leuven, 2003.

[199] R. L. Wagner and W. C. Chew. A study of wavelets for the solution of electromagnetic integral equations. *IEEE Trans. Antennas Propagat.*, 43(8):802–810, 1995.

[200] G. N. Watson. Harmonic functions associated with the parabolic cylinder. *Proc. London Math. Soc. Series 2*, 17:116–148, 1918.

[201] G. B. Whitham. *Linear and nonlinear waves*. Wiley, New York, 1974.

[202] M. V. Wickerhauser. Nonstandard matrix multiplication. Preprint, Yale University, 1990.

[203] G. M. Wing. *A primer on integral equations of the first kind*. SIAM, Philadelphia, 1991.

[204] R. Wong. *Asymptotic approximation of integrals*. SIAM, Philadelphia, 2001.

[205] J.-S. Zhao and W. C. Chew. MLFMA for solving integral equations of 2-D electromagnetic problems from static to electrodynamic. *Microwave Opt. Technol. Lett.*, 20(5):306–311, 1999.

Curriculum Vitae

Higher Education

2002–2006 Ph.D. in Engineering, Katholieke Universiteit Leuven, Leuven, Belgium
Thesis: Multiscale and hybrid methods for the solution of oscillatory integral equations

1997–2002 Burgerlijk Ingenieur in de Computerwetenschappen (Engineer in Computer Science), Katholieke Universiteit Leuven, Leuven, Belgium
Thesis: Wavelets and wavelet packets for integral equations

Teaching

2002–2005 Teaching Assistant 'Numerical Approximation and Geometric Modelling'

Co-supervision of master students

2005 Bart Vandereycken. Thesis: The three dimensional simulation of electromagnetic waves with wavelets
Rewarded by the Jos Schepens Memorial Fund.

Publications in peer reviewed journals

- D. Huybrechs, J. Simoens, and S. Vandewalle. *A note on wave number dependence of wavelet matrix compression for integral equations with oscillatory kernel*, J. Comput. Appl. Math. 172(2), pp. 233-246, December, 2004.

- D. Huybrechs, and S. Vandewalle. *Quadrature formulae for wavelet approximations of piecewise smooth and singular functions*, J. Comput. Appl. Math. 180(1), pp. 119-135, August, 2005.
- D. Huybrechs, and S. Vandewalle. *A two-dimensional wavelet packet transform for matrix compression of integral equations with highly oscillatory kernel*, J. Comput. Appl. Math., to appear.
- D. Huybrechs, and S. Vandewalle. *On the evaluation of highly oscillatory integrals by analytic continuation*, SIAM J. Numer. Anal., to appear.
- D. Huybrechs, and S. Vandewalle. *The construction of cubature rules for multivariate highly oscillatory integrals*, Math. Comp., to appear.
- D. Huybrechs, and S. Vandewalle. *A sparse discretisation for integral equation formulations of high frequency scattering problems*, SIAM J. Sci. Comput., submitted.

Publications in conference proceedings

- D. Huybrechs. *A two-dimensional wavelet packet transformation for the fast solution of highly oscillatory integral equations*, Proceedings of the 7th International Conference on Mathematical and Numerical Aspects of Wave Propagation (Abboud, T. et al, ed.), pp. 399-401. 7th International Conference on Mathematical and Numerical Aspects of Waves (WAVES05), Brown University, Providence RI, USA, June 20-24, 2005.
- D. Huybrechs. *The efficient evaluation of highly oscillatory integrals in BEM by analytic continuation*, Advances in Boundary Integral Methods (Ke Chen, ed.), pp. 20-31. 5th UK Conference on Boundary Integral Methods (UKBIM5), Liverpool University, September 12-13, 2005.

Index

admissibility condition, 92, 107
asymptotic method, 122
asymptotic order, 120

best basis algorithm, 57
block cluster tree, 106
boundary element method, 36

cluster tree, 105
column cluster basis, 109
combined field integral equation, 35
compact operator, 17
compact perturbation, 18
critical points, 152

degenerate critical point, 152
Dirac's delta function, 23
double-layer potential, 26
dual space, 17

fast multipole method, 87
Filon-type method, 123
Fredholm Alternative, 20
Fredholm integral equation, 14
Fredholm operator, 19
fundamental solution, 23

Green's function, 23
Green's identities, 23

Helmholtz equation, 9
hierarchical matrix, 106
hypersingular operator, 29

illuminated region, 188
interaction list, 93

kernel function, 23

Levin-type method, 124
Lipschitz continuous, 16
Lipschitz domain, 16
local expansions, 89

Maxwell equations, 10
multilevel fast multipole method, 93
multipole expansions, 89

nested cluster basis, 109
nonresonance condition, 152
numerical steepest descent, 125

order of an operator, 33
oscillator, 120

radar cross section, 12
regularity, 37
resonance frequency, 34
resonance points, 152, 155
Riesz' representation theorem, 17
row cluster basis, 109

scaling function, 42
sesquilinear form, 19
shadow boundary, 192
shadow region, 188
sifting property, 24

single level fast multipole method, 90
single-layer potential, 25
Sobolev space, 15
Sommerfeld radiation condition, 21
stationary point, 120
steepest descent FMM, 101
steepest descent path, 131

trace operator, 16

vanishing moments, 42
Volterra integral equation, 13

Watson's Lemma, 213
wave equation, 8
wavelet, 42
wavelet packets, 55
wavenumber, 9