STABILITY OF THE FAST MULTIPOLE METHOD FOR HELMHOLTZ EQUATION IN THREE DIMENSIONS

MARTIN NILSSON[1] *

[1] *Department of Information Technology, Scientific Computing Uppsala University, SE-75105 Uppsala, Sweden. email: martin@it.uu.se*

Abstract

Stability limits for the diagonal forms approximating the free space Green's function in Helmholtz' equation are derived. It is shown that while the original approximation of the Green's function is stable except for overflows, the diagonalized form becomes unstable due to errors from roundoff, interpolation, choice of quadrature rule and approximation of the translation operator. Numerical experiments validate the theoretical estimates.

Keywords: Fast Multipole Method, Helmholtz' equation, stability, error estimate

AMS subject classification (MSC2000): 65B10, 65G99

1 Introduction

The Fast Multipole Method (FMM) for Helmholtz' equation in three dimensions was introduced in [10]. The idea of FMM is to find a discrete diagonal representation of the free-space Green's function for Helmholtz' equation. This allows interactions between sources to be computed faster than $\mathcal{O}(N^2)$, where N is the number of sources. The method was later extended to electromagnetics in [3]. A multilevel version of FMM was introduced in [11]. It allows the interactions between sources to be computed in $\mathcal{O}(N \log N)$ arithmetic operations. The multilevel version has been used by many researchers to compute the dense matrix

*Financial support has been obtained from Parallel and Scientific Computing Institute (PSCI), which is a competence center financed by Vinnova, The Swedish Agency for Innovation Systems, and the Swedish National Aeronautical Research Program, NFFP.

vector multiplication in an iterative solution of the Method of Moments, see for instance [2, 5, 8] and references therein.

The error in FMM is controlled by the truncation number L of an infinite sum. A simple semi-empirical estimate of L was already given in [3]. Later it was found that the formula was not sufficient and a new formula was derived in [2] called the excess bandwidth formula. The relation between the error and L was also studied in [4], where it was found that the rate of convergence was fast for a certain region of L or like a geometrical series beyond that region.

It is known that the discrete diagonal form becomes unstable when L is large [3]. The instability is due to that errors other than the truncation error grow when L is increased in the geometrical convergence zone. This is due to the divergence of the spherical Hankel function for large L and a constant argument. The effects of roundoff errors on stability have been studied in [9] for two dimensions and [7] for three dimensions. As far as the author is aware no one has derived stability limits for the general case that depends on the choice of quadrature rule, interpolation scheme, approximation of translation operators and roundoff errors. This is especially important if high precision is required of the diagonal form and also in order to predict when there is a need to switch to a different approximation.

The aim of this paper is to derive guidelines for choosing L that depends on knowledge of the precision used and accuracy of the different numerical approximations. To do this, stability limits are derived for the approximations. The guidelines given here are probably accurate within a constant for any implementation of the FMM for Helmholtz equation in three dimensions. Similar results can be derived for the two dimensional case as well.

The rest of this paper is organized as follows. In section 2 the diagonal translation operators are derived. The error estimates and stability limits are derived in section 3. The theoretical estimates are validated by numerical experiments in section 4 and conclusions are given in section 5.

2 Diagonalizing the Green's function

The Fast Multipole method for Helmholtz' equation is based on two observations. The first observation is that the Green's function for Helmholtz' equation can be expanded into an infinite series using Gegenbauer's addition theorem [1]

$$\begin{aligned}\frac{e^{\imath\kappa|\mathbf{X}+\mathbf{d}|}}{|\mathbf{X}+\mathbf{d}|} &= \imath\kappa h_0^{(1)}\left(\kappa\left|\mathbf{X}+\mathbf{d}\right|\right) \\ &= \imath\kappa\sum_{l=0}^{\infty}\left(-1\right)^l\left(2l+1\right)j_l\left(\kappa d\right)h_l^{(1)}\left(\kappa X\right)P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right)\end{aligned} \tag{1}$$

Here, a boldface character denotes a vector and $x = |\mathbf{x}|$, while the notation $\hat{\mathbf{x}}$ indicates a unit length vector. The symbol κ denotes the wavenumber and $\imath$ is the imaginary unit. The function $j_l(x)$ is a spherical Bessel function, $h_l^{(1)}(x)$ is a spherical Hankel function and $P_l(x)$ is a Legendre polynomial of order l. The expansion is valid as long as $d < X$. It can be used to compute the field at a receiver point $\mathbf{x}$ from a source point $\mathbf{x}'$ in the following way. Note that

$$\begin{aligned} \mathbf{x} - \mathbf{x}' &= \mathbf{x} - \mathbf{X}_m + \mathbf{X}_m - \mathbf{X}'_m + \mathbf{X}'_m - \mathbf{x}' \\ &= (\mathbf{X}_m - \mathbf{X}'_m) + (\mathbf{x} - \mathbf{X}_m + \mathbf{X}'_m - \mathbf{x}') \end{aligned} \tag{2}$$

as in Figure 1, where $\mathbf{X}_m$ is the midpoint of a sphere close to the receiver point and $\mathbf{X}'_m$ is the midpoint of a sphere close to the source point. If $\mathbf{X}_m$ is close to $\mathbf{x}$ and $\mathbf{X}'_m$ is close to $\mathbf{x}'$, then the choice $\mathbf{d} = \mathbf{x} - \mathbf{X}_m + \mathbf{X}'_m - \mathbf{x}'$ and $\mathbf{X} = \mathbf{X}_m - \mathbf{X}'_m$ implies that $d < X$ if $|\mathbf{x} - \mathbf{X}_m + \mathbf{X}'_m - \mathbf{x}'| < |\mathbf{X}_m - \mathbf{X}'_m|$, which is always the case when $|\mathbf{x} - \mathbf{X}_m| < X/2$ and $|\mathbf{X}'_m - \mathbf{x}'| < X/2$ so that the expansion in (1) is valid.

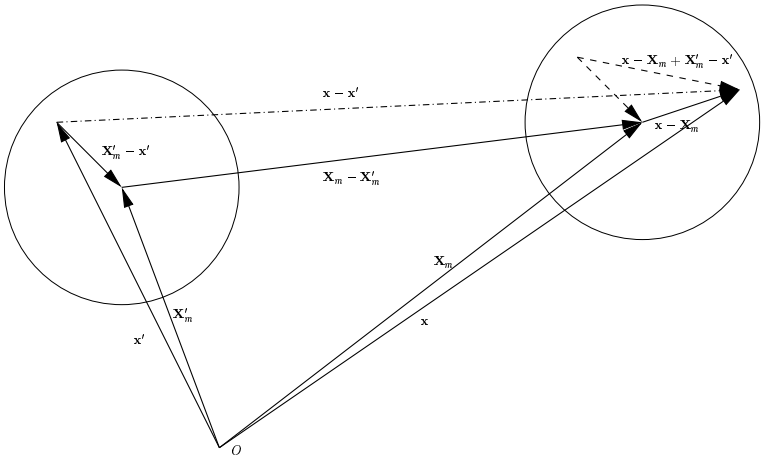

Figure 1: A graphical view of $\mathbf{x} - \mathbf{x}' = (\mathbf{X}_m - \mathbf{X}'_m) + (\mathbf{x} - \mathbf{X}_m + \mathbf{X}'_m - \mathbf{x}')$.

The second key observation is that the product $j_l(\kappa d) P_l\left(\hat{\mathbf{d}} \cdot \hat{\mathbf{X}}\right)$ can be expanded into an integral over propagating plane waves [3]

$$4\pi \imath^l j_l(\kappa d) P_l\left(\hat{\mathbf{d}} \cdot \hat{\mathbf{X}}\right) = \int_{\mathcal{S}} e^{\imath \boldsymbol{\kappa} \cdot \mathbf{d}} P_l\left(\hat{\boldsymbol{\kappa}} \cdot \hat{\mathbf{X}}\right) d^2\hat{\boldsymbol{\kappa}} \tag{3}$$

Here, the integral is taken over the unit sphere $\mathcal{S}$ and $\hat{\boldsymbol{\kappa}}$ is the direction of the plane wave. Inserting (3) into (1) yields

$$\frac{e^{\imath\kappa|\mathbf{X}+\mathbf{d}|}}{|\mathbf{X}+\mathbf{d}|} = \frac{\imath\kappa}{4\pi} \sum_{l=0}^{\infty} \imath^l (2l+1) h_l^{(1)}(\kappa X) \int_{\mathcal{S}} e^{\imath \boldsymbol{\kappa} \cdot \mathbf{d}} P_l\left(\hat{\boldsymbol{\kappa}} \cdot \hat{\mathbf{X}}\right) d^2\hat{\boldsymbol{\kappa}} \tag{4}$$

The sum in (4) is infinite. Hence, the order of integration and summation can not be changed, because the spherical Hankel function diverges when $l \to \infty$ and the argument is constant. Only a finite number of terms is needed in the sum to get a desired accuracy in an approximation of the Green's function. Once the number of terms $L+1$ is determined, the order of summation and integration is changed

$$\frac{e^{\imath\kappa|\mathbf{X}+\mathbf{d}|}}{|\mathbf{X}+\mathbf{d}|} \approx \frac{\imath\kappa}{4\pi} \int_{\mathcal{S}} e^{\imath\boldsymbol{\kappa}\cdot\mathbf{d}} \sum_{l=0}^{L} \imath^l \left(2l+1\right) h_l^{(1)}\left(\kappa X\right) P_l\left(\hat{\boldsymbol{\kappa}}\cdot\hat{\mathbf{X}}\right) d^2\hat{\boldsymbol{\kappa}} \tag{5}$$

The integral in equation (5) can be computed with an appropriately chosen quadrature formula

$$\frac{e^{\imath\kappa|\mathbf{X}+\mathbf{d}|}}{|\mathbf{X}+\mathbf{d}|} \approx \frac{\imath\kappa}{4\pi} \sum_{k=1}^{K} w_k e^{\imath\boldsymbol{\kappa}_k\cdot\mathbf{d}} \sum_{l=0}^{L} \imath^l \left(2l+1\right) h_l^{(1)}\left(\kappa X\right) P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\hat{\mathbf{X}}\right) \tag{6}$$

where w_k is the quadrature weight associated with direction $\hat{\boldsymbol{\kappa}}_k$. The translation operator is defined by

$$\mathcal{T}_k^L\left(\kappa,\mathbf{X}\right) = \frac{\imath\kappa}{16\pi^2} w_k \sum_{l=0}^{L} \imath^l \left(2l+1\right) h_l^{(1)}\left(\kappa X\right) P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\hat{\mathbf{X}}\right) \tag{7}$$

From (6) and (7) and the choices $\mathbf{X} = \mathbf{X}_m - \mathbf{X}'_m$ and $\mathbf{d} = \mathbf{x} - \mathbf{X}_m + \mathbf{X}'_m - \mathbf{x}'$ it follows that

$$\frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi\left|\mathbf{x}-\mathbf{x}'\right|} \approx \sum_{k=1}^{K} e^{\imath\boldsymbol{\kappa}_k\cdot(\mathbf{x}-\mathbf{X}_m)} \mathcal{T}_k^L\left(\kappa,\mathbf{X}_m-\mathbf{X}'_m\right) e^{\imath\boldsymbol{\kappa}_k\cdot(\mathbf{X}'_m-\mathbf{x}')} \tag{8}$$

The approximation (8) is accurate for any source close to $\mathbf{X}'_m$ and any receiver close to $\mathbf{X}_m$. The translation operator $\mathcal{T}_k^L\left(\kappa,\mathbf{X}_m-\mathbf{X}'_m\right)$ is independent of the locations of sources and receivers, since it depends on $\mathbf{X}_m$ and $\mathbf{X}'_m$ only. This approximation of the Green's function is called a diagonal form since the translation operator is a diagonal matrix.

3 The different errors in FMM

Error analysis of the Fast Multipole Method has mainly focused on the truncation errors of the diagonal form due to finite L. It is known that when $L \gtrapprox \kappa X$ the approximation (8) becomes numerically unstable. The reason is that the spherical Hankel function starts to grow exponentially when the order is larger than the argument. Previous studies on the effect of the exponential growth [7, 9] seem

to have focused on roundoff errors. Here we identify all the sources that cause instabilities in the FMM and derive stability requirements for them. In order to do this we consider the relative difference between the Green's function and the computed approximation

$$\Phi = \left(\frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi\left|\mathbf{x}-\mathbf{x}'\right|} - \sum_{k=1}^{K} \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}} \overline{\mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right)} \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}} \right) \frac{4\pi\left|\mathbf{x}-\mathbf{x}'\right|}{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}} \tag{9}$$

Here, $\overline{f(y)} = f(y)(1+\eta)$ where η is the relative error in the computed approximation of the function $f(x)$. The error Φ can be rewritten as

$$\begin{aligned}
\Phi = & \left[\left(\frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi\left|\mathbf{x}-\mathbf{x}'\right|} - \frac{\imath\kappa}{4\pi} \sum_{l=0}^{L} (-1)^l (2l+1)\, j_l(\kappa d)\, h_l^{(1)}(\kappa X)\, P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \right) \right. \\
& + \left(\frac{\imath\kappa}{4\pi} \sum_{l=0}^{L} (-1)^l (2l+1)\, j_l(\kappa d)\, h_l^{(1)}(\kappa X)\, P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \right. \\
& \quad \left. - \frac{\imath\overline{\kappa}}{4\pi} \sum_{l=0}^{L} (-1)^l (2l+1)\, j_l\left(\overline{\kappa d}\right) h_l^{(1)}\left(\overline{\kappa X}\right) P_l\left(\overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}}\right) \right) \\
& + \left(\frac{\imath\overline{\kappa}}{4\pi} \sum_{l=0}^{L} (-1)^l (2l+1)\, j_l\left(\overline{\kappa d}\right) h_l^{(1)}\left(\overline{\kappa X}\right) P_l\left(\overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}}\right) \right. \\
& \quad \left. - \sum_{k=1}^{K} e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)} \mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right) e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)} \right) \\
& + \left(\sum_{k=1}^{K} e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)} \mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right) e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)} \right. \\
& \quad \left. \left. - \sum_{k=1}^{K} \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}} \overline{\mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right)} \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}} \right) \right] \frac{4\pi\left|\mathbf{x}-\mathbf{x}'\right|}{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}
\end{aligned} \tag{10}$$

From (10) the relative errors can be identified as

$$\begin{aligned}
\Phi = & \text{ relative truncation error} \\
& + \text{relative roundoff error} \\
& + \text{relative integration error} \\
& + \text{relative function evaluation error}
\end{aligned} \tag{11}$$

Let ϕ_T be the relative truncation error, ϕ_R the relative roundoff error, ϕ_I the relative integration error and ϕ_F the relative function evaluation error so that $\Phi = \phi_T + \phi_R + \phi_I + \phi_F$. In order to have a prescribed relative error tolerance ε

the requirement is $|\Phi| \leq \varepsilon$. This is true if there exist α_T, α_R, α_I, $\alpha_F \geq 0$ such that $\alpha_T + \alpha_R + \alpha_I + \alpha_F = 1$ and

$$\phi_j \leq \alpha_j \varepsilon, \; j = T, \, R, \, I, \, F \tag{12}$$

In the Fast Multipole Method the worst error occurs when the source point $\mathbf{x}'$ and receiver point $\mathbf{x}$ are located in the corners of a box, as in Figure 2. In that case, if the side length of the box is a then $d = \sqrt{3}a$. The shortest distance between the midpoints of two boxes is $X = (n+1)\, a$. Here, n is the number of buffer boxes. A buffer box is a box which is considered too close to the source box for the approximation of the Green's function to be accurate. Thus, $|\mathbf{x} - \mathbf{x}'| = |\mathbf{X} + \mathbf{d}| \leq X + d = a\left(n + 1 + \sqrt{3}\right)$.

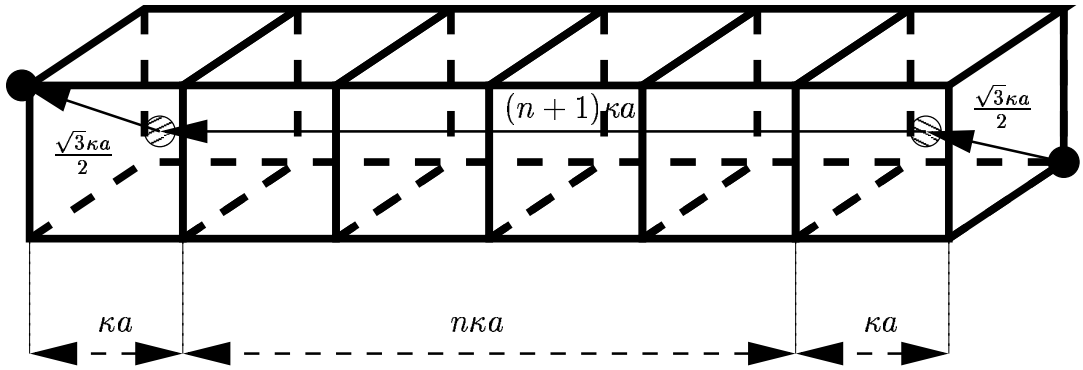

Figure 2: The worst case error in the Fast Multipole Method.

In the following sections the different relative errors ϕ_T, ϕ_R, ϕ_I and ϕ_F are analyzed. In order to keep the analysis simple we will only consider the largest errors and ignore the smaller ones. Also, arithmetic overflows and underflows are not considered. The results are summarized below so that a reader who is not interested in the technical details can skip the proofs and read section 4 instead.

The truncation errors have been analyzed before. Here, the results from [2] is repeated in (13) and (14) is derived from the results in [4]. In the interval $\sqrt{3}\kappa a \leq L < (n+1)\, \kappa a$ a lower bound on L is derived in [2]

$$L \geq \sqrt{3}\kappa a + 1.8 \left(-\log_{10}\left(\alpha_T \varepsilon\right)\right)^{\frac{2}{3}} \left(\sqrt{3}\kappa a\right)^{\frac{1}{3}} \tag{13}$$

and when $L > (n+1)\, \kappa a$ the lower bound is derived from the results [4]

$$L \geq (n+1)\, \kappa a + \frac{\log_{10}\left(\alpha_T \varepsilon\right) - \left(\frac{\left(n+1-\sqrt{3}\right)\kappa a}{1.8\left(\sqrt{3}\kappa a\right)^{\frac{1}{3}}}\right)^{\frac{3}{2}}}{\log_{10} \frac{\sqrt{3}}{n+1}} \tag{14}$$

The derivation of equation (13) in section 3.1 is repeated in this paper, since the derivation of the stability limit in section 3.5 is based on the same principles.

The stability requirement is derived in section 3.5. One conclusion is that it is sufficient that the parameters α_R, α_I and α_F satisfy

$$\alpha_R + \alpha_I + \alpha_F \geq \frac{D_R \epsilon_M + D_I \left(\epsilon_Q + C_I \epsilon_M\right) + 4\pi D_F \left(\epsilon_1 + \epsilon_2 + \epsilon_3\right)}{\varepsilon} \tag{15}$$

when $L < (n+1)\,\kappa a$. Here, D_R, D_I, D_F are constants depending on the geometry but independent of the parameters in the Fast Multipole Method. The constant ϵ_M is the machine precision. The constant ϵ_Q is the relative quadrature error defined in section 3.3 and is usually zero. The constants ϵ_1 and ϵ_3 are the relative errors in evaluating the exponentials discussed in section 3.4. When interpolation is used $\epsilon_1 + \epsilon_3 \leq C_p \epsilon_p \log N$ where ϵ_p is the largest interpolation error of all the levels and C_p is a constant. For a one level version of the Fast Multipole Method $\epsilon_1 \leq C_e \epsilon_M$ and $\epsilon_3 \leq C_e \epsilon_M$ where C_e is a constant that depends on the machine ability to evaluate the function e^{ix}. The constant ϵ_2 is the relative error in the evaluation of the translation operator. Usually, $|\epsilon_2| \leq C_{\mathcal{T}} \epsilon_M$ for some constant $C_{\mathcal{T}}$.

The second conclusion is that when $L \geq (n+1)\,\kappa a$ there is a risk that the errors start to grow exponentially and in order for the method to be numerically stable

$$L < (n+1)\,\kappa a + 1.8 \left(\log_{10} \left(\frac{(\alpha_I + \alpha_R)\,\varepsilon}{C_n C_\epsilon \kappa a} \right) \right)^{2/3} \left((n+1)\,\kappa a\right)^{1/3} \tag{16}$$

where $C_n = 2\left(n + 1 + \sqrt{3}\right)$ and $C_\epsilon = \epsilon_Q + C_I \epsilon_M + 4\pi\left(\epsilon_1 + \epsilon_2 + \epsilon_3\right)$.

Note that $\alpha_R + \alpha_I + \alpha_F \leq 0.5$, otherwise the numerical errors dominate over the truncation error due to finite L. In that case the number of buffer boxes should be increased or another approximation should be used, for instance the one in [6] or the Low frequency MLFMA in [2]. For $L \leq (n+1)\,\kappa a$ one can usually neglect $\alpha_R + \alpha_I + \alpha_F$. In that case they give a lower bound on the total error.

3.1 Truncation error

Truncation bounds for the Fast Multipole method have been analyzed in [2, 4]. While the analysis in [4] is exact up to a constant the asymptotic formula in [2] seems to be the one that is most widely used. Here, the formula in [2] is derived, since the derivation of the stability limit in section 3.5 is based on the same principles. In order to derive a formula for the number of terms needed we will use the following asymptotic expansions of the spherical Bessel function and the

spherical Hankel function [2, 9]

$$j_l(x) \approx \frac{1}{2\sqrt{f(l,x)\,x}} e^{f(l,x)-(L+1/2)\log\left(\frac{L+1/2+f(l,x)}{x}\right)} \tag{17}$$

$$h_l^{(1)}(x) \approx -\imath \frac{1}{\sqrt{f(l,x)\,x}} e^{(L+1/2)\log\left(\frac{L+1/2+f(l,x)}{x}\right)-f(l,x)} \tag{18}$$

where $f(l,x) = \sqrt{(l+1/2)^2 - x^2}$. The truncation error is given by

$$\begin{aligned} \frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi|\mathbf{x}-\mathbf{x}'|} - \frac{\imath\kappa}{4\pi}\sum_{l=0}^{L}(-1)^l(2l+1)\,j_l(\kappa d)\,h_l^{(1)}(\kappa X)\,P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \\ = \frac{\imath\kappa}{4\pi}\sum_{l=L+1}^{\infty}(-1)^l(2l+1)\,j_l(\kappa d)\,h_l^{(1)}(\kappa X)\,P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \end{aligned} \tag{19}$$

The requirement $|\phi_T| \leq \alpha_T\varepsilon$ is fulfilled if

$$\begin{aligned} |\phi_T| &= \left|\frac{\frac{\imath\kappa}{4\pi}\sum_{l=L+1}^{\infty}(-1)^l(2l+1)\,j_l(\kappa d)\,h_l^{(1)}(\kappa X)\,P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right)}{\frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi|\mathbf{x}-\mathbf{x}'|}}\right| \\ &\approx \kappa|\mathbf{x}-\mathbf{x}'|\left|(2L+3)\,j_{L+1}(\kappa d)\,h_{L+1}^{(1)}(\kappa X)\,P_{L+1}\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right)\right| \\ &\leq \kappa|\mathbf{x}-\mathbf{x}'|\left|(2L+3)\,j_{L+1}(\kappa d)\,h_{L+1}^{(1)}(\kappa X)\right| \leq \alpha_T\varepsilon \end{aligned} \tag{20}$$

If $x = \kappa a$ and the worst case error is considered then inserting (17) and (18) into (20) yields

$$\begin{aligned} \alpha_T\varepsilon &\geq \left|x\left(n+1+\sqrt{3}\right)(2L+3)\,j_{L+1}\left(\sqrt{3}x\right)h_{L+1}^{(1)}\left((n+1)\,x\right)\right| \\ &\approx \left|\frac{x\left(n+1+\sqrt{3}\right)(2L+3)}{2\sqrt{f\left(L+1,\sqrt{3}x\right)\sqrt{3}x\,f\left(L+1,(n+1)\,x\right)(n+1)\,x}}\right. \\ &\quad \times e^{f\left(L+1,\sqrt{3}x\right)-(L+3/2)\log\left(\frac{L+3/2+f\left(L+1,\sqrt{3}x\right)}{\sqrt{3}x}\right)} \\ &\quad \left.\times e^{-f(L+1,(n+1)x)+(L+3/2)\log\left(\frac{L+3/2+f(L+1,(n+1)x)}{(n+1)x}\right)}\right| \end{aligned} \tag{21}$$

When the last exponential can be controlled, the error is mainly due to the first exponential. This always the case as long as $L \leq (n+1)\,x - 3/2 \lesssim (n+1)\,x$, since $\sqrt{(L+3/2)^2 - (n+1)^2x^2}$ is complex and

$$\left|\frac{(L+3/2)+\sqrt{(L+3/2)^2-(n+1)^2x^2}}{(n+1)\,x}\right|^2 = \left|\frac{2(L+3/2)^2-(n+1)^2x^2}{(n+1)^2x^2}\right| < 1$$

(22)

In order to find the number of terms that are required we follow the steps in [2] and assume that

$$L + 3/2 = \sqrt{3}x\,(1+\delta) \tag{23}$$

Since the convergence of the series is fast when $\sqrt{3}x \leq L \leq (n+1)\,x$ [4] the number δ should be small compared to 1. The assumption on L yields

$$f\left(L+1, \sqrt{3}x\right) = \sqrt{3}x\sqrt{2\delta}\sqrt{1+\frac{\delta}{2}} \tag{24}$$

After Taylor expansion where the largest terms are kept, (21) becomes

$$\alpha_T \varepsilon \geq \left| \frac{n+1+\sqrt{3}}{\sqrt{\sqrt{\delta}\,(n+1)\,\sqrt{2\left((n+1)^2-3\right)}}} e^{-\sqrt{3}x\frac{(2\delta)^{3/2}}{3}} \right| \tag{25}$$

Take the logarithm of both sides. Then, the second term dominates over the first. Thus we have

$$\delta \geq \frac{1}{2}\left(\frac{-3\log\left(\alpha_T\varepsilon\right)}{x}\right)^{\frac{2}{3}} \approx 1.8\left(\frac{-\log_{10}\left(\alpha_T\varepsilon\right)}{x}\right)^{\frac{2}{3}} \tag{26}$$

Inserting (26) into (23) yields the required truncation number

$$L \geq \sqrt{3}\kappa a + 1.8\left(-\log_{10}\left(\alpha_T\varepsilon\right)\right)^{\frac{2}{3}}\left(\sqrt{3}\kappa a\right)^{\frac{1}{3}} \tag{27}$$

The case $L \gtrapprox (n+1)\,\kappa a$ has been analyzed in [4]. It is found that the error converges to zero as a geometrical series of ratio $\sqrt{3}/\,(n+1)$. A bound on L is then [4]

$$L \geq (n+1)\,\kappa a + \frac{7}{2} + \frac{\log_{10}\left(\alpha_T\varepsilon\right) - C}{\log_{10}\frac{\sqrt{3}}{n+1}} \tag{28}$$

for some constant C. The error at $L = (n+1)\,\kappa a$ is known from (27) and the convergence is at least as good as a geometrical series. This gives us an estimate of the constant C. We use this and derive the formula

$$L \geq (n+1)\,\kappa a + \frac{\log_{10}\left(\alpha_T\varepsilon\right) - \left(\frac{\left(n+1-\sqrt{3}\right)\kappa a}{1.8\left(\sqrt{3}\kappa a\right)^{\frac{1}{3}}}\right)^{\frac{3}{2}}}{\log_{10}\frac{\sqrt{3}}{n+1}} \tag{29}$$

for $L > (n+1)\,\kappa a$.

3.2 Roundoff error

In order to find the requirement that keep the roundoff error due to numerical approximation of the arguments below a tolerance $\alpha_R\varepsilon$ we analyze the second term in (10)

$$\begin{aligned}|\phi_R| = 4\pi\left|\mathbf{x}-\mathbf{x}'\right| \Bigg| & \frac{\imath\kappa}{4\pi}\sum_{l=0}^{L}(-1)^l\left(2l+1\right)j_l\left(\kappa d\right)h_l^{(1)}\left(\kappa X\right)P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \\ & -\frac{\imath\overline{\kappa}}{4\pi}\sum_{l=0}^{L}(-1)^l\left(2l+1\right)j_l\left(\overline{\kappa d}\right)h_l^{(1)}\left(\overline{\kappa X}\right)P_l\left(\overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}}\right)\Bigg|\end{aligned} \tag{30}$$

Let $\tilde{\boldsymbol{\epsilon}} = [\tilde{\epsilon}_1\,\tilde{\epsilon}_2\,\tilde{\epsilon}_3\,\tilde{\epsilon}_4]^T$ and define the differences $\kappa-\overline{\kappa} = \tilde{\epsilon}_1$, $\kappa d-\overline{\kappa d} = \tilde{\epsilon}_2$, $\kappa X-\overline{\kappa X} = \tilde{\epsilon}_3$ and $\hat{\mathbf{d}}\cdot\hat{\mathbf{X}} - \overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}} = \tilde{\epsilon}_4$ and assume that $|\tilde{\epsilon}_i| \leq \tilde{C}_R\epsilon_M$ for all i, where $\tilde{C}_R$ is a machine dependent constant and ϵ_M is the machine precision. Since we assume that all ϵ_i are small the mean value theorem yields

$$\begin{aligned}|\phi_R| \leq & 4\pi\left|\mathbf{x}-\mathbf{x}'\right|\left|\frac{e^{\imath\kappa|\mathbf{x}-\mathbf{x}'|}}{4\pi\left|\mathbf{x}-\mathbf{x}'\right|} - \frac{e^{\imath\overline{\kappa}\overline{|\mathbf{x}-\mathbf{x}'|}}}{4\pi\overline{|\mathbf{x}-\mathbf{x}'|}}\right| \\ & +4\pi\left|\mathbf{x}-\mathbf{x}'\right|\Bigg|\frac{\imath\kappa}{4\pi}\sum_{l=L+1}^{\infty}(-1)^l\left(2l+1\right)j_l\left(\kappa d\right)h_l^{(1)}\left(\kappa X\right)P_l\left(\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}\right) \\ & -\frac{\imath\overline{\kappa}}{4\pi}\sum_{l=L+1}^{\infty}(-1)^l\left(2l+1\right)j_l\left(\overline{\kappa d}\right)h_l^{(1)}\left(\overline{\kappa X}\right)P_l\left(\overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}}\right)\Bigg| \\ & \approx \left|\mathbf{x}-\mathbf{x}'\right|\left|\nabla\left(\frac{e^{\imath\overline{\kappa}\overline{|\mathbf{x}-\mathbf{x}'|}}}{\overline{|\mathbf{x}-\mathbf{x}'|}}\right)\cdot\tilde{\boldsymbol{\epsilon}}\right| \leq D_R\epsilon_M \leq \alpha_R\varepsilon\end{aligned} \tag{31}$$

where D_R is a constant depending on geometry and the gradient is with respect to the components. Here, we assume that $\kappa d < \kappa X$ and $\overline{\kappa d} < \overline{\kappa X}$, so that both infinite sums are convergent. From the previous section we know that both sums have values of the order ϵ_T. In [4] it was proven that the sums are convergent in the absolute sense and that the convergence rate is at least as good as a geometrical series for large l. Cancellation implies that the combined value is a second order error and can be neglected. Note that α_R can usually be neglected since it is enough that $\alpha_R \geq D_R\epsilon_M/\varepsilon \approx 0$.

3.3 Integration error

The integration error in (10) is given by

$$
\begin{aligned}
|\phi_I| = 4\pi\,|\mathbf{x}-\mathbf{x}'| \Bigg| & \frac{\imath\overline{\kappa}}{4\pi}\sum_{l=0}^{L}(-1)^l\,(2l+1)\,j_l\left(\overline{\kappa d}\right)h_l^{(1)}\left(\overline{\kappa X}\right)P_l\left(\overline{\hat{\mathbf{d}}\cdot\hat{\mathbf{X}}}\right) \\
& -\sum_{k=1}^{K} e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}\mathcal{T}_k^L\left(\overline{\kappa},\overline{\mathbf{X}_m}-\overline{\mathbf{X}'_m}\right)e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}\Bigg|
\end{aligned}
\tag{32}
$$

Equations (3) and (7) can be used to rewrite (32) as

$$
\begin{aligned}
|\phi_I| &= 4\pi\,|\mathbf{x}-\mathbf{x}'|\,\Bigg|\frac{\imath\overline{\kappa}}{4\pi}\sum_{l=0}^{L}\imath^l\,(2l+1)\,h_l^{(1)}\left(\overline{\kappa X}\right) \\
&\qquad\times\left(\int_{\mathcal{S}} e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}\cdot\overline{\hat{\mathbf{X}}}\right)-\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}}P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right)\Bigg| \\
&\leq 4\pi\,|\mathbf{x}-\mathbf{x}'|\,\Bigg|\frac{\imath\overline{\kappa}}{4\pi}\sum_{l=0}^{L}\imath^l\,(2l+1)\,h_l^{(1)}\left(\overline{\kappa X}\right) \\
&\qquad\times\left(\int_{\mathcal{S}} e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}\cdot\overline{\hat{\mathbf{X}}}\right)-\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}_k\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\overline{\hat{\mathbf{X}}}\right)\right)\Bigg| \\
&\quad+4\pi\,|\mathbf{x}-\mathbf{x}'|\,\Bigg|\frac{\imath\overline{\kappa}}{4\pi}\sum_{l=0}^{L}\imath^l\,(2l+1)\,h_l^{(1)}\left(\overline{\kappa X}\right) \\
&\qquad\times\left(\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}_k\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\overline{\hat{\mathbf{X}}}\right)-\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}}P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right)\Bigg|
\end{aligned}
\tag{33}
$$

The quadrature rule is usually chosen to integrate the first $2L$ spherical harmonics exactly [3]. In that case the first part of the error is almost zero. Let us assume that the relative error is ϵ_Q when another quadrature rule is used. The errors in the second part depends on the machine representation of the nodes $\hat{\boldsymbol{\kappa}}_k$. The mean value theorem yields

$$
\begin{aligned}
&\left|\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}_k\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\overline{\hat{\mathbf{X}}}\right)-\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}}P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right| \\
&\approx\left|\nabla\left(\sum_{k=1}^{K}w_k e^{\imath\overline{\kappa}\hat{\boldsymbol{\kappa}}_k\cdot\overline{\mathbf{d}}}P_l\left(\hat{\boldsymbol{\kappa}}_k\cdot\overline{\hat{\mathbf{X}}}\right)\right)\cdot\boldsymbol{\epsilon}\right|\leq C_I\epsilon_M,\ l=0,\ldots,L
\end{aligned}
\tag{34}
$$

for a constant C_I. The triangle inequality applied to (33) gives

$$|\phi_I| \leq \overline{\kappa}\,|\mathbf{x}-\mathbf{x}'| \sum_{l=0}^{L} \left|(2l+1)\, h_l^{(1)}\left(\overline{\kappa X}\right)\right| (\epsilon_Q + C_I \epsilon_M) \leq \alpha_I \varepsilon \tag{35}$$

Consider the worst case error. When $L < (n+1)\,\kappa a$ the size of the error in (35) can be controlled if $\alpha_I \geq D_I\,(\epsilon_Q + C_I\epsilon_M)\,/\varepsilon$, but for $L > (n+1)\,\kappa a$ the error starts to grow exponentially due to the growth of the spherical Hankel function. The latter case is analyzed in section 3.5.

3.4 Function evaluation error

The function evaluation error in (10) is given by

$$\begin{aligned} |\phi_F| = 4\pi\,|\mathbf{x}-\mathbf{x}'| \Bigg| & \sum_{k=1}^{K} e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)} \mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right) e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)} \\ & - \sum_{k=1}^{K} \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}}\, \overline{\mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right)}\, \overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}} \Bigg| \end{aligned} \tag{36}$$

Assume that the function $\overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}} = e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)}\,(1+\epsilon_{1k})$, and that the function $\overline{\mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right)} = \mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right)(1+\epsilon_{2k})$ and that the function $\overline{e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}} = e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)}\,(1+\epsilon_{3k})$. If $|\epsilon_{ij}| \leq \epsilon_i$ for all j and only the largest terms are included

$$\begin{aligned} |\phi_F| &\approx 4\pi\,|\mathbf{x}-\mathbf{x}'| \Bigg| \sum_{k=1}^{K} (\epsilon_{1k}+\epsilon_{2k}+\epsilon_{3k}) \\ &\qquad \times e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{x}}-\overline{\mathbf{X}_m}\right)} \mathcal{T}_k^L\left(\overline{\kappa}, \overline{\mathbf{X}_m} - \overline{\mathbf{X}'_m}\right) e^{\imath\overline{\boldsymbol{\kappa}_k}\cdot\left(\overline{\mathbf{X}'_m}-\overline{\mathbf{x}'}\right)} \Bigg| \\ &= |\mathbf{x}-\mathbf{x}'| \Bigg| \imath\overline{\kappa} \sum_{l=0}^{L} \imath^l\,(2l+1)\, h_l^{(1)}\left(\overline{\kappa X}\right) \\ &\qquad \times \sum_{k=1}^{K} (\epsilon_{1k}+\epsilon_{2k}+\epsilon_{3k})\, w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}} P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right) \Bigg| \end{aligned} \tag{37}$$

If the quadrature weights are positive and they integrate exactly for a constant

function the inner sum can be estimated by

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\left(\epsilon_{1k}+\epsilon_{2k}+\epsilon_{3k}\right)\, w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}} P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right| \\
&\qquad\qquad \leq \sum_{k=1}^{K}\left|\left(\epsilon_{1k}+\epsilon_{2k}+\epsilon_{3k}\right) w_k e^{\imath\overline{\kappa}\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\mathbf{d}}} P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right| \\
&\qquad\qquad \leq \left(\epsilon_1+\epsilon_2+\epsilon_3\right)\sum_{k=1}^{K}\left|w_k P_l\left(\overline{\hat{\boldsymbol{\kappa}}_k}\cdot\overline{\hat{\mathbf{X}}}\right)\right| \\
&\qquad\qquad \leq \left(\epsilon_1+\epsilon_2+\epsilon_3\right)\sum_{k=1}^{K}|w_k| = 4\pi\left(\epsilon_1+\epsilon_2+\epsilon_3\right)
\end{aligned} \tag{38}$$

since $|P_l(\cos\theta)| \leq 1$. From the triangle inequality (37) is estimated by

$$|\phi_F| \leq 4\pi\overline{\kappa}\,|\mathbf{x}-\mathbf{x}'|\sum_{l=0}^{L}\left|\imath^l\left(2l+1\right)h_l^{(1)}\left(\overline{\kappa X}\right)\right|\left(\epsilon_1+\epsilon_2+\epsilon_3\right) \leq \alpha_F\varepsilon \tag{39}$$

The worst case error can only be controlled as long as $L < (n+1)\,\kappa a$ as in (35). In that case $\alpha_F \geq 4\pi D_F\left(\epsilon_1+\epsilon_2+\epsilon_3\right)/\varepsilon$, for some constant D_F depending on geometry, is sufficient.

If a one level version is used then $\epsilon_1 \leq C_e\epsilon_M$ and $\epsilon_3 \leq C_e\epsilon_M$ where C_e is a constant that depends on the machine ability to evaluate the function $e^{\imath x}$. For a multilevel version the functions $e^{\imath\boldsymbol{\kappa}_k\cdot(\mathbf{x}-\mathbf{X}_m)}$ and $e^{\imath\boldsymbol{\kappa}_k\cdot(\mathbf{X}'_m-\mathbf{x}')}$ are computed by interpolation [11]. Therefore the errors depend on the interpolation scheme used. If only the largest terms are kept the errors are approximately $\epsilon_1+\epsilon_3 \leq C_p\epsilon_p\log N$ where ϵ_p is the largest interpolation error of all the levels and C_p is a constant.

The relative error ϵ_2 should be small since the translation operator is a sum of functions. Assume that each function can be evaluated with an error within a constant of machine precision. Unless some terms cause cancellation or the Hankel function value gives overflow, the total relative error can also be kept to within a constant of machine precision by summing the terms from the smallest to largest term. Thus the relative error $\epsilon_2 \leq C_{\mathcal{T}}\epsilon_M$ for some constant $C_{\mathcal{T}}$.

3.5 Stability requirements

If $L \geq (n+1)\,\kappa a$ the highest order spherical Hankel function start to grow exponentially. This results in a growth of the errors in (35) and (39). Using the

asymptotic form (18) the dominant worst case error is given by

$$\begin{aligned}
|\phi_I| + |\phi_F| \approx & \overline{\kappa} \left|\mathbf{x} - \mathbf{x}'\right| \left|(2L+1)\, h_L^{(1)}\left(\overline{\kappa X}\right)\right| C_\epsilon \\
\approx & \overline{\kappa} \left|\mathbf{x} - \mathbf{x}'\right| (2L+1) \\
& \times \left| \frac{e^{-f(L,(n+1)x)+(L+1/2)\log\left(\frac{L+1/2+f(L,(n+1)x)}{(n+1)x}\right)}}{\sqrt{f\left(L,(n+1)\,x\right)(n+1)\,x}} \right| C_\epsilon
\end{aligned} \tag{40}$$

where $C_\epsilon = \epsilon_Q + C_I \epsilon_M + 4\pi\left(\epsilon_1 + \epsilon_2 + \epsilon_3\right)$. Since the right hand side grows quickly with L the only possibility that the error tolerance can be achieved is if L is slightly larger than $(n+1)\,x$. Hence, a reasonable assumptions is $L + 1/2 = (n+1)\,x\,(1+\delta)$, where δ is small compared to 1 and $x = \kappa a$. Inserting this into (40) yields

$$\begin{aligned}
|\phi_I| + |\phi_F| \approx & \left| \frac{2\left(n+1+\sqrt{3}\right) x\,(n+1)\,x}{\sqrt{(n+1)\,x\sqrt{2\delta}\sqrt{1+\frac{\delta}{2}}\,(n+1)\,x}} e^{(n+1)x\frac{(2\delta)^{3/2}}{3}} \right| C_\epsilon \\
\leq & \left| \frac{2\left(n+1+\sqrt{3}\right) x}{\sqrt{\sqrt{\delta}}} e^{(n+1)x\frac{(2\delta)^{3/2}}{3}} \right| C_\epsilon \leq \left(\alpha_I + \alpha_R\right)\varepsilon
\end{aligned} \tag{41}$$

The same analysis as in Section 3.1 yields the formula

$$L < (n+1)\,\kappa a + 1.8\left(\log_{10}\left(\frac{\left(\alpha_I + \alpha_R\right)\varepsilon}{C_n C_\epsilon \kappa a}\right)\right)^{2/3} \left((n+1)\,\kappa a\right)^{1/3} \tag{42}$$

where $C_n = 2\left(n+1+\sqrt{3}\right)$. Equation (42) gives the largest possible L that can achieve a given error ε. If L is larger it is outside the stability region where the truncation error dominates and must be reduced in order to achieve a stable scheme. In that case the error that was requested can not be achieved. The best choice of L is then when $\left(\alpha_T + \alpha_I + \alpha_R\right)\varepsilon$ is minimized given by (27) and (29) and (42). The minimum occurs when $\alpha_T = \alpha_I + \alpha_R \approx 0.5$ as long as the roundoff errors are negligible.

4 Numerical experiments

The theory in section 3 is validated by numerical experiments. In the first experiment the effect of truncation and roundoff errors on (1) is investigated in Figure 3 and Figure 4. The actual relative errors for the worst case positions are compared to the relative errors predicted by the theory in (13), (14) and (15). Two cases are considered $\kappa a = 1$ and $\kappa a = 10$. For each case the number of

buffer boxes n is 1, 2, 3, and 10. It is assumed that the right hand side of (15) is approximately $10^{-15}/\varepsilon$. In the figures the line $\sqrt{3}\kappa a$ which marks the beginning of the fast convergence zone and the line $(n+1)\,\kappa a$ that marks the end of the fast convergence zone and the beginning of the geometrical convergence zone are plotted. The theoretical predictions give results that agree within a constant to the errors achieved. From the figures it is clear that increasing the number of buffer boxes gives a faster convergence rate. The reasons are that the fast convergence zone $\sqrt{3}\kappa a \leq L < (n+1)\,\kappa a$ is larger and that the convergence factor in the geometrical series is smaller. Thus, a good way to achieve smaller errors in the approximation is to increase the number of buffer boxes. As predicted by theory the effect of roundoff is that the relative errors can not be reduced below the roundoff level. Therefore, L should not be increased beyond the point where the roundoff errors begin to dominate.

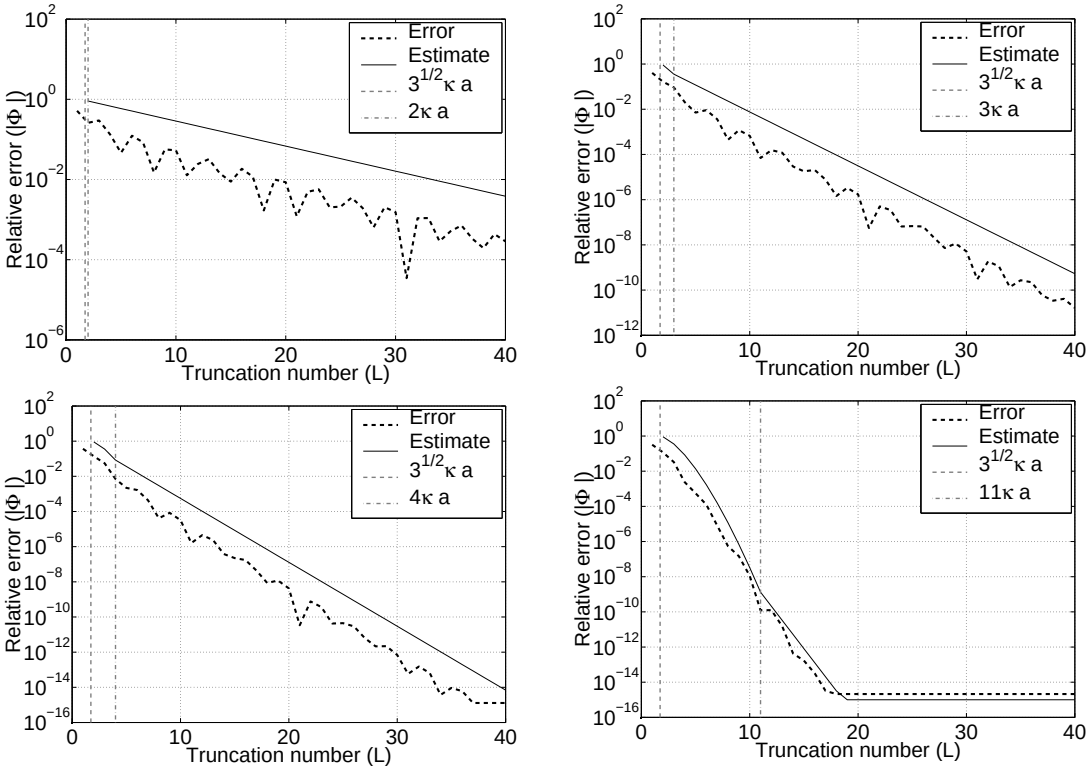

Figure 3: The relative error in (1), when $\kappa a = 1$, as a function of truncation number. The four cases are 1 buffer box (top left), 2 buffer boxes (top right), 3 buffer boxes (bottom left) and 10 buffer boxes (bottom right).

In the second experiment the stability of (8) is investigated in Figure 5 and Figure 6. The setup is the same as in the first experiment but in this case the theoretical stability limit in (16) is also needed. The constant is set to $C_n C_\epsilon \approx 3 \cdot 10^{-16} \left(n + 1 + \sqrt{3}\right)$ following from section 3.5. Also in this case the theoretical predictions agree within a constant with the computed errors. Note that the error

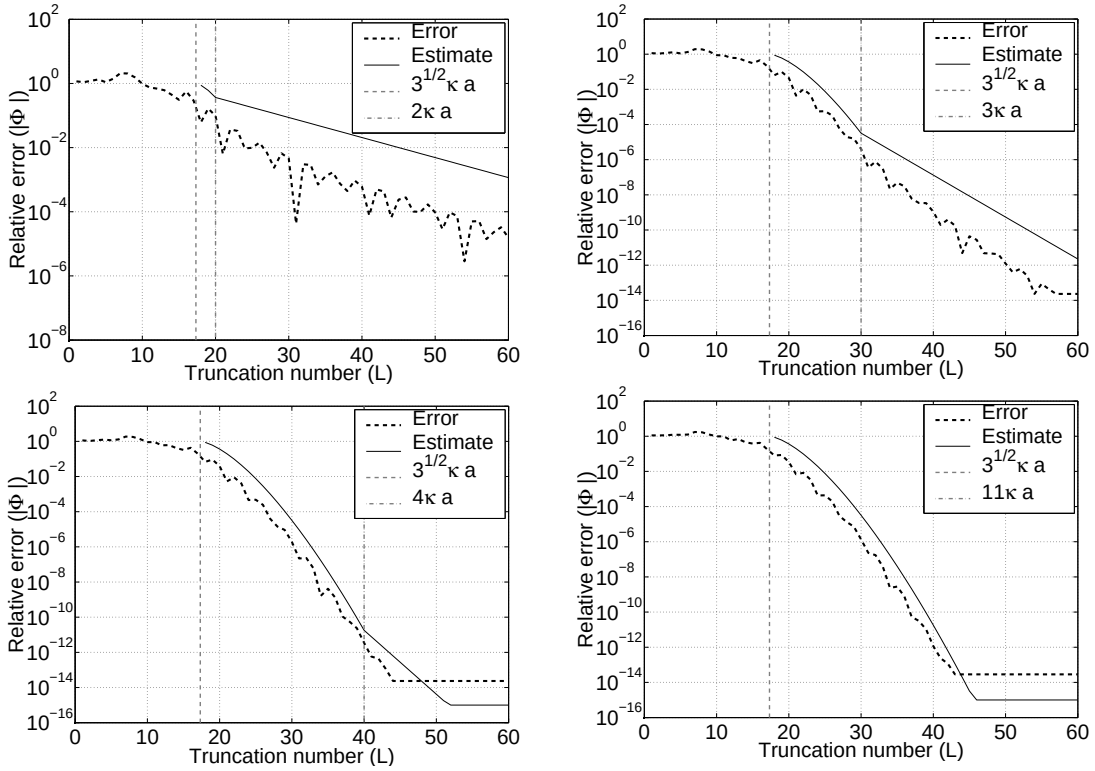

Figure 4: The relative error in (1), when $\kappa a = 10$, as a function of truncation number. The four cases are 1 buffer box (top left), 2 buffer boxes (top right), 3 buffer boxes (bottom left) and 10 buffer boxes (bottom right).

can only start to grow when $L > (n+1)\,\kappa a$ as expected. When L is smaller, either truncation error or roundoff error is the major source of error as in the previous case. There is no point in increasing L when the stability limit for L or roundoff limit is reached.

In the third experiment the effect of increasing the relative error $\epsilon_1+\epsilon_2+\epsilon_3$ on stability is demonstrated for the case $\kappa a = 10$ in Figure 7. In the first case white noise of the size $\epsilon_1 = \epsilon_3 = 10^{-4}$ computed by the MATLAB command `rand` is added to the computed exponentials. In the second case white noise of the size $\epsilon_2 = 10^{-4}$ is added to the computed translation operators. In the estimates $C_n C_\epsilon \approx 3{\cdot}10^{-4}\left(n+1+\sqrt{3}\right)$ is used. Up to constant the theory coincides with the experiments in this case as well. It is important to note that the approximations incurred by interpolation or approximating the translation operator will affect the relative errors in FMM at the end. But, if the relative error of the truncation error is larger than the accuracy in the approximation the total error is scarcely affected. However, the stability region is reduced.

The final experiment is intended to show the performance of the new error estimates in Figure 8. The method for estimating the truncation number is simple. First (13) is used to estimate the truncation number. If the estimated number

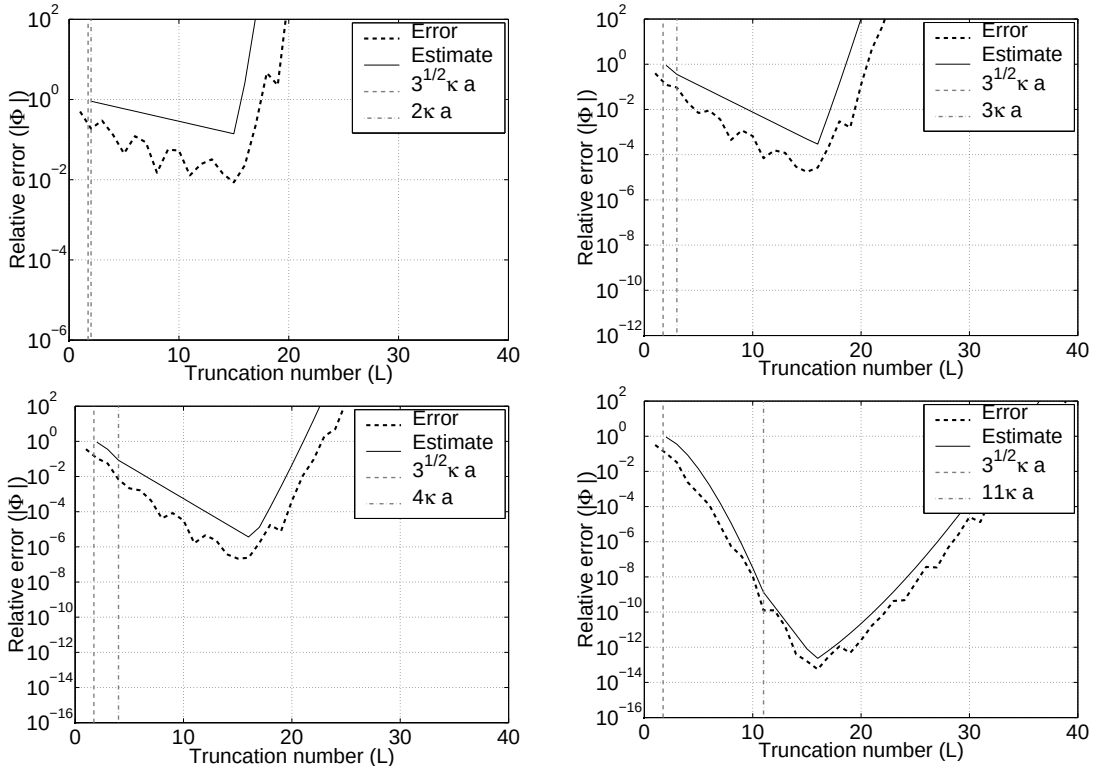

Figure 5: The relative error in (8), when $\kappa a = 1$, as a function of truncation number. The four cases are 1 buffer box (top left), 2 buffer boxes (top right), 3 buffer boxes (bottom left) and 10 buffer boxes (bottom right).

$L > (n+1)\,\kappa a$ the estimate (14) is used to compute a new truncation number. If that number does not fulfill the stability requirement (16) the truncation number is reduced until an estimated error from (14) fulfills (16) or $L \leq (n+1)\,\kappa a$. The simple method works well in these experiments. When κa is small the effect of using (14) to find the smallest possible error is visible. It is clear that the constant in (28) needs to be refined further. As κa grows the excess bandwidth formula (13) is used, which gives an accurate prediction of L up to a constant.

5 Conclusions

The theory and experiments demonstrate that the stability of FMM is affected by different errors. The main conclusion is that the choice of interpolation function, quadrature rule and approximation of translation operator becomes more important when high precision is required. For many applications the requirements on precision are low or moderate [2] and the effects reported here are not visible. The motivation for this work was to find strict limits that justify that the errors incurred in these cases can be ignored.

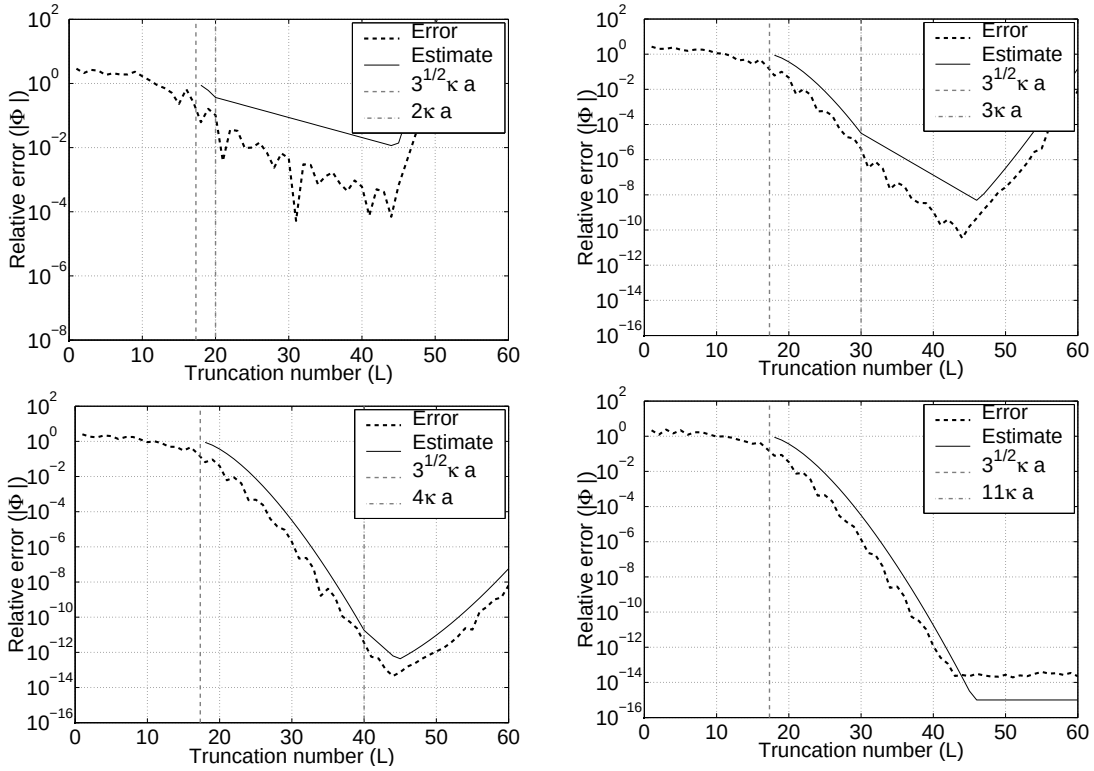

Figure 6: The relative error in (8), when $\kappa a = 10$, as a function of truncation number. The four cases are 1 buffer box (top left), 2 buffer boxes (top right), 3 buffer boxes (bottom left) and 10 buffer boxes (bottom right).

If the stability limits are reached, one has to switch to another method or increase the number of buffer boxes. The formulas in (13), (14), (15) and (16) can be used to determine when the switch should be made. For large values on κa the diagonal form given here, should be superior in terms of speed to most other approximations. Therefore, it is important to calibrate the constants that are used to estimate the errors.

References

[1] George B. Arfken and Hans J. Weber. *Mathematical Methods for Physicists.* Academic Press, Inc., 4th edition, 1995.

[2] Weng Cho Chew, Jian-Ming Jin, Eric Michielssen, and Jiming Song. *Fast and Efficient Algorithms in Computational Electromagnetics.* Artech House, Inc., Norwood, 2001.

[3] Ronald Coifman, Vladimir Rokhlin, and Stephen Wandzura. The fast multi-

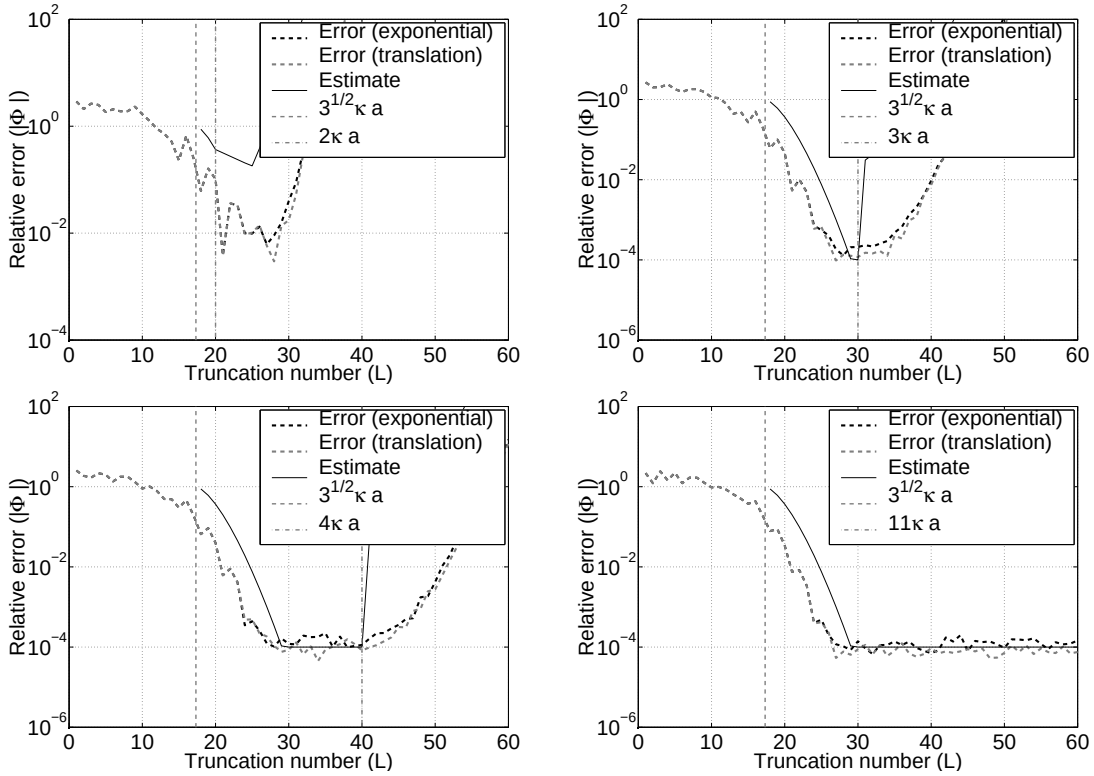

Figure 7: The relative error in (8), when $\kappa a = 10$, as a function of truncation number and the relative error in the approximation of the exponential or the translation operator is 10^{-4}. The four cases are 1 buffer box (top left), 2 buffer boxes (top right), 3 buffer boxes (bottom left) and 10 buffer boxes (bottom right).

pole method for the wave equation: A pedestrian prescription. *IEEE Transactions on Antennas and Propagation*, 35(3):7–12, June 1993.

[4] Eric Darve. The fast multipole method I: Error analysis and asymptotic complexity. *SIAM J. Numer. Anal.*, 38(1):98–128, 2000.

[5] Eric Darve. The fast multipole method: Numerical implementation. *Journal of Computational Physics*, 160(1):195–240, 2000.

[6] Leslie Greengard, Jingfang Huang, Vladimir Rokhlin, and Stephen Wandzura. Accelerating fast multipole methods for the Helmholtz equation at low frequencies. *IEEE Computational Science and Engineering*, 5(3):32–38, July–September 1998.

[7] Michael Larkin Hastriter, Schinichiro Ohnuki, and Weng Cho Chew. Error control of the translation operator in 3D MLFMA. *Microwave and Optical Technology Letters*, 37(3):184–188, May 2003.

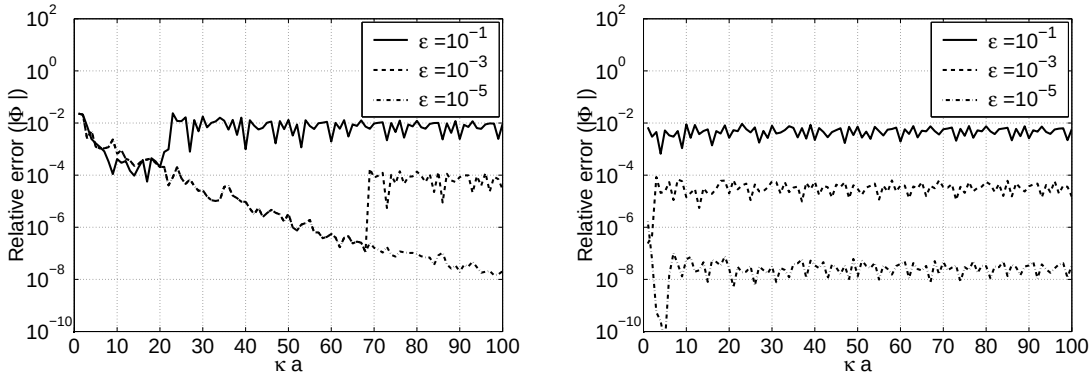

Figure 8: The relative error in (8), when a simple method is used to determine the truncation number for different κa. The two cases are 1 buffer box (left) and 3 buffer boxes (right).

[8] Martin Nilsson. A parallel shared memory implementation of the fast multipole method for electromagnetics. Technical Report 2003-049, Department of Information Technology, Scientific Computing, Uppsala University, Oct. 2003. Available at: http://www.it.uu.se/research/reports/2003-049/.

[9] Schinichiro Ohnuki and Weng Cho Chew. Numerical accuracy of multipole expansion for 2D MLFMA. *IEEE Transactions on Antennas and Propagation*, 51(8):1883–1890, August 2003.

[10] Vladimir Rokhlin. Diagonal forms of translation operators for the Helmholtz equation in three dimensions. *Applied and Computational Harmonic Analysis*, 1(1):82–93, 1993.

[11] Jiming Song and Weng Cho Chew. Multilevel fast multipole algorithm for solving combined field integral equation of electromagnetic scattering. *Microwave and Optical Technology Letters*, 10(1):14–19, September 1995.