



Contents lists available at ScienceDirect

Journal of Quantitative Spectroscopy & Radiative Transfer

journal homepage: www.elsevier.com/locate/jqsrt

Implementation and investigation of iterative solvers in the Discrete Sources Method

Roman Schuh^a, Vladimir Schmidt^{a,*}, Thomas Wriedt^b, Yuri Eremin^c^a Universität Bremen, Badgasteiner Str. 3, 28359 Bremen, Germany^b Institut für Werkstofftechnik, Badgasteiner Str. 3, 28359 Bremen, Germany^c Faculty of Applied Mathematics and Computer Science, Lomonosov State University, Lenin's Hills, 119992 Moscow, Russia

ARTICLE INFO

Available online 21 January 2011

Keywords:

Electromagnetic scattering
 Discrete Source Method
 Krylov subspace iterative solvers
 Least-squares problem

ABSTRACT

The implementation of iterative methods as solvers for the Discrete Sources Method (DSM) is presented. In this method, light scattering computation linear systems with dense and relative small matrices are generated. The linear systems are traditionally solved using the QR-decomposition method. For large particles or particles with extreme geometries even this commonly stable solver can fail. In these cases, we expect that iterative methods can provide a satisfying solution nevertheless.

We will present our investigation in two consecutive papers. Here, we study four different iterative solvers (RGMRES, BiCGStab, BiCGStab(*l*), and MinRes) considering the performance and the accuracy for typical light scattering problems. Using these iterative methods we increased the quality of a solution, especially for oblate spheroids with a higher aspect ratio. Preconditioning technique is considered in the following paper.

Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Numerical simulation of light scattering by particles in the Mie scattering regime is a modern and effective approach to investigate many physical problems like optical particle characterization. A review of available computational methods has recently been presented by Wriedt [1]. The Discrete Sources Method (DSM) is a well-established method for computations of electromagnetic scattering by axisymmetric scatterers. This method is based on a generalized point-matching method (GPM) [2], where spherical vector wave functions with multiple origins are used in order to get a good approximation of the internal and external fields. The theory of

the method is outlined in detail by Eremin et al. [3]. The DSM has been used to investigate different demanding scattering problems such as defects on surfaces [4], fibres [5], flat platelets [6], erythrocytes [7], TIRM (total internal reflection microscopy) [8,9], and nanoholes within a film [10].

One part of the numerical scheme of DSM is to solve the linear system problem. In the conventional programme this problem is solved using a direct method, the QR decomposition. This could be the subject of numerical difficulties and instabilities, especially for larger particles or particles with high aspect ratios. For very large objects, the direct method can completely fail because of high ill-conditionality of the kernel matrix and the finite precision arithmetic in computers.

As a rule, for large demanding problems iterative solvers numerically are considered more stable than the direct ones [11]. The induced iterative process can be stopped at a given accuracy of solution which is usually reached much earlier than in direct methods, especially

* Corresponding author. Tel.: +49 421 218 5418;

fax: +49 421 218 3912.

E-mail addresses: schuh@iwt.uni-bremen.de (R. Schuh),
vschmidt@iwt.uni-bremen.de (V. Schmidt),
thw@iwt.uni-bremen.de (T. Wriedt), eremin@cs.msu.ru (Y. Eremin).

for large matrices. Simplifications for the matrix-vector multiplication for sparse matrices can also significantly reduce the computational demand. Utilization of the preconditioning technique increases the range of applicability of the iterative solver.

Various efficient methods have been developed to solve large scale problems. The fast multipole method (FMM) by Greengard and Rokhlin [12] is one such method which accelerates the matrix vector multiplication. To apply FMM to DSM an iterative solver will have to be implemented. This investigation is a first step in the implementation of FMM in DSM. The extremely large, sparse, ill-conditioned kernel matrix is the case where the iterative methods mostly are preferable [13]. The numerical complexity of iterative methods is $O(n^2k)$, where n is the number of unknowns and k is the number of applied iterations, instead of $O(n^3)$ for direct methods. The numerical complexity for FMM would be even lower and corresponds to $O(n \log nk)$.

In our investigation we focus on light scattering computations for extremely large particles and particles with high aspect ratios using DSM. In this paper, we study four different iterative solvers (RGMRES, BiCGStab, BiCGStab(l), and MinRes) considering the performance and the accuracy of a solution. In the following paper, we will study different preconditioning techniques and estimate their influence on the quality of the solution.

The theory of the Discrete Sources Method is briefly outlined in Section 2. In the following sections we present different established iterative methods. In Section 5, we examine in detail four different iterative solvers in order to compare the behaviour regarding computational time, stability and accuracy. In addition, we compare the results from the iterative methods with the direct solver and a concurrent scattering computation method, the Null-Field Method with Discrete Sources (NFM-DS) [14]. This is an extension of the original Null-Field Method, also known as T-Matrix method, which was originally proposed by Waterman [15].

2. Mathematical statement

Let us start with the mathematical statement of the scattering problem. We will consider scattering in an isotropic homogeneous medium R^3 of an electromagnetic wave by a local homogeneous penetrable obstacle D_i with the smooth boundary ∂D . We assume the time dependence to be $\exp(j\omega t)$. Scattering is described by the electromagnetic fields $\{\mathbf{E}_{e,i}, \mathbf{H}_{e,i}\}$ satisfying the Maxwell equations:

$$\begin{aligned} \nabla \times \mathbf{H}_{e,i} &= jk\epsilon_{e,i}\mathbf{E}_{e,i} \\ \nabla \times \mathbf{E}_{e,i} &= -jk\mu_{e,i}\mathbf{H}_{e,i} \end{aligned} \quad \text{in } D_{e,i}, \quad D_e := R^3/\bar{D}_i, \quad (1)$$

the boundary conditions enforced on the particle surface

$$\begin{aligned} \mathbf{n}_p \times (\mathbf{E}_i(P) - \mathbf{E}_e(P)) &= \mathbf{n}_p \times \mathbf{E}^0(P) \\ \mathbf{n}_p \times (\mathbf{H}_i(P) - \mathbf{H}_e(P)) &= \mathbf{n}_p \times \mathbf{H}^0(P) \end{aligned} \quad P \in \partial D, \quad (2)$$

and Silver–Muller radiation condition at infinity

$$\lim_{r \rightarrow \infty} \left(\sqrt{\epsilon_e} \mathbf{E}_e \times \frac{\mathbf{r}}{r} - \sqrt{\mu_e} \mathbf{H}_e \right) = 0, \quad r = |M| \rightarrow \infty, \quad (3)$$

where $\{\mathbf{E}^0, \mathbf{H}^0\}$ is an exciting field, \mathbf{n}_p is the unit outward normal to ∂D , index e belongs to the external domain D_e and i to the domain inside the particle D_i , $\epsilon_{e,i}$ is the permittivity, $\mu_{e,i}$ is the permeability of media, $\text{Im } \epsilon_e, \mu_e = 0$, $\text{Im } \epsilon_i, \mu_i \leq 0$. The boundary value scattering problem is well known to have a unique solution [16].

2.1. Discrete Sources Method

In frame of the Discrete Sources Method (DSM) an approximate solution of the scattering problem is constructed as a finite linear combination of the field of dipoles and multipoles $\{z_n\}_{n=1}^N$ deposited in a supplementary domain ω_0 . We assume that an exciting field is a p-polarized plane wave with incident angle θ_0 . The scattering from the S-polarized plane wave can be derived analogously [17]. The algorithm of approximate solution construction has some differences for elongated and flat particles, that is why we will separate those two cases. In case of an elongated particle the system of lowest order multipoles distributed on the axis of symmetry z can be applied to construct an approximation solution [5]. In case of a flat particle the system of multipoles is situated in the complex plane.

Taking into account polarization of the plane wave and axial symmetry of the particle the approximate solution can be represented in the form

$$\begin{aligned} \begin{pmatrix} \mathbf{E}_{e,i}^N \\ \mathbf{H}_{e,i}^N \end{pmatrix} &= \sum_{m=0}^M \sum_{n=1}^{N_{e,i}^m} \{p_{mn}^{e,i} D_1 \mathbf{A}_{mn}^{1,e,i} + q_{mn}^{e,i} D_2 \mathbf{A}_{mn}^{2,e,i}\} \\ &+ \sum_{n=1}^{N_{e,i}^0} r_n^{e,i} D_1 \mathbf{A}_n^{3,e,i}, \end{aligned} \quad (4)$$

with differential operators D_1, D_2

$$D_1 = \begin{pmatrix} \frac{j}{k\epsilon_{e,i}\mu_{e,i}} \nabla \times \nabla \\ -\frac{1}{\mu_{e,i}} \nabla \end{pmatrix}, \quad D_2 = \begin{pmatrix} \frac{1}{\epsilon_{e,i}} \nabla \\ \frac{j}{k\epsilon_{e,i}\mu_{e,i}} \nabla \times \nabla \end{pmatrix},$$

and the vector potentials in a cylindrical coordinate system

$$\begin{aligned} \mathbf{A}_{mn}^{1,e,i} &= Y_m^{e,i}(\eta, z_n^{e,i}) \{\mathbf{e}_r \cos(m+1)\phi - \mathbf{e}_\theta \sin(m+1)\phi\}, \\ \mathbf{A}_{mn}^{2,e,i} &= Y_m^{e,i}(\eta, z_n^{e,i}) \{\mathbf{e}_r \sin(m+1)\phi + \mathbf{e}_\theta \cos(m+1)\phi\}, \\ \mathbf{A}_n^{3,e,i} &= Y_m^{e,i}(\eta, z_n^{e,i}) \mathbf{e}_z, \end{aligned} \quad (5)$$

where

$$\begin{aligned} Y_m^e(x) &= h_m^{(2)}(k_e R_{\eta\tilde{\xi}}) P_m^m(\cos\tilde{\theta}_{\tilde{\xi}}), \quad Y_m^i(x) = j_m(k_e R_{\eta\tilde{\xi}}) P_m^m(\cos\tilde{\theta}_{\tilde{\xi}}), \\ R_{\eta\tilde{\xi}} &= \sqrt{\rho^2 + (z - \tilde{\xi})^2}, \quad \sin\tilde{\theta}_{\tilde{\xi}} = \frac{\rho}{R_{\eta\tilde{\xi}}}, \\ \cos\tilde{\theta}_{\tilde{\xi}} &= \frac{z - \tilde{\xi}}{R_{\eta\tilde{\xi}}}, \quad \text{Re}(R_{\eta\tilde{\xi}}) > 0, \end{aligned} \quad (6)$$

where $h_m^{(2)}(kr)$ and $j_m(kr)$ are the spherical Hankel and Bessel functions of order m . Here $R_{\eta\tilde{\xi}}$ is a function of the complex variable $\tilde{\xi}$ and it is chosen, so that it represents a branch corresponding to the arithmetical root at the positive part of the real axis. Under these conditions the representation for the approximate solution satisfies Maxwell's equations (1) in $D_{e,i}$ and the radiation condition (3).

The electric and magnetic components of the p-polarized plane wave can be represented as follows:

$$\mathbf{E}^0 = (\mathbf{e}_x \cos \theta_0 + \mathbf{e}_z \sin \theta_0) \mathbf{e}_x \gamma \cos \theta_0, \quad \mathbf{H}^0 = -\mathbf{e}_y \gamma \cos \theta_0, \quad (7)$$

where $\gamma = \exp\{-jk\sqrt{\epsilon_0 \mu_0}(x \sin \theta_0 - z \cos \theta_0)\}$ and θ_0 is the incident angle. The exciting plane wave can be resolved into the Fourier series with respect to ϕ , using the following resolution for the plane wave:

$$\exp\{-jk_e \rho \sin \theta_0 \cos \phi\} = \sum_{m=0}^{\infty} (2 - \delta_{0m}) (-j)^m J_m(k_e \rho \sin \theta_0) \cos m \phi, \quad (8)$$

where $J_m(kr)$ is the cylindrical Bessel function.

The unknown amplitudes of discrete sources are to be determined from the boundary conditions (2). To solve this problem the General Matching-Point Technique is used. Matching of the approximate solution and external excitation over particle surface is replaced by matching over the particle generatrix $\{\eta_n\}_{n=1}^L$ for each Fourier harmonic m separately. As a consequence, the unknown vector of amplitudes $\mathbf{p}_m = \{p_{mn}^e, q_{mn}^e\}_{n=1}^{N_m^e}$ can be found as a pseudosolution of an overdetermined system of linear equations

$$\mathbf{B}_m \mathbf{p}_m = \mathbf{q}_m, \quad m = 0, \dots, M, \quad (9)$$

where \mathbf{B}_m is a rectangular matrix of dimension $4L \times 2(N_i^m + N_e^m)$ and the vector \mathbf{q}_m can be represented as the following $4L$ vector:

$$\mathbf{q}_m = (e_{m+1,l}^{0\tau}, e_{m+1,l}^{0\phi}, h_{m+1,l}^{0\tau}, h_{m+1,l}^{0\phi})^T,$$

where

$$\begin{aligned} e_m^{0\tau}(\eta) &= (-j)^m \{\alpha \cos \theta_0 [J_m(k_e \rho \sin \theta_0) - J_{m+2}(k_e \rho \sin \theta_0)] \\ &\quad - 2j\beta \sin \theta_0 J_{m+1}(k_e \rho \sin \theta_0)\} e^{-jk_e z \cos \theta_0}, \\ e_m^{0\phi}(\eta) &= (-j)^m \cos \theta_0 [J_m(k_e \rho \sin \theta_0) + J_{m+2}(k_e \rho \sin \theta_0)] e^{-jk_e z \cos \theta_0}, \\ h_m^{0\tau}(\eta) &= -(-j)^m \alpha [J_m(k_e \rho \sin \theta_0) + J_{m+2}(k_e \rho \sin \theta_0)] e^{-jk_e z \cos \theta_0}, \\ h_m^{0\phi}(\eta) &= -(-j)^m [J_m(k_e \rho \sin \theta_0) - J_{m+2}(k_e \rho \sin \theta_0)] e^{-jk_e z \cos \theta_0}. \end{aligned}$$

Here, the $(\alpha, 0, \beta)$ -vector is tangential to the generatrix at the matching point $\eta = (\rho, z)$. Similarly, amplitudes $\mathbf{p}_{-1} = \{r_n^{e,i}\}_{n=1}^{N_{e,i}^m}$ correspondent to vertical electric dipoles can be obtained from

$$\mathbf{B}_{-1} \mathbf{p}_{-1} = \mathbf{q}_{-1}, \quad (10)$$

where \mathbf{B}_{-1} has a dimension $2L \times (N_i^m + N_e^m)$ and the right-hand side vector \mathbf{q}_{-1} has length $2L$ with elements

$$\begin{aligned} e_{-1}^{0\phi}(\eta) &= [j\alpha \cos \theta_0 J_1(k_e \rho \sin \theta_0) - \beta \sin \theta_0 J_0(k_e \rho \sin \theta_0)] e^{-jk_e z \cos \theta_0}, \\ h_{-1}^{0\tau}(\eta) &= -jJ_1(k_e \rho \sin \theta_0) e^{-jk_e z \cos \theta_0}. \end{aligned}$$

2.2. Far-field

After the amplitudes of discrete sources $\{\mathbf{p}_m\}_{m=-1}^M$ have been determined, the far-field pattern can be

computed [16]:

$$\frac{\mathbf{E}(\mathbf{r})}{|\mathbf{E}^0(\mathbf{r})|} = \frac{\exp(-jk_e r)}{r} (\mathbf{e}_\theta \cdot F_\theta(\theta, \phi) + \mathbf{e}_\phi \cdot F_\phi(\theta, \phi)) + o\left(\frac{1}{r}\right), \quad r \rightarrow \infty, \quad (11)$$

where components $F_\theta(\theta, \phi)$, $F_\phi(\theta, \phi)$ can be found using asymptotic representation for Y_{mn}^e

$$\begin{aligned} F_\theta(\theta, \phi) &= j \sum_{m=0}^M \cos(m+1)\varphi (j \sin \theta)^m \sum_{n=1}^{N_m^e} \{p_{mn}^e \cos \theta + q_{mn}^e\} G_n \\ &\quad + j \sin \theta \sum_{n=1}^{N_0^e} r_n^e G_n, \end{aligned}$$

$$F_\phi(\theta, \phi) = -j \sum_{m=0}^M \sin(m+1)\varphi (j \sin \theta)^m \sum_{n=1}^{N_m^e} \{p_{mn}^e + q_{mn}^e \cos \theta\} G_n,$$

where

$$G_n = \exp\{-jk_e z_n \cos \theta\}. \quad (12)$$

2.3. Least square problem

To obtain the pseudosolution of the overdetermined system represented by Eqs. (9) and (10) they should be transformed (exactly or formally) to their 'equivalent' square form through multiplication by the conjugate transpose matrix:

$$A_m \mathbf{p}_m = \hat{\mathbf{b}}_m, \quad A_m = \mathbf{B}_m^T \mathbf{B}_m, \quad \hat{\mathbf{b}}_m = \mathbf{B}_m^T \mathbf{q}_m, \quad m \geq -1, \quad (13)$$

where A_m is a Hermitian, non-singular, positive definite matrix with the dimension $2(N_i^m + N_e^m) \times 2(N_i^m + N_e^m)$ for $m \geq 0$ and $(N_i^m + N_e^m) \times (N_i^m + N_e^m)$ for $m = -1$.

The second approach is to apply directly the QR-decomposition method, where \mathbf{B}_m are represented in the form $\mathbf{B}_m = \mathbf{Q}_m \mathbf{R}_m$, where \mathbf{Q}_m is an orthogonal matrix and \mathbf{R}_m is an upper triangular matrix, and the desired solution will be $\mathbf{p}_m = \mathbf{R}_m^{-1} \mathbf{Q}_m^* \mathbf{q}_m$.

3. Iterative approach

The main concept of the iterative approach is a contraction mapping $\phi(x) : C^n \rightarrow C^n$ including the converging iteration process

$$x_{k+1} = \phi(x_k), \quad x_0 = z, \quad (14)$$

which converges to the solution of the linear system

$$\lim_{k \rightarrow \infty} x_k = x^*, \quad Ax^* = b. \quad (15)$$

The function $\phi(x)$ is called the contraction mapping, if there is some real number $0 < k < 1$ such that $\forall x, y, \|\phi(x) - \phi(y)\| < k \|x - y\|$. In spite of the abstract conditions such functions can be quite effectively constructed. This technique leads to iteration methods [13], that hold the following advantages:

- **Stability**

They do not change the matrix elements A , but manipulate with results of multiplication A with some vectors r . This is more stable than recursive division-addition operations with the matrix A .

- *Optimal for sparse matrices*

The optimization of matrix-vector multiplications for sparse matrices can significantly reduce the calculation time.

- *Speed*

The numerical complexity of iterative methods is $O(n^2k)$, where k is the number of applied iterations, instead of $O(n^3)$ for direct methods. Therefore, a sufficient approximate solution will be reached much earlier than in direct methods, especially for large matrices.

- *Preconditioning technique*

The convergence behaviour depends on the eigenvalues spectrum of A and the number of unknowns. By multiplication (left- and/or right-sided) the matrix A with some other matrix we can achieve a significant improvement.

4. Krylov subspace

Many powerful and effective methods are based on the Krylov subspace projection approach. These methods started in the early 1950s with the introduction of the conjugate gradients methods [13]. For a given non-singular matrix A they construct the approximate solution in the so-called Krylov subspace

$$x_k \in x_0 + K^k(A; r_0), \quad K^k(A; r_0) = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\},$$

where $r_0 = b - Ax_0$ is an initial residual and x_0 is a given initial solution, k is the iteration step. Because of non-singularity of A the vectors $r_0, Ar_0, \dots, A^{k-1}r_0$ are linear independent and Krylov subspace is a k -dimension space which means that Krylov subspace methods can be considered as some kind of direct solver because the dimensionality of the subspace will increase by 1 up to n per iteration. Thus, they should give the exact solution after at least n iterations, but can give a suitable approximate solution much earlier.

By criteria 'optimality' these methods fall in three different classes:

- *The Ritz–Galerkin approach.* Construct the $x_k \in x_0 + K^k(A, r_0)$ for which the residual $r_k = b - Ax_k$ is orthogonal to the current subspace $r_k \perp K^k(A, r_0)$. The commonly used methods for symmetric (Hermitian) matrices are CG (conjugate gradients method), SYMMLQ, for non-symmetric matrices are FOM (full orthogonalisation method), CGNE (conjugate gradients for the normal error), CGNR (conjugate gradients for the normal residual).
- *The minimum residual approach.* Construct the $x_k \in x_0 + K^k(A, r_0)$ for which the Euclidean norm $\|b - Ax_k\|_2$ is minimal over the current subspace. The commonly used algorithms of the current group are GMRES (generalized minimal residual), RGMRES (restarted GMRES), FGMRES (flexible GMRES), GMRESR (the recursive variant of GMRES) and version of GMRES for symmetric (Hermitian) matrices MINRES (minimal residual).
- *The Petrov–Galerkin approach.* Construct the $x_k \in x_0 + K^k(A, r_0)$ for which the residual r_k is orthogonal to

some other suitable k -dimensional subspace. If we select $L^k = K^k(A^T, s^0)$ for some vector s^0 , then we obtain the BiCG and QMR methods and their further modifications CGS, BiCGStab, BiCGStab(l) and TFQMR, respectively.

5. Computational results

The objective of our investigation is whether iterative methods might be better suited to fulfil scattering computations using DSM. In addition, we study the convergence behaviour of iterative methods in order to find the fastest iterative method. In extreme cases, where the scattering object is very large or has a high aspect ratio, the direct solver in DSM can completely fail due to numerical limitations of the finite floating point arithmetic. In these cases, we would like to find an iterative method, which can overcome these limitations.

In our examination of DSM computations we focus on axisymmetric scattering objects, prolate and oblate spheroids with different aspect ratios. For prolate spheroids we chose four different size parameters ($kR=12.5, 25, 50, 100$) each with five aspect ratios (2,5,10,25,50), where k is the wavenumber $2\pi/\lambda$ with the wavelength λ and R is the semimajor axis (polar radius). For oblate spheroids we used four size parameters ($kR=6.25, 12.5, 25, 50$) with five aspect ratios (2,5,10,25,50), where R is the semimajor axis (equatorial radius). The refractive index in both cases is 1.6.

By using such wide ranges with respect to size parameter and aspect ratio, we can observe different scatterers, where the kernel matrix can be in a wide range from well-conditioned to ill-conditioned.

The computer used in our investigations is a Double Quad Core Xeon E5345 with a CPU clock rate of 2.33 GHz with 16 GB RAM. The operating system is a Debian 5.0.2 Linux64.

5.1. Comparison of iterative methods

Several iterative methods and their implementations in the DSM are described in detail. We decided to include four different iterative methods. Our choice includes the Restarted GMRES(m), BiCGStab, BiCGStab(l), and the Min-Res method. During preliminary investigations of various iterative methods these methods seem to be the most suitable for the considered scattering computations with DSM. In the case of extremely deformed and large particles the iterative methods QMR and TFQMR showed less efficiency. Other methods, like CG, CGNE, and CGNR oscillated strongly during the iterative process and were not stable.

For RGMRES and BiCGStab we took the implementation from the PIM library [18]. A value of 20 was chosen as restart parameter for RGMRES. For the method BiCGStab(l) for the case of complex matrices and $l=2$ we used the source code from Sleijpen [19], which can be obtained from his web page. For MinRes we used the algorithm presented by Kanzow in his book [20]. We modified the code in order to deal with complex matrices.

For comparison of the iterative methods we need a suitable parameter which corresponds to the quality of solution. Usually, the true residual (16) is used.

$$\text{True residual} = \frac{\|\hat{\mathbf{b}}_m - \mathbf{A}_m \mathbf{p}_m\|}{\|\hat{\mathbf{b}}_m\|}, \quad m > -1. \quad (16)$$

This approach is numerically straightforward and even the calculation of the true residual is part of most of the iterative methods. But it gives no information about the remaining error for the scattered field. For estimation of the remaining error in the frame of the null-field method the surface residual (17), which is computed from the matching of the fields at the boundary of the scattering particle, is better suited.

$$\text{Surface residual} = \frac{\|\mathbf{n} \times (\mathbf{E}_i - \mathbf{E}_e - \mathbf{E}^0)\|_{\partial D}}{\|\mathbf{E}^0\|_{\partial D}}. \quad (17)$$

Although, the computational effort is much larger compared to the true residual, we decided to use the surface residual in our comparisons. Both, the true and the surface residual do not necessarily correlate. This fact first showed up in preliminary computations, where the dependency between both residuals was not predictable.

In Fig. 1, we present a typical behaviour of the surface residual versus the computational time for an oblate spheroid. We used the computational time instead of the number of iterations in order to examine the performance of the methods. Due to different implementations of the methods and different number of matrix-vector-multiplications within each iteration step it is not useful to plot the surface residual versus the number of iterations. All iterative methods show a rapid decrease to an intermediate level. RGMRES(20) does not show this rapid decrease and gives a good solution after only a few iteration steps. The reason why RGMRES(20) is not the fastest method is because each iteration step takes much more time compared to the other methods. Then, RGMRES(20), BiCGStab, and BiCGStab(2) show a slow

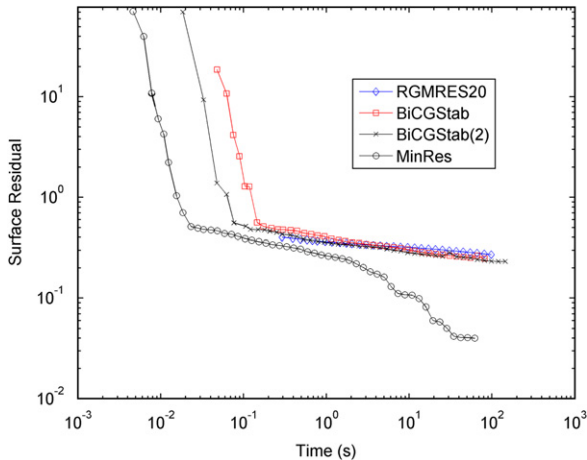


Fig. 1. Surface residual versus computational time for an oblate spheroid with an aspect ratio of 10 and the size parameter of $kR=12.5$. The size of the kernel matrix is 388×952 . All four investigated iterative methods are included.

decrease with no significant differences. It can be seen that MinRes shows the best performance especially at the end of the iterative process. There, the surface residual of MinRes values about 1/8 compared to the other methods.

In Fig. 2, a typical behaviour of the surface residual for a prolate spheroidal particle is presented. The surface residual is decreasing faster and more stable for all iterative methods compared to the oblate spheroidal particle. Again, the MinRes is outstanding in terms of performance and achievable quality of solution. The surface residual of MinRes equals 1/100–1/50 compared to the others. It seems clear that a good solution can be reached much easier for prolate than for oblate spheroids at comparable size parameters and aspect ratios.

The exact numerical values of the surface residual are shown in Section 5.3.

5.2. Computational time

In the following, we present a comparison of the performance of the iterative solvers and the QR-decomposition method used in the original DSM programme. From the theory it follows that iterative methods are faster for large and sparse matrices. The computational cost of one iteration step is $O(n^2)$, where the direct solver requires $O(n^3)$ for the whole computation. Thus, if the number of iterations is lower than n , iterative methods are faster. The sparsity of the kernel matrix allows to reduce the computational time required for the matrix-vector multiplication. The kernel matrices generated by DSM are dense, therefore, we focus on large kernel matrices in order to emphasize advantages of iterative methods. Large matrices occur with large or extremely deformed particles, where a higher number of discrete sources and matching points are needed.

In Fig. 3, we present the number of iteration steps N_{iter} versus the total computational time for a large particle. The total computational time includes the generation of

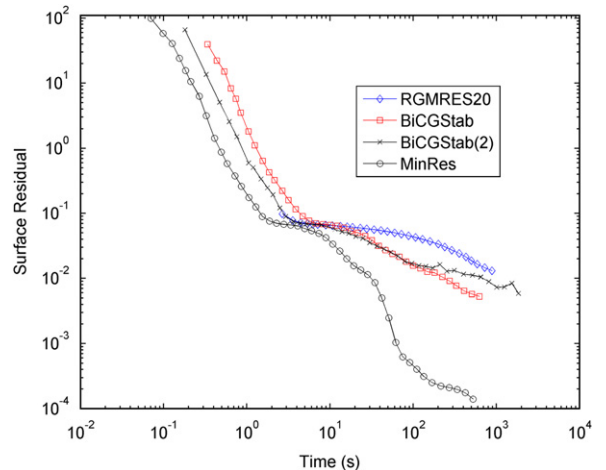


Fig. 2. Surface residual versus computational time for a prolate spheroid with an aspect ratio of 10 and the size parameter of $kR=50$. The size of the kernel matrix is 1136×2272 .

the kernel matrix, the solver, and the calculation of the scattered fields. It can be clearly seen that the time per iteration step is individually different for each of the iterative method. This corresponds to the slopes of the curves. The direct method needs 410 s for completion. Within this time, MinRes reaches about 1030 iterations, the other iterative methods pass with much lower number of iterations.

In Fig. 4, we present corresponding results for a smaller scattering object. As expected, there is no advantage of iterative methods in the frame of the total computational time. From this and several other simulations, it follows that the direct method is better suited for small particles.

5.3. Applicability area

In order to see in which area the iterative methods are preferable compared to the direct method we computed

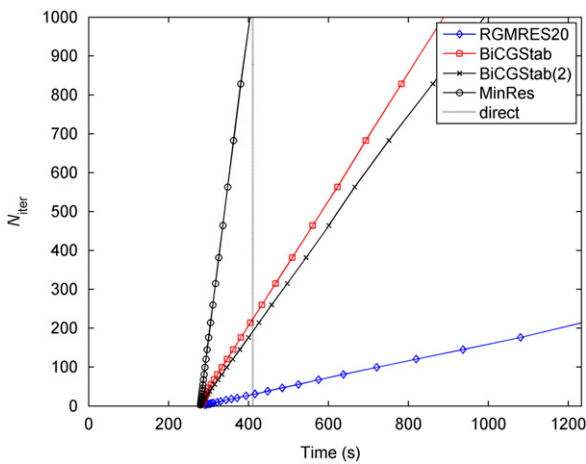


Fig. 3. Number of iterations (N_{iter}) versus total computational time for a prolate spheroid particle with an aspect ratio of 10 and a size parameter of $kR=100$. The size of the kernel matrix is 2280×4552 .

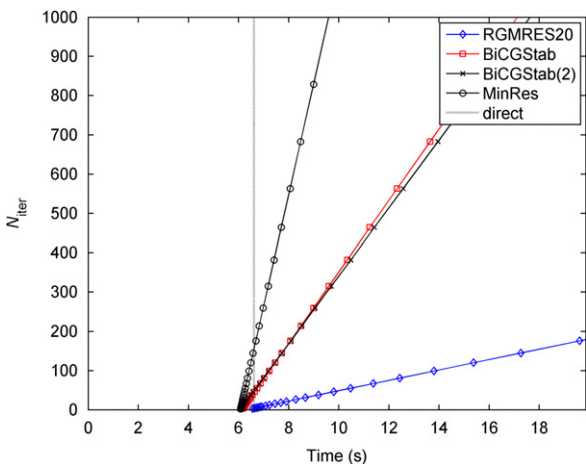


Fig. 4. Number of iterations (N_{iter}) versus total computational time for a prolate spheroid particle with an aspect ratio of 10 and a size parameter of $kR=25$. The size of the kernel matrix is 568×1136 .

the surface residual for a representative set of scattering objects. This set includes four size parameters and three aspect ratios for prolate and oblate spheroids. The maximum numbers of iterations for each method were chosen such that the computations consume about the same total time. As a default, we used the total computational time for the MinRes method with 40,000 iteration steps. For the other iterative methods the maximum number of iterations were accordingly reduced.

In Table 1, we present the surface residuals for prolate spheroids. We compare all methods, the direct and the four iterative methods for three selected different aspect ratios (2, 10, 50) and four size parameters ($kR=12.5, 25, 50, 100$). From the iterative methods the MinRes shows the best results. For most cases, the surface residuals are lower by a factor of 2 up to 10 comparing to the other iterative methods. For the aspect ratio of 2 and the smallest size parameter $kR=12.5$ it shows even better factors. For the size parameter of $kR=100$ with an aspect ratio of 2 all iterative methods could not deliver satisfying results. Apparently, all iterative methods have problems because of roundoff errors during calculation of the spherical functions in (6) for large distances $R_{\eta\zeta}$ to the source ζ .

Although in this case the surface residual amounts to about 1/5 comparing to the iterative methods, the direct method also cannot give a perfect result. In all other cases, the direct method shows perfect results. The surface residuals are mostly lower by several orders of magnitude. At least, for prolate spheroidal particles we could not find an advantage of iterative methods in the considered range of size, shape, and refractive index.

Next, we like to present some exemplary computed scattering patterns for particles with large size parameters.

Table 1

Comparison of direct and four different iterative solvers for prolate spheroids. The surface residual is given for three aspect ratios and four size parameters.

| Prolate spheroid | $kR=100$ | $kR=50$ | $kR=25$ | $kR=12.5$ |
|--------------------|----------|----------|----------|-----------|
| <i>Direct</i> | | | | |
| $ar=2$ | 5.63E-02 | 6.11E-03 | 1.51E-10 | 2.36E-11 |
| $ar=10$ | 8.60E-11 | 2.23E-10 | 2.10E-08 | 2.38E-07 |
| $ar=50$ | 2.55E-07 | 1.97E-07 | 3.27E-07 | 1.44E-04 |
| <i>RGMRES20</i> | | | | |
| $ar=2$ | 2.71E-01 | 2.77E-01 | 8.78E-02 | 1.83E-02 |
| $ar=10$ | 6.13E-02 | 1.31E-02 | 1.61E-03 | 1.38E-03 |
| $ar=50$ | 9.88E-04 | 1.08E-03 | 7.41E-04 | 5.71E-04 |
| <i>BiCGStab</i> | | | | |
| $ar=2$ | 2.69E-01 | 2.17E-01 | 6.89E-02 | 6.57E-03 |
| $ar=10$ | 5.54E-02 | 5.29E-03 | 5.31E-04 | 9.39E-04 |
| $ar=50$ | 4.83E-04 | 4.37E-04 | 3.01E-04 | 2.95E-04 |
| <i>BiCGStab(2)</i> | | | | |
| $ar=2$ | 2.69E-01 | 2.44E-01 | 7.49E-02 | 1.04E-02 |
| $ar=10$ | 5.64E-02 | 5.86E-03 | 5.56E-04 | 1.00E-03 |
| $ar=50$ | 7.16E-04 | 9.21E-04 | 2.98E-04 | 4.77E-04 |
| <i>MinRes</i> | | | | |
| $ar=2$ | 2.37E-01 | 7.38E-02 | 1.57E-02 | 1.22E-04 |
| $ar=10$ | 2.34E-02 | 1.34E-04 | 1.26E-04 | 8.92E-05 |
| $ar=50$ | 1.14E-04 | 1.75E-04 | 1.69E-04 | 1.82E-04 |

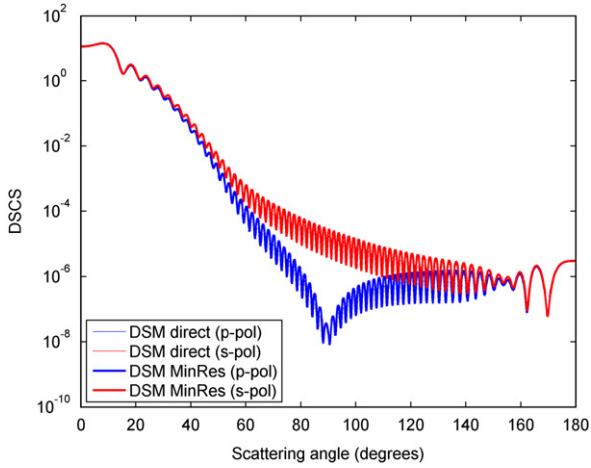


Fig. 5. The computed DSCS versus scattering angle for a prolate spheroid particle. The aspect ratio is 50. The semimajor axis R is $10\ \mu\text{m}$. With $\lambda = 2\pi/10$ this gives a size parameter of $kR=100$. The size of the kernel matrix is 2400×4800 .

In Fig. 5, differential scattering cross-section (DSCS) for s- and p-polarized incident waves is plotted. The scattering object is a prolate spheroidal particle with an aspect ratio of 50, a size parameter of $kR=100$, and a refractive index of 1.6. The incident wave propagates along the rotational axis of the particle. Both, the direct and the MinRes solutions give excellent agreement. The according surface residuals are 2.55×10^{-7} for the direct method and 1.14×10^{-4} for the MinRes method, which can be found in Table 1.

In Table 2, we present the surface residuals for oblate spheroids. Again, we compare all methods, the direct and the four iterative methods for three different aspect ratios (2,10,50) and four size parameters ($kR=6.25, 12.5, 25, 50$). It can be seen that in most cases the iterative methods produce surface residuals which are larger compared with the corresponding surface residual of the direct method. From our own experience, suitable solutions for scattering patterns can be expected for surface residuals below 1%. For oblate particles such values can only be achieved for small size parameters ($kR=6.25$). Among the iterative methods, the MinRes method shows the best results, but the differences are small.

Compared to the iterative methods, the direct method gives usually better solution up to a size parameter of $kR=25$. The direct method completely fails for the size parameter $kR=50$. Here, iterative methods show much lower surface residuals, but can also not reach a satisfying solution. In the case of the small size parameter $kR=6.25$ with an aspect ratio of 50 the MinRes method achieves a better solution in comparison to the direct method.

In order to check the computations we compare the DSCS calculated by the Null-Field Method with Discrete Sources (NFM-DS, usually referred as T-Matrix method) [14] and DSM. We compared both scattering computation for prolate and oblate spheroidal particles with size parameters and aspect ratios, where NFM-DS is applicable. There is good agreement for NFM-DS and DSM direct in the considered range. In Fig. 6, we show the results for an oblate spheroidal particle. As indicated by the surface residual (0.063%), the

Table 2

Comparison of direct and four different iterative solvers for oblate spheroids. The surface residual is given for three aspect ratios and four size parameters.

| Oblate spheroid | $kR=50$ | $kR=25$ | $kR=12.5$ | $kR=6.25$ |
|--------------------|----------|----------|-----------|-----------|
| <i>Direct</i> | | | | |
| $ar=2$ | 1.08E+01 | 1.52E-02 | 4.85E-06 | 1.37E-07 |
| $ar=10$ | 4.74E+01 | 3.00E-03 | 6.33E-04 | 1.73E-03 |
| $ar=50$ | 1.26E+02 | 8.95E-02 | 7.21E-03 | 1.77E-02 |
| <i>RGMRES20</i> | | | | |
| $ar=2$ | 5.21E-01 | 4.45E-01 | 2.20E-01 | 2.50E-02 |
| $ar=10$ | 5.71E-01 | 4.92E-01 | 2.72E-01 | 3.39E-02 |
| $ar=50$ | 5.86E-01 | 5.13E-01 | 3.05E-01 | 2.02E-02 |
| <i>BiCGStab</i> | | | | |
| $ar=2$ | 5.13E-01 | 4.34E-01 | 2.13E-01 | 1.20E-02 |
| $ar=10$ | 5.63E-01 | 4.80E-01 | 2.48E-01 | 1.24E-02 |
| $ar=50$ | 5.82E-01 | 5.07E-01 | 2.88E-01 | 2.01E-02 |
| <i>BiCGStab(2)</i> | | | | |
| $ar=2$ | 5.16E-01 | 4.38E-01 | 2.03E-01 | 1.20E-02 |
| $ar=10$ | 5.67E-01 | 4.86E-01 | 2.32E-01 | 1.23E-02 |
| $ar=50$ | 5.82E-01 | 5.11E-01 | 2.88E-01 | 2.02E-02 |
| <i>MinRes</i> | | | | |
| $ar=2$ | 4.83E-01 | 3.54E-01 | 5.46E-02 | 3.12E-03 |
| $ar=10$ | 5.33E-01 | 3.97E-01 | 4.00E-02 | 3.75E-03 |
| $ar=50$ | 5.53E-01 | 4.48E-01 | 4.82E-02 | 1.04E-02 |

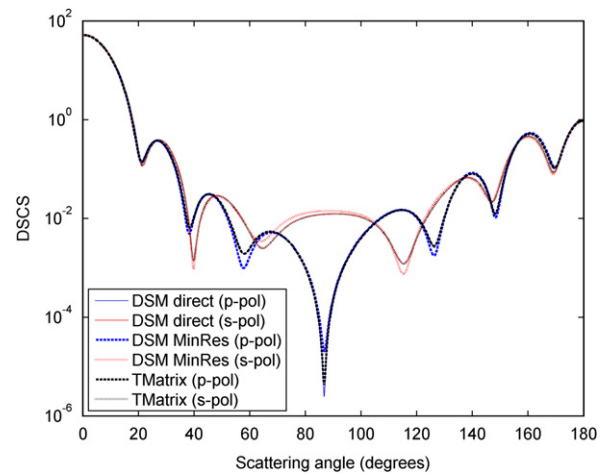


Fig. 6. The computed DSCS versus scattering angle for an oblate spheroid particle. The aspect ratio is 10. The semimajor axis R is $1.25\ \mu\text{m}$. With $\lambda = 2\pi/10$ this gives a size parameter of $kR=12.5$. The size of the kernel matrix is 388×952 .

DSM direct method and NFM-DS show excellent agreement. Since the surface residual of the MinRes method is 4% (see Table 2), there can be observed small deviations for both, the p- and the s-polarizations, compared with DSM (direct solver) and NFM-DS.

6. Summary

Contrary to direct solvers, iterative solvers are mostly preferable for sparse and large systems. Even if the DSM method produces dense and relatively small matrices, the iterative methods can also be helpful because of their higher numerical stability. The larger particle, or the

higher its aspect ratio, the more matching points and discrete sources in the DSM method are needed to solve the light scattering problem. This leads to large and ill-conditioned linear systems, that are numerically quite difficult. For such kind of particles, the iterative methods can reach a sufficient accuracy much faster than direct ones. They can also enlarge the range of applicability of the DSM method where the direct solver fail.

In this paper, we applied four different iterative methods, based on Krylov subspace projection, for DSM: RGMRES(20), BiCGStab, BiCGStab(2), and MinRes. We studied scattering for prolate and oblate spheroidal particles in the range of the size parameter of $kR=6.25$ up to 100, aspect ratios between 2 and 50, and a refractive index of 1.6. As a parameter for the quality of solution we chose the surface residual, which is computed from the matching of the fields at the boundary of the scattering particle.

A reasonable accuracy (surface residual about 1%) was reached for most prolate spheroidal particles. For oblate spheroidal particles this criteria could only be achieved for small size parameters. Among the considered iterative methods, the MinRes method showed 2–10 lower surface residual for prolate spheroids and 2–5 for oblate spheroids. The MinRes method seems to be the best suited method for DSM regarding computational time and quality of solution. Using this method the range of applicability of DSM is extended to larger particles and particles with a larger aspect ratio, especially with an oblate shape.

The ill-conditionality of the linear system is determined by the condition number of the matrix. Application of preconditioning technique allows to reduce this number and therefore will increase the numerical stability of the iterative process and will accelerate its convergence. To conclude this investigation in the next paper [21] we implement and investigate different preconditioning techniques with the MinRes method as an iterative solver for the DSM programme.

Acknowledgements

We would like to acknowledge support of this research by DFG (Deutsche Forschungsgemeinschaft).

References

- [1] Wriedt T. Light scattering theories and computer codes. *Journal of Quantitative Spectroscopy & Radiative Transfer* 2009;110:833–43.

- [2] Oguchi T, Hosoya Y. Scattering properties of oblate raindrops and cross polarization of radio waves due to rain. II—calculations at microwave and millimeter wave regions. *Journal of Radio Research Laboratories Japan* 1974;21(105):191–259.
- [3] Eremin YA, Orlov NV, Sveshnikov AG. Models of electromagnetic scattering problems based on discrete sources method. In: Wriedt T, editor. *Generalized multipole techniques for electromagnetic and light scattering*. Amsterdam, NY: Elsevier; 1999. p. 39–80.
- [4] Eremin YA, Stover JC, Orlov NV. Modeling scatter from silicon wafer features based on discrete sources method. *Optical Engineering* 1999;38(8):1296–304 doi:10.1117/1.602187. <http://link.aip.org/link/?JOE/38/1296/1>.
- [5] Eremina E, Eremin Y, Wriedt T. Extension of the discrete sources method to light scattering by highly elongated finite cylinders. *Journal of Modern Optics* 2004;51(3):423–35.
- [6] Eremina E, Wriedt T. Light scattering analysis by a particle of extreme shape via discrete sources method. *Journal of Quantitative Spectroscopy & Radiative Transfer* 2004;89(1–4):67–77.
- [7] Eremina E. Light scattering by an erythrocyte based on discrete sources method: shape and refractive index influence. *Journal of Quantitative Spectroscopy & Radiative Transfer* 2009;110(14–16): 1526–34.
- [8] Helden L, Eremina E, Riefler N, Hertlein C, Bechinger C, Eremin Y, et al. Single-particle evanescent light scattering simulations for total internal reflection microscopy. *Applied Optics* 2006;45(28): 7299–308.
- [9] Hertlein C, Riefler N, Eremina E, Wriedt T, Eremin Y, Helden L, et al. Experimental verification of an exact evanescent light scattering model for TIRM. *Langmuir* 2008;24(1):1–4.
- [10] Eremina E, Eremin Y, Grishina N, Wriedt T. Analysis of extreme light transmission through a nanohole in a metal film based on discrete sources method. *Journal of Computational and Theoretical Nanoscience* 2009;6:1–9.
- [11] Watkins D. *Fundamentals of matrix computations*. 2nd ed. New York: John Wiley & Sons Inc.; 2002.
- [12] Greengard L, Rokhlin V. A fast algorithm for particle simulations. *Journal of Computational Physics* 1987;73(2):325–48 doi: [http://dx.doi.org/10.1016/0021-9991\(87\)90140-9](http://dx.doi.org/10.1016/0021-9991(87)90140-9).
- [13] Saad Y. *Iterative methods for sparse linear systems*. 2nd ed. Philadelphia: SIAM; 2000.
- [14] Doicu A, Wriedt T, Eremin Y. Light scattering by systems of particles. In: *Null-field method with discrete sources: theory and programs*; Springer series in optical sciences, vol. 124. Berlin; NY: Springer; 2006.
- [15] Waterman PC. Symmetry, unitarity and geometry in electromagnetic scattering. *Physical Review D* 1971;3:825–39.
- [16] Colton D, Kress R. *Inverse acoustic and electromagnetic scattering theory*. Berlin: Springer-Verlag; 1992.
- [17] Doicu A, Eremin Y, Wriedt T. *Acoustic and electromagnetic scattering analysis using discrete sources*. San Diego: Academic Press; 2000.
- [18] da Cunha RD, Hopkins T. The parallel iterative methods (PIM) package for the solution of systems of linear equations on parallel computers. *Applied Numerical Mathematics* 1995;19(1–2): 33–50 doi: [http://dx.doi.org/10.1016/0168-9274\(95\)00017-0](http://dx.doi.org/10.1016/0168-9274(95)00017-0).
- [19] Sleijpen GLG, van der Vorst HA, Fokkema DR. BiCGStab(l) and other hybrid Bi-CG methods. *Numerical Algorithms* 1994;7:75–109.
- [20] Kanzow C. *Numerik linearer Gleichungssysteme. Direkte und iterative Verfahren*. Berlin: Springer; 2005.
- [21] Schmidt V, Schuh R, Wriedt T, Eremin Y. Preconditioning techniques for iterative solvers in the discrete sources method. *Journal of Quantitative Spectroscopy & Radiative Transfer* 2011;112: 1705–10, doi: [10.1016/j.jqsrt.2011.01.017](http://dx.doi.org/10.1016/j.jqsrt.2011.01.017).