Preconditioning techniques for iterative solvers in the Discrete Sources Method

Vladimir Schmidt a,*, Roman Schuh 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

Abstract

Different preconditioning techniques for the iterative method MinRes as solver for the Discrete Sources Method (DSM) are presented. This semi-analytical method is used for light scattering computations by particles in the Mie scattering regime. Its numerical schema includes a linear least-squares problem commonly solved using the QR decomposition method. This could be the subject of numerical difficulties and instabilities for very large particles or particles with extreme geometry. In these cases, we showed that iterative methods with preconditioning techniques can provide a satisfying solution.

In our previous paper, we studied four different iterative solvers (RGMRES, BiCGStab, BiCGStab(l), and MinRes) considering the performance and the accuracy of a solution. Here, we study several preconditioning techniques for the MinRes method for a variety of oblate and prolate spheroidal particles of different size and geometrical aspect ratio. Using preconditioning techniques we highly accelerated the iterative process especially for particles with a higher aspect ratio.

1. Introduction

Simulation of light scattering by particles is needed in various scientific problems. The field of application includes interstellar dust, aerosol particles, red blood cells (erythrocytes), soot particles, pigment particles, and other non-spherical particles. A review of available computational methods has recently been presented by Wriedt [1].

The Discrete Sources Method (DSM) is an established semi-analytical method for computations of electromagnetic scattering which is outlined in detail by Eremin et al. [2]. The numerical scheme of the DSM method includes the least-squares problem which is solved in the original program using the QR decomposition method.

The high ill-conditionality of the linear system increases the effect of round-off errors caused by finite precision arithmetic in computers. As a rule, the larger the particle or the more complex the shape of the particle, the more numerically difficult and instable the linear problem. For very large objects the direct solver can completely fail while the iterative method can still give a suitable solution. In order to further decrease the condition number and therefore accelerate the iterative process with small computational effort preconditioning can be applied. In each case especially for ill-conditioned problems the special features of the matrix should be studied in order to construct or choose an effective preconditioner.

In the previous paper [3], we compared different iterative methods with each other. The MinRes method was found to be the best performing method regarding computational time and quality of a solution. Here we focus on preconditioning techniques. The theory of the DSM was outlined briefly and we refer the reader to this paper. In Section 2 we present the theoretical background.
of condition number and the investigated preconditioning techniques. In Section 3 computational results are presented.

2. Iterative approach

2.1. The condition number

In our previous article we studied the iterative approach to solve the minimization problem in the DSM method. The behavior of iterative processes of four different methods (RCMRES, MinRes, BiCGStab, BiCG-Stab()) was observed in order to find that one which gives fast a suitable solution.

The convergence of the iterative process strongly depends on such properties of the matrix \( A \) like the eigenvalues spectrum \( \{ \lambda_i \} \) or the number of unknowns \( n \). On the other side, perturbations in the matrix \( A \) and the vector \( b \) can induce some changes in the solution of the linear system. These changes are being collected over any iterative process. At least, the cause of such perturbations can be finite precision arithmetic usually provided by the computer architecture.

In order to estimate the possible range of the solution, let us write the following linear equation

\[
(A + \delta A) (x + \delta x) = b + \delta b
\]

(1)

instead of the original one

\[
Ax = b,
\]

(2)

where \( \delta A = (\delta a_{ij})_{n \times n} \) and \( \delta b = (\delta b_j)_{n} \) are some small perturbations of coefficients \( b \) and \( A \). For single precision the computer can store not more than seven decimal places (for double twice more – 16) and \( \| \delta b \| / \| b \| \approx 10^{-7} \), \( \| \delta A \| / \| A \| \approx n \times 10^{-7} \). The system (1) has a unique solution \( x^* = x + \delta x \), where \( x^* \) is the solution of (2). The range of \( x^* \) is determined by the following inequality

\[
\| \delta x \| / \| x \| \leq \kappa(A) \| \delta b \| / \| b \| + \frac{\kappa(A) \| \delta A \| / \| A \|}{1 - \kappa(A) \| \delta A \| / \| A \|},
\]

(3)

where the factor \( \kappa(A) = \| A \| \| A^{-1} \| < \max | \lambda_i | / \min | \lambda_i | \) is the condition number. For high \( \kappa(A) > 10000 \) the matrix \( A \) is called ill-conditioned, the result can be very sensitive to the perturbations \( \| \delta b \| / \| b \|, \| \delta A \| / \| A \| \). If \( \| \delta A \| / \| A \| > 1 / \kappa(A) \) the matrix \( A + \delta A \) becomes even singular. By contrast, the matrix with small condition number \( \kappa(A) \) is called well-conditioned. The inequality similar to (3) can be derived for the matrix-vector multiplication \( y = Ax \):  

\[
y + \delta y = A(r + \delta r), \quad \| \delta y \| / \| y \| \leq \kappa(A) \| \delta r \| / \| r \| + \kappa(A) \| \delta A \| / \| A \|.
\]

(4)

The formulas (3) and (4) would be very interesting for understanding of the effect of perturbations as well as round-off errors. While the first inequality will be mainly interesting for direct solvers, the second one can give information about round-off errors being collected over matrix-vector multiplication. It can help us to understand how to improve the iterative process.

2.2. Preconditioning techniques

The possible way to accelerate the convergence process and to decrease round-off error effects is to narrow the spectrum of the matrix \( A \) or to lower its condition number. It leads to the so-called preconditioning technique [4]. The matrix \( A \) is multiplied by some other matrix \( P \) such that the condition number of the resulting matrix \( \hat{A} \) will be lower. By selecting the side of matrix multiplication the preconditioning techniques fall in three groups:

- **left preconditioning** \( \hat{A} = P_l A \) \( (Ax = b \Rightarrow \hat{A}x = P_l b) \)
- **right preconditioning** \( \hat{A} = AP_R \) \( (Ax = b \Rightarrow Ay = b, \ y = P_R x) \)
- **left and right preconditioning** \( \hat{A} = P_l AP_R \) \( (Ax = b \Rightarrow \hat{A}y = P_l b, \ y = P_R x) \)

where \( P_l \in R^{n \times n} \) and \( P_R \in R^{n \times n} \) are called left- and right-preconditioners, respectively. For symmetric (Hermitian) matrices the preconditioner should keep the matrix symmetry \( \hat{A} = P^T A P, \ P \in R^{n \times n} \).

There are a lot of approaches to construct an efficient preconditioner, which are in detail described by Chen et al. [5]. From a philosophical point of view the calculation of \( P_l \) and \( P_R \) should be numerically easier and more stable than the original equation solution. Unfortunately, there are no optimal left- and/or right-preconditioners, which will be effective in the general case. The choice of the optimal one should be provided for each particular case separately by considering special matrix properties like sparsity, linear dependence of rows/columns or interconnection between unknowns.

During our preliminary investigation we have chosen to study the following preconditioners:

- **Diagonal**
  
  The simplest and universal preconditioner is the singular diagonal matrix \( D = \text{diag}(d_{11}, d_{22}, \ldots, d_{nn}) \). It does not require an additional memory for calculation, nevertheless can significantly decrease the condition number. There are two approaches to fill the matrix \( D \) elements

\[
d_{ii} = (a_{ii})^{-1}, \quad d_{ii} = \left( \frac{\sum_{j=1}^{n} d_{ij}}{n} \right)^{-1}.
\]

While the first one is mostly used up to now, according to a theoretical point of view the second one gives the best reduction of \( \kappa(A) \) in the norm \( p = \infty \) by left-side multiplication \( \hat{A} = DA \) and in the norm \( p=1 \) by right-side multiplication \( \hat{A} = AD \) [6]. To keep the Hermitian symmetry of the matrix \( C \) we use the following preconditioning:

\[
\hat{A} = D^{1/2} A D^{1/2}, \quad P = D^{1/2}.
\]

(5)

For further comparisons we choose the second one and will call it the Diagonal (D) preconditioner.

- **Block-diagonal with permutations**

The second preconditioning technique we want to study is a hybrid of the block-diagonal and permutation preconditioners. In the extreme cases of particle shape and
size from the considered range, the rows and columns in the kernel matrix are usually strongly linear dependent and even nearly proportional. Let us consider the matrix $A$ as a matrix composed of $2 \times 2$ matrices $A_{ij}$. We fill each block $D_{ii}$ of the block-diagonal matrix $D$ by normalized eigenvectors of $A_{ii}$:

$$A = \begin{pmatrix} A_{11} & \ldots & A_{1n/2} \\ \vdots & \ddots & \vdots \\ A_{n/21} & \ldots & A_{n/2n/2} \end{pmatrix}, \quad D = \begin{pmatrix} D_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & D_{n/2n/2} \end{pmatrix}.$$

Thus, the matrix $D_i^{-1}A_iD_i = \text{diag}(\lambda_i^{(1)}, \lambda_i^{(2)})$ will be diagonal with eigenvalues $\lambda_i^{(j)}$, $j=1,2$ of $A_i$. Due to linear dependency of rows and columns each block $A_{ij}$ will have the similar structure after multiplication $A = D^TAD$. The matrix $D$ is labeled block-diagonal preconditioner and will be called further Block-Diagonal (D-BD1). The unknowns can be reordered by applying the permutation matrix $M$

$$\hat{A} = (MD)^T \cdot A \cdot DM, \quad P = DM,$$

such that odd unknowns will be collected separately from the even ones. The obtained matrix $\hat{A}$ will have the following structure:

$$\hat{A} = \begin{pmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{pmatrix},$$

where the matrices $\hat{A}_{11}$ and $\hat{A}_{22}$ consist of diagonal elements from $A_{11}, A_{12}, A_{21}$ of non-diagonal element. Additionally, we require that the maximum eigenvalues $\lambda_1^{(j)}$ should be included in $\hat{A}_{11}$ and the smallest ones in $\hat{A}_{22}$. Then, block-diagonal preconditioning can be again applied to the part $\hat{A}_{11}$ of the matrix $A$. This preconditioning technique will be called D-BD2. Analogously, a D-BD preconditioner can be constructed by providing this schema $l$-times.

During preliminary investigations, we studied other preconditioning techniques based on the incomplete matrix factorization like $IC, ILU[p], \text{and ILUT}$ [5]. It turned out that their numerical schemes were either not effective regarding the computational time or unstable for the extreme cases of the considered range of size parameter and aspect ratio.

3. Computational results

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 $\lambda$ and $R$ is the semi-major 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 semi-major 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.

3.1. Comparison of different preconditioning techniques

Several iterative methods and different preconditioning techniques for DSM are described in detail in [3]. In an earlier investigation of various considered iterative methods for DSM, MinRes showed the best solutions over a wide range of particle shape and size [3]. Therefore, in this work we use the MinRes method based on the algorithm presented by Kanzow in his book [6]. We modified the code in order to deal with complex matrices. We consider three different preconditioning techniques and the pure MinRes method without preconditioning. We denote them as Diagonal (D), Block-Diagonal (D-BD1), and twice recurrently applied Block-Diagonal (D-BD2).

For comparison of the iterative methods we use the surface residual which is better suited to estimate the quality of scattering solution. In the frame of the null-field method the surface residual (7) is computed from the matching of the fields at the boundary of the scattering particle:

$$\text{Surface Residual} = \| n \times (E_r-E_r^p) \|_{\partial Q} / \| E^p \|_{\partial Q}.$$  \hspace{1cm} (7)

3.2. Oblate spheroidal particles

In Fig. 1 we present the surface residual versus computational time for an oblate particle with an aspect ratio of 2 and a size parameter of $kr=6.25$. One can see, that the iterative process without preconditioning starts with high values of the surface residual but shows a rapid

![Fig. 1. Surface residual versus computational time for an oblate spheroidal particle with an aspect ratio of 2 and the size parameter of $kr=6.25$ using different preconditioning techniques. The size of the kernel matrix is $260 \times 648$.]
The surface residual is significantly lower in the beginning for all considered preconditioning techniques but the decrease is lower. Thus, after a short time all methods show comparable values. In the following, the iterative process is qualitatively comparable for all methods which show a mainly monotonically decreasing behavior. As expected, D-BD1 and D-BD2 are the best preconditioning techniques compared to the other ones. D-BD2 shows a slightly better result compared to D-BD1. Extraordinarily, Diagonal (D) shows the biggest surface residual after some time. For all preconditioning techniques the surface residual can begin to increase if the iterative process will not be interrupted. This effect can be observed in Fig. 1 for times above 10 s.

In Fig. 2 we present the surface residual versus computational time for an oblate particle with an aspect ratio of 25 and a size parameter of \(kR=12.5\). The iterative behavior for all preconditioning techniques is similar compared to the one for the size parameter of \(kR=6.25\). After a fast decrease for the case without preconditioning all methods show a slow iterative process up to a computational time of 10 s. Then, the methods show slightly different decreasing rates. Diagonal (D) gives the lowest decreasing rate where the preconditioning techniques D-BD1 and D-BD2 show the highest ones.

In Table 1 we present the minimal surface residuals for oblate spheroidal particles over iterative processes with all considered preconditioning techniques, in addition to the direct method. The comparison includes three different aspect ratios (2, 10, 50) and four size parameters \((kR=6.25, 12.5, 25, 50)\). It can be seen, that for large particles \((kR=25, 50)\) the MinRes method with all preconditioning techniques can not give satisfying results. The direct method shows still acceptable surface residuals for \(kR=25\), but completely fails for the size parameter \(kR=50\). For the case of \(kR=6.25\) there can be seen the advantage of preconditioning techniques. For the aspect ratios 10 and 50 the corresponding surface residuals are lower for all preconditioning techniques compared to the direct method. The use of preconditioning results in an improvement of the surface residual from about 10% (aspect ratio = 50) up to 80% (aspect ratio = 2). With regard to the efficiency for oblate spheroidal particles, we can put the considered techniques in the following order. Diagonal (D), no preconditioning, D-BD1, and D-BD2, where D-BD2 is the most effective.

### Table 1

<table>
<thead>
<tr>
<th>Oblate spheroid</th>
<th>(kR=50)</th>
<th>(kR=25)</th>
<th>(kR=12.5)</th>
<th>(kR=6.25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>(ar=2)</td>
<td>(ar=10)</td>
<td>(ar=50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.08E+01</td>
<td>4.74E+01</td>
<td>1.26E+02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.52E-02</td>
<td>3.00E-03</td>
<td>8.95E-02</td>
<td>7.21E-03</td>
</tr>
<tr>
<td>MinRes no precond.</td>
<td>(ar=2)</td>
<td>(ar=10)</td>
<td>(ar=50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.83E-01</td>
<td>5.33E-01</td>
<td>5.35E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.54E-01</td>
<td>3.97E-01</td>
<td>4.48E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.46E-02</td>
<td>4.00E-02</td>
<td>4.82E-02</td>
<td>1.04E-02</td>
</tr>
<tr>
<td>MinRes D-BD1</td>
<td>(ar=2)</td>
<td>(ar=10)</td>
<td>(ar=50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.79E-01</td>
<td>5.31E-01</td>
<td>5.58E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.54E-01</td>
<td>4.02E-01</td>
<td>4.60E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.01E-02</td>
<td>4.25E-02</td>
<td>1.30E-01</td>
<td></td>
</tr>
<tr>
<td>MinRes D-BD2</td>
<td>(ar=2)</td>
<td>(ar=10)</td>
<td>(ar=50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.75E-01</td>
<td>5.29E-01</td>
<td>5.44E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.33E-01</td>
<td>3.55E-01</td>
<td>4.25E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.05E-02</td>
<td>1.62E-02</td>
<td>4.54E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.54E-04</td>
<td>1.25E-03</td>
<td>9.51E-03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(ar=2)</td>
<td>(ar=10)</td>
<td>(ar=50)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.73E-01</td>
<td>5.27E-01</td>
<td>5.42E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.30E-01</td>
<td>3.42E-01</td>
<td>4.17E-01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.04E-02</td>
<td>1.59E-02</td>
<td>4.43E-02</td>
<td>9.50E-03</td>
</tr>
</tbody>
</table>

Fig. 2. Surface residual versus computational time for an oblate spheroidal particle with an aspect ratio of 25 and the size parameter of \(kR=12.5\) using different preconditioning techniques. The size of the kernel matrix is \(1444 \times 3592\).
considered preconditioning techniques, in addition to the direct method. The comparison includes three different aspect ratios (2, 10, 50) and four size parameters (\(kR=12.5\), 25, 50, 100). The iterative method MinRes without preconditioner shows already good results for nearly all considered cases. The use of preconditioning further reduces the surface residual up to 97%. The highest improvements can be found for size parameters to \(kR=12.5\). For \(kR=25\) and 50 the improvements still can be up to 57%. As shown in Fig. 4 and from the values of the surface residual in Table 2 the use of preconditioning for large particles (\(kR=100\)) can reduce the quality of solution. With regard to the efficiency for prolate spheroidal particles, we can put the considered techniques in the following order: No preconditioning, Diagonal (D), D-BD1, and D-BD2, where the latter two are very close.

### 4. Summary

The efficiency of iterative solvers is strongly determined by the conditionality of the linear system. The lower the condition number of the kernel matrix, the faster the convergence of the iterative process and the lower the values of the surface residual which can be reached. Additionally, the influence of round-off errors is reduced for lower condition numbers. Usage of a preconditioning technique allows to reduce the condition number and therefore accelerate the iterative process and ameliorate the solution of the linear system. Because there are no universal and good preconditioning technique for ill-conditioned problems, each approach should be studied separately for a given problem.

In the previous paper [3], we found that the iterative method MinRes as a solver for the Discrete Sources Method (DSM) shows the best behavior compared to several other iterative methods regarding computational time and quality of solution. Here, we applied different preconditioning techniques to the MinRes method. The examined preconditioning techniques are Diagonal (D), Block-Diagonal (D-BD1), and twice recurrently applied Block-Diagonal (D-BD2). We studied scattering computations for prolate and oblate spheroidal particles in the range of a size parameter of \(kR=6.25\) up to 100, aspect ratios between 2 and 50, and a refractive index of 1.6. As parameter for the quality of the solution we chose the surface residual, which is computed from the matching of the fields at the boundary of the scattering particle.

Applying the preconditioning technique the surface residual becomes 2–3 order lower in the beginning of the iterative process. Then, this difference becomes much smaller. The block-diagonal preconditioners D-BD1 and D-BD2 reduce the surface residual up to 97% for small prolate spheroidal particles. For oblate spheroidal particles the

---

**Fig. 3.** Surface residual versus computational time for a prolate spheroidal particle with an aspect ratio of 10 and the size parameter of \(kR=25\) using different preconditioning techniques.

**Fig. 4.** Surface residual versus computational time for a prolate spheroidal particle with an aspect ratio of 50 and the size parameter of \(kR=100\) using different preconditioning techniques.

**Table 2**

Comparison of direct and MinRes solver with different preconditioning techniques for prolate spheroids. The surface residual is given for three aspect ratios and four size parameters.

<table>
<thead>
<tr>
<th>Prolate spheroid</th>
<th>(kR=100)</th>
<th>(kR=50)</th>
<th>(kR=25)</th>
<th>(kR=12.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>(5.63E-02)</td>
<td>(6.11E-03)</td>
<td>(1.51E-10)</td>
<td>(2.36E-11)</td>
</tr>
<tr>
<td>(ar=2)</td>
<td>(8.60E-11)</td>
<td>(2.23E-10)</td>
<td>(2.10E-08)</td>
<td>(2.38E-07)</td>
</tr>
<tr>
<td>(ar=10)</td>
<td>(2.55E-07)</td>
<td>(1.97E-07)</td>
<td>(3.27E-07)</td>
<td>(1.44E-04)</td>
</tr>
<tr>
<td>No precondition</td>
<td>(2.37E-01)</td>
<td>(7.38E-02)</td>
<td>(1.57E-02)</td>
<td>(1.22E-04)</td>
</tr>
<tr>
<td>MinRes D-BD1</td>
<td>(2.87E-02)</td>
<td>(1.07E-04)</td>
<td>(1.22E-04)</td>
<td>(5.09E-05)</td>
</tr>
<tr>
<td>MinRes D-BD2</td>
<td>(1.58E-04)</td>
<td>(1.80E-04)</td>
<td>(1.38E-04)</td>
<td>(9.85E-05)</td>
</tr>
</tbody>
</table>
reduction can be up to 80% compared to MinRes without preconditioning.

To summarize this two part paper, the iterative methods together with a suitable preconditioning technique can extend the range of applicability of DSM to larger particles and more complex shapes with a larger aspect ratio. This is feasible because of higher stability of the iterative solver. The further improvement in term of the DSM method could be more complex iterative methods such as BiCGStab(l > 2) [7] or QMRCGStab [8] and preconditioning techniques such as incomplete factorizations or wavelet preconditioners [5] or acceleration of matrix-vector multiplication using the Fast Multipole Method.

**Acknowledgments**

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

**References**


