Contents

Chapter 1

Introduction

The processes of scattering and absorption of light determines how we perceive the objects and substances in our environment. Scattering processes are responsible for many phenomena in the atmosphere. For example, the blue color of the sky and the spectrum of colors in a rainbow are due to the scattering of light from gas and water molecules in the air.

In this work our motivation has been the scattering of light in astronomy. Scattering processes occur when light from stars propagates through interplanetary dust clouds or when light is reflected off solar system bodies. In the case of atmosphereless bodies in the solar system, such as the Moon, there are some old optical observations that haven't yet been fully explained [28]. One phenomenon is that for example the brightness of the Moon increases rapidly when the Moon becomes a full Moon, or in other words, the Moon is near the opposition. This so called *opposition effect* cannot be explained without considering the micron-scale structure of the dust on the surface on the Moon.

Another observation is the reversal of linear polarization near the opposition. The degree of linear polarization is given by the ratio

$$P = \frac{I_\perp - I_\|}{I_\perp + I_\|}, \tag{1.1}$$

where $I_\perp$ is the intensity component of the wave perpendicular to and $I_\|$ parallel to the scattering plane. The scattering plane is defined by the source of radiation (e.g. the Sun), the object that scatters the light (e.g. the Moon), and the observer (e.g. on the Earth). For atmosphereless solar system bodies, the degree of linear polarization is negative near the opposition and changes sign to positive at an angle of approximately 20°. This reversal angle is the same for a variety of bodies.

It is probable that the optical properties of the Moon can be derived from the optical properties of the dust grains on its surface. Small irregular porous particles scatter light in a way that can explain both the opposition effect and the reversal of linear polarization [28].

Here we are interested how a single particle scatters light. For a general introduction to light scattering, see [4]. The size of the particles considered here is up to 10 times the wavelength. For larger particles, the computational methods become too expensive and we have to rely on geometric optics which is most easily applied to particles of regular shape. On the other hand, if particles are very small compared to the wavelength, the shape of the particle has little effect on the scattering process. Our interest is in particles whose dimensions are of the same order of magnitude than the wavelength.

The scattering of light is described by a volume integral equation over the volume of the particle. Actually, this equation gives the field inside the particle when the particle is illuminated with light. The scattered field can be easily computed once the incident field and the internal field are known. The scattering of a particle can be approximated by replacing the particle with interacting electric dipoles. This so called discrete-dipole approximation and the integral equation formalism are explained in Chapter 2 and their relation is shown. Numerical techniques for integral equations are reviewed. The stochastic geometries that are used to model the scatterer are also described in this chapter.

The computational problem in the scattering calculations is the solution of a large dense system of linear equations where the coefficient matrix is complex symmetric. The main part of this thesis considers the efficient solution of such equations.

Dense matrices often arise from the discretization of integral equations. As a contrast, discretization of partial differential equations leads to sparse matrices so that the matrix-vector products are easy to compute. On the other hand, iterative methods may converge slowly for such problems; preconditioning is often mandatory. In dense matrices the problem is quite the opposite. Iterative methods converge quickly while the most computationally intensive part is the computation of the matrix-vector product.

Iterative methods are generally used with sparse coefficient matrices. They can also be successfully applied to dense coefficient matrices. Iterative methods can easily outperform direct methods such as Gaussian elimination that require on the order of n^3 operations for a linear system of n equations. In Chapter 3, several iterative methods are tested on the dense system of equations arising from the scattering calculations. The spectrum of some coefficient matrices is computed and plotted. The performance of iterative methods can in some cases be quite accurately predicted from the distribution of the eigenvalues of the matrix.

The computationally most intensive task in the solution of dense linear systems is the computation of matrix-vector products. The straightforward way to compute the product of a matrix and a vector takes n^2 operations, where n is the number of equations. In Chapter 4 we present two fast ways to compute the matrix-vector product.

In regular geometries the matrix-vector product reduces to a convolution that can be calculated with the fast Fourier transform. The other algorithm is the so called fast multipole method that has been used to compute the interactions of particles for example in N-body simulations under the gravitational or electromagnetic forces. The matrix-vector multiplication actually means that we have to compute the interactions of the dipoles, that is, to compute the total electric field radiated by all the other dipoles at each dipole position. For this interaction, the fast multipole method is derived and implemented. A comparison of the performance of the various methods for the matrix-vector product is also presented.

Finally, in Chapter 5 we conclude what has been achieved in the thesis and we also list the points where future research would be beneficial. Part of the results reported here have also been published in [34].

Chapter 2

The scattering equations

The scattering of light from arbitrary-shaped particles is governed by an integral equation over the volume of the particle. On the other hand, the scattering process can be approximated by modeling the particle as a collection of electric dipoles. These two views will be considered in this chapter and it is shown that the dipole approximation can be derived from the integral equation formalism.

We will assume that incident radiation is a monochromatic electromagnetic plane wave that propagates with wave vector $\mathbf{k}$ and with wave number $k = |\mathbf{k}|$. A harmonic time dependence $exp(-i\omega t)$ is assumed, where ω is the angular frequency of the incident radiation. Thus the electric field of the plane wave is given by

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_1 e^{i\mathbf{k}\cdot\mathbf{r} - i\omega t}. \tag{2.1}$$

In the rest of this thesis, we will leave out the time dependence, which means that we are effectively working with Fourier transformed quantities. We will only consider elastic or coherent scattering, where the scattered radiation has the same frequency as the incident field. The relations between k, ω, and the wavelength λ are

$$\begin{aligned} \lambda &= 2\pi c/\omega, \\ k &= 2\pi/\lambda = \omega/c, \end{aligned}$$

where c is the speed of light.

2.1 The discrete-dipole approximation

In the *discrete-dipole approximation* (DDA), the scatterer is modeled as group of interacting electric dipoles. The dipole moments $\mathbf{p}_i$ and thus the electric fields $\mathbf{E}_i$

radiated by the dipoles are unknown. The method is also called the coupled dipole method by some authors [24, 19].

An oscillating electric dipole with dipole moment $\mathbf{p}$ situated at the origin gives rise to an electric field given by [23]

$$\mathbf{E}(\mathbf{r}) = k^2(\hat{\mathbf{r}} \times \mathbf{p}) \times \hat{\mathbf{r}}\frac{e^{ikr}}{r} + (3\hat{\mathbf{r}}(\hat{\mathbf{r}} \cdot \mathbf{p}) - \mathbf{p})\left(\frac{1}{r^3} - \frac{ik}{r^2}\right) e^{ikr}. \tag{2.2}$$

Here $r = |\mathbf{r}|$ and $\hat{\mathbf{r}} = \mathbf{r}/r$. When we take into account that $(\hat{\mathbf{r}} \times \mathbf{p}) \times \hat{\mathbf{r}} = \mathbf{p} - \hat{\mathbf{r}}(\hat{\mathbf{r}} \cdot \mathbf{p})$, the previous equation can be written in tensor notation as

$$\mathbf{E}(\mathbf{r}) = k^3 \left(\frac{e^{ikr}}{(kr)^3}(k^2r^2 + ikr - 1)\mathbf{1} + \frac{e^{ikr}}{(kr)^3}(-k^2r^2 - 3ikr + 3)\hat{\mathbf{r}}\hat{\mathbf{r}}\right) \cdot \mathbf{p}, \tag{2.3}$$

where $\mathbf{1}$ is the unit dyadic and $\hat{\mathbf{r}}\hat{\mathbf{r}}$ is the symmetric dyadic formed from the radial unit vectors.

At each dipole position $\mathbf{r}_i$, the electric field is determined by the incident field $\mathbf{E}_i^0 = \mathbf{E}^0 \exp(i\mathbf{k} \cdot \mathbf{r}_i)$ plus the field radiated by the other dipoles. Here we have neglected the magnetic dipole radiation. The dipole moment $\mathbf{p}$ is proportional to the total electric field at the dipole position, $\mathbf{p} = \alpha\mathbf{E}$, where α is the polarizability. The total electric field at the dipole positions is determined by the system of equations

$$\mathbf{E}_i = \mathbf{E}_i^0 + k^3\alpha \sum_{j \neq i}^{N_{dip}} \mathbf{T}_{ij} \cdot \mathbf{E}_j, \quad i = 1, \ldots, N_{dip}. \tag{2.4}$$

The dipole interaction tensor (or dyadic) is given by

$$\begin{aligned} \mathbf{T}_{ij} &= \frac{e^{i\rho_{ij}}}{\rho_{ij}^3}(\rho_{ij}^2 + i\rho_{ij} - 1)\mathbf{1} + \frac{e^{i\rho_{ij}}}{\rho_{ij}^3}(-\rho_{ij}^2 - 3i\rho_{ij} + 3)\hat{\mathbf{r}}_{ij}\hat{\mathbf{r}}_{ij}, \\ \rho_{ij} &= k|\mathbf{r}_i - \mathbf{r}_j|, \end{aligned} \tag{2.5}$$

where $\hat{\mathbf{r}}_{ij} = (\mathbf{r}_i - \mathbf{r}_j)/|\mathbf{r}_i - \mathbf{r}_j|$.

In order to couple the polarizability of the dipole to the index of refraction, we assume that each dipole represents a spherical volume with radius r_0. We define a dimensionless size parameter as $x_0 = 2\pi r_0/\lambda = kr_0$. Note that in Equation (2.5) all the dimensions are of the form kr so that the scattering process only depends on the ratio of the dimensions of the particle to the wavelength. The interdipole distance parameter ρ_{ij} is given in terms of the radius and the size parameter as $\rho_{ij} = x_0|\mathbf{r}_i - \mathbf{r}_j|/r_0$. The coefficient $k^3\alpha$ depends on the index of refraction m as [28]

$$k^3\alpha \approx x_0^3\frac{m^2 - 1}{m^2 + 2} + \frac{3}{5}x_0^5\frac{(m^2 - 2)(m^2 - 1)}{(m^2 + 2)^2} + i\frac{2}{3}x_0^6\left(\frac{m^2 - 1}{m^2 + 2}\right)^2. \tag{2.6}$$

Here we are assuming the the index of refraction is constant over the particle. Because cubical volume elements are approximated by spheres, the effective index of refraction used in the calculation and the actual index of refraction of the particle need to be connected using semi-empirical effective medium theories [28].

The computational challenge in the DDA method is to solve the system of linear equations (2.4). If the particle consists of N_{dip} dipoles, the linear equation is of the size $3N_{dip} \times 3N_{dip}$.

2.2 Integral equation for scattering

A more general approach than the DDA is to formulate the integral equation for scattering. Let V be the space that the particle occupies and m be the index of refraction that is allowed to vary within the particle. The scattering integral equation reads [2]

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_{\text{inc}}(\mathbf{r}) + k^3 \int_V (m(\mathbf{r}')^2 - 1)\mathbf{G}(\mathbf{r}, \mathbf{r}') \cdot \mathbf{E}(\mathbf{r}')\, d^3\mathbf{r}', \tag{2.7}$$

where $\mathbf{G}$ is the dyadic Green's function

$$\mathbf{G}(\mathbf{r}, \mathbf{r}') = \left(\mathbf{1} + \frac{\nabla\nabla}{k^2}\right) g(|\mathbf{r} - \mathbf{r}'|) \tag{2.8}$$

and

$$g(r) = \frac{e^{-ikr}}{kr}. \tag{2.9}$$

The differentiation can be either with respect to $\mathbf{r}$ or $\mathbf{r}'$. The explicit form of $\mathbf{G}$ is

$$\mathbf{G}(\mathbf{r}, \mathbf{r}') = g(r)\left(\left(1 + \frac{i}{kr} - \frac{1}{k^2r^2}\right)\mathbf{1} - \left(1 + \frac{3i}{kr} - \frac{3}{k^2r^2}\right)\hat{\mathbf{r}}\hat{\mathbf{r}}\right), \tag{2.10}$$

where $r = |\mathbf{r} - \mathbf{r}'|$. The incident field is typically a plane wave of the form

$$\mathbf{E}_{\text{inc}}(\mathbf{r}) = \mathbf{E}_1 e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{2.11}$$

Equation (2.7) is a linear singular integral equation of the second kind for the unknown electric field $\mathbf{E}(\mathbf{r})$. The kernel of the integral equation is singular at $\mathbf{r} = \mathbf{r}'$. The singularity is of the order $1/r^3$ and is thus not integrable. The integral equation is not a Fredholm integral equation because the kernel is not compact.

The singularity is most naturally handled by computing the integral in the sense of a Cauchy principal value integral, i.e. by excluding a vanishingly small volume V_ϵ around the singularity [43]. The geometry of this "principal volume" influences the integral equation. For cubical and spherical principal volumes the integral equation reads

$$\mathbf{E}(\mathbf{r}) - (m^2(\mathbf{r}) - 1)\mathbf{E}(\mathbf{r})/3 = \mathbf{E}_{\text{inc}}(\mathbf{r}) + k^3 \lim_{\epsilon \to 0} \int_{V - V_\epsilon} (m(\mathbf{r}')^2 - 1)\mathbf{G}(\mathbf{r}, \mathbf{r}') \cdot \mathbf{E}(\mathbf{r}')\, d^3\mathbf{r}'. \tag{2.12}$$

The formulation of the scattering problem as an integral equation and the handling of the singularity has also been discussed in [33].

Another formulation in terms of an integro-differential equation is given in [19]:

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_{\text{inc}}(\mathbf{r}) + k^3 \int_V g(|\mathbf{r} - \mathbf{r}'|) \left(\mathbf{1} + \frac{\nabla\nabla}{k^2}\right) \cdot \left((m(\mathbf{r}')^2 - 1)\mathbf{E}(\mathbf{r}')\right) d^3\mathbf{r}'. \tag{2.13}$$

2.3 Numerical methods for integral equations

The integral equation (2.7) is of the form

$$x(s) = y(s) + \int_V K(s,t)x(t)\, dt, \tag{2.14}$$

for the unknown $x(s)$. Define the linear operator $\mathcal{K}$ by

$$(\mathcal{K}x)(s) = \int_V K(s,t)x(t)\, dt. \tag{2.15}$$

Now the integral equation can be written as

$$(1 - \mathcal{K}x)(s) = y(s). \tag{2.16}$$

The scattering integral equations can be solved by the Ritz-Galerkin method, also called the method of moments.

In the Galerkin method, the unknown function is expanded in terms of basis functions h_i. If set of basis functions $\{h_i\}$ is a complete set in the appropriate function space, the solution can be approximated with arbitrary precision. Initially we will require that the basis functions are orthonormal. The solution candidate has the form

$$\tilde{x}(s) = \sum_{i=1}^{N} a_i h_i(s), \tag{2.17}$$

where a_i are unknown coefficients to be determined. We shall form the residual function

$$r = y - (1 - \mathcal{K})\tilde{x}. \tag{2.18}$$

We would like the residual to be “as small as possible”. One way to accomplish this is to require that the residual be orthogonal to the basis functions with respect to the inner product defined by

$$(f, g) = \int_V f(s)g(s)\,ds. \tag{2.19}$$

The residual would be forced to zero if we could use the whole complete set of basis functions. In practice we of course use a finite number of basis functions. The Galerkin equations read

$$\int_V h_i(s)(y(s) - (1 - \mathcal{K}\tilde{x})(s))\,ds = 0, \quad i = 1, \ldots, N. \tag{2.20}$$

This leads to a dense system of linear equations for the unknown expansion coefficients $a = \begin{pmatrix} a_1 & \ldots & a_N \end{pmatrix}$:

$$La = y, \tag{2.21}$$

where

$$L_{ij} = \int_V h_i(s)h_j(s)\,ds - \int_V \int_V h_i(s)K(s,t)h_j(t)\,dsdt, \tag{2.22}$$

$$y_i = \int_V h_i(s)y(s)\,ds. \tag{2.23}$$

In practical calculations, the set of basis functions has a finite number of members. They do not form a complete set of functions, but they will approximate the solution to an arbitrary precision if the number of functions is increased and the size of the support of the basis functions is diminished. The basis functions need not be orthogonal, the same equations (2.21) are satisfied by non-orthogonal functions.

2.4 Discretization of the scattering integral equations

A common practice in scattering calculations is to use piecewise constant basis functions so that the electric field is constant within each computational cell [27, 22, 24]. Furthermore, the integrals (2.22) and (2.23) are computed using a simple one-point quadrature so that the value of the integral is the value of the integrand at the center times the volume of the computational cell. For this reason, some authors also call the method of moments the digitized Green's function algorithm [19].

The diagonal terms L_{ii}, also called the self-terms, need special attention, because the Green's function has a singularity at $\mathbf{r} = \mathbf{r}'$. The integrals associated with the diagonal terms can be estimated by analytically calculating the integrals over a spherical region where a small spherical volume has been excluded from the integral [24].

The system of linear equations from the method of moments is [24]

$$\left(1-(m(\mathbf{r}_i)^2-1)\left(k^3M-\frac{1}{3}\right)\right)\mathbf{E}_i=\mathbf{E}_i^0+k^3V_1\sum_{\substack{j=1\\j\neq i}}^{N}(m(\mathbf{r}_j)^2-1)\mathbf{T}_{ij}\cdot\mathbf{E}_j, \quad (2.24)$$

where $i=1,\ldots,N$, V_1 is the volume of the computational box, $\mathbf{T}_{ij}$ is given by Equation (2.5), and M is given by

$$M=\frac{2}{3k^3}\left(\left(1-ikb(3/4\pi)^{1/3}\right)e^{ikb(3/4\pi)^{1/3}}-1\right). \quad (2.25)$$

Here b is the length of side of the computational box [24].

The discrete-dipole approximation can be derived from the integral equation formalism. If one requires that the residual (2.18) is zero at the centers of individual computational domains, one gets an equation similar the DDA equations (2.4). This technique is known as the collocation method or point-matching technique.

The principal difference with the DDA and the method of moments or Galerkin method is the handling of the self-term in the discretization. Apart from the diagonal term, the systems of linear equations arising from the method of moments and the DDA are the same. One can say that the discretizations presented here are the strong form of the method of moments and the weak form of the discrete-dipole approximation [24]. Actually, the DDA with effective-medium theories and the method of moments are very closely related, as can be seen in Equations (2.4) and (2.24).

The method of moments can be refined by using more accurate integration methods in numerical integration. Especially the integrals over neighboring computational cells should be computed with more precision. Also the self-integral could be approximated more accurately. The constant basis functions can be replaced with higher order functions, e.g. with vector spherical harmonics [33]. The DDA is appealing because of its simplicity as it can be derived directly from dipole interactions. However, the integral equation formalism has several benefits. The inclusion of more accurate self-terms gives more precision in the scattering calculations [22]. In the integral equation formalism the index of refraction appears naturally in the equations, whereas in the DDA one has to use effective-medium theories and other approximations to connect the index of refraction of the whole particle to the dipole polarizabilities.

Although the integral equation formalism does have some advantages over the DDA, we will concentrate on the efficient solution of the linear equations arising from the DDA. The systems of linear equations from these two methods are rather similar so that the solution methods developed for the DDA can be readily applied to the volume integral formalism. The DDA was chosen mainly because the author has more experience in using it. In the future, we expect to work also with the integral equation formalism and compare the results to those obtained using the DDA. In the case of a spherical scatterer, the results can be verified against the Mie theory [4].

2.5 Computational geometry

In our calculations, the three main geometries for the dipoles were a pseudospherical cluster of dipoles, a sparse hierarchical pseudosphere, or a particle generated by the diffusion-limited aggregation process.

The pseudospheres are generated by taking all the dipoles in a regular lattice that are within a given radius from the origin. A number of particles can be stochastically removed from the particle. An example for such a geometry is given in Fig 2.1.

The hierarchical pseudospheres are pseudospheres where each basic unit is itself a pseudosphere. In general, the pseudospheres in both these hierarchy levels can be porous ones, where a number of particles have been removed. An example of a full pseudospherical cluster of small pseudospheres is given in Fig 2.2.

The diffusion-limited aggregation (DLA) process builds fractal clusters by a three-dimensional random walk [29]. We start with a single seed particle. Each new particle is injected far away from the cluster of particles that has been generated so far. The new particle undergoes a random walk. It either hits the cluster and sticks to it or wanders so far away that it can be considered as lost. A sample cluster generated with the DLA process is given in Figure 2.3.

In the scattering calculations, one has to solve the scattering equations for two incident polarization states that are parallel and perpendicular to the scattering plane. The computed scattering quantities have to be averaged because one relatively small scatterer does not produce physically feasible results. One typically needs on the order of 100–200 orientational averages.

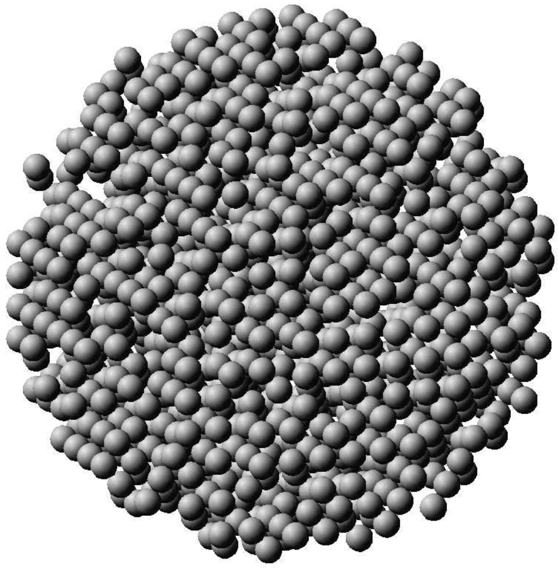

Figure 2.1: An example of a pseudosphere where a number of particles have been removed.

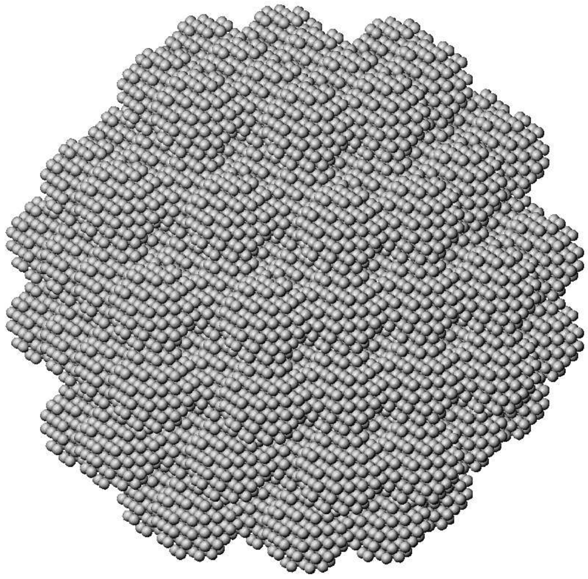

Figure 2.2: A pseudospherical cluster of 136 pseudospheres of 136 dipoles each.

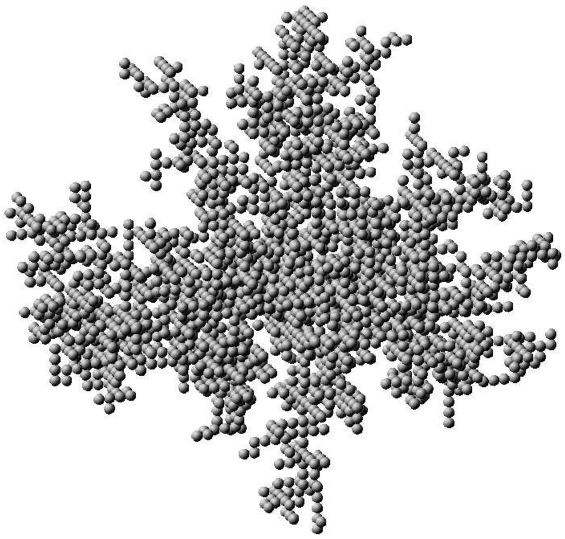

Figure 2.3: A cluster of dipoles generated by the diffusion-limited aggregation process.

Chapter 3

Solution of dense systems of linear equations

Many computational problems in science and technology lead to linear systems of equations with sparse coefficient matrices. These problems are often described by partial differential equations. Discretizing a partial differential equation naturally leads to equations where only the nearest-neighbour interactions are present thus forming a sparse linear system.

In contrast, most dense linear systems are derived from the discretization of integral equations or from other systems where global interactions are present. Dense matrices are of course more expansive to work with than sparse matrices. Anyhow, dense linear systems do offer some benefits over sparse systems, as will be shown below.

Because of the global nature of interactions in the discretizations of integral equations, the resulting linear systems tend to be much better conditioned than corresponding linear systems arising in the discretization of partial differential equations. When discretization is refined in partial differential equations, the conditioning gets even worse. The conditioning suggests that iterative methods should perform quite well with integral equations whereas preconditioning techniques are almost mandatory with partial differential equations.

Systems of linear equations can be solved either with direct methods based on Gaussian elimination, or with iterative methods. It is widely accepted that iterative methods are generally more powerful than direct methods, especially in three-dimensional problems.

To solve a dense linear system, direct linear solvers require $\mathcal{O}(n^3)$ arithmetic operations and $\mathcal{O}(n^2)$ memory locations, where n is the number of unknowns. The LAPACK project has produced efficient dense solvers that are optimized on many computer platforms from vector supercomputers to parallel computers. But even with this powerful technology, solution of dense linear systems with direct method is generally limited to systems with less than 10 000 unknowns. The computer power grows rapidly with the size of the linear system; in order to double the number of unknowns one needs eight times the computer power. Also the amount on central memory becomes a bottleneck at some point.

The performance of iterative solver for dense linear systems is determined by the number of iterations and by the number of operations that are needed to apply the coefficient matrix to a given vector. In dense problems, the number of iterations is typically a slowly growing function of the size of the system. Going from direct to iterative solvers can as such reduce the computation time by many orders of magnitude.

A modern recipe for solving dense linear algebra systems is given by Edelman in his survey [10]: use a preconditioned iterative method which accesses the matrix only by matrix-vector multiplication and look for fast approximate methods for computing the matrix-vector product. The fast methods can speed up the calculations by many orders of magnitude. They also consume much less memory allowing one to solve systems with hundreds of thousands of unknowns.

Typically iterative methods are used with sparse matrices, but they can be very successful also with dense matrices. Actually, for very large dense problems, iterative methods are the only plausible solution method.

In this chapter we will first review some iterative methods that can be used with the dense systems arising from electromagnetic scattering calculations. The iterative method QMR will be described in more detail, as it was chosen to be the iterative method in the production calculations. The spectrum of the linear equations arising from the scattering calculations will be shown and the effects of the structure of the spectrum on the convergence are discussed. At the end of this chapter, various iterative methods are tested. The computation of the matrix-vector product is discussed in the next chapter.

Although we are working with the linear equations arising from the DDA formulation of electromagnetic scattering, the techniques described here can be applied to other systems of dense linear equations arising from the discretization of integral equations.

3.1 Iterative methods

Iterative methods produce a sequence of iterates that are hopefully converging towards the solution of the linear system. The linear system to be solved is denoted by

$$Ax = b. \tag{3.1}$$

The simplest iterative methods, such as the Jacobi and Gauss-Seidel iterations, are based on the splitting of the coefficient matrix and have the form

$$Mx^{k+1} = Nx^k + b, \tag{3.2}$$

where $A = M - N$.

Here we shall be working with the so called Krylov subspace methods, that access the coefficient matrix only by matrix-vector multiplication. A Krylov subspace generated by r and A is defined as

$$K_k(r, A) := \operatorname{span}\{r, Ar, \ldots, A^{k-1}r\}. \tag{3.3}$$

All Krylov subspace methods generate iterates of the form

$$x^k \in x^0 + K_k(r^0, A), \tag{3.4}$$

where r^0 is the initial residual $b - Ax^0$ with initial guess x^0.

One design goal for Krylov subspace methods is to minimize some norm of the residual $r^k = b - Ax^k$:

$$\|r^k\| = \min_{x \in x^0 + K_k(r^0, A)} \|b - Ax\|. \tag{3.5}$$

Another possibility is to require that the residual fulfills a Galerkin-type condition:

$$s^H r^k = 0 \quad \forall\, s \in \mathcal{S}_k, \tag{3.6}$$

where $\mathcal{S}_k$ is a subspace of dimension k and ^{H} denotes the complex transpose.

The classical conjugate gradient method, that will be described below, is optimal in the following sense: its iterates satisfy a minimization property, namely

$$\|b - Ax^k\|_{A^{-1}} = \min_{x \in x^0 + K_k(r^0, A)} \|b - Ax\|_{A^{-1}} \tag{3.7}$$

and the iterates x^k can be computed from a simple three-term recurrence.

The conjugate gradient algorithm only works with Hermitian positive definite matrices (or real symmetric positive definite matrices). For non-Hermitian matrices, optimal methods are difficult to obtain. To this end, Faber and Manteuffel have proved a theorem for non-Hermitian matrices [12, 13], which we state here in the form given in [16].

Theorem 1 *Except for a few anomalies, ideal CG-like methods whose iterates are characterized by a minimal residual property (3.5) or an orthogonal residual property (3.6) and which can be implemented based on short vector recursions, exist only for matrices of the special form*

$$A = e^{i\theta}(T + \sigma I), \tag{3.8}$$

where T is Hermitian, $\theta \in \mathbb{R}$, $\sigma \in \mathbb{C}$.

We will now describe some iterative methods that can be used to solve the complex symmetric (but non-Hermitian) system of linear equations arising from the discrete-dipole approximation. The methods will be given here in a form that is suitable for complex matrices. The complex transpose will be denoted by ^{H} and the complex inner product by $(a, b) = a^H b$. A review of iterative solution of linear solution can be found in [16]. The methods are also presented in [1, 42]. The emphasis of the presentation of algorithms is on the clarity of the algorithm; in the actual computer implementations we use temporary variables to hold various matrix-vector products and inner products.

3.1.1 CG

The conjugate gradient method (CG) is a widely used Krylov-subspace method. It can be derived from the optimization problem: find the minimum of the function

$$f(x) = \frac{1}{2} x^H A x - b^H x + c, \tag{3.9}$$

where A is a Hermitian positive definite matrix. The minimum of the function f will be found in a point where the gradient is zero, or

$$\nabla f(x) = Ax - b = 0 \tag{3.10}$$

Thus by minimizing the function $f(x)$ the system of linear equations $Ax = b$ is solved at the same time.

Many Krylov subspace method produce new iterates minimizing the value of the objective function in a given direction p^k from the old iterate:

$$x^{k+1} = x^k + \alpha_k p^k, \tag{3.11}$$

where α_k is chosen so that $f(x^{k+1})$ is minimized.

In the conjugate gradient algorithm the search directions p^k are chosen so that they are conjugate with respect to the matrix A, that is

$$(p^i, Ap^j) = 0 \quad \text{when} \quad i \neq j. \tag{3.12}$$

$$
\begin{array}{|l|}
\hline
\\
\quad r^0 = b - Ax^0 \\
\quad p^0 = r^0 \\
\quad k = 0 \\
\quad \textbf{repeat} \\
\qquad \alpha_k = \dfrac{(r^k, r^k)}{(p^k, Ap^k)} \\
\qquad x^{k+1} = x^k + \alpha_k p^k \\
\qquad r^{k+1} = r^k - \alpha_k Ap^k \\
\qquad \beta_k = \dfrac{(r^{k+1}, r^{k+1})}{(r^k, r^k)} \\
\qquad p^{k+1} = r^{k+1} + \beta_k p^k \\
\qquad k = k + 1 \\
\quad \textbf{until } \|r^k\| < \epsilon \\
\\
\hline
\end{array}
$$

Table 3.1: The conjugate gradient algorithm (CG).

The conjugate gradient algorithm is given in Table 3.1. The operations during each iteration are a matrix vector product Ap^k, inner products (r^k, r^k), and vector updates $x^{k+1} = x^k + \alpha_k p^k$.

An upper limit to the convergence of the conjugate gradient method is given in terms of the condition number $\kappa(A)$. The condition number is defined by $\kappa(A) = \|A\|\|A^{-1}\|$ and it specifies how the solution of the linear system is affected by perturbations of the matrix and the right-hand side vector. The CG error behaves like

$$\|x - x^k\|_A \leq 2\|x - x^0\|_A \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k . \tag{3.13}$$

Here x is the exact solution of the linear system and the norm is defined by $\|x\|_A = \sqrt{(x, Ax)}$. The actual convergence of CG depends on the whole spectrum of A. For example, if the eigenvalues of A are highly clustered, CG will converge in only a few iterations.

The conjugate gradient algorithm only works if the coefficient matrix is Hermitian positive definite. Thus the conjugate gradient algorithm as such cannot be applied to the linear system arising from the DDA method.

$$
\begin{aligned}
& r^0 = b - Ax^0 \\
& p^0 = A^H r^0 \\
& k = 0 \\
& \textbf{repeat} \\
& \qquad \alpha_k = \frac{(Ar^k, Ar^k)}{(Ap^k, Ap^k)} \\
& \qquad x^{k+1} = x^k + \alpha_k p^k \\
& \qquad r^{k+1} = r^k - \alpha_k A p^k \\
& \qquad \beta_k = \frac{(Ar^{k+1}, Ar^{k+1})}{(Ar^k, Ar^k)} \\
& \qquad p^{k+1} = Ar^{k+1} + \beta_k p^k \\
& \qquad k = k + 1 \\
& \textbf{until } \|r^k\| < \epsilon
\end{aligned}
$$

Table 3.2: The conjugate gradient method applied to the normal equations (CGN-R).

3.1.2 CGNR

One way to use the conjugate gradient algorithm with non-Hermitian matrices is to apply it to the normal equations

$$A^H A x = A^H b. \tag{3.14}$$

Here the coefficient matrix $A^H A$ is Hermitian positive definite provided that A is regular. When conjugate gradient is applied to (3.14), the method is called CGNR. If CG is applied to normal equations

$$A A^H y = b, \quad x = A^H y, \tag{3.15}$$

the method is called CGNE [16, 42].

The major drawback in using the CGNR (or CGNE) algorithm is that the condition number of the linear system is essentially squared in the formation of the normal equations: $\kappa(A^H A) = \kappa(A)^2$. In the particular case of the DDA equations, the normal equations are still usable as the original linear system has a very low condition number. The CGNR and CGNE algorithms require two matrix-vector products per iteration.

The CGNR algorithm is given in Table 3.2.