Physics Open 3 (2020) 100019

Contents lists available at ScienceDirect

Physics Open

journal homepage: www.journals.elsevier.com/physics-open

An overview of the null-field method. II: Convergence and numerical stability

Adrian Doicu [a,*], Michael I. Mishchenko [b]

[a] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Methodik der Fernerkundung (IMF), Oberpfaffenhofen, 82234, Germany
[b] NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY, 10025, USA

ARTICLE INFO

Handling editor: Pinar Menguc

ABSTRACT

In this paper we provide an analysis of the convergence and numerical stability of the null-field method with discrete sources. We show that (i) if the null-field scheme is numerically stable then we can decide whether or not convergence can be achieved; (ii) if the null-field scheme is numerically unstable then we cannot draw any conclusion about the convergence issue; and (iii) the numerical stability is closely related to the property of a tangential system of radiating discrete sources to form a Riesz basis. Our numerical analysis indicates that for prolate spheroids and localized vector spherical wave functions, the null-field scheme is numerically unstable (this system of vector functions does not form a Riesz basis), while for distributed vector spherical wave functions, the numerical instability is not so pronounced (this system of discrete sources almost possesses the property of being a Riesz basis). We also describe an analytical method for computing the surface integrals in the framework of the conventional null-field method with localized vector spherical wave functions which increases the stability of the numerical scheme.

1. Introduction

In the first part of this series [1], we formulated the null-field scheme with discrete sources, as an approach aiming at constructing an approximate solution to the transmission boundary-value problem in electromagnetic scattering. In this paper we analyze the convergence and numerical stability of the method. Before proceeding we summarize some results of Ref. [1] that are relevant in our analysis.

Let D_i be a bounded three-dimensional domain with a smooth closed boundary S, and simply connected exterior D_s. Furthermore, let $\hat{\mathbf{n}}$ be the outward pointing unit normal vector to S, and ε_t and μ_t the (constant) electric permittivity and magnetic permeability in the domain $D_t, t = s,i$, respectively. The wavenumber in D_t is $k_t = k_0\sqrt{\varepsilon_t\mu_t}$, where k_0 is the wavenumber in free space. In the null-field method with discrete sources we consider the vector functions $\mathfrak{M}_\alpha^q(k_t\mathbf{r})$ and $\mathfrak{N}_\alpha^q(k_t\mathbf{r})$, for $q = 1,3$ and $t = s,i$, with the properties (i) $\nabla\times\mathfrak{M}_\alpha^q = k_t\mathfrak{N}_\alpha^q$ and $\nabla\times\mathfrak{N}_\alpha^q = k_t\mathfrak{M}_\alpha^q$; (ii) $\mathfrak{M}_\alpha^1$ and $\mathfrak{N}_\alpha^1$ are finite at the origin; and (iii) $\mathfrak{M}_\alpha^3$ and $\mathfrak{N}_\alpha^3$ satisfy the radiation condition. The vector functions $\mathfrak{M}_\alpha^q(k_t\mathbf{r})$ and $\mathfrak{N}_\alpha^q(k_t\mathbf{r})$ stand for the localized, multiple, and distributed vector spherical wave functions, distributed vector Mie potentials, and distributed magnetic and electric dipoles; the significance of the multi-index α for each system of discrete sources is explained in Ref. [1]. In terms of discrete sources, the infinite set of null-field equations for the (total) tangential fields $\mathbf{e}$ and $\mathbf{h}$, reads as.

$$\int_S\left[(\mathbf{e}-\mathbf{e}_0)\cdot\mathfrak{M}_{\bar{\alpha}}^{\,3}(k_s\,\cdot)+\mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}(\mathbf{h}-\mathbf{h}_0)\cdot\mathfrak{N}_{\bar{\alpha}}^{\,3}(k_s\,\cdot)\right]\mathrm{d}S=0, \tag{1}$$

$$\int_S\left[(\mathbf{e}-\mathbf{e}_0)\cdot\mathfrak{N}_{\bar{\alpha}}^{\,3}(k_s\,\cdot)+\mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}(\mathbf{h}-\mathbf{h}_0)\cdot\mathfrak{M}_{\bar{\alpha}}^{\,3}(k_s\,\cdot)\right]\mathrm{d}S=0, \tag{2}$$

$$\int_S\left[\mathbf{e}\cdot\mathfrak{M}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)+\mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\mathbf{h}\cdot\mathfrak{N}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\right]\mathrm{d}S=0, \tag{3}$$

$$\int_S\left[\mathbf{e}\cdot\mathfrak{N}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)+\mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\mathbf{h}\cdot\mathfrak{M}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\right]\mathrm{d}S=0, \tag{4}$$

for $\alpha = 1,2,\ldots$. The entire formalism is based on the completeness and linear independence of the systems of tangential vector functions

$$\left\{\begin{bmatrix}\hat{\mathbf{n}}\times\mathfrak{M}_\alpha^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\hat{\mathbf{n}}\times\mathfrak{N}_\alpha^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\hat{\mathbf{n}}\times\mathfrak{M}_\alpha^1(k_i\,\cdot)\end{bmatrix},\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\hat{\mathbf{n}}\times\mathfrak{N}_\alpha^1(k_i\,\cdot)\end{bmatrix}\,\middle|\,k_i\notin\Lambda(D_i)\right\}_{\alpha=1}^{\infty} \tag{5}$$

Given his role as Editor-in-Chief of this journal, M. Mischenko had no involvement in the peer-review of articles for which he was an author and had no access to information regarding their peer-review. Full responsibility for the peer-review process for this article was delegated to another Editor.

* Corresponding author.
E-mail address: adrian.doicu@dlr.de (A. Doicu).

https://doi.org/10.1016/j.physo.2020.100019
Received 13 February 2020; Accepted 23 March 2020
Available online 29 April 2020

and

$$\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^{3,1}(k_{s,i}\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_{s,i}}{\varepsilon_{s,i}}}\,\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^{3,1}(k_{s,i}\,\cdot)\end{bmatrix},\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^{3,1}(k_{s,i}\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_{s,i}}{\varepsilon_{s,i}}}\,\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^{3,1}(k_{s,i}\,\cdot)\end{bmatrix}\right\}_{\alpha=1}^{\infty} \tag{6}$$

in the product space $\mathfrak{T}^2(S)=T^2(S)\times T^2(S)$, where $T^2(S)$ is the Hilbert space of all square integrable tangential vector functions, **0** is the zero vector, and the superscripts 3 and 1 correspond to the subscripts *s* and *i*. Once the surface fields **e** and **h** are determined, the scattered field as well as the (electric) far-field pattern are computed outside a sphere enclosing the particle by means of the expansions

$$\mathbf{E}_s(\mathbf{r})=\sum_{\alpha=1}\left[f_\alpha\mathbf{M}_\alpha^3(k_s\mathbf{r})+g_\alpha\mathbf{N}_\alpha^3(k_s\mathbf{r})\right], \tag{7}$$

$$\mathbf{E}_{s\infty}(\widehat{\mathbf{r}})=\frac{1}{k_s}\sum_{\alpha=1}[f_\alpha\tilde{\mathbf{m}}_\alpha(\widehat{\mathbf{r}})+\mathrm{j}g_\alpha\tilde{\mathbf{n}}_\alpha(\widehat{\mathbf{r}})], \tag{8}$$

where.

$$\begin{bmatrix}f_\alpha\\ g_\alpha\end{bmatrix}=\mathrm{j}k_s^2\int_S\left\{\mathbf{e}\cdot\begin{bmatrix}\mathbf{N}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\\ \mathbf{M}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\end{bmatrix}\right.$$

$$\left.+\mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\mathbf{h}\cdot\begin{bmatrix}\mathbf{M}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\\ \mathbf{N}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\end{bmatrix}\right\}\mathrm{d}S,\quad\alpha=1,2,\ldots, \tag{9}$$

and

$$\tilde{\mathbf{m}}_\alpha(\widehat{\mathbf{r}})=\tilde{\mathbf{m}}_{mn}(\widehat{\mathbf{r}})=(-\mathrm{j})^{n+1}\mathbf{m}_{mn}(\widehat{\mathbf{r}}), \tag{10}$$

$$\tilde{\mathbf{n}}_\alpha(\widehat{\mathbf{r}})=\tilde{\mathbf{n}}_{mn}(\widehat{\mathbf{r}})=(-\mathrm{j})^{n+1}\mathbf{n}_{mn}(\widehat{\mathbf{r}}), \tag{11}$$

with $\mathbf{m}_{mn}(\widehat{\mathbf{r}})$ and $\mathbf{n}_{mn}(\widehat{\mathbf{r}})$ being the normalized spherical harmonic vectors.

For a numerical implementation, the infinite set of null-field equations (1)–(4) is truncated at some order *N*, and the tangential fields **e** and **h** are approximated by the system of tangential vector functions (5). Using the orthogonality relation

$$\int_S\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}\cdot\begin{bmatrix}\mathfrak{M}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\\ \mathfrak{N}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\end{bmatrix}+\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix}\cdot\begin{bmatrix}\mathfrak{N}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\\ \mathfrak{M}_{\bar{\alpha}}^{\,1}(k_i\,\cdot)\end{bmatrix}\right\}\mathrm{d}S$$
$$=\begin{bmatrix}0\\ 0\end{bmatrix}, \tag{12}$$

for any $\alpha,\beta=1,2,\ldots$, we found that the approximants

$$\begin{bmatrix}\mathbf{e}_N\\ \mathbf{h}_N\end{bmatrix}=\sum_{\beta=1}^{N}\left\{c_\beta^N\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix}\right.$$

$$\left.+d_\beta^N\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}\right\}, \tag{13}$$

satisfy the truncated systems of null-field equations (1) and (2). Inserting the above representation into the truncated systems of null-field equations, and taking into account the expansion of the incident field in terms of regular vector spherical wave functions

$$\mathbf{E}_0(\mathbf{r})=\sum_{\alpha=1}^{\infty}\left[a_\alpha\mathbf{M}_\alpha^1(k_s\mathbf{r})+b_\alpha\mathbf{N}_\alpha^1(k_s\mathbf{r})\right], \tag{14}$$

we obtained the following matrix equation for the expansion coefficients $\{c_\beta^N,d_\beta^N\}_{\beta=1}^N$:

$$\mathrm{Q}_{31N}(k_s,k_i)\begin{bmatrix}(c_\beta^N)_{\beta=1}^N\\ (d_\beta^N)_{\beta=1}^N\end{bmatrix}=\mathrm{Q}_{31N}^0(k_s,k_s)\begin{bmatrix}(a_\alpha)_{\alpha=1}^N\\ (b_\alpha)_{\alpha=1}^N\end{bmatrix}. \tag{15}$$

The approximate scattered field and far-field pattern are then given by

$$\mathbf{E}_{sN}(\mathbf{r})=\sum_{\alpha=1}\left[f_\alpha^N\mathbf{M}_\alpha^3(k_s\mathbf{r})+g_\alpha^N\mathbf{N}_\alpha^3(k_s\mathbf{r})\right], \tag{16}$$

$$\mathbf{E}_{s\infty N}(\widehat{\mathbf{r}})=\frac{1}{k_s}\sum_{\alpha=1}\left[f_\alpha^N\tilde{\mathbf{m}}_\alpha(\widehat{\mathbf{r}})+\mathrm{j}g_\alpha^N\tilde{\mathbf{n}}_\alpha(\widehat{\mathbf{r}})\right], \tag{17}$$

with

$$\begin{bmatrix}f_\alpha^N\\ g_\alpha^N\end{bmatrix}=\mathrm{j}k_s^2\int_S\left\{\mathbf{e}_N\cdot\begin{bmatrix}\mathbf{N}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\\ \mathbf{M}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\end{bmatrix}\right.$$
$$\left.+\mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\mathbf{h}_N\cdot\begin{bmatrix}\mathbf{M}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\\ \mathbf{N}_{\bar{\alpha}}^{\,1}(k_s\,\cdot)\end{bmatrix}\right\}\mathrm{d}S,\quad\alpha=1,2,\ldots. \tag{18}$$

while, for computational reasons, the series (16) and (17) are approximated by their partial sums. i.e.,

$$\mathbf{E}_{sNM}(\mathbf{r})=\sum_{\alpha=1}\left[f_\alpha^N\mathbf{M}_\alpha^3(k_s\mathbf{r})+g_\alpha^N\mathbf{N}_\alpha^3(k_s\mathbf{r})\right], \tag{19}$$

$$\mathbf{E}_{s\infty NM}(\widehat{\mathbf{r}})=\frac{1}{k_s}\sum_{\alpha=1}\left[f_\alpha^N\tilde{\mathbf{m}}_\alpha(\widehat{\mathbf{r}})+\mathrm{j}g_\alpha^N\tilde{\mathbf{n}}_\alpha(\widehat{\mathbf{r}})\right]. \tag{20}$$

The null-field scheme corresponds to the choice $M=N$, in which case, the above development enabled us to introduce the T matrix of the particle

$$\begin{bmatrix}(f_\alpha^N)_{\alpha=1}^N\\ (g_\alpha^N)_{\alpha=1}^N\end{bmatrix}=\mathrm{T}_N\begin{bmatrix}(a_\alpha)_{\alpha=1}^N\\ (b_\alpha)_{\alpha=1}^N\end{bmatrix},$$

by

$$\mathrm{T}_N=\mathrm{Q}_{11N}(k_s,k_i)[\mathrm{Q}_{31N}(k_s,k_i)]^{-1}\mathrm{Q}_{31N}^0(k_s,k_s).$$

The explicit expressions of the matrices Q_{31N}, Q_{31N}^0, and Q_{11N} are given in Ref. [1]. When localized vector spherical wave functions are used as discrete sources, the null-field scheme coincides with the T-matrix scheme of Waterman.

2. Convergence and numerical stability

Almost all of the studies published on the null-field method are of practical (numerical) nature, and little progress has been made on answering the fundamental questions concerning the convergence of the algorithm. Till now it is not known.

1. to which particle shapes and sizes the method is applicable;

2. which field quantities are convergently approximated, i.e., the far-field pattern, the field outside a circumscribing sphere, the surface fields and
3. what type of convergence can be expected, i.e., pointwise, uniform, or in the least-squares sense.

Denoting by $\mathbf{u} = [\mathbf{e}, \mathbf{h}]^T$ the solution to the infinite set of null-field equations (1)–(3))–(4), where T denotes transpose, and by $\mathbf{u}_N = [\mathbf{e}_N, \mathbf{h}_N]^T$ its approximation given by Eq. (13), a justification of the null-field method requires positive answers to the following questions.

Viability. Is the matrix Q_{31N} *nonsingular?*

Convergence. Does the sequence $\mathbf{u}_N$ *converge to* $\mathbf{u}$ *as* $N \to \infty$*?*

Numerical stability. Does the condition number of the matrix Q_{31N} *have a "sufficiently small" upper bound?*

The numerical stability of the null-field method is a more stringent requirement than the viability of the method. The reason is that a non-singular matrix with a large condition number is difficult to invert numerically, and in this case, the numerical results will strongly depend on the roundoff errors and the errors in the data. The convergence of the surface fields is a very strong result because, in view of the estimates (38) and (39) in Ref. [1], it implies the uniform convergence of the approximate far-field patterns $\mathbf{E}_{s\infty NN}$ on the unit sphere Ω. Actually, the convergence issue is related to the following decision question.

Decision. For particles for which unstable results are obtained, does the algorithm diverge for those shapes or are the numerical instabilities so strong that an approximate solution cannot be accurately computed?

For proving the convergence, it is not sufficient to know that one or more systems of functions are complete and linearly independent; we need to know that such a system possesses the much stronger property of being a basis. It should be pointed out that, as mentioned in Ref. [1], we have to distinguish between the "completeness property" and the "basis property", i.e., between a (convergent) infinite-series expansion and a finite sum of a complete system intended as an approximation.

Only in a few papers dealing with acoustic scattering, some fundamental results about the viability, convergence, and numerical stability have been established.

1. Kristensson et al. [2] (see also, the reorganization of Ramm in Ref. [3]), presented a convergence proof of the surface field through a more general formulation than that of Waterman. Specifically, they considered the general exterior Dirichlet and Neumann radiation problems (not just the scattering problems) and allowed more flexibility in the choices of the trial and test functions. This formulation does specialize to cover the (first) Waterman algorithm when the trial and test functions are constructed from the spherical wave functions. However, the convergence theorem established in Ref. [2] does not apply to this case. The reason is that the hypotheses require at least that the test-function sequence form a basis (in fact, a Riesz basis) for $L^2(S)$ (the Hilbert space of all square integrable scalar functions on S), whereas the spherical wave functions form a basis for $L^2(S)$ only when S is a sphere centered at the pole of the spherical solutions.
2. Dallas [4,5] identified a weaker sense in which the radiating spherical wave functions do form bases, namely, with respect to an inner product that is intimately connected with the far-field patterns of the radiating solutions to the Helmholtz equation. For the exterior Neumann radiation problem, Dallas provided an operator condition that guarantees (i) the viability of the algorithm, (ii) the mean-square convergence of the far-field patterns of the approximations generated from the (second) Waterman scheme on the unit sphere, and (iii) the boundedness of the sequence of matrix condition numbers. Furthermore, he proved that the operator condition holds at least when the scattering obstacle is *ellipsoidal*.

In Appendix 1 we extend the convergence analysis of Kristensson et al. [2] to the electromagnetic scattering by a dielectric particle. The strategy that we follow consists in assuming a basis property, proving the desired convergence result, and then attempting to identify particle geometries (surfaces) for which the basis property (and so, the convergence proof) holds. The conclusions of this analysis can be summarized as follows.

1. If for a given particle geometry, the null-field scheme is numerically stable, then by means of the condition (57) we can decide whether or not the sequence $\mathbf{u}_N$ converges to $\mathbf{u}$ as $N \to \infty$.
2. If for a given particle geometry, the null-field scheme is numerically unstable we cannot say anything about the convergence of the null-field method; the decision question cannot answered. This situation is typical of Fredholm integral equations of the first kind when the operator corresponding to the null-field equations is bounded and injective, but has an unbounded inverse.
3. The null-field scheme is numerically stable if and only if the tangential system of radiating discrete sources forms a Riesz basis.

In practice, the stability and the convergence can be checked numerically.

1. A simple test of the numerical stability involves the computation of the condition number $\kappa(\mathrm{Q}_{31N})$ of the matrix Q_{31N}. If $\kappa(\mathrm{Q}_{31N})$ is bounded with respect to N (the degree of approximation), the null-field method is numerically stable; otherwise, the scheme is numerically unstable. It should be pointed out that, because the calculation of the condition number $\kappa(\mathrm{A})$ of a matrix A is an expensive computational process, we may calculate an upper bound for $\kappa(\mathrm{A})$ as [6]

$$\kappa(\mathrm{A}) < \frac{2}{|\det(\mathrm{A})|}\left(\frac{||\mathrm{A}||_F^2}{2N}\right)^N,$$

where A is a $2N \times 2N$ complex matrix,

$$||\mathrm{A}||_F = \sqrt{\sum_{i,j=1}^{2N} |a_{ij}|^2}$$

is the Frobenius norm of $\mathrm{A} = (a_{ij})$, and the determinant $\det(\mathrm{A})$ can be computed, for example, by means of the LU factorization (when the inverse of Q_{31N} is also computed by means of the LU factorization).

2. As the convergence of the tangential fields appears to be a too strong result, we may define the convergence criterion of the null-field method in terms of the far-field pattern, i.e., the question about the convergence of the null-field method is formulated as

Convergence. Does $\mathbf{E}_{s\infty NN}$ *converge to* $\mathbf{E}_{s\infty}$ *as* $N \to \infty$ *uniformly on* Ω*?*

The far-field pattern is delivered by all T-matrix codes, and so, the convergence of this quantity can be easily checked. If convergence is achieved, then from the result $\lim_{N\to\infty} z_N = z \Rightarrow \lim_{N\to\infty} |z_N|^2 = |z|^2$ for $z_N, z \in \mathbb{C}$, we obtain the convergence of the differential scattering cross section, i.e.,

$$\sigma_N(\hat{\mathbf{r}}) = \left|\mathbf{E}_{s\infty NN}(\hat{\mathbf{r}})\right|^2$$

$$\to \left|\mathbf{E}_{s\infty}(\hat{\mathbf{r}})\right|^2 = \sigma(\hat{\mathbf{r}}) \quad \text{as} \quad N \to \infty \quad \text{uniformly on} \quad \Omega, \tag{21}$$

which in turn implies the convergence of the scattering cross section, i.e.,

$$C_{\mathrm{sct}N} = \frac{1}{|\mathbf{E}_0|^2}\int_\Omega |\mathbf{E}_{s\infty NN}|^2 \, \mathrm{d}^2\hat{\mathbf{r}} \to \frac{1}{|\mathbf{E}_0|^2}\int_\Omega |\mathbf{E}_{s\infty}|^2 \, \mathrm{d}^2\hat{\mathbf{r}} = C_{\mathrm{sct}} \quad \text{as} \quad N \to \infty.$$

A simpler and more pragmatic criterion is to claim that the null-field method converges if the differential scattering cross section converges, i.e., the question about the convergence of the null-field method is

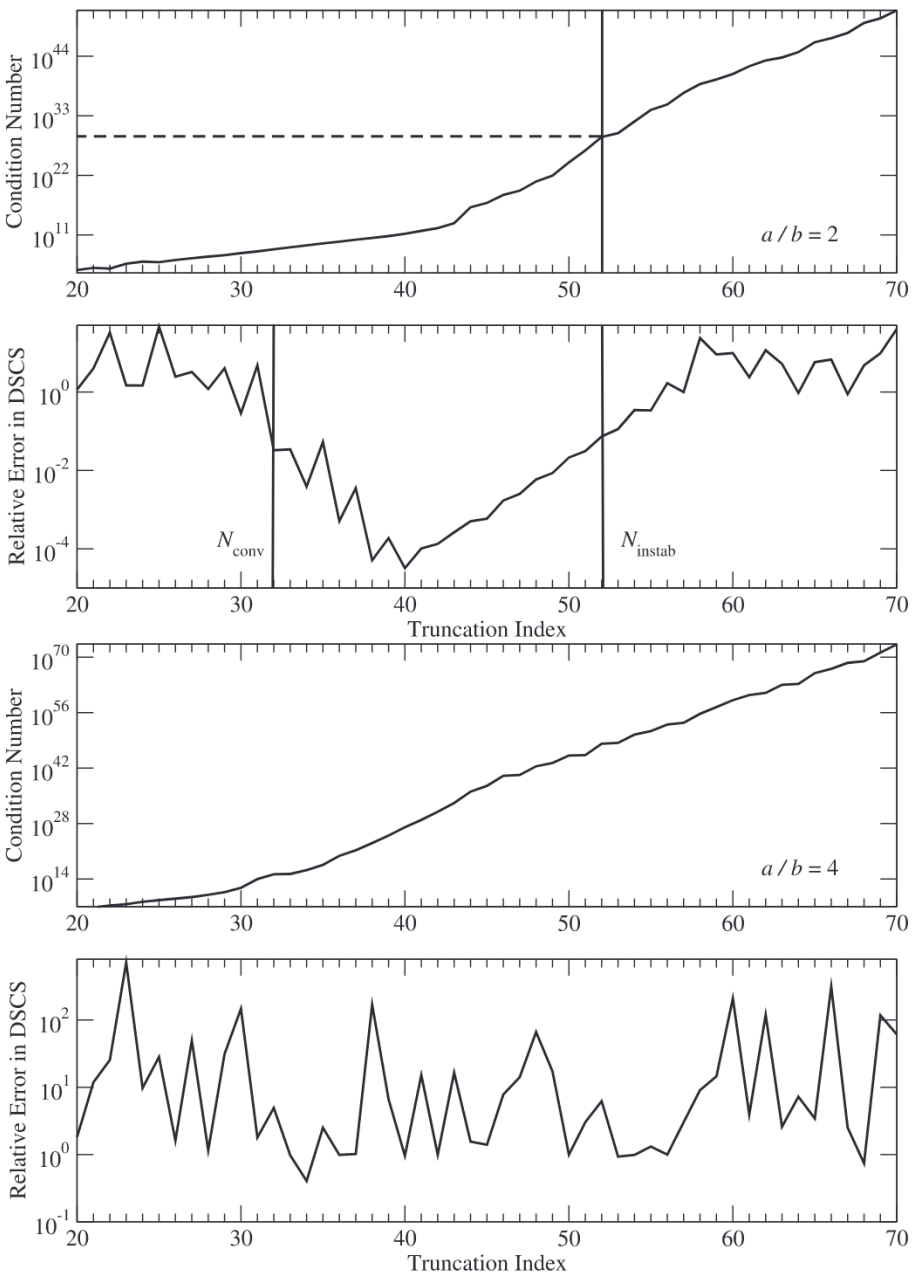

Fig. 1. The condition number of the matrix Q_{31N} and the relative error in the differential scattering cross section σ_N versus the truncation index N for a prolate spheroid with $m_r = 1.5$ and $k_s a = 20$. The results correspond to localized vector spherical wave functions and are computed in double precision. The two plots in the upper panel correspond to $a/b = 2$, where and a and b are the semi-major and the semi-minor axis of the spheroid, respectively, while the two plots in the lower panel correspond to $a/b = 4$.

formulated as

Convergence. Does $\sigma_N(\hat{\mathbf{r}})$ converge to $\sigma(\hat{\mathbf{r}})$ as $N \to \infty$ uniformly on Ω?

A convergence test of the differential scattering cross section is considered in the T-matrix code developed by Barber and Hill [7], i.e., the differential scattering cross section is assumed to converge if it converges within a prescribed tolerance at 10 scattering angles uniformly chosen in $[0, \pi]$.

A comment regarding the convergence of the conventional null-field method can be made here. All numerical experiments performed with the null-field method with localized vector spherical wave functions suggested that for spheroidal particles, the method is numerically unstable, but the far-field pattern converges as long as the numerical instability does not considerably influence the results. In this context, it seems that the conclusion reached by Dallas in the acoustic case is also valid in the electromagnetic case, i.e., the null-field method with localized vector spherical wave functions converges for spheroidal particles. Thus, denoting by N_{conv} the truncation index for which the far-field pattern converges within a prescribed tolerance and by N_{instab} the truncation index after which the numerical instability worsens the results, a reliable approximate solution can be obtained when $N_{\text{conv}} < N_{\text{instab}}$; otherwise, erratic results are obtained. At this time, our efforts to extend the approach of Dallas to the electromagnetic scattering by a dielectric particle have failed, and more work is required to solve this problem. A first step toward this goal is to consider the simpler problem of electromagnetic scattering by a perfectly conducting particle, i.e., the direct electromagnetic scattering boundary-value problem.

In Figs. 1–3 we illustrate the variation of the condition number $\kappa(Q_{31N})$ of the matrix Q_{31N} and the relative error in the differential scattering cross section σ_N with respect to the truncation index N. As discrete sources we consider localized and distributed vector spherical wave functions. The localized vector spherical wave functions are given by

$$\mathfrak{M}_{\alpha}^{q}(k\mathbf{r}) = \mathbf{M}_{mn}^{q}(k\mathbf{r}) = \frac{1}{\sqrt{2\pi n(n+1)}k}\nabla \times \left[u_{mn}^{q}(k\mathbf{r})\mathbf{r}\right], \tag{22}$$

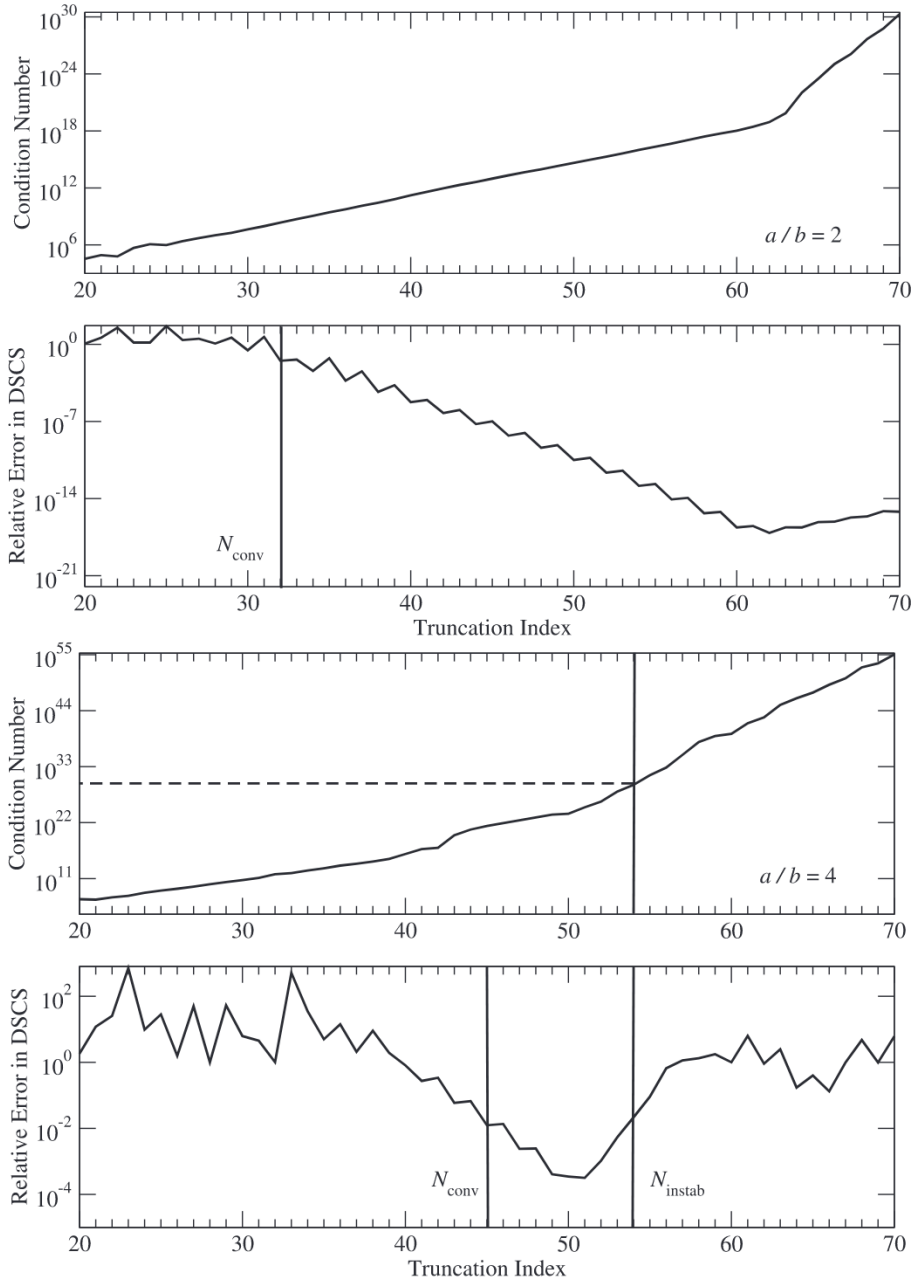

Fig. 2. The same as in Fig. 1, but the results corresponding to localized vector spherical wave functions are computed in extended precision.

$$\mathfrak{N}^q_\alpha(k\mathbf{r}) = \mathbf{N}^q_{mn}(k\mathbf{r}) = \frac{1}{k}\nabla \times \mathbf{M}^q_{mn}(k\mathbf{r}), \quad q = 1, 3, \tag{23}$$

where $\alpha = (m, n)$ and $\overline{\alpha} = (-m, n)$ for $n = 1, 2, \ldots$ and $m = -n, \ldots, n$, while the distributed vector spherical wave functions are given by

$$\mathfrak{M}^q_\alpha(k\mathbf{r}) = \mathbf{M}^q_{m,|m|+l}[k(\mathbf{r} - z_n\widehat{\mathbf{z}})], \tag{24}$$

$$\mathfrak{N}^q_\alpha(k\mathbf{r}) = \mathbf{N}^q_{m,|m|+l}[k(\mathbf{r} - z_n\widehat{\mathbf{z}})], \quad q = 1, 3, \tag{25}$$

where $\{z_n\}_{n=1}^{\infty}$ is a dense set of points situated on a segment $\Gamma_z \subset D_i$ of the z-axis, $\widehat{\mathbf{z}}$ is the unit vector in the direction of the z-axis, $l = 1$ if $m = 0$ and $l = 0$ if $m \neq 0$, $\alpha = (m, n)$ and $\overline{\alpha} = (-m, n)$ for $n = 1, 2, \ldots$ and $m = -n, \ldots, n$. The particle is a prolate spheroid with refractive index $m_r = 1.5$, size parameter $k_s a = 20$, and eccentricity a/b, where a and b are the semi-major and the semi-minor axis of the spheroid, respectively. Two values for the eccentricity are considered, namely $a/b = 2$ and $a/b = 4$. The results are computed with (i) localized vector spherical wave functions in double precision, (ii) localized vector spherical wave functions in extended precision, and (iii) distributed vector spherical wave functions in double precision. We identify N_{conv} with the first value of the truncation index N for which the convergence criterion of Barber and Hill [7] is satisfied, and N_{instab} with the first value of the truncation index N for which this convergence criterion ceases to be satisfied. For prolate spheroids, the following conclusions can be drawn.

1. In the case of localized vector spherical wave functions, the condition number $\kappa(\mathrm{Q}_{31N})$ grows without bound, regardless of the precision used. We conclude that this null-field scheme is numerically unstable, and so, that this system of vector functions does not form a Riesz basis. Convergence is achieved whenever $N_{\text{conv}} < N_{\text{instab}}$, that is, as long as the numerical instability does not have a significant impact on the results. Observe that for double precision calculation and $a/b = 4$,

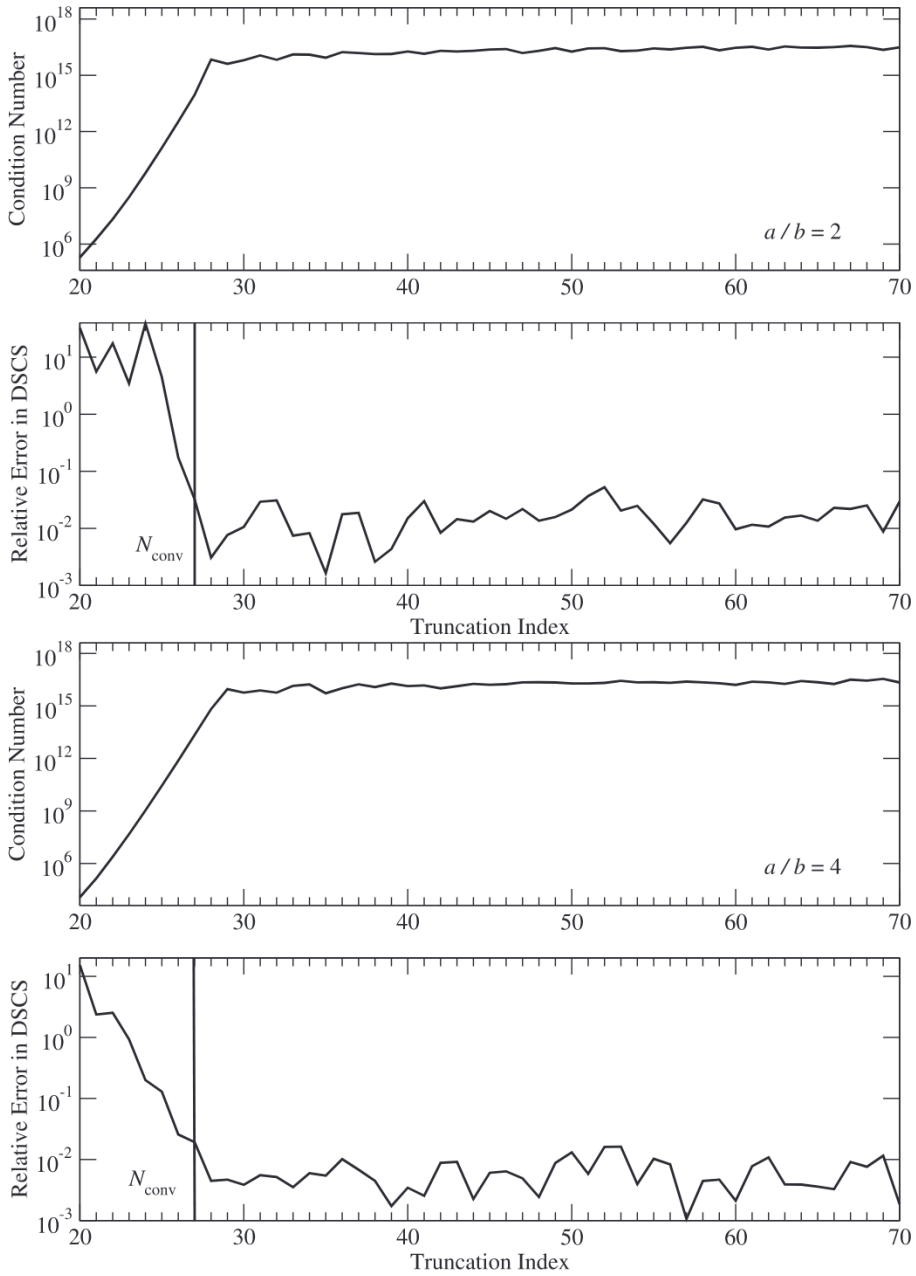

Fig. 3. The same as in Fig. 1, but the results correspond to distributed vector spherical wave functions and are computed in double precision.

no reliable approximate solution can be computed; presumably, the situation $N_{\text{instab}} > N_{\text{conv}}$ occurs.

2. In the case of distributed vector spherical wave functions, the condition number $\kappa(\mathrm{Q}_{31N})$ is significantly smaller than that corresponding to localized sources and increases very slowly with respect to N (especially when the particle eccentricity increases). The numerical instability is not so pronounced, and it seems that from a numerical point of view, this system of discrete sources *almost* possesses the property of being a Riesz basis. Convergence is achieved even for double precision calculation.

   Thus, for this type of scatterers, the superiority of distributed over localized vector spherical wave functions is evident.

As a final comment we mention that the numerical problems associated with the conventional null-field method are due to (i) the numerical instability of the inversion process and (ii) a loss of precision in computing the Q_{31N}-matrix elements. Several results in overcoming these problems deserve to be mentioned.

1. To increase the numerical stability of the matrix inversion, the orthogonalization approach which exploits the unitarity property of the T matrix [8,9], a special form of the LU-factorization method [10], the Gauss elimination method with backsubstitution [11], the block matrix inversion method [12], and the perturbation approach for the Q_{31N}-matrix inversion [13] have been proposed.
2. Somerville et al. [14], following the results established by Waterman in the acoustic case [15], showed that in the case of spheroids, the numerical computation of the integrals of the Q_{31N}-matrix elements may suffer a significant loss of precision due to exact cancellations of large parts of the integrands. The sources of this problematic behavior are some particular terms in the Laurent series expansions of the integrand. Later on, Somerville et al. [16] reformulated the integrals such that these problematic terms are removed, and designed a numerically stable implementation of the null-field method for T-matrix calculation.
3. Petrov et al. [17] developed the so called shape matrix (or Sh-matrix) method as an effective approach for averaging particle ensembles

over their size parameter and refractive index. The Q_{31N}-matrix elements are expressed through analytical relations in terms of the Sh-matrix elements, which depend only on the particle shape. On the other hand, the Sh-matrix elements are determined analytically for many types of particle [18–21], and the resulting analytical solutions speed up the calculations and make them more stable.

4. An increase of the accuracy in computing and inverting the matrix Q_{31N} can be achieved by using extended- and multiple-precision floating-point variables [14,22].

Relying on an ingenious and simplified representation of the Q_{31N}-matrix elements due to Somerville et al. [23], we designed an analytical method for computing these matrix elements. The technical details of this approach are given in Appendix 2**.** As in the case of the shape matrix method, we found that the stability of the resulting analytical solutions is substantially increased.

3. Discussion

Waterman had an extraordinary practical intuition and always preferred physical instead of mathematical arguments. Here are two examples.

1. As a consequence of the completeness of the radiating spherical wave functions, the infinite system of null-field equations is uniquely solvable for all values of the wavenumber. In fact, this was one of the aims in Ref. [24], i.e., to formulate a moment problem which is free of interior-eigenvalue type of instability. However, in order to approximate the total field, Waterman selected sequences of trial functions constructed from the regular spherical wave functions which fail to be complete at the wavenumbers corresponding to the pertinent interior eigenvalues (see for example, Eq. (5)). While the Appendix of Ref. [24] indicates that Waterman was aware of this lack of completeness, he nevertheless retained this choice because the numerical results seemed to indicate that the defect produced no instabilities.
2. The null-field scheme of Waterman can be regarded as Galerkin–Petrov projection scheme, in which the trial and test functions are different. Consider now the infinite set of null-field equations (1)–(4), written in an equivalent form as

$$\int_S \left\{ \left[\widehat{\mathbf{n}}\times(\mathbf{e}-\mathbf{e}_0)\right]\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{M}_{\bar{\alpha}}^{3}(k_s\cdot)\right] + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\left[\widehat{\mathbf{n}}\times(\mathbf{h}-\mathbf{h}_0)\right]\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{N}_{\bar{\alpha}}^{3}(k_s\cdot)\right]\right\} \mathrm{d}S = 0, \tag{26}$$

$$\int_S \left\{ \left[\widehat{\mathbf{n}}\times(\mathbf{e}-\mathbf{e}_0)\right]\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{N}_{\bar{\alpha}}^{3}(k_s\cdot)\right] + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\left[\widehat{\mathbf{n}}\times(\mathbf{h}-\mathbf{h}_0)\right]\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{M}_{\bar{\alpha}}^{3}(k_s\cdot)\right]\right\} \mathrm{d}S = 0 \tag{27}$$

$$\int_S \left\{ (\widehat{\mathbf{n}}\times\mathbf{e})\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{M}_{\bar{\alpha}}^{1}(k_i\cdot)\right] + \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}(\widehat{\mathbf{n}}\times\mathbf{h})\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{N}_{\bar{\alpha}}^{1}(k_i\cdot)\right]\right\} \mathrm{d}S = 0, \tag{28}$$

$$\int_S \left\{ \left(\widehat{\mathbf{n}}\times\mathbf{e}\right)\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{N}_{\bar{\alpha}}^{1}(k_i\cdot)\right] + \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\left(\widehat{\mathbf{n}}\times\mathbf{h}\right)\cdot\left[\widehat{\mathbf{n}}\times\mathfrak{M}_{\bar{\alpha}}^{1}(k_i\cdot)\right]\right\} \mathrm{d}S = 0 \tag{29}$$

2 for $\alpha = 1, 2, \ldots$, and approximate **e** and **h** by the system of tangential vector functions (6), i.e.,

$$\begin{bmatrix}\widehat{\mathbf{n}}\times\mathbf{e}_N\\ \widehat{\mathbf{n}}\times\mathbf{h}_N\end{bmatrix} = \sum_{\beta=1}^{N}\left\{ c_\beta^N \begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^3(k_s\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^3(k_s\cdot)\end{bmatrix} + d_\beta^N \begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^3(k_s\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^3(k_s\cdot)\end{bmatrix} + \tilde{c}_\beta^N \begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\cdot)\end{bmatrix} + \tilde{c}_\beta^N \begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\cdot)\end{bmatrix}\right\}.$$

The resulting null-field scheme, in which the trial and test functions are the same, is the Galerkin projection scheme, and according to the discrete approximation theorem [25], the sequence $\mathbf{u}_N = [\mathbf{e}_N, \mathbf{h}_N]^T$ converges to $\mathbf{u} = [\mathbf{e}, \mathbf{h}]^T$ as $N \to \infty$. It deserves mention that in this case, the approximants $\mathbf{e}_N$ and $\mathbf{h}_N$ have no physical meaning; they cannot be interpreted as the tangential components of the internal electromagnetic field. Moreover, although the scheme is convergent, it is severely numerically unstable. As a result, even for simple particle geometries, an approximate solution cannot be accurately computed (we are in the case $N_{\mathrm{conv}} > N_{\mathrm{instab}}$). On the other hand, the numerical performance of the Waterman scheme, for which we cannot prove convergence, but in which the approximants have a physical meaning, are much better.

In conclusion, Waterman gave us a power tool for analyzing the acoustic and electromagnetic scattering by nonspherical particles, but, unfortunately, after more than 50 years we are still not able to explain mathematically why this method works so well in practice.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Adrian Doicu: Conceptualization, Methodology, Writing - original draft. **Michael I. Mishchenko:** Conceptualization, Methodology, Writing - review & editing.

Acknowledgments

Adrian Doicu acknowledges the financial support from DLR programmatic (S5P KTR 2 472 046) for the S5P algorithm development. Michael I. Mishchenko was supported the NASA Remote Sensing Theory program managed by Lucia Tsaoussi and the NASA Radiation Sciences Program managed by Hal Maring.

Appendix 1

In this appendix we extend the convergence analysis of Kristensson et al. [2] to electromagnetic scattering by a dielectric particle. The definitions of the function spaces that are relevant in our analysis are given in Appendix 1 of Ref. [1], while some basic results from functional analysis are summarized in Appendix 4 of Ref. [1].

Setting $\mathbf{u} = [\mathbf{e}, \mathbf{h}]^T$, we express the infinite set of null-field equations (1)–(2) and (3)–(4) in an operator form as

$$\begin{bmatrix} \mathcal{A}_s \\ \mathcal{A}_i \end{bmatrix} \mathbf{u} = \begin{bmatrix} \mathrm{f} \\ 0 \end{bmatrix}, \tag{30}$$

Here, the operators $\mathcal{A}_s : \mathfrak{T}^2(S) \to \mathbf{l}^2$ and $\mathcal{A}_i : \mathfrak{T}^2(S) \to \mathbf{l}^2$, where $\mathbf{l}^2 = l^2 \times l^2$ and l^2 is the Hilbert space of square-summable sequences, are given by

$$\mathcal{A}_s : \mathbf{u} = \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix} \in \mathfrak{T}^2(S) \to \mathrm{j}k_s^2 \begin{bmatrix} \left(\int_S \left[\mathbf{e} \cdot \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \mathbf{h} \cdot \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \\ \left(\int_S \left[\mathbf{e} \cdot \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \mathbf{h} \cdot \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \end{bmatrix} = -\mathrm{j}k_s^2 \begin{bmatrix} \left(\left\langle \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix}, \overline{\begin{bmatrix} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) \end{bmatrix}} \right\rangle_{2,S} \right)_{\alpha=1}^{\infty} \\ \left(\left\langle \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix}, \overline{\begin{bmatrix} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) \end{bmatrix}} \right\rangle_{2,S} \right)_{\alpha=1}^{\infty} \end{bmatrix} \in \mathbf{l}^2 \tag{31}$$

and

$$\mathcal{A}_i : \mathbf{u} = \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix} \in \mathfrak{T}^2(S) \to \mathrm{j}k_i^2 \begin{bmatrix} \left(\int_S \left[\mathbf{e} \cdot \mathfrak{R}\overline{\mathfrak{a}}^1(k_i \cdot) + \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}} \mathbf{h} \cdot \mathfrak{M}\overline{\mathfrak{a}}^1(k_i \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \\ \left(\int_S \left[\mathbf{e} \cdot \mathfrak{M}\overline{\mathfrak{a}}^1(k_i \cdot) + \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}} \mathbf{h} \cdot \mathfrak{R}\overline{\mathfrak{a}}^1(k_i \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \end{bmatrix} = -\mathrm{j}k_i^2 \begin{bmatrix} \left(\left\langle \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix}, \overline{\begin{bmatrix} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{R}\overline{\mathfrak{a}}^1(k_i \cdot) \\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{M}\overline{\mathfrak{a}}^1(k_i \cdot) \end{bmatrix}} \right\rangle_{2,S} \right)_{\alpha=1}^{\infty} \\ \left(\left\langle \begin{bmatrix} \mathbf{e} \\ \mathbf{h} \end{bmatrix}, \overline{\begin{bmatrix} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{M}\overline{\mathfrak{a}}^1(k_i \cdot) \\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}} \widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathfrak{R}\overline{\mathfrak{a}}^1(k_i \cdot) \end{bmatrix}} \right\rangle_{2,S} \right)_{\alpha=1}^{\infty} \end{bmatrix} \in \mathbf{l}^2, \tag{32}$$

respectively, and

$$f = \mathrm{j}k_s^2 \begin{bmatrix} \left(\int_S \left[\mathbf{e}_0 \cdot \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \mathbf{h}_0 \cdot \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \\ \left(\int_S \left[\mathbf{e}_0 \cdot \mathfrak{M}\overline{\mathfrak{a}}^3(k_s \cdot) + \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}} \mathbf{h}_0 \cdot \mathfrak{R}\overline{\mathfrak{a}}^3(k_s \cdot) \right] \mathrm{d}S \right)_{\alpha=1}^{\infty} \end{bmatrix}, \tag{33}$$

where the bar notation means complete conjugate. Note that the last relations in Eqs. (31) and (32) follow from the identity $\mathbf{a} = -\widehat{\mathbf{n}} \times \widehat{\mathbf{n}} \times \mathbf{a}$ for $\mathbf{a} \in \mathfrak{T}^2(S)$.

Consider the null-space of the operator $\mathcal{A}_i$, i.e.,

$$\tilde{\mathfrak{T}}^2(S) = \mathcal{N}(\mathcal{A}_i) = \left\{ \mathbf{v} \in \mathfrak{T}^2(S) \;\middle|\; \mathcal{A}_i \mathbf{v} = 0 \right\},$$

which, under the assumption that $\mathcal{A}_i$ is bounded, is a closed subspace of $\mathfrak{T}^2(S)$, and so, a Hilbert space with the induced scalar product.

Assertion 1 *The system of tangential vector functions*

$$\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^3(k_s\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\,\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^3(k_s\,\cdot)\end{bmatrix},\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^3(k_s\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\,\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^3(k_s\,\cdot)\end{bmatrix}\right\}_{\alpha=1}^{\infty} \tag{34}$$

is complete in $\tilde{\mathfrak{T}}^2(S)$.

Proof. The second relation in Eq. (32) shows that for all $\tilde{\mathbf{v}}=[\mathbf{e},\mathbf{h}]^T\in\tilde{\mathfrak{T}}^2(S)$, we have

$$\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix},\overline{\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^1(k_i\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^1(k_i\,\cdot)\end{bmatrix}}\right\rangle_{2,S}=0, \tag{35}$$

$$\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix},\overline{\begin{bmatrix}\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{M}_\alpha^1(k_i\,\cdot)\\ \mathrm{j}\sqrt{\frac{\mu_i}{\varepsilon_i}}\,\widehat{\mathbf{n}}\times\widehat{\mathbf{n}}\times\mathfrak{N}_\alpha^1(k_i\,\cdot)\end{bmatrix}}\right\rangle_{2,S}=0,\quad\alpha=1,2,\ldots. \tag{36}$$

These equations and the fact the system of tangential vector functions (6) is complete in $\mathfrak{T}^2(S)$, imply that the system of tangential vector functions (34) is complete in $\tilde{\mathfrak{T}}^2(S)$. Indeed, for $\tilde{\mathbf{v}}\in\tilde{\mathfrak{T}}^2(S)$, Eqs. (35) and (36) and the closure relations for the system of tangential vector functions (34) coincide with the closure relations for the system of tangential vector functions (6); hence $\tilde{\mathbf{v}}=0$. □

Consider now the subspace.

$$\mathfrak{T}_N^2(S)=T_N^2(S)\times T_N^2(S)$$

$$=\mathrm{span}\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix},\right.$$

$$\left.\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix},\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix}\right\}_{\beta=1}^{N}. \tag{37}$$

The completeness of the system of tangential vector functions (5) implies that the sequence of subspaces $\mathfrak{T}_N^2(S)$ is limit dense in $\mathfrak{T}^2(S)$, that is,

$$\left|\left|\mathbf{v}-P_N\mathbf{v}\right|\right|_{2,S}\to 0\quad\text{as}\quad N\to\infty, \tag{38}$$

where $P_N\mathbf{v}\in\mathfrak{T}_N^2(S)$ is the orthogonal projection of $\mathbf{v}\in\mathfrak{T}^2(S)$ onto $\mathfrak{T}_N^2(S)$. Defining

$$\tilde{\mathfrak{T}}_N^2(S)=\mathrm{span}\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix},\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}\right\}_{\beta=1}^{N} \tag{39}$$

and, say,

$$\widehat{\mathfrak{T}}_N^2(S)=\mathrm{span}\left\{\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ \mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix},\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ \mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}\right\}_{\beta=1}^{N}, \tag{40}$$

and taking into account that the tangential vector functions generating $\tilde{\mathfrak{T}}_N^2(S)$ and $\widehat{\mathfrak{T}}_N^2(S)$ are linear combinations of the tangential vector functions generating $\mathfrak{T}_N^2(S)$, i.e.,

$$\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ \pm\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix}=\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix}\pm\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix},$$

$$\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ \pm\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}=\begin{bmatrix}\widehat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ \mathbf{0}\end{bmatrix}\pm\begin{bmatrix}\mathbf{0}\\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\,\widehat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix},$$

we find

$$\mathfrak{T}_N^2(S) = \tilde{\mathfrak{T}}_N^2(S) \oplus \widehat{\mathfrak{T}}_N^2(S). \tag{41}$$

The subspace $\tilde{\mathfrak{T}}_N^2(S)$ can be characterized as follows.

1. From the orthogonality relation (12) and the definition of the operator $\mathcal{A}_i$ as given by the first relation in Eq. (32), we see that

$$\mathcal{A}_i \begin{bmatrix} \widehat{\mathbf{n}} \times \mathfrak{M}_\beta^1(k_i \cdot) \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\, \widehat{\mathbf{n}} \times \mathfrak{N}_\beta^1(k_i \cdot) \end{bmatrix} = 0 \quad \text{and} \quad \mathcal{A}_i \begin{bmatrix} \widehat{\mathbf{n}} \times \mathfrak{N}_\beta^1(k_i \cdot) \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\, \widehat{\mathbf{n}} \times \mathfrak{M}_\beta^1(k_i \cdot) \end{bmatrix} = 0$$

for any $\beta = 1, 2, \ldots$. Therefore, $\tilde{\mathfrak{T}}_N^2(S) \subset \tilde{\mathfrak{T}}^2(S)$, *i.e., for any* $\tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S)$, *we have* $\mathcal{A}_i \tilde{\mathbf{v}}_N = 0$.

2. Let Q_N be the orthogonal projection operator in $\mathbf{l}^2$, defined by the formula

$$Q_N \mathrm{f} = \mathrm{f}_N = [\tilde{a}_1, \ldots, \tilde{a}_N, \tilde{b}_1, \ldots, \tilde{b}_N]^T \in \mathbf{l}_N^2,$$

where $\mathbf{l}_N^2 = Q_N \mathbf{l}^2$ and

$$\mathrm{f} = [\tilde{a}_1, \ldots, \tilde{a}_N, \tilde{a}_{N+1}, \ldots, \tilde{b}_1, \ldots, \tilde{b}_N, \tilde{b}_{N+1}, \ldots]^T.$$

Obviously, the equation $\mathcal{A}_i \tilde{\mathbf{v}}_N = 0$ implies $Q_N \mathcal{A}_i \tilde{\mathbf{v}}_N = 0$ for any $\tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S)$. Thus, $\tilde{\mathfrak{T}}_N^2(S)$ *is the null-space of the operator* $Q_N \mathcal{A}_i$, i.e.,

$$\tilde{\mathfrak{T}}_N^2(S) = \mathcal{N}(Q_N \mathcal{A}_i) = \left\{ \tilde{\mathbf{v}}_N \in \mathfrak{T}_N^2(S) \mid Q_N \mathcal{A}_i \tilde{\mathbf{v}}_N = 0 \right\}$$

The next result states that the sequence of subspaces $\tilde{\mathfrak{T}}_N^2(S)$ is limit dense in $\tilde{\mathfrak{T}}^2(S)$.

Assertion 2 *Assume that*

(1) *the operator* $\mathcal{A}_i$ *is bounded in* $\mathfrak{T}^2(S)$, *and*
(2) *there exists* $c_{1i} > 0$ *such that*

$$Q_N \mathcal{A}_i \widehat{\mathbf{v}}_{N2} \geq c_{1i} \left|\left| \widehat{\mathbf{v}}_N \right|\right|_{2,S} \tag{42}$$

for all $\widehat{\mathbf{v}}_N \in \widehat{\mathfrak{T}}_N^2(S)$ *and all N.*

Then, the sequence of subspaces $\tilde{\mathfrak{T}}_N^2(S)$ *is limit dense in* $\tilde{\mathfrak{T}}^2(S)$.

Proof. By assumption, the operator $\mathcal{A}_i$ is bounded in $\mathfrak{T}^2(S)$, and therefore, there exists $c_{2i} \geq 0$ such that

$$\left|\left| \mathcal{A}_i \mathbf{v} \right|\right|_2 \leq c_{i2} \left|\left| \mathbf{v} \right|\right|_{2,S} \quad \text{for all} \quad \mathbf{v} \in \mathfrak{T}^2(S). \tag{43}$$

Take some $\tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S) \subset \mathfrak{T}^2(S)$ and let P_N be the orthogonal projection operator from $\mathfrak{T}^2(S)$ onto $\mathfrak{T}_N^2(S)$. Because $\mathfrak{T}_N^2(S)$ is limit dense in $\mathfrak{T}^2(S)$, we have

$$\left|\left| \tilde{\mathbf{v}} - P_N \tilde{\mathbf{v}} \right|\right|_{2,S} \to 0 \quad \text{as} \quad N \to \infty. \tag{44}$$

From Eq. (41), we see that $P_N \tilde{\mathbf{v}} \in \mathfrak{T}_N^2(S)$ can be written as

$$P_N \tilde{\mathbf{v}} = \tilde{\mathbf{v}}_N + \widehat{\mathbf{v}}_N \tag{45}$$

with $\tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S) = \mathcal{N}(Q_N \mathcal{A}_i)$ and $\widehat{\mathbf{v}}_N \in \widehat{\mathfrak{T}}_N^2(S)$. Then, using the relations

$$Q_N \mathcal{A}_i \tilde{\mathbf{v}}_N = 0, \tag{46}$$

$$\mathcal{A}_i \tilde{\mathbf{v}} = 0, \tag{47}$$

we get

$$
\begin{aligned}
||\widehat{\mathbf{v}}_N||_{2,S} &\overset{(42)}{\leq} \frac{1}{c_{1i}}||Q_N\mathcal{A}_i\widehat{\mathbf{v}}_N||_2 \\
&\overset{(45)}{=} \frac{1}{c_{1i}}||Q_N\mathcal{A}_i\left(P_N\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right)||_2 \\
&\overset{(46)}{=} \frac{1}{c_{1i}}||Q_N\mathcal{A}_iP_N\tilde{\mathbf{v}}||_2 \\
&\leq \frac{1}{c_{1i}}||\mathcal{A}_iP_N\tilde{\mathbf{v}}||_2 \\
&\overset{(47)}{=} \frac{1}{c_{1i}}||\mathcal{A}_i(P_N\tilde{\mathbf{v}}-\tilde{\mathbf{v}})||_2 \\
&\overset{(43)}{\leq} \frac{c_{2i}}{c_{1i}}||P_N\tilde{\mathbf{v}}-\tilde{\mathbf{v}}||_{2,S}.
\end{aligned}
\tag{48}
$$

Finally, letting $\tilde{P}_N$ be the orthogonal projection operator from $\tilde{\mathfrak{T}}^2(S)$ onto $\tilde{\mathfrak{T}}^2_N(S)$, and taking into account that

$$
||\tilde{\mathbf{v}}-\tilde{P}_N\tilde{\mathbf{v}}||_{2,S} \leq ||\tilde{\mathbf{v}}-\tilde{\mathbf{w}}_N||_{2,S} \quad \text{for any} \quad \tilde{\mathbf{w}}_N \in \tilde{\mathfrak{T}}^2_N(S),
\tag{49}
$$

we obtain

$$
\begin{aligned}
&\left|\left|\tilde{\mathbf{v}}-\tilde{P}_N\tilde{\mathbf{v}}\right|\right|_{2,S} \\
&\leq \left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S} \\
&\leq \left|\left|\tilde{\mathbf{v}}-(\tilde{\mathbf{v}}_N+\widehat{\mathbf{v}}_N)\right|\right|_{2,S}+\left|\left|\widehat{\mathbf{v}}_N\right|\right|_{2,S} \\
&= \left|\left|\tilde{\mathbf{v}}-P_N\tilde{\mathbf{v}}\right|\right|_{2,S}+\left|\left|\widehat{\mathbf{v}}_N\right|\right|_{2,S} \\
&\overset{(48)}{\leq} \left(1+\frac{c_{2i}}{c_{1i}}\right)\left|\left|P_N\tilde{\mathbf{v}}-\tilde{\mathbf{v}}\right|\right|_{2,S} \\
&\overset{(44)}{\to} 0 \quad \text{as} \quad N\to\infty,
\end{aligned}
$$

and the assertion is proved. □

Let $\tilde{\mathcal{A}}_s$ be the restriction of $\mathcal{A}_s$ to $\tilde{\mathfrak{T}}^2(S)$, regarded as

$$
\tilde{\mathcal{A}}_s := \mathcal{A}_s|_{\tilde{\mathfrak{T}}^2(S)} : \tilde{\mathfrak{T}}^2(S) \to \mathcal{R}(\tilde{\mathcal{A}}_s),
$$

where

$$
\mathcal{R}(\tilde{\mathcal{A}}_s) = \left\{\mathrm{f}\in\mathbf{l}^2 \middle| \mathrm{f}=\mathcal{A}_s\tilde{\mathbf{v}} \quad \text{for some} \quad \tilde{\mathbf{v}}\in\tilde{\mathfrak{T}}^2(S)\right\}
$$

is the range of $\tilde{\mathcal{A}}_s$. In this case, the solution of the operator equation (30) is equivalent with the solution of the operator equation

$$
\tilde{\mathcal{A}}_s\tilde{\mathbf{u}} = \mathrm{f}.
\tag{50}
$$

According to Hadamard, the operator equation (50) is called well-posed provided that.

1. for any $\mathrm{f}\in\mathcal{R}(\tilde{\mathcal{A}}_s)$, a solution $\tilde{\mathbf{u}}$ exists, i.e., the operator $\tilde{\mathcal{A}}_s$ is surjective;
2. the solution $\tilde{\mathbf{u}}$ is unique, i.e., the operator $\tilde{\mathcal{A}}_s$ is injective; and
3. the solution is stable with respect to perturbations in f, in the sense that if $\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}_0=\mathrm{f}_0$ and $\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}=\mathrm{f}$, then $\tilde{\mathbf{u}}\to\tilde{\mathbf{u}}_0$ whenever $\mathrm{f}\to\mathrm{f}_0$, i.e., the inverse operator $\tilde{\mathcal{A}}_s^{-1}$ is bounded.

If one of the Hadamard conditions is violated, the problem is said to be ill-posed. In our case, we know from the second relation in Eq. (31) and the completeness of the system of tangential vector functions (34) in $\tilde{\mathfrak{T}}^2(S)$ that the operator $\tilde{\mathcal{A}}_s$ is injective.

In practice, the third Hadamard condition has a significant importance; the violation of this condition creates serious numerical problems because small errors in the data can be dramatically amplified in the solution. The stability of the solution can be verified by using the Bounded Inverse Theorem,

according to which, the bounded operator $\tilde{\mathcal{A}}_s$ has a bounded inverse $\tilde{\mathcal{A}}_s^{-1}$ on its range $\mathcal{R}(\tilde{\mathcal{A}}_s)$ if and only one of the following assumptions is satisfied:

1. $\tilde{\mathcal{A}}_s$ is bijective;
2. $\tilde{\mathcal{A}}_s$ is injective and has a closed range $\mathcal{R}(\tilde{\mathcal{A}}_s)$; and
3. $\tilde{\mathcal{A}}_s$ is bounded from below, i.e., there exists $c_{1s} > 0$ such that

$$c_{1s}||\tilde{\mathbf{v}}||_{2,S} \leq ||\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}||_2 \quad \text{for all} \quad \tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S) \tag{51}$$

More precisely, we have the following result.

Assertion 3 *The operators $\tilde{\mathcal{A}}_s$ and $\tilde{\mathcal{A}}_s^{-1}$ are both bounded if and only if the complete system of tangential vector functions (34) is a Riesz basis of $\tilde{\mathfrak{T}}^2(S)$.*

Proof. Before proceeding we note that by Assertion 1, the system of tangential vector functions (34) is complete in $\tilde{\mathfrak{T}}^2(S)$. According to the Bounded Inverse Theorem, the operators $\tilde{\mathcal{A}}_s$ and $\tilde{\mathcal{A}}_s^{-1}$ are both bounded if and only there exist $c_{1s}, c_{2s} > 0$ such that (cf. Eq. (51))

$$c_{1s}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S} \leq \left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2 \leq c_{2s}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S} \quad \text{for all} \quad \tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S). \tag{52}$$

Taking into account that for $\tilde{\mathbf{v}} = [\mathbf{e}, \mathbf{h}]^T$, we have (for real k_s)

$$\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_{2,S}^2 = k_s^4 \sum_{\alpha=1}^{\infty}\left\{\left|\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix}, \overline{\begin{bmatrix}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{N}_{\tilde{\alpha}}^3(k_s\,\cdot)\\ \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{M}_{\tilde{\alpha}}^3(k_s\,\cdot)\end{bmatrix}}\right\rangle_{2,S}\right|^2 + \left|\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix}, \overline{\begin{bmatrix}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{M}_{\tilde{\alpha}}^3(k_s\,\cdot)\\ \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{N}_{\tilde{\alpha}}^3(k_s\,\cdot)\end{bmatrix}}\right\rangle_{2,S}\right|^2\right\}, \tag{53}$$

the conditions (52) become

$$c_{1s}^{'2}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S}^2 \leq \sum_{\alpha=1}^{\infty}\left\{\left|\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix}, \overline{\begin{bmatrix}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{N}_{\tilde{\alpha}}^3(k_s\,\cdot)\\ \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{M}_{\tilde{\alpha}}^3(k_s\,\cdot)\end{bmatrix}}\right\rangle_{2,S}\right|^2 + \left|\left\langle\begin{bmatrix}\mathbf{e}\\ \mathbf{h}\end{bmatrix}, \overline{\begin{bmatrix}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{M}_{\tilde{\alpha}}^3(k_s\,\cdot)\\ \\ \mathrm{j}\sqrt{\frac{\mu_s}{\varepsilon_s}}\hat{\mathbf{n}}\times\hat{\mathbf{n}}\times\mathfrak{N}_{\tilde{\alpha}}^3(k_s\,\cdot)\end{bmatrix}}\right\rangle_{2,S}\right|^2\right\} \tag{54}$$

$$\leq c_{2s}^{'2}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S}^2 \quad \text{for all} \quad \tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2\left(S\right),$$

where $c_{1,2s}^{'} = c_{1,2}/k_s^2$. These inequalities hold if and only if the system of tangential vector functions (34) is a Riesz basis of $\tilde{\mathfrak{T}}^2(S)$. □

Thus, if the system of tangential vector functions (34) is a Riesz basis of $\tilde{\mathfrak{T}}^2(S)$ then the operators $\tilde{\mathcal{A}}_s$ and $\tilde{\mathcal{A}}_s^{-1}$ are both bounded, and so, in view of the Bounded Inverse Theorem, $\tilde{\mathcal{A}}_s$ is bijective. As a result, the Hadamard conditions are satisfied, and the operator equation (50) is well-posed. If this is not the case, $\tilde{\mathcal{A}}_s^{-1}$ is unbounded, and the solution to the operator equation (50) is not stable.

We come now to the null-field scheme for the operator equation (50). This consists in the computation of the coefficients $\{c_\beta^N, d_\beta^N\}_{\beta=1}^N$ from the projection equation

$$Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}_N = Q_N\mathsf{f}, \tag{55}$$

where $\tilde{\mathbf{u}}_N \in \tilde{\mathfrak{T}}_N^2(S)$ is given by

$$\tilde{\mathbf{u}}_N = \begin{bmatrix}\mathbf{e}_N\\ \mathbf{h}_N\end{bmatrix} = \sum_{\beta=1}^{N}\left\{c_\beta^N\begin{bmatrix}\hat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\\ \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\hat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\end{bmatrix} + d_\beta^N\begin{bmatrix}\hat{\mathbf{n}}\times\mathfrak{N}_\beta^1(k_i\,\cdot)\\ \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\hat{\mathbf{n}}\times\mathfrak{M}_\beta^1(k_i\,\cdot)\end{bmatrix}\right\} \tag{56}$$

The explicit representation of Eq. (55) is (cf. Eq. (15))

$$\mathrm{Q}_{31N}(k_s, k_i)\begin{bmatrix}(c_\beta^N)_{\beta=1}^N\\ (d_\beta^N)_{\beta=1}^N\end{bmatrix} = \begin{bmatrix}(\tilde{a}_\alpha)_{\alpha=1}^N\\ \\ (\tilde{b}_\alpha)_{\alpha=1}^N\end{bmatrix},$$

where

$$\begin{bmatrix} (\tilde{a}_\alpha)_{\alpha=1}^N \\ (\tilde{b}_\alpha)_{\alpha=1}^N \end{bmatrix} = \mathrm{Q}_{31N}^0(k_s,k_s) \begin{bmatrix} (a_\alpha)_{\alpha=1}^N \\ (b_\alpha)_{\alpha=1}^N \end{bmatrix} \in \mathbf{l}^2.$$

A justification of the null-field method requires positive answers to the following questions.

Viability. Is the matrix Q_{31N} *nonsingular?*

Convergence. Does the sequence $\mathbf{u}_N$ *converge to* $\mathbf{u}$ *as* $N \to \infty$?

Numerical stability. Does the condition number of the matrix Q_{31N} *have a "sufficiently small" upper bound?*

Because a nonsingular matrix with a large condition number is difficult to invert numerically, it is apparent that the numerical stability of the null-field method is a more stringent requirement than the viability of the method. With respect to the numerical stability we note that if the solution to the operator equation (50) is not stable, i.e., $\tilde{\mathcal{A}}_s^{-1}$ is unbounded, then the null-field scheme is not numerically stable, i.e., the condition number of the matrix Q_{31N} increases without bound when N increases.

In the following we consider the convergence issue, and in order to simplify the analysis, we assume that the matrix Q_{31N} is nonsingular and that Assertion 2 is satisfied, i.e., the sequence of subspaces $\tilde{\mathfrak{T}}_N^2(S)$ is limit dense in $\tilde{\mathfrak{T}}^2(S)$. The next result gives *sufficient conditions* for the convergence of the null-field method.

Assertion 4 *Let* $\tilde{\mathbf{u}}$ *and* $\tilde{\mathbf{u}}_N$ *be the unique solutions of Eqs.* (50) *and* (55)*, respectively. Assume that*

(1) *the operator* $\tilde{\mathcal{A}}_s$ *is bounded in* $\tilde{\mathfrak{T}}^2(S)$*,and*

(2) *there exists* $c_{1s} > 0$ *such that*

$$||Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N||_2 \geq c_{1s}\left|\left|\tilde{\mathbf{v}}_N\right|\right|_{2,S}, \tag{57}$$

for all $\tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S)$ *and all N.*

Then,

$$\left|\left|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_N\right|\right|_{2,S} \to 0 \quad \text{as} \quad N \to \infty. \tag{58}$$

If moreover.

(3) *the condition*

$$\lim_{M\to\infty} \mathbf{E}_{s\infty NM}(\hat{\mathbf{r}}) = \mathbf{E}_{s\infty N}(\hat{\mathbf{r}}) \quad \textit{is uniformly in } N \textit{ and } \hat{\mathbf{r}} \in \Omega \tag{59}$$

satisfied

then

$$\lim_{N\to\infty} \mathbf{E}_{s\infty NN}(\hat{\mathbf{r}}) = \mathbf{E}_{s\infty}(\hat{\mathbf{r}}) \textit{ uniformly on } \Omega, \tag{60}$$

and

$$\lim_{N\to\infty} f_\alpha^N = f_\alpha, \quad \lim_{N\to\infty} g_\alpha^N = g_\alpha \quad \text{uniformly in } \alpha, \tag{61}$$

where $\mathbf{E}_{s\infty}(\hat{\mathbf{r}})$*,* $\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})$*, and* $\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})$ *are given by Eq. (8), (17), and (20), respectively.*

Proof. Let $\tilde{\mathbf{u}} = [\mathbf{e},\mathbf{h}]^T \in \tilde{\mathfrak{T}}^2(S)$ and $\tilde{\mathbf{u}}_N = [\mathbf{e}_N,\mathbf{h}_N]^T \in \tilde{\mathfrak{T}}_N^2(S)$ be the unique solutions of the equations (50) and (55), respectively. The proof is organized as follows.

1. First, we prove the convergence result (58). The boundedness of the operator $\tilde{\mathcal{A}}_s$, i.e.,

$$\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2 \leq c_{2s}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S} \quad \text{for all} \quad \tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S), \tag{62}$$

implies

$$\left|\left|Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2 \leq \left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2 \leq c_{2s}\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S}^2 \quad \text{for all} \quad \tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S) \tag{63}$$

Hence, from Eqs. (57) and (63), we obtain

$$c_{1s}||\tilde{\mathbf{v}}_N||_{2,S} \leq ||Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N||_2 \leq c_{2s}||\tilde{\mathbf{v}}_N||_{2,S} \quad \text{for all} \quad \tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2. \tag{64}$$

Because by assumption, the sequence of subspaces $\tilde{\mathfrak{T}}_N^2(S)$ is limit dense in $\tilde{\mathfrak{T}}^2(S)$, we have, for any $\tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S)$,

$$\left|\left|\tilde{\mathbf{v}} - \tilde{P}_N\tilde{\mathbf{v}}\right|\right|_{2,S} \to 0 \quad \text{as} \quad N \to \infty, \tag{65}$$

where $\tilde{P}_N\tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}_N^2(S)$ is the orthogonal projection of $\tilde{\mathbf{v}} \in \tilde{\mathfrak{T}}^2(S)$ onto $\tilde{\mathfrak{T}}_N^2(S)$. The projection of Eq. (50) onto $\mathfrak{l}_N^2$, i.e.,

$$Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{u}} = Q_N\mathsf{f}, \tag{66}$$

yields

$$Q_N\tilde{\mathcal{A}}_s[\tilde{P}_N\tilde{\mathbf{u}} + (I - \tilde{P}_N)\tilde{\mathbf{u}}] = Q_N\mathsf{f}. \tag{67}$$

From Eqs. (55) and (67), we obtain

$$Q_N\tilde{\mathcal{A}}_s[\tilde{P}_N\tilde{\mathbf{u}} + (I - \tilde{P}_N)\tilde{\mathbf{u}}] = Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}_N, \tag{68}$$

and further,

$$Q_N\tilde{\mathcal{A}}_s(\tilde{\mathbf{u}}_N - \tilde{P}_N\tilde{\mathbf{u}}) = Q_N\tilde{\mathcal{A}}_s(I - \tilde{P}_N)\tilde{\mathbf{u}}. \tag{69}$$

Then, taking into account that $\tilde{\mathbf{u}}_N - \tilde{P}_N\tilde{\mathbf{u}} \in \tilde{\mathfrak{T}}_N^2(S)$, we get

$$\begin{aligned}
\left|\left|\tilde{\mathbf{u}}_N - \tilde{P}_N\tilde{\mathbf{u}}\right|\right|_{2,S} &\le \frac{1}{c_{1s}}||Q_N\tilde{\mathcal{A}}_s(\tilde{\mathbf{u}}_N - \tilde{P}_N\tilde{\mathbf{u}})||_2 \\
&\overset{(69)}{=} \frac{1}{c_{1s}}||Q_N\tilde{\mathcal{A}}_s(I - \tilde{P}_N)\tilde{\mathbf{u}}||_2 \\
&\le \frac{c_{2s}}{c_{1s}}\left|\left|\tilde{\mathbf{u}} - \tilde{P}_N\tilde{\mathbf{u}}\right|\right|_{2,S} \\
&\overset{(65)}{\to} 0 \quad \text{as} \quad N \to \infty,
\end{aligned} \tag{70}$$

and consequently,

$$\begin{aligned}
\left|\left|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_N\right|\right|_{2,S} &\le \left|\left|\tilde{\mathbf{u}} - \tilde{P}_N\tilde{\mathbf{u}}\right|\right|_{2,S} + \left|\left|\tilde{\mathbf{u}}_N - \tilde{P}_N\tilde{\mathbf{u}}\right|\right|_{2,S} \\
&\overset{(65),\ (70)}{\to} 0 \quad \text{as} \quad N \to \infty.
\end{aligned} \tag{71}$$

2. We prove now the uniform convergence of $\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})$ to $\mathbf{E}_{s\infty}(\hat{\mathbf{r}})$ as $N \to \infty$. Employing the same arguments as in the derivation of the estimate (130) in Appendix 3 of Ref. [1], we find

$$\left|\mathbf{E}_{s\infty N}(\hat{\mathbf{r}}) - \mathbf{E}_{s\infty}(\hat{\mathbf{r}})\right| \le C_{e\infty}\left(\left|\left|\mathbf{e}_N - \mathbf{e}\right|\right|_{2,S} + \left|\left|\mathbf{h}_N - \mathbf{h}\right|\right|_{2,S}\right), \tag{72}$$

for all $\hat{\mathbf{r}} \in \Omega$, where $\mathbf{E}_{s\infty}(\hat{\mathbf{r}})$ and $\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})$, are given by Eq. (8) and

$$\mathbf{E}_{s\infty N}(\hat{\mathbf{r}}) = \frac{\mathrm{j}k_s}{4\pi}\int_S \left\{\hat{\mathbf{r}} \times \mathbf{e}_N(\mathbf{r}') + \sqrt{\frac{\mu_s}{\varepsilon_s}}\hat{\mathbf{r}} \times [\mathbf{h}_N(\mathbf{r}') \times \hat{\mathbf{r}}]\right\} \mathrm{e}^{-\mathrm{j}k_s\hat{\mathbf{r}}\cdot\mathbf{r}'}\, \mathrm{d}S(\mathbf{r}'), \tag{73}$$

respectively. From the convergence result (cf. Eq. (71))

$$\left|\left|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_N\right|\right|_{2,S}^2 = \left|\left|\mathbf{e}_N - \mathbf{e}\right|\right|_{2,S}^2 + \left|\left|\mathbf{h}_N - \mathbf{h}\right|\right|_{2,S}^2 \to 0 \quad \text{as} \quad N \to \infty,$$

we then get

$$\lim_{N\to\infty} \mathbf{E}_{s\infty N}(\hat{\mathbf{r}}) = \mathbf{E}_{s\infty}(\hat{\mathbf{r}}) \quad \text{uniformly} \quad \text{on} \quad \Omega. \tag{74}$$

Since $\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})$ possesses the series representation (17) and $\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})$, given by Eq. (20), is the partial sum of this series, it follows that

$$\lim_{M\to\infty}\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})=\mathbf{E}_{s\infty N}(\hat{\mathbf{r}}) \quad \text{uniformly on } \Omega. \tag{75}$$

Now, for $\hat{\mathbf{r}}\in\Omega$, the assumption (59), which is stronger than the convergence result (75), implies that for $\varepsilon>0$, there exists $M_0=M_0(\varepsilon)$ such that for $M\geq M_0$, we have

$$|\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})-\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})|<\frac{\varepsilon}{2} \quad \text{for all } N.$$

For the same $\varepsilon>0$, there exists $N_0=N_0(\varepsilon)$ such that for $N\geq N_0$, we have (cf. Eq. (74))

$$|\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})-\mathbf{E}_{s\infty}(\hat{\mathbf{r}})|<\frac{\varepsilon}{2}.$$

Hence, for $N\geq\max(N_0,M_0)$ and $M\geq\max(N_0,M_0)$, we find.

$$|\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})-\mathbf{E}_{s\infty}(\hat{\mathbf{r}})|\leq|\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})-\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})|$$

$$+|\mathbf{E}_{s\infty N}(\hat{\mathbf{r}})-\mathbf{E}_{s\infty}(\hat{\mathbf{r}})|$$

$$=\varepsilon.$$

This means that the double sequence $\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})$ converges to $\mathbf{E}_{s\infty}(\hat{\mathbf{r}})$, i.e.,

$$\lim_{N,M\to\infty}\mathbf{E}_{s\infty NM}(\hat{\mathbf{r}})=\mathbf{E}_{s\infty}(\hat{\mathbf{r}}) \quad \text{uniformly on } \Omega, \tag{76}$$

and in particular, in the case $N=M$, that Eq. (60) is satisfied.

3. Coming to the last result of the assertion, we first note that Eq. (74) gives

$$\lim_{N\to\infty}\left|\left|\mathbf{E}_{s\infty N}-\mathbf{E}_{s\infty}\right|\right|_{2,\Omega}=0. \tag{77}$$

From Eqs. (8) and (17), and the orthogonality of the normalized spherical harmonic vectors on the unit sphere Ω, we find that the expansion coefficients of the scattered field are given by

$$f_\alpha=f_{mn}=(-1)^{n+1}k_s\langle\mathbf{E}_{s\infty},\overline{\tilde{\mathbf{m}}_{-mn}}\rangle_{2,\Omega},$$

$$g_\alpha=g_{mn}=(-1)^{n}\mathrm{j}k_s\langle\mathbf{E}_{s\infty},\overline{\tilde{\mathbf{n}}_{-mn}}\rangle_{2,\Omega},$$

and

$$f_\alpha^N=f_{mn}^N=(-1)^{n+1}k_s\langle\mathbf{E}_{s\infty N},\overline{\tilde{\mathbf{m}}_{-mn}}\rangle_{2,\Omega},$$

$$g_\alpha^N=g_{mn}^N=(-1)^{n}\mathrm{j}k_s\langle\mathbf{E}_{s\infty N},\overline{\tilde{\mathbf{n}}_{-mn}}\rangle_{2,\Omega},$$

respectively. The result (77) together with the Cauchy–Schwarz inequality $|\langle\mathbf{a},\mathbf{b}\rangle_{2,\Omega}|\leq||\mathbf{a}||_{2,\Omega}||\mathbf{b}||_{2,\Omega}$ then yields

$$\lim_{N\to\infty}f_\alpha^N=f_\alpha,\quad\lim_{N\to\infty}g_\alpha^N=g_\alpha$$

uniformly with respect to α, provided that

$$||\tilde{\mathbf{m}}_\alpha||_{2,\Omega}\leq c_m,\quad||\tilde{\mathbf{n}}_\alpha||_{2,\Omega}\leq c_n.$$

This finishes the proof of the assertion. □

The converse result can be formulated as follows.

Assertion 5 *Assume that the operator $\tilde{\mathcal{A}}_s^{-1}$ is bounded. Then the convergence result (58) implies the condition (57).*

Proof. From $\left|\left|\tilde{\mathbf{u}}-\tilde{\mathbf{u}}_N\right|\right|_{2,S}\to 0$ as $N\to\infty$, we get

$$\left|\left|\tilde{\mathcal{A}}_s^{-1}\mathrm{f}-(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_N\mathrm{f}\right|\right|_{2,S}\to 0 \quad \text{as} \quad N\to\infty \tag{78}$$

for any $\mathrm{f}\in\mathcal{R}(\tilde{\mathcal{A}}_s)$. This result together with the boundedness of $\tilde{\mathcal{A}}_s^{-1}$ on its range, i.e., $\left|\left|\tilde{\mathcal{A}}_s^{-1}\mathrm{f}\right|\right|_{2,S}\le C\left|\left|\mathrm{f}\right|\right|_2$ for some $C\ge 0$, yield.

$$\left|\left|(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_N\mathrm{f}\right|\right|_{2,S}\le\left|\left|\tilde{\mathcal{A}}_s^{-1}\mathrm{f}-(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_N\mathrm{f}\right|\right|_{2,S}+\left|\left|\tilde{\mathcal{A}}_s^{-1}\mathrm{f}\right|\right|_{2,S}$$

$$\le\left|\left|\tilde{\mathcal{A}}_s^{-1}\mathrm{f}-(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_N\mathrm{f}\right|\right|_{2,S}+C\left|\left|\mathrm{f}\right|\right|_2; \tag{79}$$

hence, in view of Eq. (34), the operator $(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_N$ is bounded. As a consequence, we obtain.

$$\left|\left|\tilde{\mathbf{u}}_N\right|\right|_{2,S}=\left|\left|(Q_N\tilde{\mathcal{A}}_s)^{-1}(Q_N\tilde{\mathcal{A}}_s)\tilde{\mathbf{u}}_N\right|\right|_{2,S}$$

$$=\left|\left|(Q_N\tilde{\mathcal{A}}_s)^{-1}Q_NQ_N\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}_N\right|\right|_{2,S}$$

$$\le C\left|\left|Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{u}}_N\right|\right|_2, \tag{80}$$

showing that the condition (57) is satisfied. □

From Assertion 4 we see that the boundedness of $\tilde{\mathcal{A}}_s$ and the condition (57) imply the (strong) convergence of the tangential fields, which in turn, implies the uniform convergence of the far-field pattern on the unit sphere and of the scattered field coefficients. These assumptions are also sufficient conditions for the stability of the solution to Eq. (50).

Assertion 6 *Assume that the operator $\tilde{\mathcal{A}}_s$ is bounded in $\tilde{\mathfrak{T}}^2(S)$ and the condition (57) is satisfied. Then, $\tilde{\mathcal{A}}_s^{-1}$ is bounded.*

Proof. For any $\tilde{\mathbf{v}}\in\tilde{\mathfrak{T}}^2(S)$, we have

$$\lim_{N\to\infty}\left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S}\to 0, \tag{81}$$

where $\tilde{\mathbf{v}}_N=\tilde{P}_N\tilde{\mathbf{v}}\in\tilde{\mathfrak{T}}_N^2(S)$ is the orthogonal projection of $\tilde{\mathbf{v}}\in\tilde{\mathfrak{T}}^2(S)$ onto $\tilde{\mathfrak{T}}_N^2(S)$. Because $\tilde{\mathcal{A}}_s$ is bounded, we also have

$$\lim_{N\to\infty}\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}-\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N\right|\right|_{2,S}\to 0. \tag{82}$$

Consequently, we get

$$\left|\left|\tilde{\mathbf{v}}\right|\right|_{2,S}\le\left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S}+\left|\left|\tilde{\mathbf{v}}_N\right|\right|_{2,S}$$

$$\le\left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S}+\frac{1}{c_{1s}}\left|\left|Q_N\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N\right|\right|_2$$

$$\le\left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S}+\frac{1}{c_{1s}}\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N\right|\right|_2$$

$$\le\left|\left|\tilde{\mathbf{v}}-\tilde{\mathbf{v}}_N\right|\right|_{2,S}+\frac{1}{c_{1s}}\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N-\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2+\frac{1}{c_{1s}}\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2$$

$$\le\frac{1}{c_{1s}}\left|\left|\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}\right|\right|_2 \quad \text{as} \quad N\to\infty,$$

showing that $\tilde{\mathcal{A}}_s$ is bounded from below. Finally, from the Bounded Inverse Theorem we deduce that the bounded operator $\tilde{\mathcal{A}}_s$ has a bounded inverse $\tilde{\mathcal{A}}_s^{-1}$ on its range $\mathcal{R}(\tilde{\mathcal{A}}_s)$. □

The above results can be summarized as follows.

Stability Case. If the operators $\tilde{\mathcal{A}}_s$ and $\tilde{\mathcal{A}}_s^{-1}$ are bounded, the operator equation (50) is well-posed. Furthermore, if the condition (57) is satisfied, then in view of Assertion 4, the null-field method converges; otherwise, by Assertion 5, the null-field method diverges.

Instability Case. If the inverse operator $\tilde{\mathcal{A}}_s^{-1}$ is unbounded, the operator equation (50) is ill-posed. From Assertion 6 we deduce that the condition (57) is not satisfied. At this stage of our argumentation we reach an impasse because we cannot say anything about the convergence of the null-field method. Indeed, because $\tilde{\mathcal{A}}_s^{-1}$ is unbounded we cannot conclude from the unrealized condition (57) in Assertion 5 that the null-field method diverges.

All we can say in the case of instability is that the null-field method *may or may not converge*.

The main assumptions employed in the analysis can be checked numerically.

1. The null-field scheme is stable, i.e., the inverse operator $\tilde{\mathcal{A}}_s^{-1}$ is bounded, if and only if the condition number $\kappa(\mathrm{Q}_{31N})$ of the matrix $\mathrm{Q}_{31N}$is bounded when N increases.
2. According to Assertion 3, the operators $\tilde{\mathcal{A}}_s$ and $\tilde{\mathcal{A}}_s^{-1}$ are both bounded if and only if the complete system of tangential vector functions (34) is a Riesz basis of $\tilde{\mathfrak{T}}^2(S)$. Therefore, we may test the numerical stability by checking the Riesz basis property of this system of discrete sources. Now, the result stated in Appendix 4 of Ref. [1] "*A complete system* $\{\psi_i\}_{i=1}^{\infty}$ *forms a Riesz basis of a Hilbert space H if and only if the inequalities*

$$c_1 \sum_{i=1}^{N} |a_i|^2 \leq || \sum_{i=1}^{N} a_i \psi_i ||_H^2 \leq c_2 \sum_{i=1}^{N} |a_i|^2 \tag{83}$$

hold for any constants a_i *and for any N, where the positive constants* c_1 *and* c_2 *do not depend on N and* a_i" implies that the system of tangential vector functions (34) is a Riesz basis of $\tilde{\mathfrak{T}}^2(S)$if and only if there exist positive constants L_1 and L_2 such that

$$0 < L_1 \leq \lambda_{\min}\left(\mathrm{G}_N^3\right) < \lambda_{\max}\left(\mathrm{G}_N^3\right) \leq L_2 < \infty \quad \text{for all } N, \tag{84}$$

where $\lambda_{\min}(\mathsf{A})$ and $\lambda_{\max}(\mathsf{A})$ are the minimal and the maximal eigenvalues of the matrix A, and G_N^3 is the Gramm matrix of the system (34). Note that, because the tangential vector functions (34) are linearly independent, G_N^3 is a symmetric and positive definite matrix, and so, its eigenvalues coincide with its singular values.

3. Consider the condition (57), i.e.,

$$c_{1s} ||\tilde{\mathbf{v}}_N||_{2,S} \leq ||Q_N \tilde{\mathcal{A}}_s \tilde{\mathbf{v}}_N||_2 \quad \text{for all } \tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S) \quad \text{and all } N.$$

For

$$\tilde{\mathbf{v}}_N = \sum_{\beta=1}^{N} \left\{ c_\beta^N \begin{bmatrix} \widehat{\mathbf{n}} \times \mathfrak{M}_\beta^1(k_i \cdot) \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\, \widehat{\mathbf{n}} \times \mathfrak{N}_\beta^1(k_i \cdot) \end{bmatrix} + d_\beta^N \begin{bmatrix} \widehat{\mathbf{n}} \times \mathfrak{N}_\beta^1(k_i \cdot) \\ -\mathrm{j}\sqrt{\frac{\varepsilon_i}{\mu_i}}\, \widehat{\mathbf{n}} \times \mathfrak{M}_\beta^1(k_i \cdot) \end{bmatrix} \right\}, \tag{85}$$

we find

$$\left|\left|Q_N \tilde{\mathcal{A}}_s \tilde{\mathbf{v}}_N\right|\right|_2^2 = \left[\left(\overline{c_{\beta'}^N}\right)_{\beta'=1}^{N} \left(\overline{d_{\beta'}^N}\right)_{\beta'=1}^{N} \right] \mathrm{Q}_{31N}^{\dagger} \mathrm{Q}_{31N} \begin{bmatrix} \left(c_\beta^N\right)_{\beta=1}^{N} \\ \left(d_\beta^N\right)_{\beta=1}^{N} \end{bmatrix}, \tag{86}$$

and

$$\left|\left|\tilde{\mathbf{v}}_N\right|\right|_{2,S}^2 = \left[\left(\overline{c_{\beta'}^N}\right)_{\beta'=1}^{N} \left(\overline{d_{\beta'}^N}\right)_{\beta'=1}^{N} \right] \mathrm{G}_N^1 \begin{bmatrix} \left(c_\beta^N\right)_{\beta=1}^{N} \\ \left(d_\beta^N\right)_{\beta=1}^{N} \end{bmatrix}, \tag{87}$$

where G_N^1 is the Gramm matrix of the complete and linear independent system of tangential vector functions (85). The condition (57) is equivalent to the matrix inequality

$$C_1 \mathrm{G}_N^1 \leq \mathrm{Q}_{31N}^{\dagger} \mathrm{Q}_{31N} \quad \text{for all } N \text{ and } C_1 = c_{1s}^2. \tag{88}$$

This inequality is satisfied if, for example, there exist positive constants l_1 and L_2 such that

$$\mathrm{Q}_{31N}^{\dagger} \mathrm{Q}_{31N} \geq l_1 \mathrm{I}_N \quad \text{and} \quad \mathrm{G}_N^1 \leq L_2 \mathrm{I}_N \tag{89}$$

for all N, meaning that

$$\lambda_{\min}\left(\mathrm{Q}_{31N}^{\dagger} \mathrm{Q}_{31N}\right) \geq l_1 > 0 \quad \text{and} \quad \lambda_{\max}\left(\mathrm{G}_N^1\right) \leq L_2 < \infty \tag{90}$$

for all N. Because $\lambda_k(\mathrm{Q}_{31N}^{\dagger} \mathrm{Q}_{31N}) = \sigma_k^2(\mathrm{Q}_{31N})$, where $\sigma_k(\mathrm{Q}_{31N})$, $k = 1, \ldots, 2N$ are the singular values of the matrix Q_{31N}, and the eigenvalues of G_N^1 coincide with its singular values, the above conditions can be written as

$$\sigma_{\min}(\mathrm{Q}_{31N}) \geq s_1 > 0 \quad \text{and} \quad \sigma_{\max}\left(\mathrm{G}_N^1\right) \leq S_2 < \infty \tag{91}$$

for all N and some positive constants s_1 and S_2. Thus, if the conditions (91) are satisfied, then the condition (57) is also satisfied.

The main conclusion of our analysis is that we can decide whether or not the null-field scheme converges only when the scheme is numerically stable. In this regard we make a final remark.

Comment. The boundedness of the operator $\tilde{\mathcal{A}}_s$ and the condition (57) imply (cf. Eq. (64))

$$c_{1s}\left|\left|\tilde{\mathbf{v}}_N\right|\right|_{2,S} \leq \left|\left|\mathcal{Q}_N\tilde{\mathcal{A}}_s\tilde{\mathbf{v}}_N\right|\right|_2 \leq c_{2s}\left|\left|\tilde{\mathbf{v}}_N\right|\right|_{2,S} \quad \text{for all} \quad \tilde{\mathbf{v}}_N \in \tilde{\mathfrak{T}}_N^2(S) \quad \text{and all} \quad N. \tag{92}$$

Employing the same arguments as above we find that the conditions (92) are equivalent to the matrix inequalities

$$C_1\mathrm{G}_N \leq \mathrm{Q}_{31N}^{\dagger}\mathrm{Q}_{31N} \leq C_2\mathrm{G}_N \quad \text{for all} \quad N \quad \text{and} \quad C_{1,2} = c_{1,2s}^2. \tag{93}$$

These are satisfied if the conditions

$$0 < s_1 \leq \sigma_{\min}(\mathrm{Q}_{31N}) < \sigma_{\max}(\mathrm{Q}_{31N}) \leq s_2 < \infty \quad \text{and} \tag{94}$$

$$0 < S_1 \leq \sigma_{\min}\left(\mathrm{G}_N^1\right) < \sigma_{\max}\left(\mathrm{G}_N^1\right) \leq S_2 < \infty \tag{95}$$

for all N and some positive constants s_1, s_2, S_1, and S_2, are fulfilled. Denoting by $\kappa(\mathrm{A}) = \sigma_{\max}(\mathrm{A})/\sigma_{\min}(\mathrm{A})$ the condition number of the matrix A, we end up with

$$\kappa(\mathrm{Q}_{31N}) = \frac{\sigma_{\max}(\mathrm{Q}_{31N})}{\sigma_{\min}(\mathrm{Q}_{31N})} \leq \frac{s_2}{s_1} \quad \text{and} \tag{96}$$

$$\kappa\left(\mathrm{G}_N^1\right) = \frac{\sigma_{\max}\left(\mathrm{G}_N^1\right)}{\sigma_{\min}\left(\mathrm{G}_N^1\right)} \leq \frac{S_2}{S_1} \tag{97}$$

for all N. Therefore, the condition number of the matrix Q_{31N} gives not only information about the stability of the method but also on the realization of the conditions (92), which in view of Assertion 4, imply the convergence of the null-field scheme.

Appendix 2

In this appendix we present an analytical method for computing the Q_{31N}-matrix elements in the framework of the null-field method with localized vector spherical wave functions.

For an axisymmetric particle, Somerville et al. [23] showed that the block-matrix elements of the matrix Q_{31N} can be expressed as

$$Q_{31Nmnk}^{11} = 2\pi \mathrm{j} c_n c_k \left[-\mathrm{m_r} L_{mnk}^1 + L_{mnk}^3 + \frac{1}{\mathrm{m}_r}\left(L_{mnk}^2 - L_{mnk}^4\right)\right], \tag{98}$$

$$Q_{31Nmnk}^{12} = 2\pi c_n c_k \frac{1-\mathrm{m_r^2}}{\mathrm{m_r}} K_{mnk}^1, \tag{99}$$

$$Q_{31Nmnk}^{21} = 2\pi c_n c_k \frac{\mathrm{m_r^2}-1}{\mathrm{m_r}} K_{mnk}^2, \tag{100}$$

$$Q_{31Nmnk}^{22} = 2\pi \mathrm{j} c_n c_k \left(-L_{mnk}^1 + \frac{1}{\mathrm{m_r}} L_{mnk}^3 + L_{mnk}^2 - L_{mnk}^4 \right), \tag{101}$$

where $m = 0, 1, \ldots M_{\mathrm{rank}}$, $n, k = \max(1, |m|), \ldots, N_{\mathrm{rank}}$, M_{rank} and N_{rank} are the maximum azimuthal mode and expansion order, respectively, and

$$K_{mnk}^1 = m \int_0^{\pi} \xi_n(x)\psi_k{}'(\mathrm{m_r}x) P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) \frac{\partial x}{\partial \theta}\, \mathrm{d}\theta, \tag{102}$$

$$K_{mnk}^2 = m \int_0^{\pi} \xi_n{}'(x)\psi_k(\mathrm{m_r}x) P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) \frac{\partial x}{\partial \theta}\, \mathrm{d}\theta, \tag{103}$$

$$L_{mnk}^1 = \int_0^{\pi} \xi_n(x)\psi_k(\mathrm{m_r}x) \left[\sin\theta \tau_n^{|m|}(\theta)\right] P_k^{|m|}(\cos\theta) \frac{\partial x}{\partial \theta}\, \mathrm{d}\theta, \tag{104}$$

$$L_{mnk}^2 = \int_0^{\pi} \xi_n(x)\psi_k(\mathrm{m_r}x) \left[\sin\theta \tau_k^{|m|}(\theta)\right] P_n^{|m|}(\cos\theta) \frac{\partial x}{\partial \theta}\, \mathrm{d}\theta, \tag{105}$$

$$L_{mnk}^3 = \int_0^{\pi} \psi_k{}'(\mathrm{m_r}x) P_k^{|m|}(\cos\theta) \left\{ \xi_n{}'(x) \left[\sin\theta \tau_n^{|m|}(\theta)\right] \frac{\partial x}{\partial \theta}\right.$$

$$-n(n+1)\xi_n(x)P_n^{|m|}(\cos\theta)\sin\theta\Big\}\,\mathrm{d}\theta \tag{106}$$

$$L_{mnk}^4 = \int_0^\pi \xi_n{}'(x)P_n^{|m|}(\cos\theta)\Big\{m_\mathrm{r}\psi_k{}'(\mathrm{m_r}x)\Big[\sin\theta\tau_k^{|m|}(\cos\theta)\Big]\frac{\partial x}{\partial\theta}$$

$$-k(k+1)\psi_k(\mathrm{m_r}x)P_k^{|m|}(\cos\theta)\sin\theta\Big\}\,\mathrm{d}\theta \tag{107}$$

In Eqs. 102–107, $\psi_n = \psi_n(\mathrm{m_r}x) = \mathrm{m_r}xj_n(\mathrm{m_r}x)$ are the regular Riccati–Bessel functions, $\xi_n = \xi_n(x) = xh_n(x)$ are the Riccati–Hankel functions of the first kind, $\mathrm{m_r}$ is the relative refractive index of the particle, $x = x(\theta) = k_s r(\theta)$, where $r(\theta)$ describes the generatrix in polar coordinates, $\psi_n{}'(x) = \mathrm{d}\psi_n(x)/\mathrm{d}x$, $\xi_n{}'(x) = \mathrm{d}\xi_n(x)/\mathrm{d}x$, and

$$c_n = \frac{1}{\sqrt{2\pi n(n+1)}}. \tag{108}$$

Note that for an axisymmetric particle with mirror symmetry, i.e., $r(\theta) = r(\pi-\theta)$, we have

$$K_{mnk}^i = 0 \quad \text{for} \quad n+k = \quad \text{even} \quad \text{and} \quad i = 1,2,$$

and

$$L_{mnk}^i = 0 \quad \text{for} \quad n+k = \quad \text{odd} \quad \text{and} \quad i = 1,2,3,4.$$

A loss of precision may occur during the computation of the integrals (102)–(107) by a numerical scheme [14]. To reveal this event we consider the integral term K_{mnk}^1 because the main concepts are fully represented in this case (as $\sin\theta\tau_k^{|m|}(\theta)$ can be expressed in terms of $P_{n+1}^{|m|}(\cos\theta)$ and $P_{n-1}^{|m|}(\cos\theta)$, the terms containing $\sin\theta\tau_k^{|m|}(\theta)$ have a similar expression). Using the decomposition $\xi_n = \psi_n + \mathrm{j}\chi_n$ and the representation $\xi_n\psi_k{}' = \psi_n\psi_k{}' + \mathrm{j}\chi_n\psi_k{}'$, where $\chi_n(x) = xy_n(x)$ are the irregular Riccati–Bessel functions, we express K_{mnk}^1 as $K_{mnk}^1 = K_{mnk}^{1(1)} + \mathrm{j}K_{mnk}^{1(3)}$, where the integrands of $K_{mnk}^{1(1)}$ and $K_{mnk}^{1(3)}$ contain the terms $\psi_n\psi_k{}'$ and $\chi_n\psi_k{}'$, respectively. Actually, only the computation of the integral term $K_{mnk}^{1(3)}$ is problematic. Therefore, we focus on the integral term $K_{mnk}^{1(3)}$, but in order to avoid an abundance of notation we still denote this term by K_{mnk}^1.

To compute K_{mnk}^1, we use the series expansions of ψ_n and χ_n; these yield

$$\psi_n{}'(x) = \sum_{s=0}\frac{1}{s!}\left(-\frac{1}{2}\right)^s \alpha_{sn}x^{2s+n}, \tag{109}$$

$$\chi_n(x) = -\sum_{s=0}\frac{1}{s!}\left(-\frac{1}{2}\right)^s \beta_{sn}x^{2s-n}, \tag{110}$$

where α_{sn} and β_{sn} are given, respectively, by

$$\alpha_{sn} = \frac{2s+n+1}{(2n+2s+1)!!}, \tag{111}$$

$$\beta_{sn} = \begin{cases} \beta_{sn}^- = (-1)^s(2n-2s-1)!!, & s \le n-1 \\ \beta_{sn}^+ = (-1)^n/(2s-2n-1)!!, & s \ge n \end{cases} \tag{112}$$

with the convention $(-1)!! = 1$. Using these results, truncating the series at the index $\overline{N}$, and moreover, assuming the representation

$$x(\theta) = k_s r(\theta) = X\bar{r}(\theta), \tag{113}$$

where $X = k_s l$ is the size parameter, l a characteristic length of the particle, and $\bar{r}(\theta)$ a (dimensionless) representation of the particle shape in polar coordinates, we obtain

$$K_{mnk}^1 = K_{mnk}^{1-} + K_{mnk}^{1+}, \tag{114}$$

where

$$K_{mnk}^{1-} = -m\mathrm{m_r^k}X\sum_{q=0}\left(-\frac{1}{2}\right)^q \gamma_{qnk}^-(\mathrm{m_r})\frac{1}{X^{n-k-2q}}I_{qnk}^-, \tag{115}$$

$$I^{-}_{qnk} = \int_0^{\pi} \frac{1}{\bar{r}^{n-k-2q}(\theta)} P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) \left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right) \sin\theta \, \mathrm{d}\theta, \tag{116}$$

$$\gamma^{-}_{qnk}(\mathrm{m_r}) = \sum_{s=0} \mathrm{m_r}^{2(q-s)} \Gamma^{-}_{qnk,s}, \tag{117}$$

$$\Gamma^{-}_{qnk,s} = \frac{\alpha_{q-s,k}\beta^{-}_{sn}}{s!(q-s)!}, \tag{118}$$

and

$$K^{1+}_{mnk} = -m\mathrm{m_r}^{k} X \sum_{q=\max(0,q_0+1)}^{2\bar{N}} \left(-\frac{1}{2}\right)^{q} \gamma_{qnk}(\mathrm{m_r}) X^{2q+k-n} I^{+}_{qnk}, \tag{119}$$

$$I^{+}_{qnk} = \int_0^{\pi} \bar{r}^{2q+k-n}(\theta) P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) \left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right) \sin\theta \, \mathrm{d}\theta, \tag{120}$$

$$\gamma_{qnk}(\mathrm{m_r}) = \sum_{s=\max(0,q-\bar{N})}^{\min(q,\bar{N})} \mathrm{m_r}^{2(q-s)} \Gamma_{qnk,s}, \tag{121}$$

$$\Gamma_{qnk,s} = \frac{\alpha_{q-s,k}\beta_{sn}}{s!(q-s)!}, \tag{122}$$

with

$$q_0 = q_0(n,k) = \begin{cases} (n-k-1)/2, & \text{if } n-k = \text{odd} \\ (n-k-2)/2, & \text{if } n-k = \text{even} \end{cases}. \tag{123}$$

Here and in the following we omit to indicate the dependency of I^{-}_{qnk} and I^{+}_{qnk} on the azimuth mode m and use the finite-sum convention $\sum_{s=s_{\min}}^{s_{\max}}(\cdot) = 0$ if $s_{\max} < s_{\min}$.

The term K^{1-}_{mnk}, which is nonzero for $n > k$, contains negative powers of $\bar{r}$, while the term K^{1+}_{mnk} contains positive powers of $\bar{r}$. For large values of $n - k > 0$, and in particular, for large size parameters and/or strongly deformed particles, the integrand of K^{1-}_{mnk}, obtained by inserting Eq. (116) into Eq. (115), i.e.,

$$K^{1-}_{mnk} = -m\mathrm{m_r}^{k} X \int_0^{\pi} F^{-}_{nk}(\theta) P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) \left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right) \sin\theta \, \mathrm{d}\theta, \tag{124}$$

$$F^{-}_{nk}(\theta) = \sum_{q=0} \left(-\frac{1}{2}\right)^{q} \gamma^{-}_{qnk}(\mathrm{m_r}) \frac{1}{[X\bar{r}(\theta)]^{n-k-2q}}, \tag{125}$$

oscillates around zero and its magnitude varies significantly across the range of integration. As a result, a loss of precision can occur during the summation step of a Gauss–Legendre quadrature method. To overcome this problem we propose an analytical method for computing K^{1}_{mnk}, and in fact, of the integrals I^{-}_{qnk} and I^{+}_{qnk}.

The method relies.

1. on the addition theorem for the associated Legendre functions

$$P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) = \sum_{p=|n-k|;2}^{n+k} a(m,n|-m,k|p) P_p(\cos\theta), \tag{126}$$

where

$$a(m,n|-m,k|p) = \int_0^{\pi} P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta) P_p(\cos\theta) \sin\theta \, \mathrm{d}\theta \tag{127}$$

are the Gaunt coefficients and the notation $\sum_{p=|n-k|;2}^{n+k}$ means that the index p increases from $|n-k|$ to $n+k$ in steps of 2; and.

2. the Legendre polynomial expansions of powers of the shape function, i.e,

$$\frac{1}{\bar{r}^{n-k-2q}(\theta)}\left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right)=\sum_{p=0}\rho^{-}_{qnk,p}P_p(\cos\theta), \tag{128}$$

$$\bar{r}^{2q+k-n}(\theta)\left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right)=\sum_{p=0}\rho^{+}_{qnk,p}P_p(\cos\theta). \tag{129}$$

The result is

$$I^{-}_{qnk}=\sum_{p=|n-k|;2}\rho^{-}_{qnk,p}a(m,n|-m,k|p), \tag{130}$$

$$I^{+}_{qnk}=\sum_{p=|n-k|;2}\rho^{+}_{qnk,p}a(m,n|-m,k|p). \tag{131}$$

Note that the Legendre expansions (128) and (129) are valid for any shape function $\bar{r}\in C^1([0,\pi])$. Some peculiarities of the computational algorithm are listed below.

1. To speed up the calculation, the expansion coefficients of the shape function

$$\rho^{-}_{s,p}=\int_0^{\pi}\frac{1}{\bar{r}^s(\theta)}\left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right)P_p(\cos\theta)\sin\theta\,\mathrm{d}\theta,\quad s=1,2,\ldots,\overline{N}, \tag{132}$$

and

$$\rho^{+}_{s,p}=\int_0^{\pi}\bar{r}^s(\theta)\left(\frac{1}{\sin\theta}\frac{\mathrm{d}\bar{r}}{\mathrm{d}\theta}\right)P_p(\cos\theta)\sin\theta\,\mathrm{d}\theta,\quad s=0,\ldots 5\overline{N}-1, \tag{133}$$

defined through Eqs. (128) and (129), respectively, are computed in *extended precision* in a preprocessing step, stored in a database, and used as input parameters of the algorithm.

2. The coefficients

$$\Gamma^{-}_{qnk,s}=\frac{\alpha_{q-s,k}\beta^{-}_{sn}}{s!(q-s)!}$$

$$=\frac{(-1)^s}{s!(q-s)!}\frac{2(q-s)+k+1}{(2k+2q-2s+1)!!}(2n-2s-1)!! \tag{134}$$

in Eq. (118) are computed for $s=0,\ldots q$ by using the upward recurrence relation

$$\Gamma^{-}_{qnk,s+1}=-\frac{2(q-s)+k-1}{2(q-s)+k+1}\frac{2(k+q-s)+1}{2(n-s)-1}\frac{q-s}{s+1}\Gamma^{-}_{qnk,s}, \tag{135}$$

for $s=0,\ldots,q-1$ with

$$\Gamma^{-}_{qnk0}=\frac{(2q+k+1)(2n-1)!!}{q!(2k+2q+1)!!}. \tag{136}$$

The coefficients $\Gamma_{qnk,s}$ in Eq. (121) are expressed as

$$\Gamma_{qnk,s}=\begin{cases}\Gamma^{-}_{qnk,s}, & \text{if } s\le n-1\\ \Gamma^{+}_{qnk,s}, & \text{if } s\ge n\end{cases}, \tag{137}$$

where.

$$\Gamma^{+}_{qnk,s}=\frac{\alpha_{q-s,k}\beta_{sn}}{s!(q-s)!}$$

$$=\frac{(-1)^n}{s!(q-s)!}\frac{2(q-s)+k+1}{(2k+2q-2s+1)!!}\frac{1}{(2s-2n-1)!!}. \tag{138}$$

For $s=0,\ldots q$, the coefficients $\Gamma_{qnk,s}$ are calculated as follows.

(a) If $q \leq n-1$, we have $\Gamma_{qnk,s} = \Gamma^{-}_{qnk,s}$, and we compute $\Gamma_{qnk,s}$ by using the upward recurrence (135) for $s = 0,\ldots,q-1$.

(b) If $q \geq n$, we compute $\Gamma_{qnk,s}$ separately for the ranges $s = 0,\ldots,n-1$ and $s = n,\ldots,q$. In the range $s = 0,\ldots,n-1$, we have $\Gamma_{qnk,s} = \Gamma^{-}_{qnk,s}$, and we compute $\Gamma_{qnk,s}$ by using the upward recurrence (135) for $s = 0,\ldots,n-2$. In the range $s = n,\ldots,q$, we have $\Gamma_{qnk,s} = \Gamma^{+}_{qnk,s}$, and we compute $\Gamma_{qnk,s}$ by using the upward recurrence

$$\Gamma_{qnk,s+1} = \frac{2(q-s)+k-1}{2(q-s)+k+1}\,\frac{2(k+q-s)+1}{2(s-n)+1}\,\frac{q-s}{s+1}\Gamma_{qnk,s}, \tag{139}$$

for $s = n,\ldots,q-1$ with

$$\Gamma_{qnk,n} = \frac{(-1)^n}{n!(q-n)!}\,\frac{2(q-n)+k+1}{(2k+2q-2n+1)!!}. \tag{140}$$

3. The Gaunt coefficients $a(m,n|-m,k|p)$ are computed by using the downward recurrence relation

$$\varsigma_{p+1}a(\cdot,p) = \sqrt{\frac{2p+5}{2p+1}}(4m^2+\varsigma_{p+2}+\varsigma_{p+3})a(\cdot,p+2)$$

$$-\sqrt{\frac{2p+9}{2p+1}}\varsigma_{p+4}a(\cdot,p+4) \tag{141}$$

with the starting values.

$$a(m,n|-m,k|n+k) = (-1)^m\sqrt{\frac{2(2n+1)(2k+1)}{2(n+k)+1}}$$

$$\times\frac{(2n-1)!(2k-1)!}{(2n+2k-1)!}\sqrt{\frac{(n-m)!(k-m)!}{(n+m)!(k+m)!}}$$

$$\times\frac{(n+k-1)!(n+k)!}{(n-m)!(n-1)!(k-m)!(k-1)!}, \tag{142}$$

$$a(\cdot,n+k-2) = \frac{\sqrt{(2n+2k-3)(2n+2k+1)}}{(2n-1)(2k-1)(n+k)}$$

$$\times\left[nk-m^2(2n+2k-1)\right]a(\cdot,n+k), \tag{143}$$

where $a(\cdot,p)$ stands for $a(m,n|-m,k|p)$ and

$$\varsigma_p = \frac{\left[p^2-(n+k+1)^2\right]\left[p^2-(n-k)^2\right]}{4p^2-1}. \tag{144}$$

The three-term recurrence formula (141) is due to Bruning and Lo [26], and provides accurate numerical results for all low- and high-degree coefficients.

To exemplify the loss of accuracy in computing K^{1-}_{mnk} by a quadrature method in double precision, we assume that the shape function is the superellipse (Lame′ curve)

$$\overline{r}(\theta) = \left(\cos^{n_{\rm p}}\theta + e^{n_{\rm p}}\sin^{n_{\rm p}}\theta\right)^{-1/n_{\rm p}}, \tag{145}$$

where $e = a/b$ is the eccentricity, and a and b are the semi-major and the semi-minor axis of the superellipse, respectively. The case $n_{\rm p} = 2$ corresponds to a spheroid, the cases $n_{\rm p} = 4$ and $n_{\rm p} = 6$ correspond to a cylinder with rounded corners, and the case $n_{\rm p} \to \infty$ corresponds to a cylinder. In Table 1 we illustrate the values of K^{1-}_{mnk} and K^{1+}_{mnk}. The calculations are performed by using the analytical and the Gauss–Legendre quadrature method. Note that the Gauss–Legendre method is applied to K^{1-}_{mnk} given by Eqs. (124) and (125), while for K^{1+}_{mnk} we use

$$K^{1+}_{mnk} = -m\mathrm{m}^k_{\rm r}X\int_0^{\pi}F_{nk}(\theta)P_n^{|m|}(\cos\theta)P_k^{|m|}(\cos\theta)\left(\frac{1}{\sin\theta}\frac{{\rm d}\overline{r}}{{\rm d}\theta}\right)\sin\theta\,{\rm d}\theta, \tag{146}$$

Table 1
The values of K^{1-}_{mnk} and K^{1+}_{mnk} computed in double and extended precision. The values of K^{1-}_{mnk} are computed by using the analytical method ($(K^{1-}_{mnk})_{\text{analytic}}$) and the Gauss–Legendre quadrature method ($(K^{1-}_{mnk})_{\text{quadrat}}$). The values of K^{1+}_{mnq} calculated by means of the analytical and the Gauss–Legendre quadrature method are the same, and therefore, they are not listed separately. The parameters of the calculation are $\mathrm{m_r} = 1.5$, $k_s a = 20$, $k_s b = 5$, i.e., $e = 4$, $m = 1$, $\overline{N} = 60$, and $n = 40$.

K_{mnk}	k	Double Precision	Extended Precision
$(K^{1-}_{mnk})_{\text{analytic}}$	1	−7.552759e + 15	−7.552759e + 15
$(K^{1-}_{mnk})_{\text{quadrat}}$		−2.785579e + 15	−7.552759e + 15
K^{1+}_{mnk}		6.347593e + 13	6.347593e + 13
$(K^{1-}_{mnk})_{\text{analytic}}$	3	2.497338e + 16	2.497338e + 16
$(K^{1-}_{mnk})_{\text{quadrat}}$		9.000000e + 15	2.497338e + 16
K^{1+}_{mnk}		−2.209145e + 14	−2.209145e + 14
$(K^{1-}_{mnk})_{\text{analytic}}$	5	−3.957203e + 16	−3.957203e + 16
$(K^{1-}_{mnk})_{\text{quadrat}}$		−3.369987e + 16	−3.957203e + 16
K^{1+}_{mnk}		3.802838e + 14	3.802838e + 14
$(K^{1-}_{mnk})_{\text{analytic}}$	7	4.557751e + 16	4.557751e + 16
$(K^{1-}_{mnk})_{\text{quadrat}}$		4.740872e + 16	4.557751e + 16
K^{1+}_{mnk}		−4.837075e + 14	−4.837075e + 14
$(K^{1-}_{mnk})_{\text{analytic}}$	9	−4.230531e + 16	−4.230531e + 16
$(K^{1-}_{mnk})_{\text{quadrat}}$		−4.079553e + 16	−4.230531e + 16
K^{1+}_{mnk}		4.930445e + 14	4.930445e + 14
$(K^{1-}_{mnk})_{\text{analytic}}$	11	3.298516e + 16	3.298516e + 16
$(K^{1-}_{mnk})_{\text{quadrat}}$		3.226520e + 16	3.298516e + 16
K^{1+}_{mnk}		−4.055114e + 14	−4.055114e + 14

$$F_{nk}(\theta) = \sum_{q=\max(0,q_0+1)}^{2\overline{N}} \left(-\frac{1}{2}\right)^q \gamma_{qnk}(\mathrm{m_r}) [X\overline{r}(\theta)]^{2q+k-n}. \tag{147}$$

Also note that due to the mirror symmetry, K^1_{mnk} is zero for even values of $n-k$. The results show that.

1. for large (odd) values of $n-k>0$, a significant loss of accuracy occurs when K^{1-}_{mnk} is computed by the quadrature method in double precision;
2. no loss of accuracy occurs when K^{1-}_{mnk} is computed by the analytical method in double precision; and
3. no loss of accuracy occurs when K^{1+}_{mnk} is computed by either the analytical or the quadrature method in double precision.

Taking into account that the analytical method is much faster than the quadrature method, we may conclude that the analytical method is at the same time accurate and efficient.

We conclude our analysis with some comments.

1. In Refs. [14,15] it was shown that for spheroids, some terms I^-_{qnk} which contribute to K^{1-}_{mnk} , according to Eq. (115), vanish. Let us extend this result to the shape function (145). By straightforward calculation, we find

$$\frac{1}{\overline{r}^{n_\mathrm{p}(s+1)+1}} \left(\frac{1}{\sin\theta}\frac{\mathrm{d}\overline{r}}{\mathrm{d}\theta}\right)$$

$$= \left(\cos^{n_\mathrm{p}}\theta + e^{n_\mathrm{p}}\sin^{2q}\theta\right)^s \left(\cos^{n_\mathrm{p}-2}\theta - e^{n_\mathrm{p}}\sin^{n_\mathrm{p}-2}\theta\right)\cos\theta$$

$$= \mathcal{P}_{n_\mathrm{p}(s+1)-1}(\cos\theta) \quad \text{for} \quad s=0,1,\ldots \quad \text{and} \quad n_\mathrm{p}=2u \quad \text{with} \quad u\geq 1, \tag{148}$$

where $\mathcal{P}_n(x)$ stands for a polynomial of degree n in x. The expansion of the polynomial $\mathcal{P}_{n_\mathrm{p}(s+1)-1}$ in terms of the Legendre polynomials is

$$\mathcal{P}_{n_\mathrm{p}(s+1)-1}(\cos\theta) = \sum_{p=0}^{n_\mathrm{p}(s+1)-1} \varrho_{s,p} P_p(\cos\theta), \tag{149}$$

yielding

$$\int_0^\pi \frac{1}{\overline{r}^{n_\mathrm{p}(s+1)+1}(\theta)} \left(\frac{1}{\sin\theta}\frac{\mathrm{d}\overline{r}}{\mathrm{d}\theta}\right) P_n^{|m|}(\cos\theta) P_k^{|m|}(\cos\theta)\sin\theta\,\mathrm{d}\theta$$

$$= \sum_{p=\min\left(n_\mathrm{p}(s+1)-1,|n-k|\right);2} \varrho_{s,p}\, a(m,n|-m,k|p). \tag{150}$$

Therefore, for values of $s \geq 0$ satisfying $n_{\rm p}(s+1)-1 < |n-k|$, that is,

$$n_{\rm p}(s+1)+1 \leq |n-k|+1, \tag{151}$$

the contribution of the term corresponding to

$$\frac{1}{\bar{r}^{n_{\rm p}(s+1)+1}}\left(\frac{1}{\sin\theta}\frac{{\rm d}\bar{r}}{{\rm d}\theta}\right)$$

is zero. In other words, the terms with powers $-e$ of the form

$$e = n_{\rm p}(s+1)+1, \quad s \geq 0, \tag{152}$$

and satisfying

$$n_{\rm p}+1 \leq e \leq |n-k|+1 \tag{153}$$

will not contribute to K^{1-}_{mnk}. This theoretical result should be expressly used in the computation of K^{1-}_{mnk}. In the favorable case of spheroids, we have $n_{\rm p}=2$ and according to Eqs. (152) and (153), the terms $(\cdot)/\bar{r}^3$, $(\cdot)/\bar{r}^5$, $(\cdot)/\bar{r}^7, \ldots$, will not contribute to K^{1-}_{mnk}. Furthermore, since $K^{i}_{mnk}=0$ for $n+k=$ even, it follows that all terms $(\cdot)/\bar{r}^2$, $(\cdot)/\bar{r}^3$, $(\cdot)/\bar{r}^4, \ldots$, will not contribute to K^{1-}_{mnk}. Eliminating these terms from the series expansions, Somerville et al. [16] implemented a numerically stable algorithm for T-matrix calculation in the case of electromagnetic scattering by spheroidal particles.

2. In order to calculate all integral terms K^{i}_{mnk} and L^{i}_{mnk}, we have to compute and store the Legendre polynomial expansions of the functions

$$\bar{r}^s(\theta) \quad \text{and} \quad \bar{r}^s(\theta)\left(\frac{1}{\sin\theta}\frac{{\rm d}\bar{r}}{{\rm d}\theta}\right)$$

for say, $s=-\overline{N},\ldots,5\overline{N}$. In fact, it is sufficient to compute only the expansions of the functions

$$\bar{r}^s(\theta) \quad \text{and} \quad \frac{1}{\bar{r}(\theta)}\left(\frac{1}{\sin\theta}\frac{{\rm d}\bar{r}}{{\rm d}\theta}\right),$$

because the remainder terms can be computed by using a derivative rule for Legendre polynomial expansions. This rule states that given a finite sum representation for a function $f(\theta)$, i.e.,

$$f(\theta) = \sum_{p=0}^{N} a_p P_p(\cos\theta), \tag{154}$$

the result

$$P_p{}'(\mu) = \sum_{q=p_0;2}^{p-1} \sqrt{(2p+1)(2q+1)}\,P_q(\mu), \tag{155}$$

where

$$p_0 = \begin{cases} 1, & \text{for } p=2u \\ 0, & \text{for } p=2u+1 \end{cases} \tag{156}$$

and $\mu=\cos\theta$, yields the following finite sum representation for the derivative function:

$$\frac{1}{\sin\theta}\frac{{\rm d}f}{{\rm d}\theta} = -\sum_{p=0}^{N-1} b_p P_p(\cos\theta), \tag{157}$$

where for $N=2K$, the expansion coefficients are given by

$$b_{2s-1} = \sum_{k=s} a_{2k}\sqrt{(4k+1)(4s-1)}, \quad s=1,\ldots,K, \tag{158}$$

$$b_{2s} = \sum_{k=s} a_{2k+1}\sqrt{(4k+3)(4s+1)}, \quad s=0,\ldots,K-1, \tag{159}$$

and for $N = 2K + 1$, by

$$b_{2s-1} = \sum_{k=s} a_{2k} \sqrt{(4k+1)(4s-1)}, \quad s = 1, \ldots, K, \tag{160}$$

$$b_{2s} = \sum_{k=s} a_{2k+1} \sqrt{(4k+3)(4s+1)}, \quad s = 0, \ldots, K. \tag{161}$$

3. The integrals I^{-}_{qnk} and I^{+}_{qnk} which enter in the expressions of K^{1-}_{mnk} and K^{1+}_{mnk}, respectively, depend only on the shape of the particle and not on the size and refractive index. Therefore, the above method is very effective for averaging particle ensembles over their size parameter and refractive index. From this point of view, the method seems to be similar to the shape-matrix method developed by Petrov et al. [17–21] and based on power series representations for the spherical Bessel and Neumann functions, and the multiplication theorem

$$j_n(X\bar{r}) = X^n \sum_{k=0} \frac{(-1)^k (X^2-1)^k}{k!} \left(\frac{\bar{r}}{2}\right)^k j_{n+k}(\bar{r}),$$

$$y_n(X\bar{r}) = \frac{1}{X^{n+1}} \sum_{k=0} \frac{(X^2-1)^k}{k!} \left(\frac{\bar{r}}{2}\right)^k y_{n-k}(\bar{r}). \tag{162}$$

The differences between the two approaches are that in our approach we (i) use the representations for the Q_{31}-matrix elements given by Somerville et al. [23] rather than the conventional representations; and (ii) compute the integrals over θ by using the addition theorem for the associated Legendre functions, rather than by presumably representing the associated Legendre functions in terms of trigonometric functions (because this computational step is not explicitly described in Refs. [17–21], we believe that this is the case).

References

[1] Doicu A, Mishchenko MI. An overview of the null-field method. I: formulation and basic results. Submitted to J Quant Spectrosc Radiat Transfer.
[2] G. Kristensson, A.G. Ramm, S. Ström, Convergence of the T-matrix approach in scattering theory, II, J. Math. Phys. 24 (1983) 2619–2631.
[3] A.G. Ramm, Scattering by Obstacles, Reidel, Dordrecht, 1986.
[4] A.G. Dallas, Basis Properties of Traces and Normal Derivatives of Spherical-Separable Solutions of the Helmholtz Equation. Technical Report No. 2000-6, Department of Mathematical Sciences, University of Delaware, 2000.
[5] A.G. Dallas, On the Convergence and Numerical Stability of the Second Waterman Scheme for Approximation of the Acoustic Field Scattered by a Hard Obstacle. Technical Report No. 2000-7, Department of Mathematical Sciences, University of Delaware, 2000.
[6] H.W. Guggenheimer, A.S. Edelman, C.R. Johnson, A simple estimate of the condition number of a linear system, Coll. Math. J. 26 (1995) 2–5.
[7] P.W. Barber, S.C. Hill, Light Scattering by Particles: Computational Methods, World Scientific, Singapore, 1990.
[8] P.C. Waterman, Symmetry, unitarity and geometry in electromagnetic scattering, Phys Rev D 3 (1971) 825–839.
[9] A. Lakhtakia, V.K. Varadan, V.V. Varadan, Scattering by highly aspherical targets: EBCM coupled with reinforced orthogonalization, Appl. Optic. 23 (1984) 3502–3509.
[10] M.I. Mishchenko, L.D. Travis, A.A. Lacis, Scattering, Absorption and Emission of Light by Small Particles, Cambridge University Press, Cambridge, 2002.
[11] A. Moroz, Improvement of Mishchenko's T-matrix code for absorbing particles, Appl. Optic. 44 (2005) 3604–3609.
[12] D. Petrov, Y. Shkuratov, G. Videen, Optimized matrix inversion technique for the T-matrix method, Opt. Lett. 32 (2007) 1168–1170.
[13] M. Kahnert, T. Rother, Modeling optical properties of particles with small-scale surface roughness: combination of group theory with a perturbation approach, Optic Express 19 (2011) 11138–11151.
[14] W.R.C. Somerville, B. AuguiError! Not a valid embedded object, E.C. Le Ru, Severe loss of precision in calculations of T-matrix integrals, J. Quant. Spectrosc. Radiat. Transfer 113 (2012) 524–535.
[15] P.C. Waterman, The T-matrix revisited, J. Opt. Soc. Am. A (2007) 2257–2267.
[16] W.R.C. Somerville, B. AuguiError! Not a valid embedded object, E.C. Le Ru, A new numerically stable implementation of the T-matrix method for electromagnetic scattering by spheroidal particles, J. Quant. Spectrosc. Radiat. Transfer 123 (2013) 153–168.
[17] D. Petrov, E. Synelnyk, Y. Shkuratov, G. Videen, The T-matrix technique for calculations of scattering properties of ensembles of randomly oriented particles with different size, J. Quant. Spectrosc. Radiat. Transfer 102 (2006) 85–110.
[18] D. Petrov, G. Videen, Y. Shkuratov, M. Kaydash, Analytic T-matrix solution of light scattering from capsule and bi-sphere particles: applications to spore detection, J. Quant. Spectrosc. Radiat. Transfer 108 (2007) 81–105.
[19] D. Petrov, Y. Shkuratov, G. Videen, Sh-matrices method applied to light scattering by finite circular cylinders, J. Quant. Spectrosc. Radiat. Transfer 109 (2008) 1474–1495.
[20] D. Petrov, Y. Shkuratov, G. Videen, Application of Sh-matrix method to light scattering by prolate and oblate spheroids, J. Optic. 12 (2010), 095701.
[21] D. Petrov, Y. Shkuratov, G. Videen, An analytical approach to electromagnetic wave scattering from particles of arbitrary shapes, J. Quant. Spectrosc. Radiat. Transfer 112 (2011) 1636–1645.
[22] M.I. Mishchenko, L.D. Travis, T-matrix computations of light scattering by large spheroidal particles, Optic Commun. 109 (1994) 16–21.
[23] W.R.C. Somerville, B. Auguie, E.C. Le Ru, Simplified expressions of the T-matrix integrals for electromagnetic scattering, Optic Express 36 (2011) 3482–3484.
[24] P.C. Waterman, New formulation of acoustic scattering, J. Acoust. Soc. Am. 45 (1969) 1417–1429.
[25] A. Doicu, Y. Eremin, T. Wriedt, Acoustic and Electromagnetic Scattering Analysis Using Discrete Sources, Academic Press, London, 2000.
[26] J.H. Bruning, Y.T. Lo, Multiple scattering of EM waves by spheres. Part I. Multipole expansion and ray-optical solutions, IEEE Trans. Antenn. Propag. 19 (1971) 378–390.